University of Stavanger

Faculty of Science and Technology

MASTER'S THESIS

Study program/ Specialization: Computer Science	Spring semester, 2017. Open / Restricted access
Writer: Dhanya Therese Jose	………………………………………… (Writer's signature)
Faculty supervisor: Dr. Tomasz Wiktor Wlodarczyk	
Thesis title: **Big Data Analytics** **Case Study- Yelp Dataset**	
Credits (ECTS): 30	
Key words: Big data analysis, Change point analysis, Sentiment analysis, Yelp dataset	Pages: 58 Stavanger, 15/06/2017 Date/year

Front page for master thesis
Faculty of Science and Technology
Decision made by the Dean October 30th 2009

Big Data Analytics
Case Study – Yelp Dataset

Dhanya Therese Jose

Faculty of Science and Technology

University of Stavanger

June 2017

Abstract

In recent years, organizations have changed their work culture in which the business and IT leaders work together with the organizational data in order to make decisions and planning. The handling of these big data was always a challenging taking for IT people as it involved large and complex information, which cannot be handled by conventional tools. For the present study on big data analytics, yelp dataset is taken as a case study. Yelp is a website which publishes crowd-sourced reviews about local businesses and provides opportunity to business owners to improve their services and helps the users to choose best business amongst available. However, it is not possible for the business owners to go through all the user reviews and make important decisions for the improvement of their business. Here comes the importance of big data analytics.

There have been many researchers in the past who worked with yelp dataset and produced very good results with the data. However, many of those studies were focussed on prediction algorithms. In the present study, an attempt is made to interpret the yelp review data using two different data processing techniques; change point analysis and sentiment analysis. Our approach is aimed to provide the owners a more realistic interpretation of the yelp data and finally make some important decisions on the improvement of the business.

The relevant businesses for the present study are obtained based on certain criteria, in order to have a better applicability of the analysis methods. The businesses which have adequate number of reviews and highest fluctuation in the business star ratings are chosen for the study. The change point algorithm is used to obtain the period of fluctuation in the star rating over the past years. In order to ensure optimum number of change points obtained, various parameters used in the change point algorithm is determined based on a sensitivity study. The change points obtained indicated the time where there is a noticeable deviation in the business star ratings. From the present study, it is observed that the number of change points obtained strongly depends on the penalty function used in the algorithm.

Further in the study, sentiment analysis is performed on the review text data corresponding to the same business and star rating data used in change point analysis. Sentiment analysis is meant for text data processing, in which the overall polarity of the text is obtained based on the positive and negative words and phrases used in the text data. In the present study, the polarity of the review text data is obtained using sentiment analysis. Sentiment analysis is performed using Textblob text processing in python. It was observed that there is an overall agreement with the sentiment score of the review text and business ratings. The correlation between sentiment score and change points obtained for the selected businesses were further investigated. There was clear deviation in the sentiment score whenever there is a change point obtained. The possible reasons for the deviation in the star ratings were made based on reviewing the positive and negative noun phrases in the business review text data.

Keywords: change point analysis, sentiment analysis, Yelp dataset, business ratings

Acknowledgements

I would like to thank Dr. Tomasz Wiktor Wlodarczyk, my supervisor for his valuable advises and contributions. The thesis would have never been possible without his help. He was always available whenever, I needed help. He made me to keep my track on the work and further improvement of the project.

I would like to extend my gratitude to my family, especially my husband Jithin Jose. I started my thesis few weeks after my delivery. It was so hard in the beginning to manage work and maternity duties. He helped me to keep balance on both. I never felt stressed during those days. He always motivated me for the completion of the thesis. I also like to thank my little angel Gizel Marie, being a calm, understanding bundle of joy. Also, my parents who were there for my help and support during the early days.

Last but not the least, I would like to thank entire family in India and friends in Norway and almighty for making my project successful.

Dhanya Therese Jose

University of Stavanger

Contents

List of Figures

List of Tables

List of Listings

1. Introduction

More than 40% of the world population uses internet these days compared to 1% in 1995. This made a huge difference in the data world. With advanced technologies, world is leading to a system which relies on real time data. The storing and retrieving of the data became much easy these days. It is capable to generate the information every second of the time and it is a big challenge to analyse such enormous amount of data. There are many organisations whose day today work is directly connected with these information. Handling these data efficiently is a challenging task for IT leaders. Here comes the relevance of the big data analytics. For instance, the weather forecasting based on the real-time measurement weather data. The forecasting station will receive the weather information at a specific location every minute of the hour and this information will be stored in required formats. The user has to read the data and use necessary information required for weather forecasting. These data are so enormous as the station is getting measurements from different locations at the same time. These kinds of bulk information cannot be handled by conventional methods. Big data analytics plays a big role in the manipulation of these kind of data. There are many applications of the big data analytics including oil and gas field, business planning. etc. For the present project, yelp dataset is chosen as a case study.

1.1 Related works

As we have taken yelp dataset for the case study, there are many advantages. The yelp services provide these data freely for the user and invite programmers to participate in yelp dataset challenge in which the participant can come up with an algorithm which can predict the business rating efficiently based on the given dataset and produce reasonable comparison with the upcoming release of the dataset. There are many related works available based on yelp dataset. However, the motivation of most of works are somehow related to yelp dataset challenge.

Fan and Khademi [1] used a combination of three feature generation methods as well as four machine-learning models to find the best prediction of star ratings for the businesses. [2] used yelp dataset and investigated potential factors that may affect business performance. They have found that the review sentiment is one of the main factor affecting review ratings and hence need to be further investigated for accurate prediction. [3] also performed business rating prediction based on sentiment analysis. He has also compared the strength and weakness of different sentiment analysis models. [4] done similar work, predicting star rating based on sentiment analysis of business review data. Most of these studies were focussed on the star rating prediction.

Change point analysis is a powerful tool for determining whether a change has taken place in a time series data. It is capable of detecting subtle changes missed by control charts [5]. [6] developed an R package capable of doing detailed change point detection analysis. It included most of the advanced change point algorithms such as binary segmentation and PELT. [7] applied change point analysis for post market surveillance. They have used this method to perform trend analysis of the medical product sales data. [8] used change point analysis for detecting changes in the incidence of emergency department visit in US hospitals due to influenzas like illness. Change point analysis along with Early Aberration Reporting System (EARS) is found to be effective in detecting illness from emergency department data more effectively than conventional methods. However, the applicability of these methods on yelp data need to be investigated further.

1.2 Motivation and goals

Big data analytics involves the processing of large quantity of diverse data and uncover the correlation and trends in the data in order to obtain useful information. There is always a challenge to process such kind of datasets as the amount of data to be handled is very high and the engineer need to use the computational resources efficiently in order to avoid large analysis time. The present project revolves around the big data analysis of a selected data set, namely, yelp dataset. Various information is read from the dataset and the necessary analysis of the data is performed to obtain certain useful interpretations. Most of the previous studies on yelp dataset were revolved around predictive algorithms. In the present study, an attempt is made to introduce other data analysis methods like change point analysis and sentiment study on the yelp data in order to reach some interpretation on the businesses which can be useful to the owners for the improvement in the performance of the business.

The primary objectives of the project can be divided in two parts. First is to understand local businesses around the world based on yelp dataset and study relevant parameters such as business rating and customer reviews. In this study, relevant businesses are obtained based on certain selection criteria. Five businesses are obtained with respect to largest standard deviation in star rating across the time period and another five based on smallest standard deviations in the rating. In the second part of the study, further analysis of the data is performed based on change point analysis and sentiment analysis on the business star rating and reviews of the selected businesses, respectively. Interpretations were made based on both the studies.

The scope of the thesis is dedicated to research and approaches on big data with reference to the selected dataset. The author used the reference dataset to perform a real interpretation of the data based on the analysis. Various functions to read and analyse the data is introduced in the study and a detail sensitivity study is performed on various parameters used in these functions.

1.3 Organization of the thesis

The thesis is organized in the following way.

- Chapter 2 discuss background of the present study
- Chapter 3 discuss the methodologies used in the present study

- Chapter 4 discuss the results and interpretation based on the present study.
- Chapter 5 discuss on the summary of the present study.

The workflow in the present study is represented in the Figure 1 Workflow.

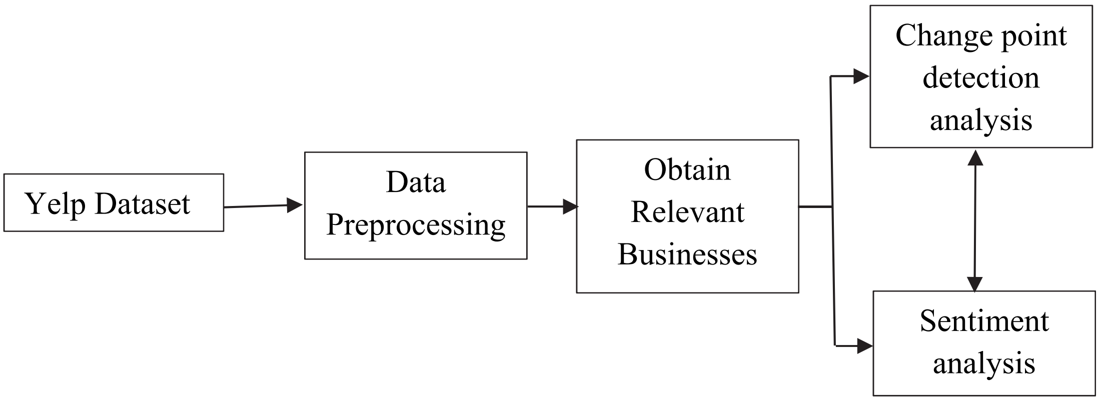

Figure 1 Workflow

2. Background

2.1 Yelp Dataset

For the present study the yelp academic dataset provided by Yelp corporation is used. Yelp connects people with local businesses and the dataset provides rich data about customer's experiences at each businesses via reviews, tips, check-in and business attributes during a period between 2004 and 2017. The scope of local businesses in the chosen dataset is mostly in Canada, USA and some parts in Germany and UK. Yelp provides a way for users to explore, rate and review the businesses they visit. Businesses can highlight their products and services that will attract users to them and finally rate the business. Yelp dataset contains a vast variety of businesses, like restaurants, bars, cafes, local events, doctors, pharmacies, hotels and so on. Users having accounts can also add their friends to yelp. Users can give a star rating from 1 to 5 for a business, and can also write a text review which clarifies the rating. These ratings are very useful for users who are exploring local business, and help them in judging which one would be the best for them. These features of yelp make it a highly recommend system. Each business has an overall rating which is just an average of the star ratings for all the reviews that the business has reviewed. Users can also vote for reviews written by other users.

2.2 Python

Python is a high-level object oriented programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed [9].

Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makes this simple approach very effective.

2.3 Pandas in Python

pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive [10]. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language. It is already well on its way toward this goal.

pandas is well suited for many different kinds of data:

- Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet
- Ordered and unordered (not necessarily fixed-frequency) time series data.
- Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels
- Any other form of observational / statistical data sets. The data actually need not be labeled at all to be placed into a pandas data structure

In pandas there are mainly two type of data structures, namely Series, which is a 1-dimensional and DataFrame which 2-dimensional. Both these data structures covers most of the data applications in most of the engineering and non-engineering field. Pandas dataframe is also compatible with other users such as R. pandas is built on top of NumPy and is intended to integrate well within a scientific computing environment with many other 3rd party libraries.

Many of these principles are here to address the shortcomings frequently experienced using other languages / scientific research environments. For data scientists, working with data is typically divided into multiple stages: munging and cleaning data, analyzing / modeling it, then organizing the results of the analysis into a form suitable for plotting or tabular display. pandas is the ideal tool for all of these tasks.

2.4 R- Language

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R [11].

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and display either on-screen or on hardcopy, and a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

2.5 Change point detection

Change point detection is the name given to the problem of estimating the point at which the statistical properties of a sequence of observations change. Detecting such changes is important in many different application areas. Detection of change point have relevance based on type of data being analysed. In our present study, we are analysing the fluctuations in the star rating over a time span for relevant businesses. Hence, we can find a relevance of doing change point detection in our study. For instance, if the change point detection analysis obtained a change point at year 2012 for the star ratings of a certain business A, which implies that there is a high fluctuation in the star rating for business A in 2012. In the further study, the user can go inside the reviews over that period and find the specific reason for such a fluctuation.

It is based on our requirement that we chose whether to identify single change point or multiple change points. In the present study we use multiple change point detection inorder to all the fluctuations in the star ratings are captured. There are mainly three multiple change point detection algorithms such as Binary Segmentation [12], Segment Neighbourhoods [13] and the Pruned Exact Linear Time (PELT) [6]. The most common approach to identify multiple change points in the literature is to minimize both cost function for a segment and penalty to guard against over fitting.

Different multiple change point algorithms are briefly explained below,

Binary Segmentation- Binary Segmentation first applies a single change point test statistic to the entire data. If a change point is identified the data is split into two at the change point location. The single change point procedure is repeated on the two new data sets, before and after the change. If change points are identified in either of the new data sets, they are split further. This process continues until no change points are found in any parts of the data. Binary segmentation is thus an approximate algorithm but is computationally fast as it only considers a subset of the 2n−1 possible solutions. The computational complexity of the algorithm is O(n log n), but this speed can come at the expense of accuracy of the resulting change points.

Segment Neighbourhood-The segment neighbourhood algorithm was proposed by [13] and further explored in [14]. The algorithm minimizes cost function for a segment and penalty to guard against over fitting exactly using a dynamic programming technique to obtain the optimal segmentation (for m + 1) change points reusing the information that was calculated for m change points. This reduces the computational complexity from O(2n) for a naive search to O(Qn2) where Q is the maximum number of change points to identify. Whilst this algorithm is exact, the computational complexity is considerably higher than that of binary segmentation.

Pruned Exact Linear Time-The binary segmentation and segment neighbourhood algorithms would appear to indicate a trade-off between speed and accuracy however this need not be the case. The PELT algorithm proposed by [6] is similar to that of the segment neighbourhood algorithm in that it provides an exact segmentation. However, due to the construction of the PELT algorithm, it can be shown to be more computationally efficient, due to its use of dynamic programming and pruning which can result in an O(n) search algorithm subject to certain assumptions being satisfied, the majority of which are not particularly onerous. Indeed, the main assumption that controls the computational time is that the number of change points increases linearly as the data set grows, i.e., change points are spread throughout the data rather than confined to one portion.

2.6 Sentimental Analysis

Sentiment analysis is a text processing method used to determine the sentiment of a text data with the help of Natural Language Processing (NLP), artificial intelligence and computer linguistics. The text data can be any useful information such as reviews, comments. etc. In general, the sentiment analysis returns the tone of the text based on the polarity of the words and phrases used in the text information. This method make use of a known text database, which contains the polarity of positive and negative words commonly used in writings. By comparing the input text data with the known database, overall sentiment of the text is returned. However, the implementation of the text processing can be different in different sentiment analysis models. The sentiment score is a measure of the positivity and negativity of the text input. For example, consider an online review written by someone about a particular hotel business. The sentiment analysis can return whether the user is happy about the business or not based on the overall sentiment score of the review text.

The sentiment analysis finds its application when there is a large amount of text data to be handled, like an entire book or social media comments. However, this method is not fully accurate. In some cases in which there is a sarcastic text, the actual sentiment should be negative. However, the sentiment analysis will rate the text with a positive polarity, as the words in the text are positive. Sentiment analysis is difficult to validate because in many scenarios "ground truth" is not available. But in the case of yelp review data, both the user reviews and star ratings are available. A combination of both can overcome these issues. There are many models available for sentiment analysis. All models have their strength and weaknesses.

3. Methodology

3.1 Structure of the Yelp dataset

The Yelp dataset is a single zip-compressed file, composed of five compressed json files. Every file contains a 'type' field, which implies whether it is a business, a user, a review, a check-in or a tip. The fields are separated by comma. The size of business file: 114.5 MB, review file: 3.46GB, user file: 1.18GB, tip file: 182.2MB, check in file: 46.2MB

Yelp dataset contains 144072 businesses and 1029432 users with 4153149 reviews and 946599 tips. The dataset includes businesses in four different countries: Edinburgh, U.K.; Karlsruhe, Germany, Montreal and Waterloo, Canada; Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison, U.S., making it a very versatile dataset. Following is a description of all the fields in each of the data types.

Business

Table 1 Fields of business file

" address": (localized address),↵
"business_id": (encrypted business id),↵
"categories": [(localized category names)],
"city": (city),
"hours": { (day_of_week): {open": (HH:MM),
"close": (HH:MM)}},
"is_open": True / False (corresponds to closed, **not** business hours),
"latitude": latitude,↵
"longitude": longitude,
"name": (business name),↵
"neighborhoods": [(neighborhood names)], ↵
"postal_code": (postal code of the location),↵
"review_count": review count,↵

"stars": (star rating, rounded to half-stars),
"state": (state),₠"type": "business"

For each business there is a business id, address including latitudes and longitudes. It also has the number of reviews that have ever been written for the business, and an average star rating across all the reviews. Yelp also stores other information as attributes such as hours, parking, wheelchair accessibility, ambience etc.

Review

Table 2 Fields of review file

"type": "review",
"business_id": (encrypted business id),
"votes": {(vote type): (count)},
"date": (date, formatted like "2012-03-14"),
"review_id":(encrypted review id),
"stars": (star rating, rounded to half-stars),
"text": (review text),
"user_id": (encrypted user id),

Each review consists of a star rating and review text, possibly justifying the star rating. Each review can also get votes from other users, they can vote if they find the particular review is cool, funny, or useful.

User

Table 3 Fields of user file

"average_stars": (floating point average, like" 4.31),
"compliment": {(compliment type): (number of compliment)},
"votes": {(vote type): (count)},
"elite": [(years_elite)],
"fans":(fans count),

"friends":[(friends user_id’s)],
"name":(name),
"review_count": (review count),
"type": "user",
"user_id": (encrypted user id),
" yelping_since": (date)

Yelp contains a strong user network, and stores information about them like their name, the number of reviews they have written, how long they have been using Yelp for, their friends those who uses Yelp, user’s fans count

Check in

Table 4 Fields of check-in file

"business_id": (encrypted business id),
"time": { (no. of checkins in time periods)
"type": "checkin"

This gives an aggregated view of all the checkins for a business for every hour of the day, for every day of the week, and gives a great idea about what are the busiest times for the business.

Tip

Table 5 Fields of tip file

"business_id": (encrypted business id),
"date": (date, formatted like "2012-03-14"),
"likes": (count),
"text": (tip text),
"type": "tip",
"user_id": (encrypted user id),

Tips stores random comments that users leave about a business, they are different from reviews in that they don't have a star rating, and are just quick indications for others.

3.2 Reading the Dataset in Pandas

For reading the dataset in json format, a function is defined. The user defined function is shown in this section. The dataset file in json format is read line by line instead of reading all line together due to the size of the dataset. However, this method will draw some additional time due to the reading routine. The read lines are assigned to dataframe for the convenience of further processing of the data Business and review files in the yelp dataset are used in the current work.

```
#Reading the data line by line

def load_data(filepath):
    d = []
    with open(filepath) as file:
        d = pd.DataFrame.from_dict(json.loads(line.rstrip())
for line in file)
```

Listing 1 Reading the data line by line

3.3 Preprocessing of the data

For any data, it is important to preprocess the data before analysis. The preprocessing is done on the review dataset. As a first step, data is sorted in such a way that all the review dates of each business are in ascending order. It is important as in our study we are performing a time series analysis. Due to the size of the dataset, analysing the dataset demands large computational requirements. Hence, we tried to reduce the fields in the dataset, retaining useful information. In the dataset, there is a field "votes" which have three different fields which designates three different values, namely, funny, useful and cool. However, in our study we reduce the three spaces in vote field to one, which indicates the total number of votes obtained for a particular review.

```
#Reducing the vote field to single column

review_df=review_df.sort_values(by=['business_id','date'],
ascending=[True, True])
review_df['votes']=review_df['funny']+review_df['cool']+
review_df['useful']
review_df=review_df.drop([col for col in
['funny','useful','cool'] if col in review_df], axis=1)
```

Listing 2 Reducing the vote field to single column

3.4 Data compensation

In the present work authors have performed preliminary analysis of the data and found certain inconsistencies in the data with respect to the objectives of the work. In the present work, we are performing a time series analysis and hence it is important to have consistent data with respect to time. However, in the yelp dataset the review data are somewhat scattered. For example, when we analyze the monthly average star rating over a period of time, there are some missing data for certain months. These inconstancies will affect the data analysis. There is a need to fill the missing data (star rating) with most probably values. caseIn the present case, we filled the missing fields in the star ratings with the values from the most adjacent field. This yield a consistent data time series without changing much of the information in the data. However, for the review text data, this method not implemented and the missing fields are kept as it is.

3.5 Change point detection

As mentioned in the previous chapter, change point detection is a method for analyzing the point of fluctuations in a time series data. The change point detection algorithm detects multiple points of fluctuations depending on the data. In the present study change point detection is relevant as we are focused on the changes in the business rating over a period of time.

In the present study, R programming language is used for change point detection. One of the key challenges in change point analysis is the ability to detect multiple changes within a given time series or sequence. The change point package has been developed to provide users with a choice of multiple change points search in conjunction with a given change point method. The change point package 'cpt' is called in the R language for change point analysis. It implements various mainstream and specialised change point methods for finding single and multiple change points within data. Many popular non-parametric and frequentist methods are also included. Functions included in this package are cpt.mean(), cpt.var() and cpt.meanvar(), in which cpt stands for change point and mean, var and meanvar stands for the criteria use for change point detection. In cpt.mean, the change point is detected based on the mean of the time series data. In our study we have chosen cpt.mean as it is the most simplified method and it is capable of capturing change points in the present data. Moreover, we are focussed on the mean deviation in the star ratings of the business over time. cpt.var and cpt.meanvar are based on the variance and mean- variance of the time series data which is suitable for more complicated data types.

Among the three major change point detection algorithms, for the present study we have used binary segmentation (BinSeg) method as it is relatively fast. PELT algorithm is more accurate than BinSeg and is more suited for complicated data. For example, if we need to find a stoke in ECG graph, PELT algorithm is recommended over BinSeg as it accurately estimate the point where there is deviation in the ECG. However, in our study our data is the star rating over a period of time, which is relatively simple and BinSeg algorithm is sufficient for present data. In this method, given time series data is divided into two segments based on the fluctuation in the mean of the data. Further divisions in the segments are performed depending on the fluctuations. Change points are detected based on the divisions. However, there are many factors which affects the number of change points obtained. It is very important to obtain optimum number of change points from the data.

```
#Change point detection in R

cpt.mean(data, penalty,pen.value,method,Q,test.stat,class,
param.estimates,minseglen)
```

Listing 3 Change point detection in R language

In the change point package in R, there are different parameters which need to be defined for accurately estimating the change points. The various parameters used in the ‘cpt.mean’ function are listed below,

Data	The object or variable in which the change point is to be detected.
Penalty	Choice of the penalty function. It can take the values "None", "SIC", "BIC", "MBIC", AIC", "Hannan-Quinn", "Asymptotic", "Manual" and "CROPS"
pen.value	It is the parameter value for penalty function
method	Choice of "AMOC", "PELT", "SegNeigh" or "BinSeg"
Q	The maximum number of changepoints to search for using the "BinSeg" method.
test.stat	Type of distribution of the data. For example, Normal distribution.
class	If TRUE object class is returned.
param.estimates	If TRUE and class=TRUE then parameter estimates are returned
minseglen	Positive integer giving the minimum segment length

It is very clear from the cpt.mean usage that there are many parameters, which need to be defined for proper estimation of the change points. For the present study, we did not use any complicated penalty functions such as SIC or BIC. Instead, we used manual method and its impact on the analysis is investigated. In the manual penalty function, we need to give pen.value, which decide the accuracy of the number of change points detected.

3.2.1. Optimum value for 'pen.value' using Elbow method

In order to obtain optimum value for the penalty function in manual mode, we used elbow graph method, in which the number of change points detected for different values of the pen.value is obtained. The point where there is sudden shift in the number of change point with the increase in the pen.value is taken as the optimum value for the analysis. The advantage of this method is that it avoid the chances of detecting any noise related change points in the data and the user will have proper control on the optimum number of change points to be detected.

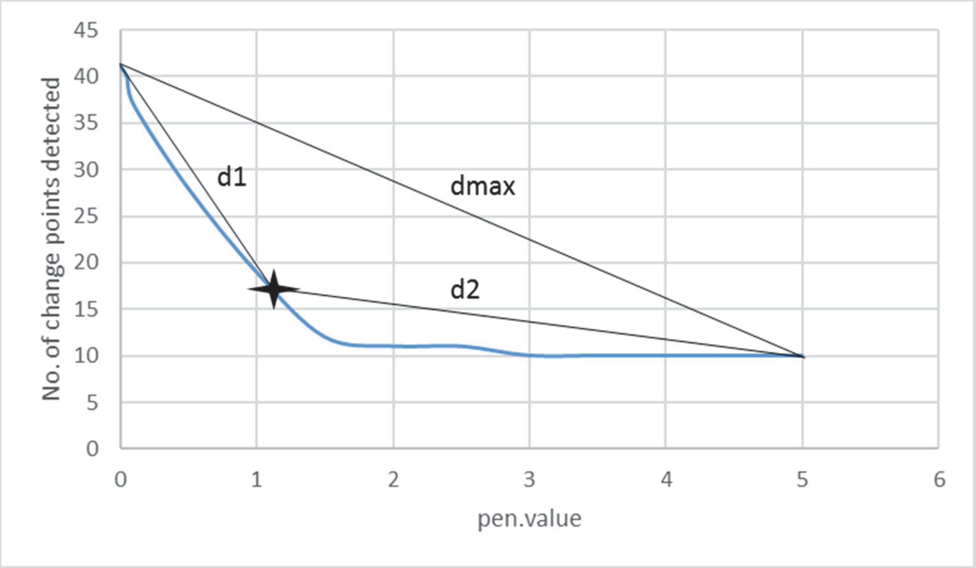

Figure 2 Steps involved in elbow point detection algorithm

The algorithm used for obtaining elbow point from the elbow graph is shown in Figure 2. The various steps involved are,

1. Obtain the pen.value Vs number of change point graph
2. For each point on the elbow curve, obtain the triangle with sides dmax (connecting two end point of the elbow graph),d1 (connecting one end point and the point of the elbow graph) and d2 (connecting other end point and the point of the elbow graph). The length of dmax, d1 and d2 are obtained using the coordinates of the triangle.
3. The angle between the d1 and d2 is obtained by applying the cosine rule as follows,
$$\cos\theta = (\mathrm{d1}^{2} + \mathrm{d2}^{2} - \mathrm{dmax}^{2})/(2 * \mathrm{d1} * \mathrm{d2})$$
4. The angle is obtained for all points on the elbow graph. The elbow point is point which has the minimum angle.

The code in R for obtaining the elbow point is shown below,

```
#Obtaining the elbow point in R
# Inputs a vector and the number of elbow points to be found
elbowpoints <- function(x)
{
dvec = c()
# Normalize the vector
x = x/max(x)
L = length(x)# The distance of the endpoints
dmax    =    dist(rbind(c(1/L,    x[1]),    c(1,    x[L])),
method="euclidean")
# Find the point with maximum distance (minimum angle)
for (i in 1:L)
{
  d1 = dist(rbind(c(1/L, x[1]), c(i/L,x[i])),
       method="euclidean")
  d2 = dist(rbind(c(i/L, x[i]), c(1, x[L])),
       method="euclidean")
  dvec = c(dvec, abs((d1^2 + d2^2 - dmax^2)/(2*d1*d2)))
}
return (order(dvec)[1:L])
}
```

Listing 4 Obtaining the elbow point in R language

3.6 Sentiment Analysis

Sentiment analysis is often used to quantify business opinions in social medias in a more effective way. These algorithms give sentiment scores based on the polarity of the relevant words in the reviews written by the users. In the present work sentiment analysis is used to verify the change points obtained for the businesses. The variation in the sentiment score can have direct correlation with the change points detected. However, the correlation is also a function of the credibility of the reviews written. For a customer review credibility is not fully guaranteed. So in the present case we use sentiment analysis along with change point detection method to obtain the fluctuations in the star rating.

In the present study we have used Textblob, which is a text processing library in Python language for sentiment analysis. TextBlob is a simpler, more human interface for natural language processing. Textblob heavily depends on NLTK and pattern module by CLIPS. It works on finding the polarity of words in the text and averages them all together for longer text..The various features of Textblob includes the following:

- Noun phrase extraction
- Part-of-speech tagging

- Sentiment analysis
- Classification (Naive Bayes, Decision Tree)
- Language translation and detection powered by Google Translate
- Tokenization (splitting text into words and sentences)
- Word and phrase frequencies
- Parsing
- n-grams
- Word inflection (pluralization and singularization) and lemmatization
- Spelling correction
- Add new models or languages through extensions
- WordNet integration

In the present study, we are using sentiment analysis and noun phrase extraction.

```
tb_value = []
stars_tb = []
d=[]
for ind,review in islice(FilteredDf.iterrows(),#no:of rows):

    details = TextBlob(review['text'])
    tb_value.append(details.sentiment.polarity)
    d.append(review['date'])

p_stars_tb = pd.DataFrame()
p_stars_tb['senti_value'] = tb_value
p_stars_tb['date']=d
```

Listing 5 Obtaining sentiment score(polarity)

4. Results and Discussion

4.1 Identifying relevant businesses

In Yelp dataset, 144072 unique business ids are there. It is not necessary to take all the businesses. We consider only frequently reviewed businesses or active businesses. This is because, in the present study we are mainly focussed on the star ratings of businesses over a period, which is not suitable for small or recently started businesses. If the number of reviews are too less, it will affect the quality of the results. It is necessary to filter out passive businesses from the dataset before we start the analysis.

For the present study a certain criterion is chosen for selecting relevant businesses. The criteria must be satisfied throughout the study.

The criteria for selecting relevant business are,

- Top five and bottom business whose standard deviation in the business star rating during the period 2004 to 2016 are highest and lowest.

- The number of reviews for the selected businesses must be above 100.

4.1.1 Scaling effect

Figure 3 shows the histogram plot of the review counts and number of businesses. From the histogram, it can be noted that there are around 130000 businesses that have reviews less than 100. As we are aimed to perform a time series study on the reviews and star ratings of the businesses, this is not sufficient.

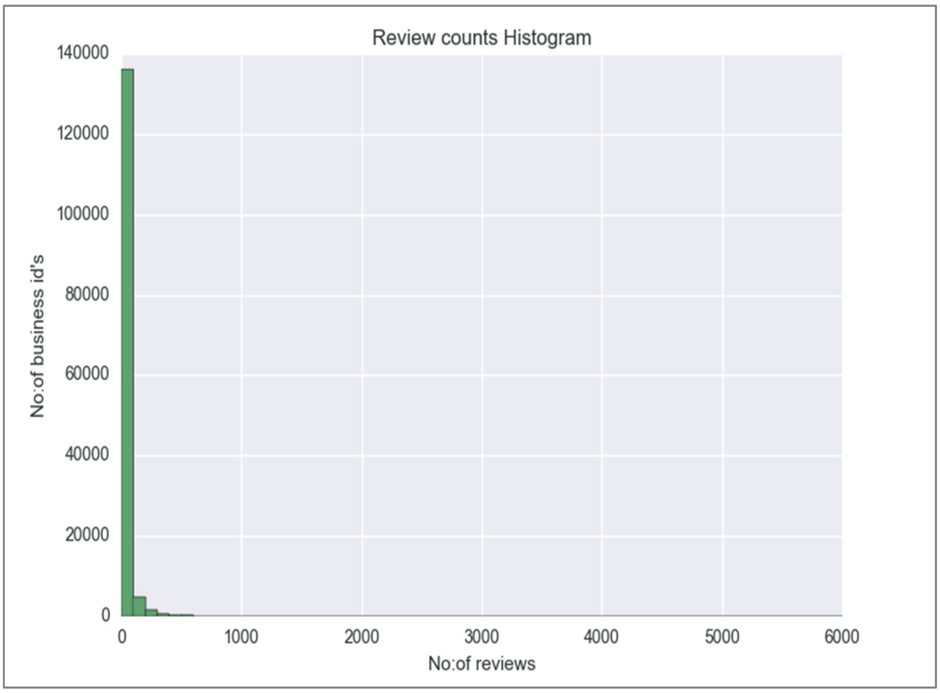

Figure 3 Histogram of review counts vs number of business ids.

Even though we assumed businesses with review counts more than 100 as established businesses in the selection criteria, there are even decent number of businesses whose review counts are more than 1000. Hence, we subdivide the second criteria into two cases as follows,

- Case 1- Businesses with review counts more than 1000.
- Case 2- Businesses with review counts more than 100.

Table 6 Total number of businesses for each case

Review count	No: of businesses
Case 1 (>1000)	177
Case 2 (>100)	7846

In case1 we considered businesses greater than 1000 review counts. The number of businesses in this case is 177. For the ease of understanding, an annual average of the star rating is plotted in Figure 6 with top 5 businesses with higher standard deviation in the review rating. In the same way, Figure 4 shows the top five businesses with lowest standard deviation. All our further studies will be based on these chosen businesses.

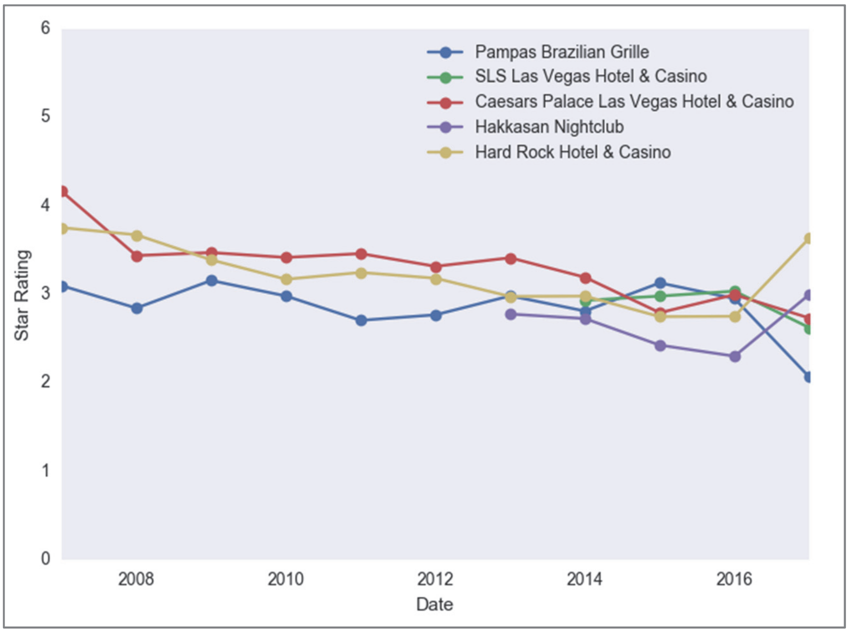

Figure 4 Top 5 businesses with highest standard deviation in business ratings (Annual average)-Case1

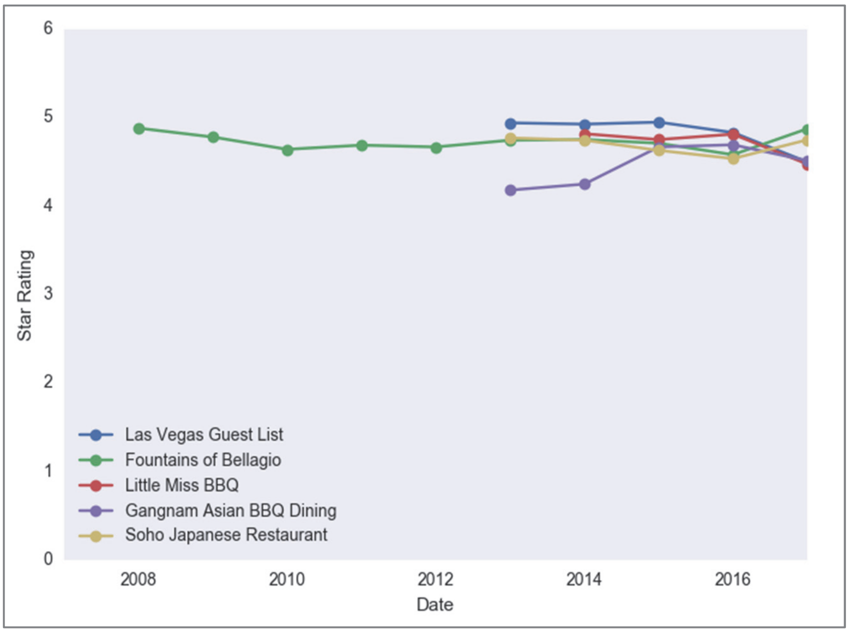

Figure 5 Top 5 businesses with lowest standard deviation in business ratings (Annual average)-Case1

While considering the businesses with review count greater than 100 (case 2), there are 7934 businesses in the dataset. It was expected that there will be variation in the results with review count 1000 and 100 as we have taken standard deviation as the criteria for identifying the businesses. Figure 5 shows top 5 businesses with higher standard deviation in the review rating and Figure 6 shows top five businesses with lowest standard deviation. In figure 6, as it is the set of businesses with lowest standard deviation according to each businesses stars, standard deviation is zero, so the points are overlapped. It is observed that the cases with lowest standard deviation are relatively new businesses.

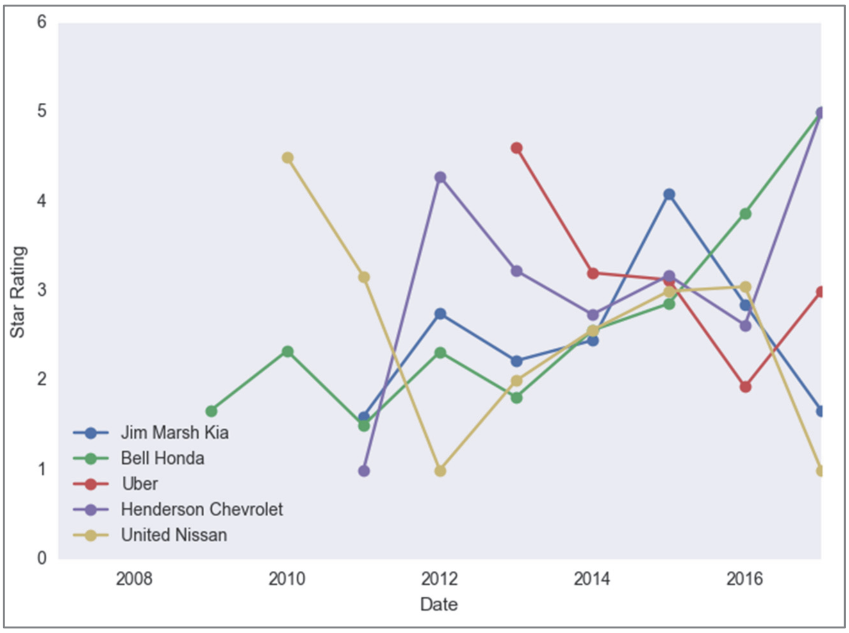

Figure 6 Top 5 businesses with highest standard deviation in business ratings (Annual average)- Case2

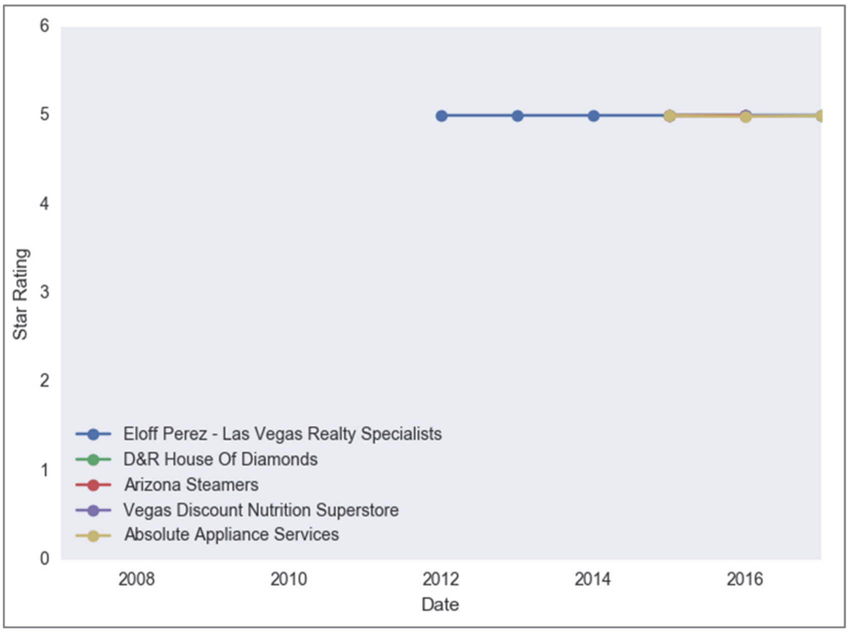

Figure 7 Top 5 businesses with lowest standard deviation in business ratings (Annual average)-Case2

From the above observation it is clear that, case 2 is more interesting than case 1. Moreover, the identified businesses are totally different in both the cases. In case 1, there we only few number of businesses compared to case 2. Also, it was observed that the standard deviation in business ratings were less for case1 than case 2. Due to all the above-mentioned reasons, we selected case 2 as the selection criteria for the further studies. The relevant businesses are selected based on the selection criteria as shown in Table 7.

Table 7 Businesses selected based on the selection criteria

No	Top Five with Highest Std. Deviation	Top Five with lowest Std. Deviation
1	Jim Marsh Kia	Eloff Perez
2	Bell Honda	D & R House of Diamonds
3	Uber	Arizona Steamers
4	Henderson Chevrolet	Vegas Discount Nutrition Superstore
5	United Nissan	Absolute Appliance Services

It is interesting to note that the businesses with highest standard deviation in the star ratings are all automobile businesses located in United States. One of the main reason for that is our present selection criteria. Yelp was popular in United States in the earlier times. Recently it is stretched into other territories. As our criteria stops the businesses with less than 100 reviews, may of the businesses which came to the dataset in recent times will be dropped. Moreover, automobile businesses are always popular in US business sector.

4.2 Change Point detection

Change point detection is a method of analysing fluctuation in a time series data. For the present study, change point package in R programming language is used for change point detection. The binary segmentation algorithm is used along with a manual penalty function. The penalty function value can be chosen as per the user requirement. However, for the present study the penalty function value is chosen using elbow point method which estimates optimum number of change points.

The change point analysis is performed on the star ratings of the businesses obtained in the previous section. However, the businesses with least variation in star ratings were not used for the change point study.

4.2.1 Business: Jim Marsh Kia

Jim Marsh Kia is a car dealer in Las Vegas which deals with sale of new and used cars. We have the star ratings of Jim Marsh Kia from year 2011 to 2016. Frequency of data in each year is set to 12, which means there is a minimum of one star rating every month of the year. However, in the present dataset there are some months in which no reviews were recorded, which results in inconsistent data for the change point study. In the present study, we used some filling method for missing data. The missing values in the dataset are filled with the neighbour values. Figure 8 shows the plot of star rating time series for Jim Marsh Kia. From Figure 8 it is evident that there are lot of fluctuations in the star rating in the period between 2011 to 2017.

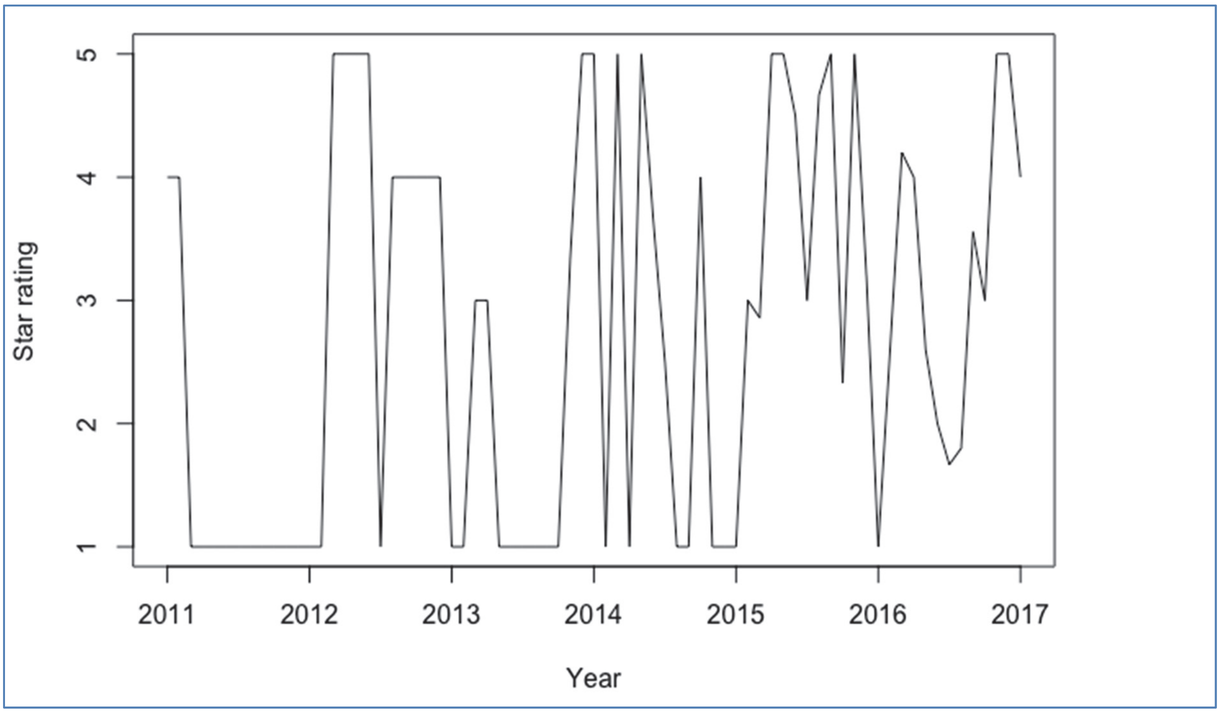

Figure 8 Star rating time series for Jim Marsh Kia.

As discussed in the previous chapter the binary segmentation method used for the present study uses a penalty function value based on the elbow method. Figure 9 shows the 'elbow graph' for Jim Marsh Kia business.

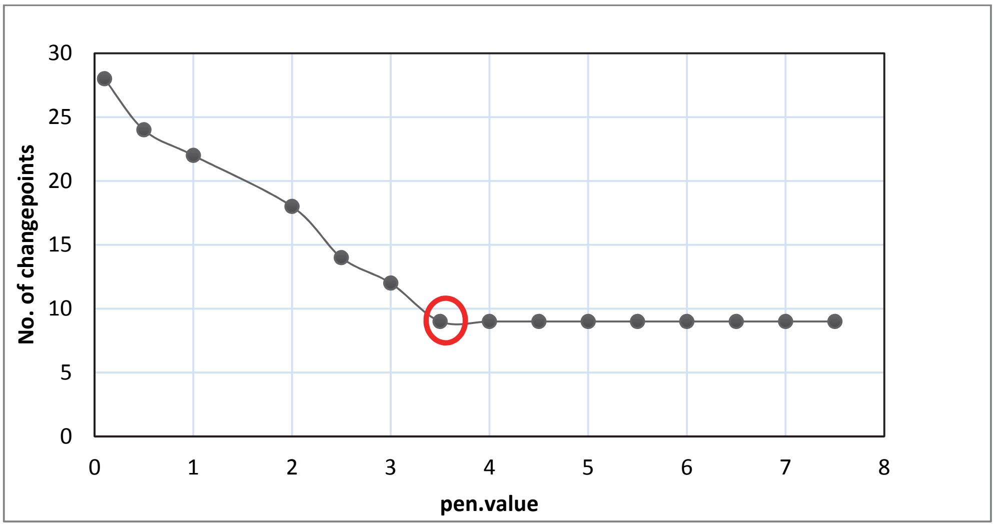

Figure 9 Elbow plot for Jim Marsh Kia's star rating

In the graph, there is a considerable decrease in the number of change points detected when the pen.value is 3.5. Hence, we can take the value 3.5 as the elbow point of the graph. This value should be used as the pen.value in the change point detection. Table 3 shows the final output from the change point analysis.

Table 8 Change points detected based on BinSeg algorithms- Jim Marsh Kia

BinSeg Algorithm , pen.value=3.5	
Criteria	**Change points detected based on monthly performance**
Mean	(Feb-11, Feb-12, Jun-12, Dec-12, Oct-13, Jun-14, Mar-15, Sep-15, Oct-16)

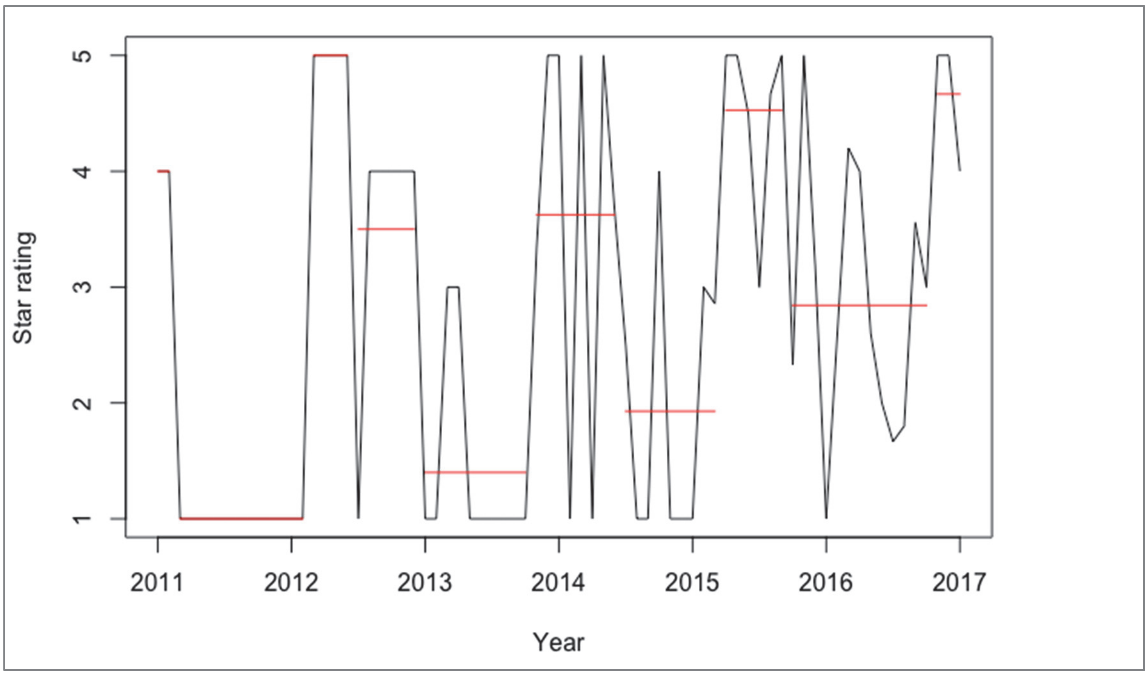

Figure 10 Change points of Jim Marsh Kia's star rating

Figure 10 shows the change points detected along the star rating time series for Jim Marsh Kia. The red band show the regions where the average rating is constant and the breaks in the red line indicated the change point as listed in Table 8. However, in the time series there are instances where there is a single review which rates the business much different from the nearest reviews. These points in the time series are outliers and should not be considered as change points. For the present study, the algorithm and the penalty function values used will takes care of these outlier reviews ensuring optimum number of change points detected.

There is a sudden drop in the rating of the business between the period Feb 2011 to Feb 2012. However, the ratings increased to good level in the following years until 2013. It is interesting to notice that the fluctuation in the star rating follows up and down scheme. Also it can be noted that the average business rating over the period 2011 to 2017, of the business is increased. This indicated the good performance of the business over time.

4.2.2 Bell Honda

Bell Honda is also car dealers located in Phoenix. In the dataset we have the star ratings of Bell Honda from year 2009 to 2016. Frequency of data in each year is set to 12, indicating monthly average of the star ratings. With the data, plotted the time series of star rating is shown in Figure 11. From the star rating time series it is observed that there is less deviations in the star rating until 2012. The fluctuations in the star ratings increases after year 2012. The reason for such a large fluctuation can be due to the increase in the number of reviews for the business in recent times.

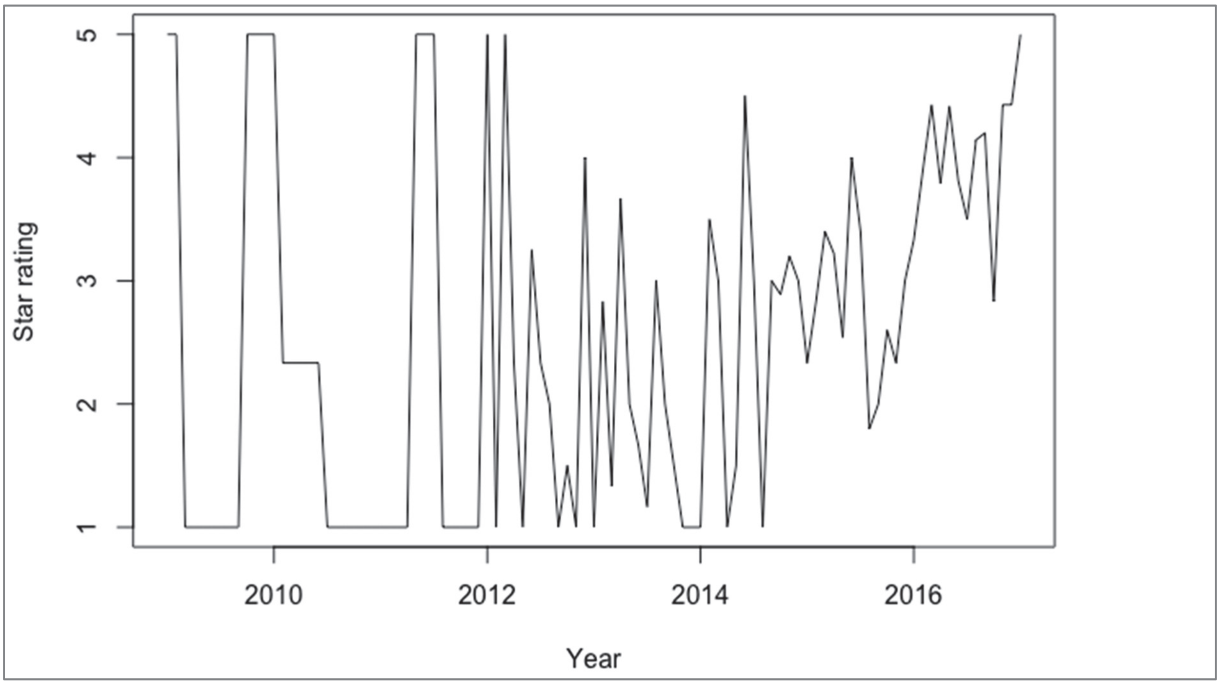

Figure 11 Bell Honda: star rating in time series

Table 9 Change points detected based on BinSeg algorithms- Bell Honda

BinSeg Algorithm, pen.value=1.5	
Criteria	**Change points detected based on monthly performance**
Mean	(Feb-09, Sep-09, Jan-10, June-11, April-11, July-11, Dec-11, Feb-12, March-12, May-14, July-14, Dec-15)

The pen.value used for the change point detection is obtained using the elbow method as explained for Jim Marsh Kia. The pen.value obtained for this case is 1.5. Figure 12 and Table 9 shows the change points detected for Bell Honda data. There are 12 change points. The outlier points in the star ratings are ignored. In such cases an average star rating is considered for obtaining the change points.

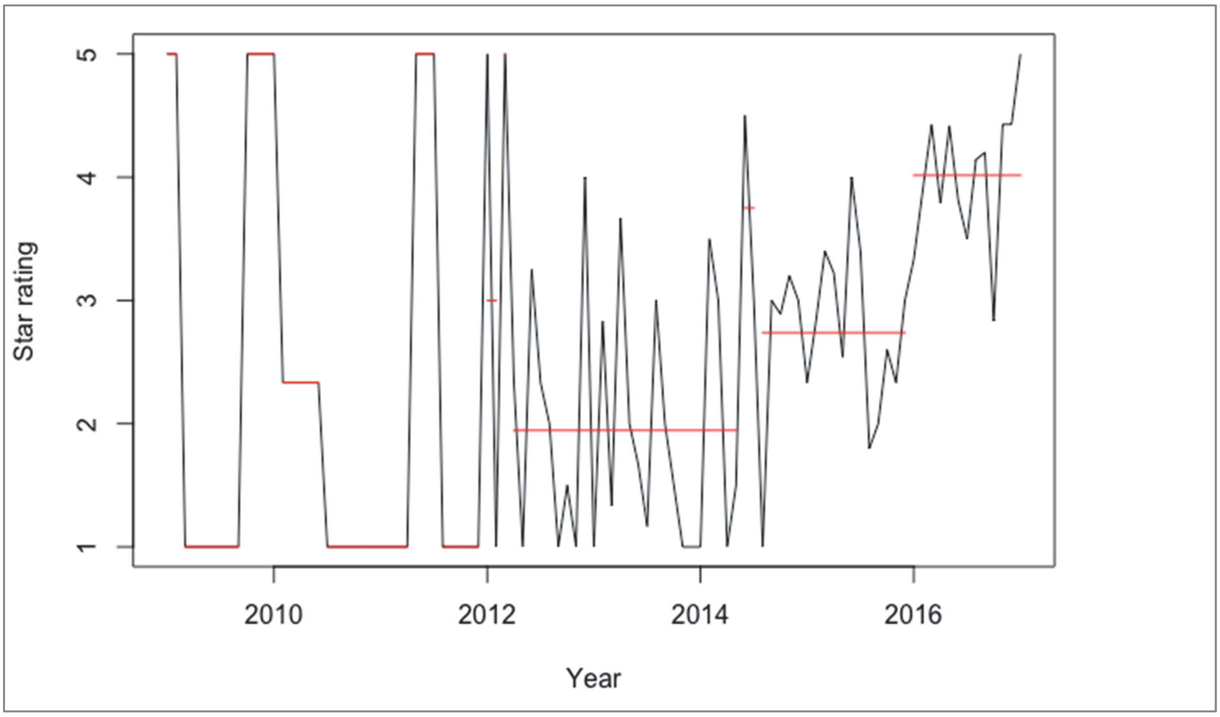

Figure 12 Change points of Bell Honda's star rating

4.2.3 Business: Uber

Uber is one of the famous taxi transportation. It became very popular these days. We have the star rating of Phoenix city's Uber from year 2013 to 2016. Frequency of data in each year is set to 12 as like before. With the data, plotted the time series of star rating as shown in Figure 13.

Table 10 Change points detected based on BinSeg algorithms- Uber

BinSeg Algorithm , pen.value=4.5	
	Change points detected based on monthly performance
Mean	(July-14, Sep-14, Oct-15)

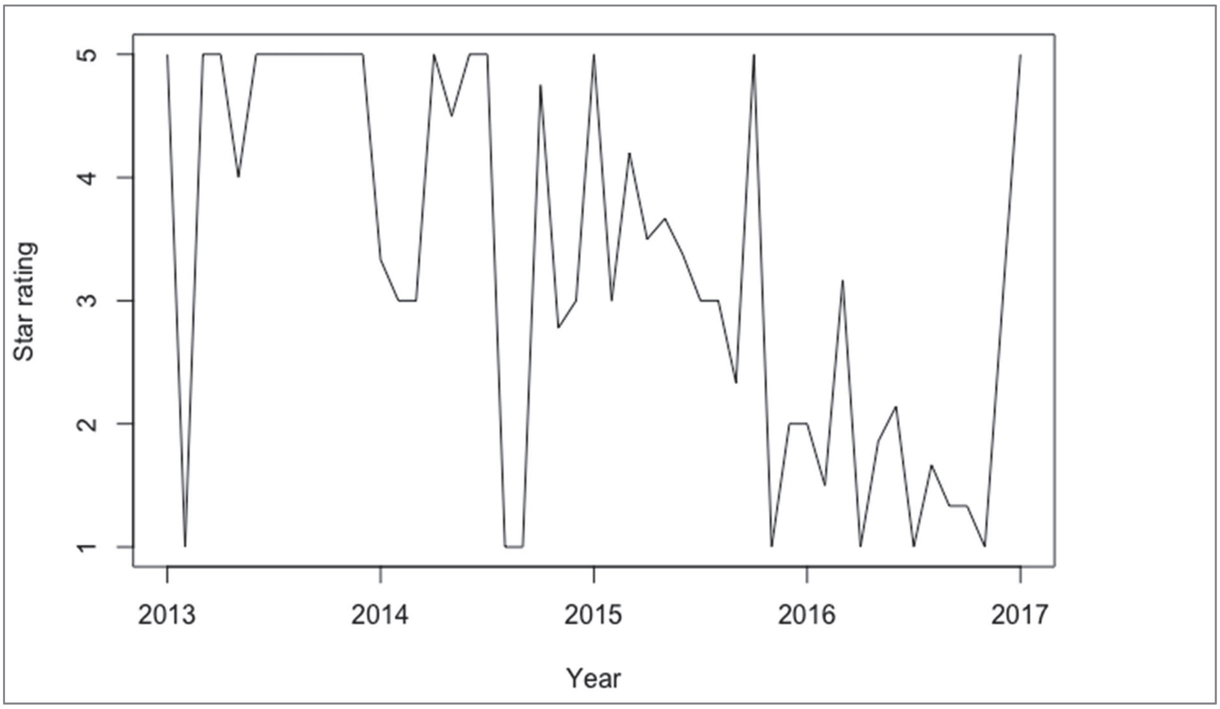

Figure 13 Uber: star rating in time series

The pen.value obtained for this case is 4.5. Figure 14 and Table 10 shows the change points detected for Uber. There are three change points. It can be noted that the overall performance of the business is dropping over the years. However, Uber details are added into the database recently.

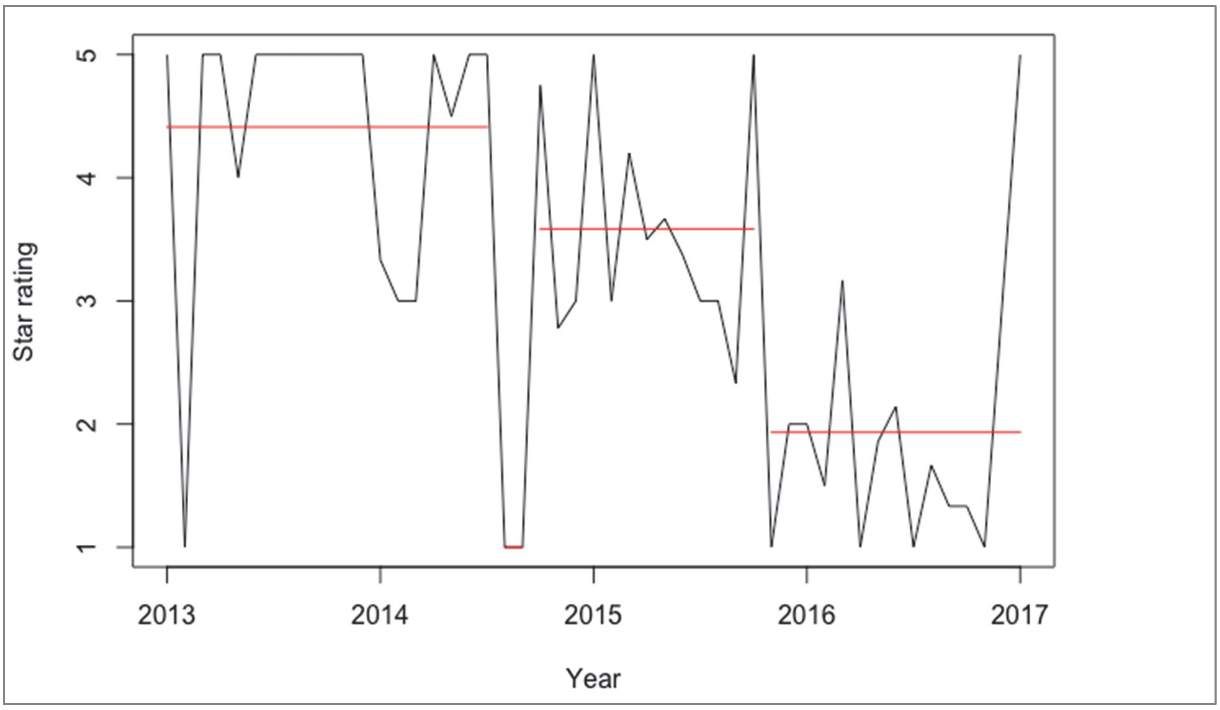

Figure 14 Change points of Uber's star rating

4.2.4 Business: Henderson Chevrolet

Henderson Chevrolet are also car dealers in Henderson city. We have the star rating for the business over a period of 2011 to 2017. The time series of the business star rating over this period is shown in Figure 15. The frequency of the star rating is set to 12, similar to the previous cases.

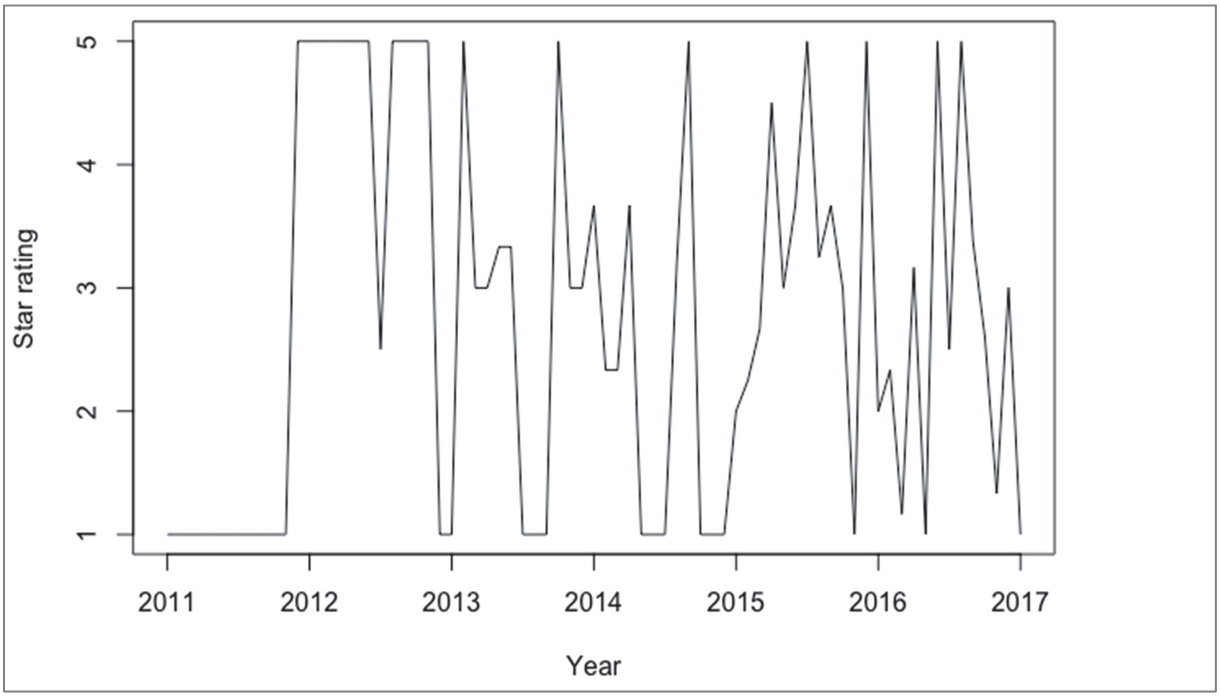

Figure 15 Henderson Chevrolet: star rating in time series

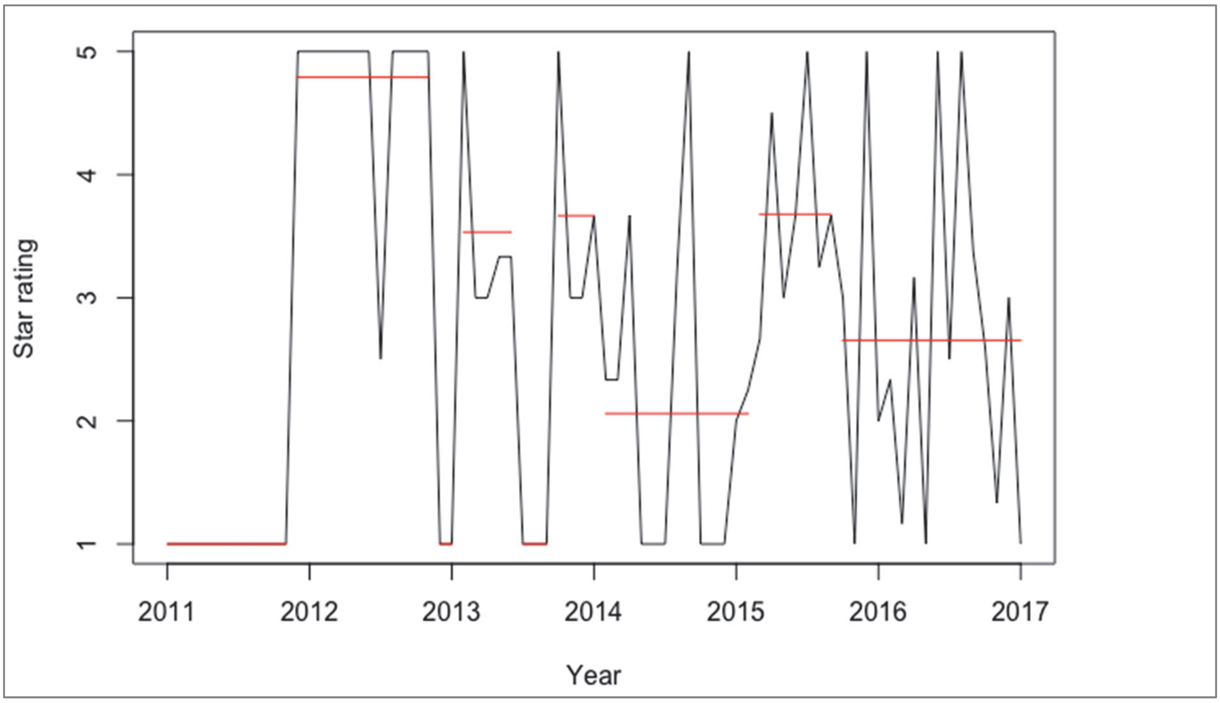

Figure 16 Change points of Henderson Chevrolet's star rating

Table 11 Change points detected based on BinSeg algorithms- Henderson Chevrolet

BinSeg Algorithm , pen.value=3.5	
	Change points detected based on monthly performance
Mean	(Nov-11, Nov-12, Jan-13, June-13, Sep-13, Jan-14, Feb-15, Sep-15)

Figure 16 and Table 11 shows the change points obtained for Chevrolet star ratings. There are 8 change points detected for this case. There is a lot of variation in the average star rating for the business between the period 2011 and 2017.

4.2.5 Business: United Nissan

United Nissan are car dealers in Las Vegas. In the dataset, we have business rating of United Nissan over a period of 2010 to 2017. Figure 17 shows the star rating time series of United Nissan. There are much fluctuations in the star rating between 2014 and 2017 compared to period 2010 and 2014.

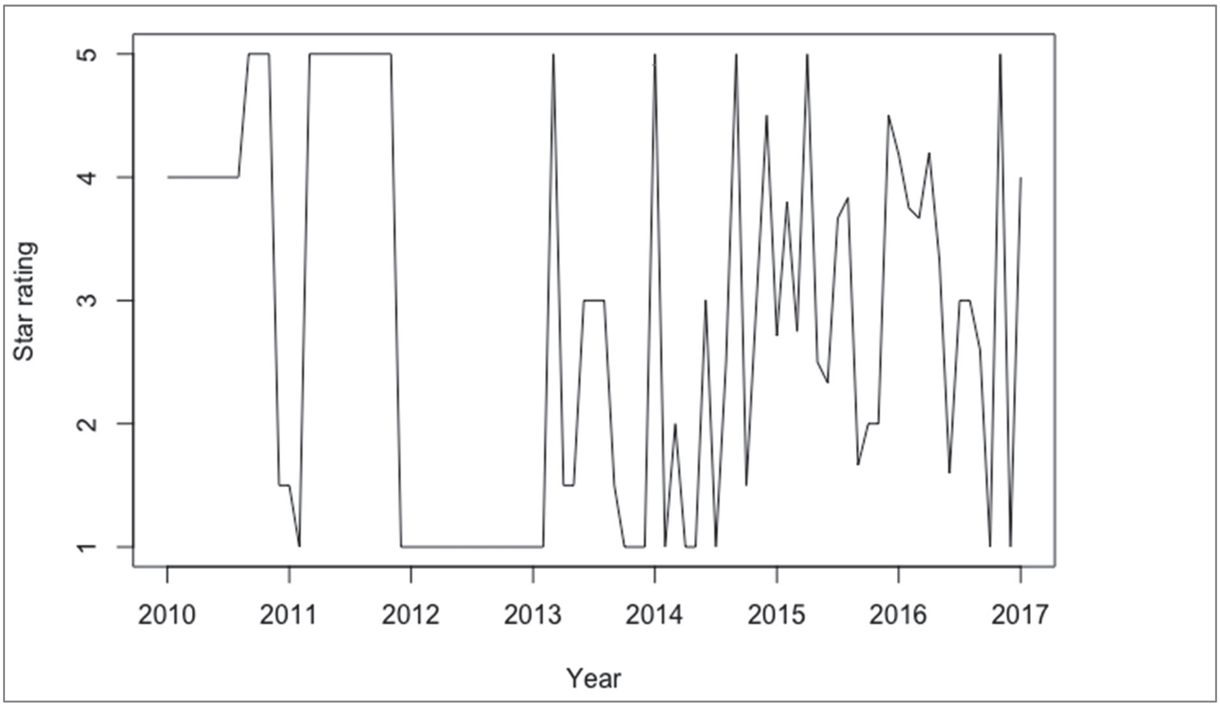

Figure 17 United Nissan: star rating in time series

Table 12 Change points detected based on BinSeg algorithms- United Nissan

BinSeg Algorithm , pen.value=2	
	Change points detected based on monthly performance
Mean	(Aug-10, Nov-10, Feb-11, Nov-11, Feb-13, Aug-13, July-14, Aug-15, Nov-15, May-16, Oct-16)

Figure 18 and Table 12 shows the change points obtained for United Nissan. The pen.value obtained using elbow method for this case is 2 . There are 11 change points obtained for this case using binary segmentation algorithm.

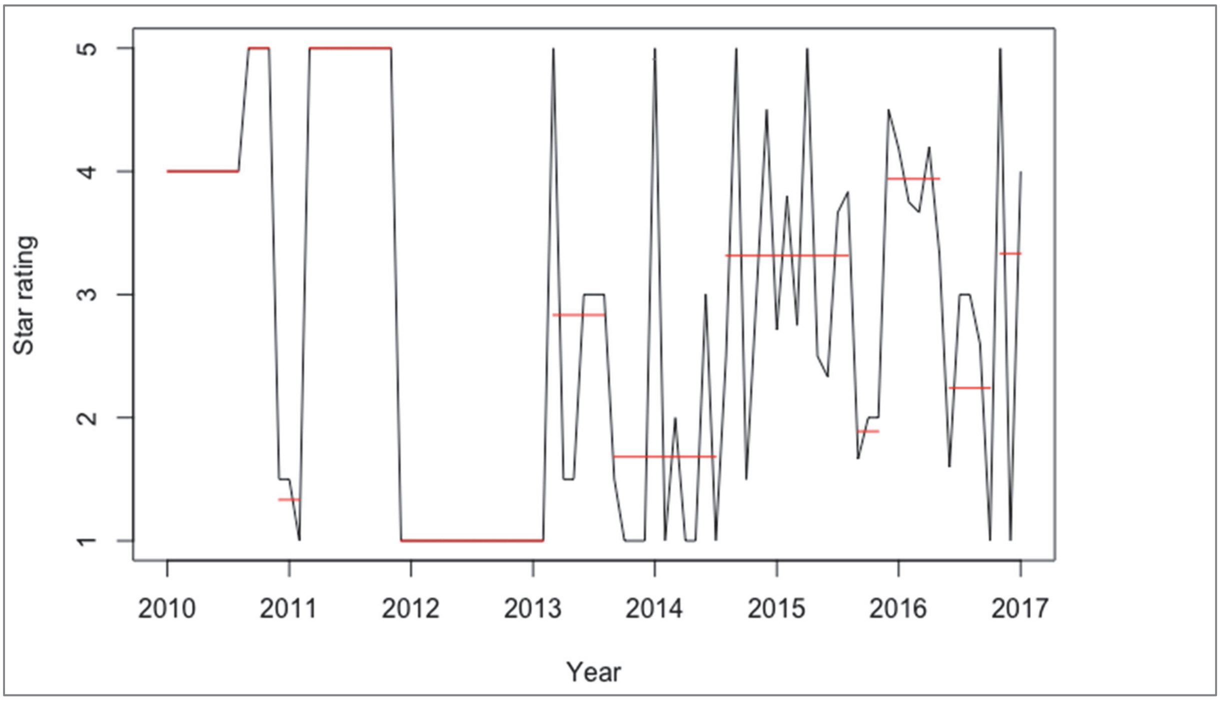

Figure 18 Change points of United Nissan's star rating

4.3 Sentiment Analysis

In this section sentiment analysis is performed on the selected businesses and the correlation of the sentiment score with the fluctuation in the star rating is investigated. It is expected that the fluctuation in the business star rating is also reflected in the review written by the same user. However, in the reality, this is not the case. The accuracy in the reviews written is specific to the customer's will. Sometimes the customer gives average rating and very nasty comments. In such cases we cannot expect a good correlation between the star rating and the sentiment score, which is based on the positive and negative polarity of the reviews available. These kind of uncertainties will be much reflected in cases where there are limited number of review texts.

The advantage of sentiment analysis is that by reviewing the polarity of the words and phrases for a particular business one can decide on the overall recommendation for the business instead of reading the whole review text. Most of the rating prediction algorithms are based on these sentiment analyses. For the present study, we are not using any prediction method. However, we are trying to correlate the sentiment score obtained for the relevant business with the star rating and the change points detected. Figure 19 shows the correlation between average sentiment score versus the business star ratings. There is a reasonable agreement on star rating and sentiment score. Higher the sentiment score indicates higher the star rating. This also indicates the suitability of using sentiment analysis with the present data. Figure 20 and 21 shows the word cloud for ten selected businesses, for five star rating and one star rating, respectively. The word cloud indicates the most frequent words in the review text.

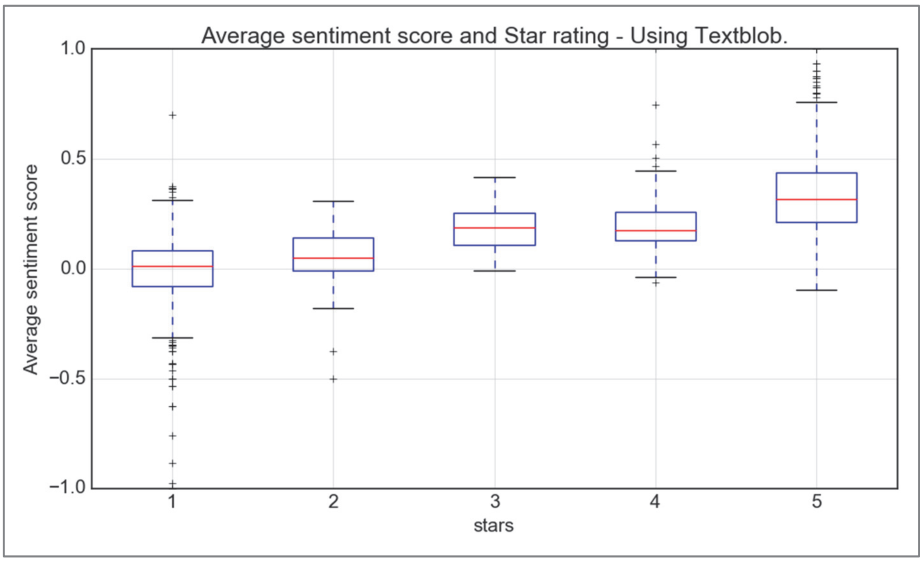

Figure 19 Average sentiment score and star rating for top five business with highest standard deviation and top five with lowest standard deviation in star ratings.

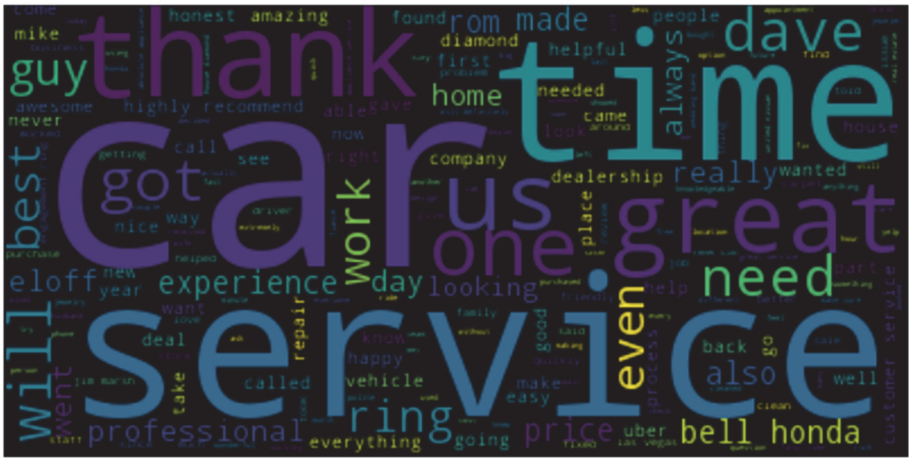

Figure 20 Word cloud for the relevant business reviews with star rating 5.

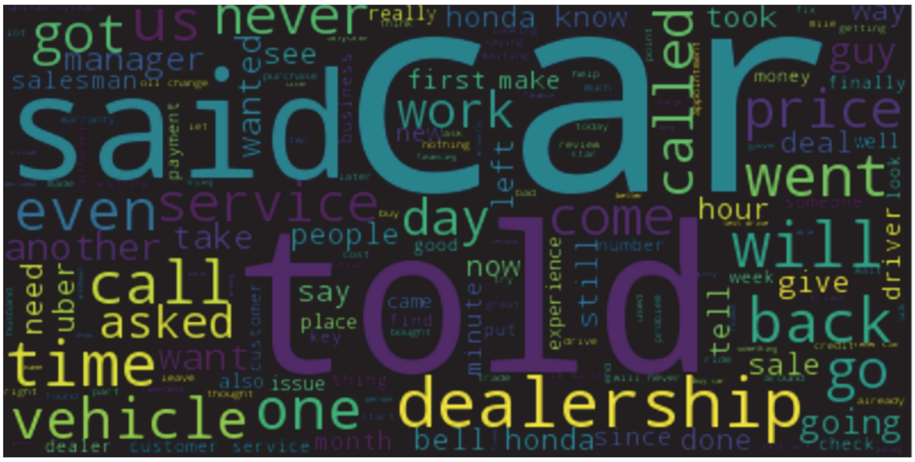

Figure 21 Word cloud for the relevant business reviews with star rating 5.

4.3.1 Business: Jim Marsh Kia

For Jim Marsh Kia business, the average sentiment score and business rating is shown in Figure 22. There is an overall agreement with star rating and sentiment score. However, some inconsistencies are observed for star rating 3. The sentiment score is higher for star rating 3 than 4. This may be reflected in the accuracy of the interpretation.

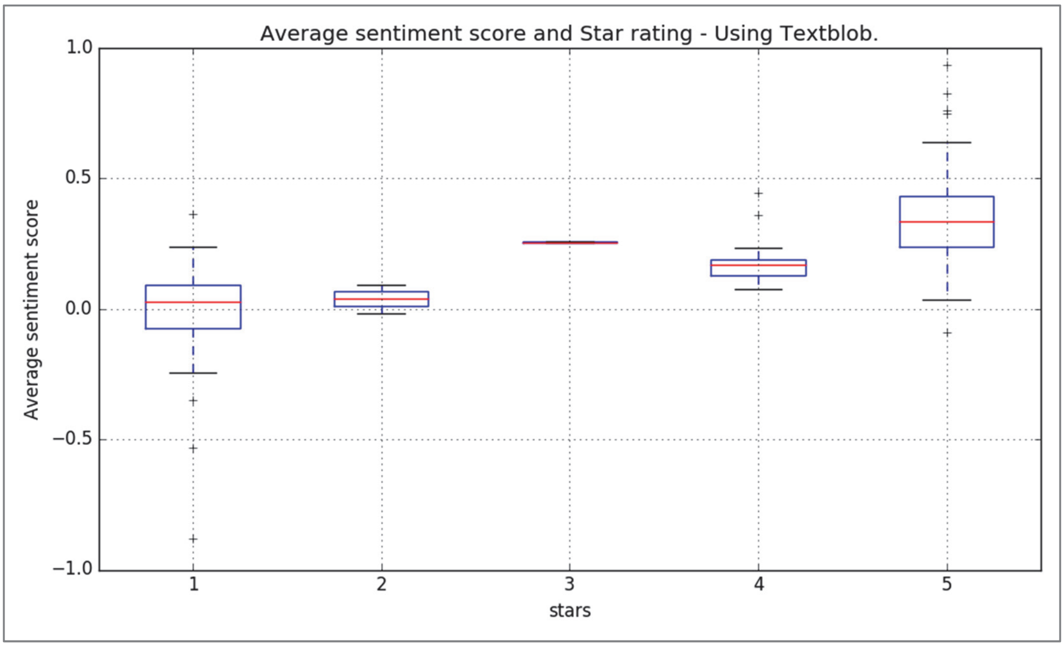

Figure 22 Average sentiment score and star rating for Jim Marsh Kia

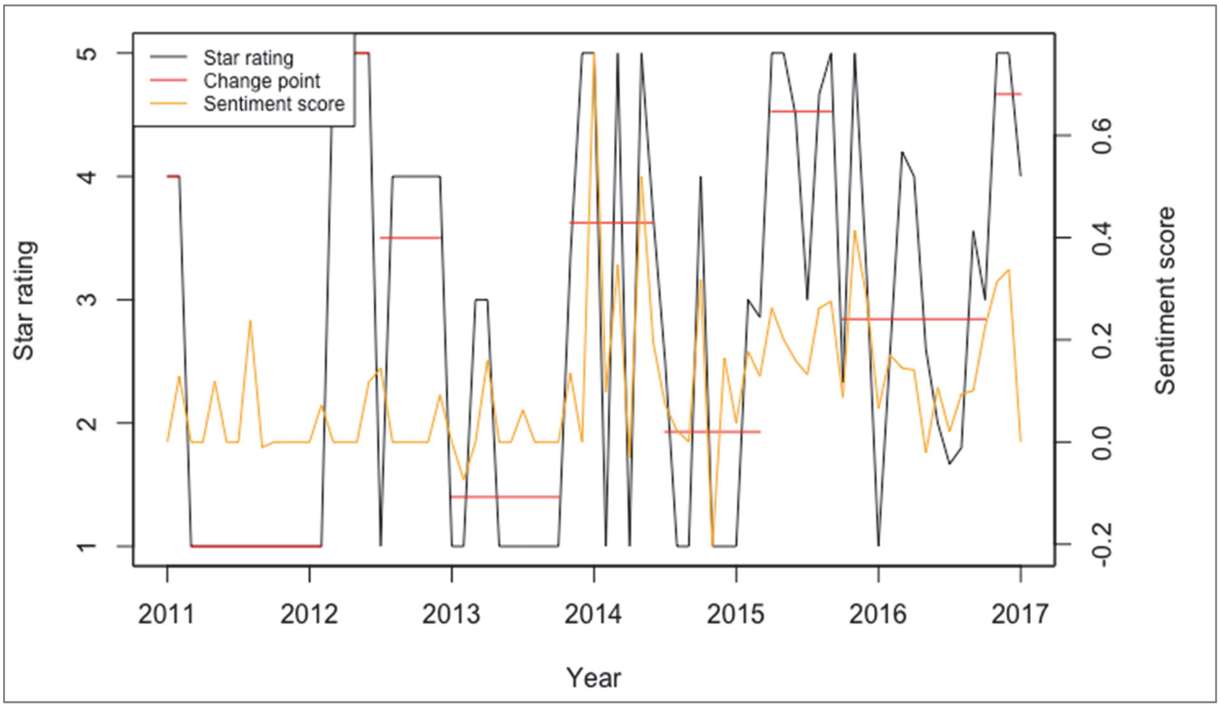

Figure 23 Sentiment score time series and star rating time series for Jim Marsh Kia along with the change points obtained.

Figure 23 shows the overlapping plot of sentiment score and the star rating over the period 2011 and 2017. In the change point analysis, we used data compensation to fill the months which doesn't have any star rating data. However, this method is not suitable for sentiment analysis as the reviews need to be copied to fill the missing data. Hence for the sentiment analysis, we have kept the missing information as it is. In such cases the sentiment score obtained will be neutral indicated by zero sentiment score.

It is interesting to check the correlation between sentiment score and change points detected. As the sentiment score is a reflection of the customer satisfaction, we can expect a fluctuation in the sentiment score at the location of change points. For instance, the star rating and review in between period Oct 2013 to March 2015 is considered. As per the change point detection analysis there is a change point at the June 2014. From Figure 22, it is clear that there is a high fluctuation in the sentiment score at the change point. The average sentiment score is higher in between the period Oct 2013 to June 2014 and lower for the period June 2014 to March 2015. In order to further look into this, the positive noun phrase in the first part and negative noun phrases in the second part of the period is taken as shown in Table 13 and 14. From the top 5 noun phrases during the period Oct 2013 to June 2014, it is clear that the car dealer company had a good impression among the customers in terms of price, service and transactions. These positive things are reflected in the higher sentiment score during this period. Conversely, a sudden decline in the business rating is observed during the period June 2014 to March 2015. As indicated by the noun phrases with lowest polarity, there were issues related to credit transactions, bad staff behavior, etc. which brought inconvenience to the customers. From the present study, it can be also concluded that the accuracy of the star rating and sentiment score is highly dependent on the number of review. More reviews indicate higher accuracy in the sentiment score. The less correlation of the star rating and sentiment score during the period 2011 to 2013 indicated the same.

Table 13 Positive noun phrases during the period Oct 2013 to June 2014

No	Positive Noun Phrases
1	Pleasant time
2	Good price
3	Great experience
4	Whole transaction effortless
5	Excellent job

Table 14 Negative noun phrases during the period June 2014 to March 2015

No	Negative Noun Phrases
1	Address error
2	Miserable job

3	Credit issues
4	Stupid female
5	Stupid tax

4.3.2 Business: Bell Honda

For the business Bell Honda, the overall correlation between the star ratings and the sentiment score is shown in Figure 24. There is a good agreement with the star rating and sentiment score in this case, especially for higher star ratings. However, for a star rating 3, there are some inconsistencies.

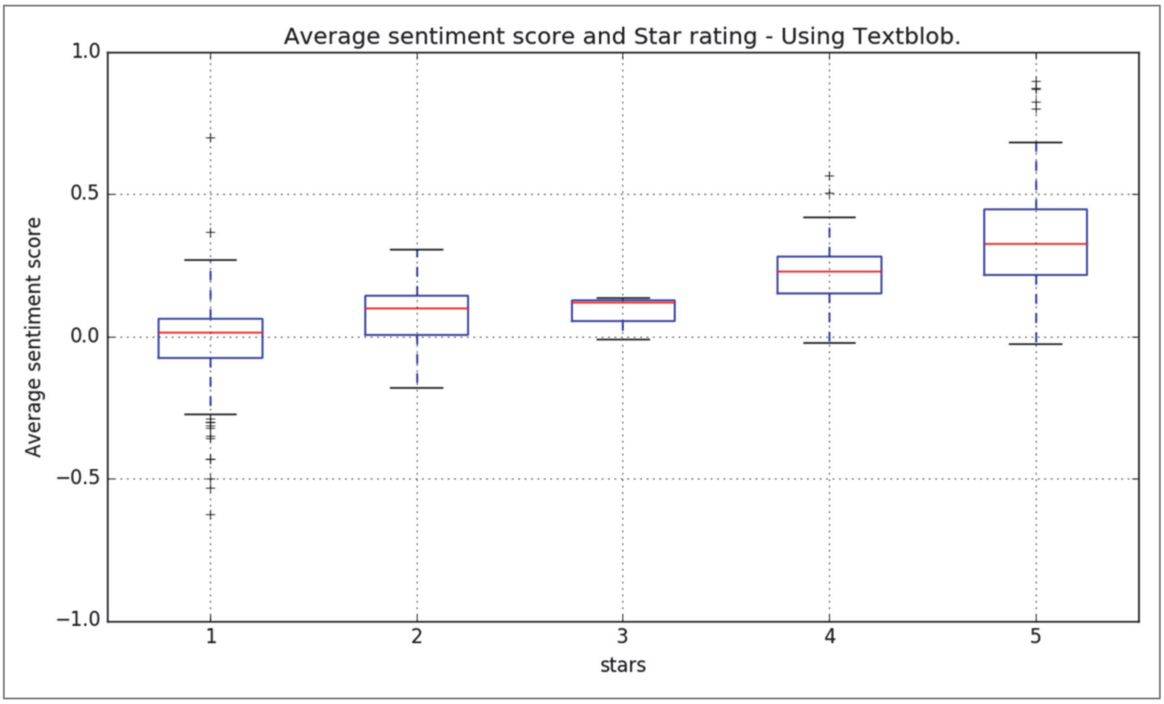

Figure 24 Average sentiment score and star rating for Bell Honda

The overlapping plot of star rating and sentiment score (Figure. 25) showed reasonable correlation. However, as mentioned for Jim Marsh Kia, the missing reviews in certain months are indicated by neutral sentiment score and proper conclusion based on the sentiment analysis cannot be drawn in such cases. For an instance, we have considered star ratings and reviews for Bell Honda during the period July 2014 and Dec 2016 is considered. According to the change point analysis, during this period change point is observed Dec 2015. If the average sentiment score is considered during the period Jul 2014 to Dec 2015 and Dec 2015 to Dec 2016, the former have lower score compared to the latter. This variation in the sentiment score is reflected in the star ratings and hence the estimated change point. Table 15 and 16 shows the negative noun phrases in the first part of the period. The negative noun phrases are the indicators for lower business ratings. However, for the second part, due the great service and sincere staffs it increases the business rating.

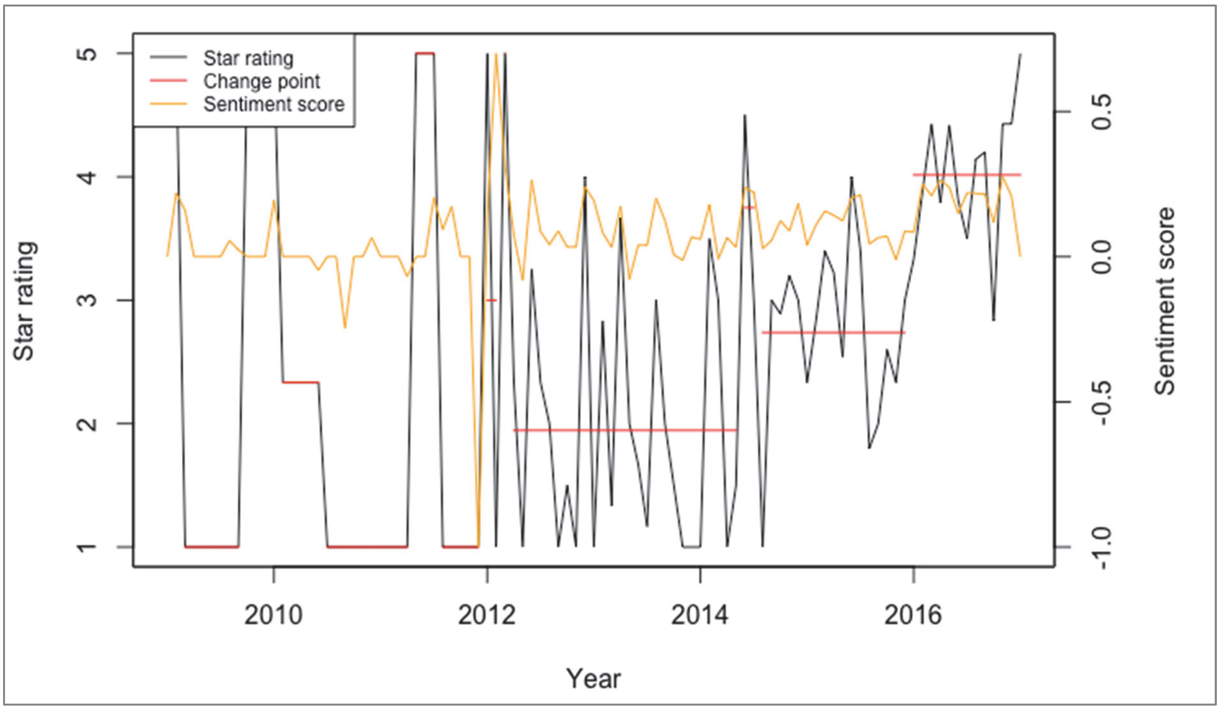

Figure 25 Sentiment score time series and star rating time series for Bell Honda along with the change points obtained

Table 15 Negative noun phrases during the period July 2014 to Dec 2015

No	Negative Noun Phrases
1	Expensive things
2	Safety issues
3	Bad situation
4	Bad feeling
5	Bad transmission

Table 16 Positive noun phrases during the period Dec 2015 to Dec 2016

No	Positive Noun Phrases
1	Amazing job
2	Super helpful
3	Sincere people
4	Great service
5	Good surveys

4.3.3 Business: Uber

Similar study is performed for Uber as shown in Figure 26 and 27. There is good correlation in the overall sentiment score and the business ratings. As shown in Figure 27, lot of fluctuations in the sentiment score is observed over time. All the fluctuations in the sentiment score does not indicate the change points. However, fluctuation in sentiment score is observed for the most of the estimated change points.

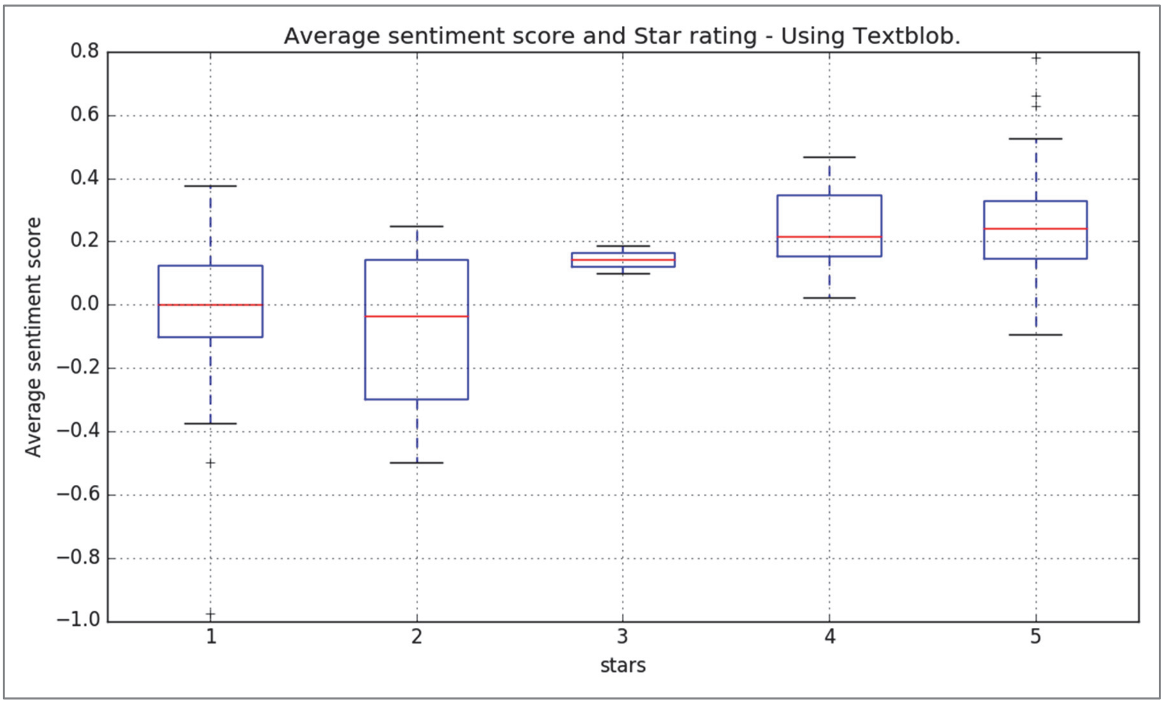

Figure 26 Average sentiment score and star rating for Uber

As a case, in for Uber the business star rating and reviews during the period Sep 2014 and Dec 2016 is considered. According to the change point analysis, during this period change point is observed Oct 2015. If the average sentiment score is considered during the period Sep 2014 to Oct 2015 and Oct 2015 to Dec 2016, the former has higher score compared to the latter. This variation in the sentiment score is reflected in the star ratings and hence the estimated change point. Table 17 and 18 shows the positive noun phrases in the first part of the period. The positive noun phrases are the indicators for higher business ratings. However, for the second part, due the bad experience faced by some customers in terms of the employees and poor service resulted in the dip in business rating.

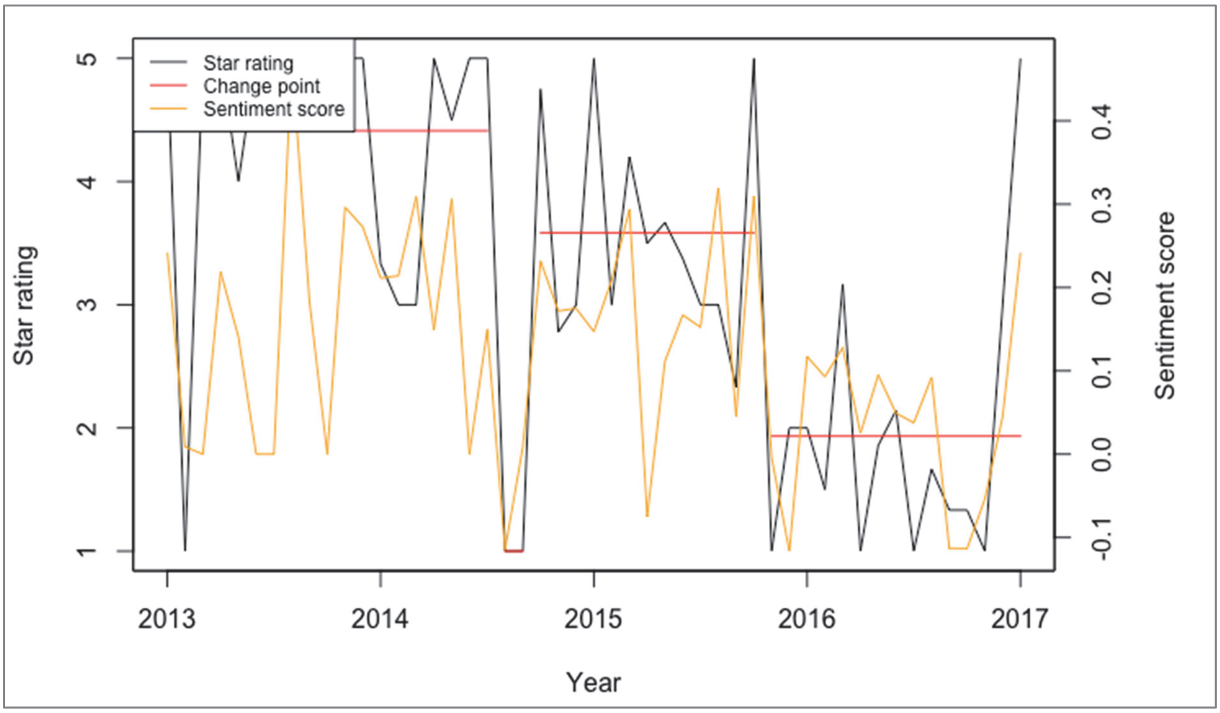

Figure 27 Sentiment score time series and star rating time series for Uber along with the change points obtained.

Table 17 Positive noun phrases during the period Sep 2014 to Oct 2015

No	**Positive Noun Phrases**
1	Great Service
2	Safe option
3	Friendly Drivers
4	Cheap
5	Great app

Table 18 Negative noun phrases during the period Oct 2015 to Dec 2016

No	**Negative Noun Phrases**
1	Poor Service
2	Bad experience
3	Cancelling
4	Drunk people
5	Unnecessary stress

4.3.4 Business: Henderson Chevrolet

The correlation between average sentiment score and star rating for Henderson Chevrolet is shown in Figure 28. Similar to other cases there is a reasonable agreement in sentiment score and star ratings.

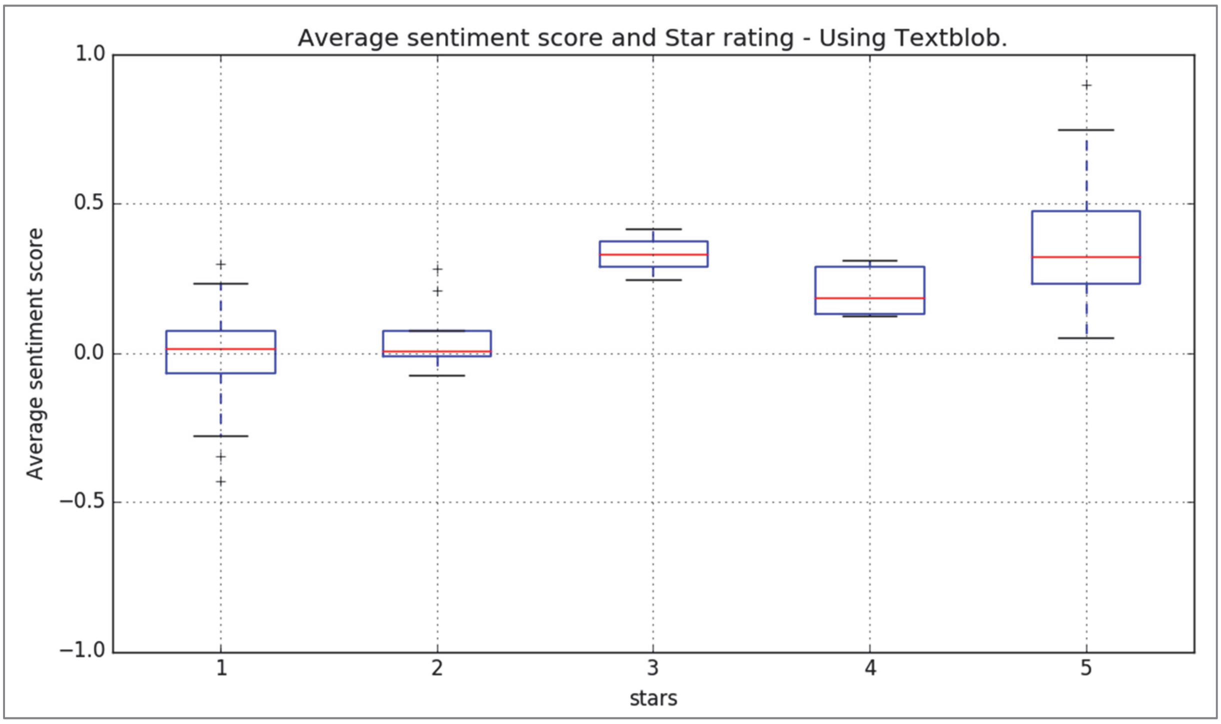

Figure 28 Average sentiment score and star rating for Henderson Chevrolet

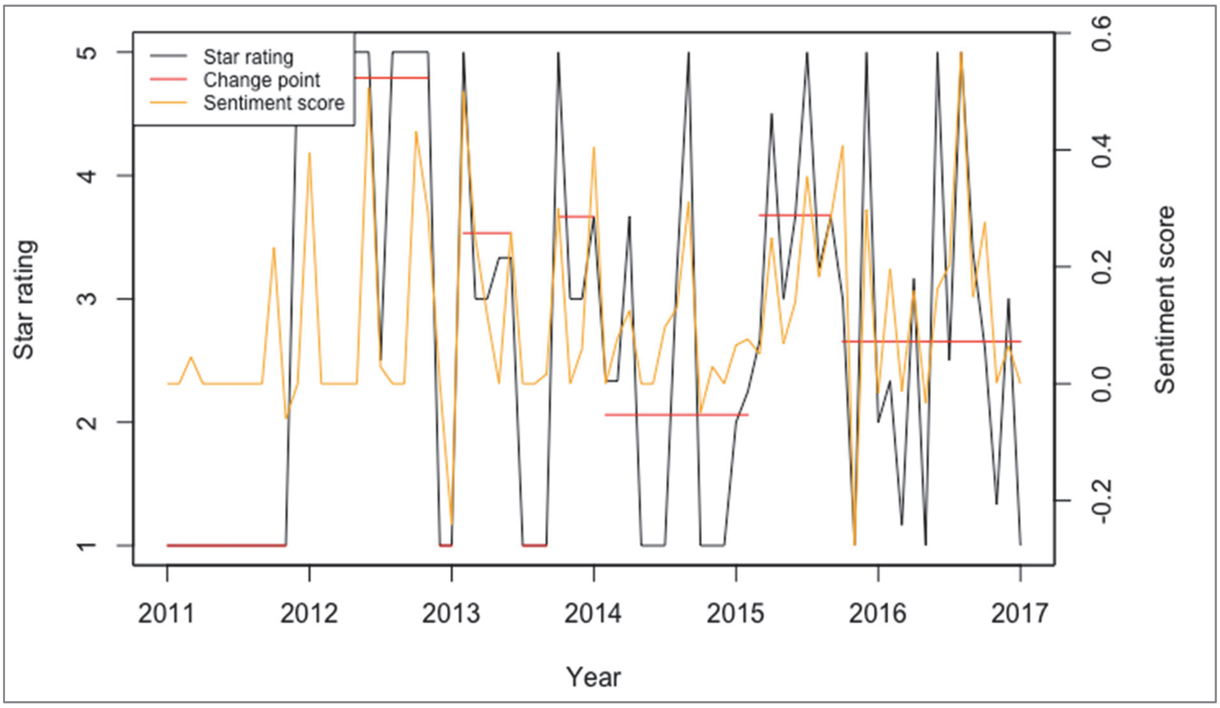

Figure 29 Sentiment score time series and star rating time series for Henderson Chevrolet along with the change points obtained

Figure 29 shows the correlation between star rating and sentiment score between the years 2011 and 2017. It can be seen from the graph that the obtained change points are in accordance with the fluctuations in the sentiment score. This affirms the accuracy of the change points obtained. For instance, we have chosen the star ratings and reviews between the period Jan 2014 and Sep 2015. Table 19 and 20 shows the negative noun phrases and Positive phrases during the first part and second part of the considered years, respectively. The employer excellence and performance of the business is reflected in the higher business ratings during the period Feb 2015 to Sep 2015. Conversely, the decline in the service quality and customer service resulted in a bad rating during the period Jan 2014 to Feb 2015.

Table 19 Negative noun phrases during the period Jan 2014 to Feb 2015

No	Negative Noun Phrases
1	Worst
2	Horrific customer service
3	Poor customer service
4	Small interrogation
5	Bad survey

Table 20 Positive noun phrases during the period Feb 2015 to Sep 2015

No	Positive Noun Phrases
1	Fair Price
2	Excellent service
3	Great sales person
4	Good reliable work
5	Fantastic Staff

4.3.5 Business: United Nissan

Figure 30 shows the overall sentiment score based on all the reviews for the United Nissan. There is a good correlation between sentiment score and star rating for this case as well. The time series of sentiment score for the business over the period 2010 to 2017. It can be observed from Figure 31 that there is a good correlation between change points obtained and the sentiment score. However, in certain cases this correlation is not observed.

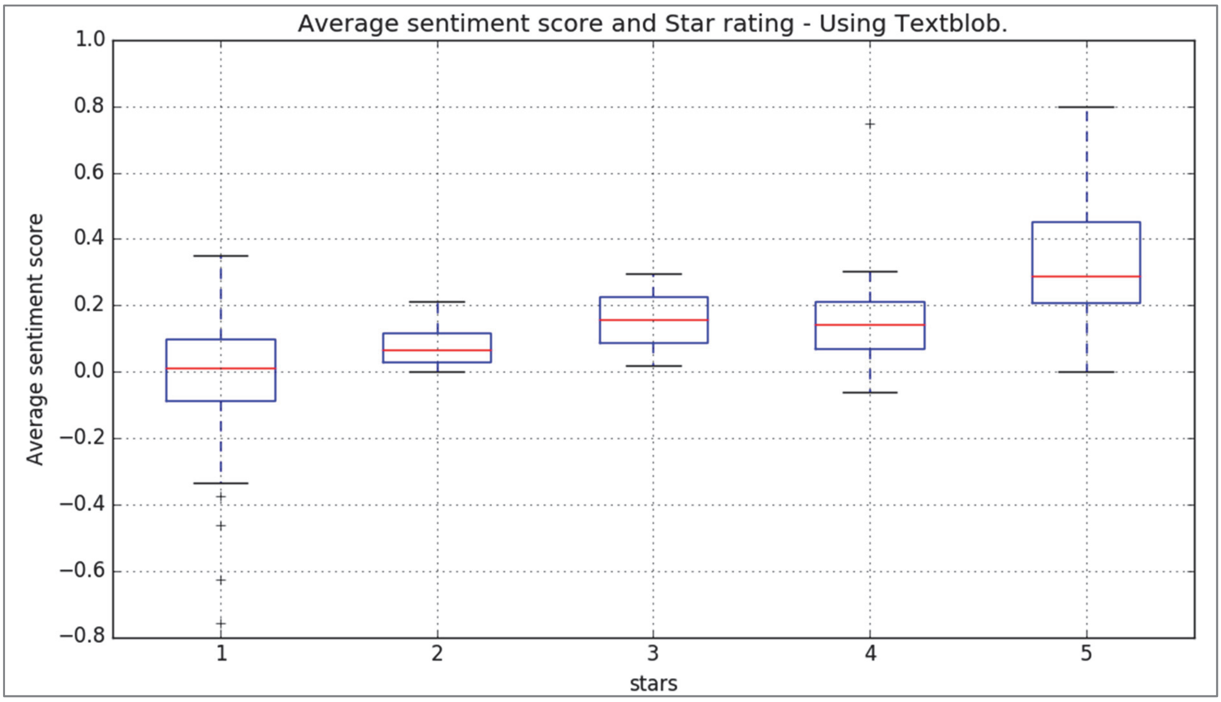

Figure 30 Average sentiment score and star rating for United Nissan

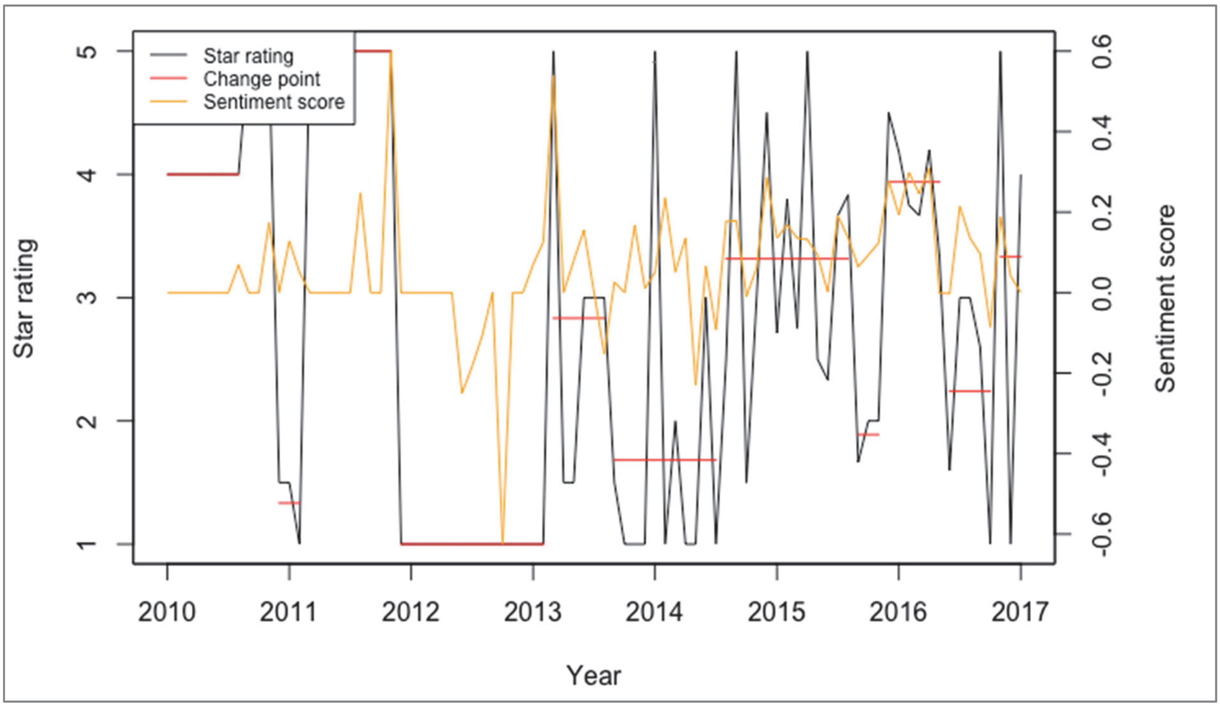

Figure 31 Sentiment score time series and star rating time series for United Nissan along with the change points obtained.

Table 21 Negative noun phrases during the period Feb 2011 to Nov 2011

No	Positive Noun Phrases
1	Excellent customer service
2	Amazing service

3	Great people
4	Great car
5	Exceptional service

Table 22 Positive noun phrases during the period Nov 2011 to Feb 2013

No	Negative Noun Phrases
1	Big trouble
2	Terrible place
3	Shady
4	Worst Nissan
5	Weak handshake

Table 21 and 22 shows the positive and negative noun phrases during the first part and second part of the considered years, respectively. In order to verify the correlation between obtained change point and sentiment score, the star ratings and reviews for the period Feb 2011 to Feb 2013 is considered. A higher rating is observed for the period Feb 2011 to Nov 2011, which is indicated in Figure 31 by a higher sentiment score. Moreover, the positive noun phrases shown in Table 16 shows the good performance of the business in terms of high quality and good service. However, the business rating suddenly dropped between the period Nov 2011 to Feb 2013, due to the bad performance of the business.

4.3.6 Business: Absolute Appliance Services

For all five businesses that having lowest standard deviation in their star rating, no change point is detected. This is because among five businesses, four having 5 star rating for all the reviews. Only for one business named Absolute Appliance Services having a very small deviation, 4.9. This value is because we are calculating the average star for each month. Figure 32 shows the overall analysis done on Absolute Appliance Services. From the figure it is clear that only a very small deviation is there in the entire monthly average and that change is not a relevant change in order to detect in change point analysis. The obtained sentiment score proves that only positive reviews are there for this particular business, as there is no score below zero.

Table 23 shows few positive noun phrases obtained during the entire period of Absolute Appliance Service. The table gives a clear picture about the customer's satisfaction with the particular business. Both price and the services given by this business is excellent.

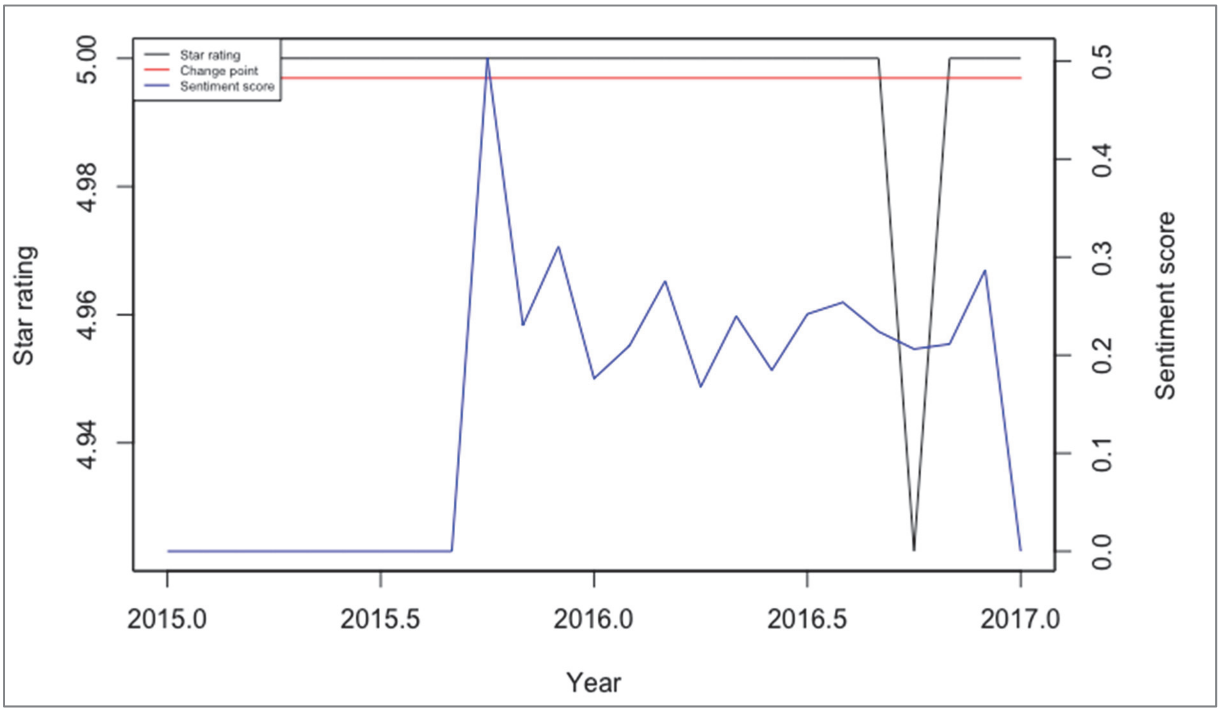

Figure 32 Sentiment score time series and star rating time series for Absolute Appliance Services is along with the change points obtained.

Table 23 Positive noun phrases during entire period

No	Positive Noun Phrases
1	Excellent response time
2	Great price
3	Great clean job
4	Fantastic job
5	Great technician

4.4 Summary

For the businesses with highest standard deviation in the business star ratings, a clear fluctuation in the ratings over the years is noted. The change points obtained using change point method indicate those fluctuating points in the time series data. The elbow point method used to obtain the optimum change points neglected most of the outlier points in the time series data. From the sentiment analysis performed on the business review text, it is clear that there is a direct correlation between sentiment score and the star ratings of the business. The average sentiment score of the business over the period was used to verify the change points obtained. There is a clear deviation in the sentiment score is observed whenever there is a change point. In order to further understand the relevance of the change point analysis and sentiment analysis, the noun phrases with highest and lowest polarity in the regions of change points were obtained and interpretations made regarding the performance of the business.

Conclusions

In this thesis, big data analytics is performed using Yelp dataset as a case study. The data is pre-processed and various data interpretation methods such as change point detection and sentiment analysis is implemented. Yelp is a social media platform where the user can rate a business based on his/her experience. For the present study, Yelp dataset is chosen as it is freely available and easy to interpret.

As the objective of the present study is to perform change point detection and sentiment analysis, the reviews and star rating in the Yelp dataset is used. The data is read from the dataset using Python programming. Before performing the analysis, some pre-processing and data compensation is done in order to make the data consistent for the further analysis. In the Yelp dataset there are large number of businesses. However, the relevant businesses for the present study is obtained based on some selection criteria. Only the businesses which have more than 100 number of reviews been considered for the present study. Among them, top five businesses which have highest and lowest standard deviation in star rating is chosen. Further study is performed on these businesses only.

Change point detection method is used to obtain the fluctuation points in a time series data. It is relevant for the present study in order to obtain the points where there is deviation in the business star ratings. Change point library in R language is used here to obtain the change points in the time series data. Sensitivity study on parameters involved in the change point algorithm is performed and optimum number of change points are obtained for the selected businesses. The change points obtained for each business designates period in which there is a deviation in the star rating of the business. However, it is not possible to make more interpretation of the data with change point analysis. Hence is recommended to perform further studies on the data.

Further in the study, the sentiment analysis is performed on the reviews recorded for these businesses. The businesses data used for the sentiment study is same as for change point analysis. The sentiment analysis returns the positive or negative polarity of the text data. In the present study, sentiment analysis is performed on the review text data for the selected businesses. Based on this, the average sentiment score is obtained for the business over a period of time. A very good correlation between the sentiment score and change point obtained is observed. Whenever there is a change point detected in the star rating time series, corresponding fluctuation in the sentiment score is observed. The noun phrases with most polarity score in the review text in those regions were obtained to interpret the reasons for fluctuations in the star rating of the business. This method worked for most of the cases. However, the sentiment analysis is not fully accurate. Sometimes the review text written may not reflect the actual star rating given for the business. In such cases, there is less correlation between sentiment score and the change points detected. It is also observed that the accuracy of sentiment analysis increases when there is large number of review text data. This is considered as one limitation of the present methodology.

Further work

Based on the present study it was observed that the star ratings and reviews written by different uses have different pattern. For example, a user who is very happy with the food served at the hotel and not satisfied by the service can give high star rating for the business and gave very bad reviews about the services. Hence those reviews can have high rating with low sentiment score. There can be instances were certain users always give low rating even for good businesses. These all are some setback of the present dataset. However, there are many methods to overcome these problems. One is to normalize the star rating based on the pattern of ratings given by different users. Another method is to cross verify the ratings for the businesses based on prediction algorithms.

References

[1] M. Fan and M. Khademi. “Predicting a business star in yelp from its reviews text alone,” arXiv:1401.0864, 2014.

[2] K. Carbon, K. Fujii, and P. Veerina. “Applications of Machine Learning to Predict Yelp Ratings,” Stanford Univ., Stanford, CA, 2014.

[3] J. Jong. “Predicting Rating with Sentiment Analysis,” Stanford Univ., Stanford, CA, 2011.

[4] C. Li and J. Zhang. “Prediction of Yelp Review Star Rating using Sentiment Analysis,” Stanford Univ., Stanford, CA, 2014.

[5] WA. Taylor. "Change-Point Analysis: A Powerful New Tool For Detecting Changes," 2000. WEB: http://www.variation.com/cpa/tech/changepoint.html.

[6] R. Killick, I. Eckley. “Changepoint: An R package for changepoint analysis. R package version 0.6.1,” 2012

[7] Z. Xu, TA. Kass-Hout, C. Anderson-Smits and G. Gray, “Signal detection using change point analysis in postmarket surveillance,” Pharmacoepidemiol Drug Saf, 24, 663–668. doi: 10.1002/pds.3783, 2015.

[8] TA. Kass-Hout, Z. Xu, P. McMurray. “Application of change point analysis to daily influenza-like illness emergency department visits,” Journal of the American Medical Informatics Association : JAMIA. 2012;19(6):1075-1081. doi:10.1136/amiajnl-2011-000793, 2011

[9] Python use guide. Can be retrieved from https://www.python.org/doc/essays/blurb/

[10] Pandas for python guide. Can be retrieved from http://pandas.pydata.org/pandas-docs/stable/

[11] R language user guide. Can be retrieved from https://www.r-project.org/about.html

[12] LL. Cavalli-Sforza, AWF. Edwards. “Phylogenetic analysis. Models and estimation procedures,” American Journal of Human Genetics, 19(3 Pt 1):233-257, 1967.

[13] IE. Auger and CE. Lawrence. “Bltn Mathcal Biology”, 51: 39. doi:10.1007/BF02458835, 1989.

[14] J. Bai, and P. Pierre. "Estimating and Testing Linear Models with Multiple Structural Changes." Econometrica 66, no. 1 : 47-78. doi:10.2307/2998540,1998.

Appendix 1

Python code :

# Reading data from json file

```
def load_data(filepath):
    d = []
    with open(filepath) as file:
        d = pd.DataFrame.from_dict(json.loads(line.rstrip()) for line in
file)
    return d

print("reading data from json")
review_df=load_data('/Yelp/yelp_academic_dataset_review.json')
```

# Pre-processing

```
all_business_id=[]
review_df=review_df.sort_values(by=['business_id', 'date'],
ascending=[True, True])
all_business_id=review_df.business_id.unique()
review_df['votes']=review_df['funny']+review_df['cool']+review_df['useful']
review_df=review_df.drop([col for col in ['funny','useful','cool'] if col
                                                              in review_df],
axis=1)
counter = Counter(review_df['business_id'])
columns = ['b_id','count','std_d']
df = pd.DataFrame(columns=columns)
df['b_id']=counter.keys()
df['count']=counter.values()
df=df.sort_values(by=['count'], ascending=[False])
df.reset_index(drop=True, inplace=True)
```

# Histogram

```
plt.hist(df['count'], bins=[0,50,100,200,300,400])
plt.title("Review counts Histogram")
plt.ylabel("No:of business id's")
plt.xlabel("No:of reviews")
fig = plt.gcf()
```

# Finding standard deviation of each business id

```
columns = ['b_id','std_d']
df3 = pd.DataFrame(columns=columns)
print("finding sd of ids with review count more than #")
for k in range(0,7934):
    test=review_df.loc[review_df['business_id'].isin(df.iloc[k])]
    l = len(test.stars)
    if l <= 1 :
              sd = 0
    else:
              sd=np.std(test.stars)
    df3.loc[k] = [df.iloc[k].b_id, sd]
```

# First 5 businesses according to Standard Deviation

```
df3=df3.sort_values(by=['std_d'], ascending=[False])
df3.reset_index(drop=True, inplace=True)
```

# Lowest 5 businesses according to Standard Deviation

```
df3=df3.sort_values(by=['std_d'], ascending=[True])
df3.reset_index(drop=True, inplace=True)
```

# Finding monthly average star rating for each business

```
columns = ['b_id','date','avg_star','name', 'text']
df = pd.DataFrame(columns=columns)
df[['date']] = df[['date']].astype(int)
c1=np.arange(1,7)
c2=np.arange(1,13)
k=0
for i in  all_business_id:
    name= business_df[business_df.business_id == i].name.item()
    for j in newdf[(newdf.business_id == i)].year.unique():
      for h in c2:
        avg_star=np.average(newdf[(newdf.business_id == i) & (newdf.year ==
j)&
                 (newdf.month ==h)].stars)
        text= newdf[(newdf.business_id == i) & (newdf.year == j) &
(newdf.month ==
              h)].text.values
        df.loc[k] = [i,j*100+h,avg_star, name,text]
        k =k+1
df['date'] = df['date'].apply(lambda x: pd.to_datetime(str(x),
format='%Y%m'))
df=df[df['date'] < '2017-01-01']
df.to_csv('/Yelp/r_100.csv')
```

# Generating Word Cloud for reviews with rating stars 1 /5

```
words=" ".join(df[(df.stars== #1 or 5)&(df.year<2017)].text.str.lower())
wordcloud = WordCloud(stopwords=STOPWORDS,).generate(words)
plt.imshow(wordcloud)
plt.axis('off')
plt.show()
```

# Sentiment Analysis

```
mask = newdf.business_id== #business_id
FilteredDf = newdf[['stars', 'text']].loc[mask]
FilteredDf.count()
```

```
tb_value = []
stars_tb = []
for ind,review in islice(FilteredDf.iterrows(),#no:of rows):
    details = TextBlob(review['text'])
    tb_value.append(details.sentiment.polarity)
    stars_tb.append(review['stars'])
p_stars_tb = pd.DataFrame()
p_stars_tb['stars'] = stars_tb
p_stars_tb['senti_value'] = tb_value
ax = p_stars_tb.boxplot(by=['stars'], figsize=(10,6))
ax.get_figure().suptitle("")
ax.set_title('Average sentiment score and Star rating - Using Textblob.')
ax.set_xlabel('stars')
ax.set_ylabel('Average sentiment score')
```

# Sentiment analysis: finding positive or negative noun_phrases

```
if details.sentiment.polarity < 0:
    pos.append(review['text'].replace('\n', ' '))
    for item in details.noun_phrases:
        print (item,details.sentiment.polarity )
```

R programming :

# Function for finding Elbow point

```
elbowpoints <- function(x)
  {
  dvec = c()
  # Normalize the vector
  x = x/max(x)
  L = length(x)
  # The distance of the endpoints
  dmax = dist(rbind(c(1/L, x[1]), c(1, x[L])), method="euclidean")
  # Find the point with maximum distance (minimum angle)
  for (i in 1:L)
    {
    d1 = dist(rbind(c(1/L, x[1]), c(i/L, x[i])),method="euclidean")
    d2 = dist(rbind(c(i/L, x[i]), c(1, x[L])), method="euclidean")
    #if (d1 == 0 | d2 == 0) {next}      # Avoid singularities
    dvec = c(dvec, abs((d1^2 + d2^2 - dmax^2)/(2*d1*d2)))
    }
  return (order(dvec)[1:L])
}
```

# Change point detection

```
library(changepoint)
MyData <- read.csv(file="/Yelp/r_100.csv", header=TRUE, sep=",")
MyData <- data.frame(MyData)
pdata <- MyData[MyData$b_id == '#business id', ]
library(zoo)
pdata <-na.locf(pdata, fromLast = TRUE)
pdata
date1=strtoi(substring(min(pdata$date),1,4))
date2=strtoi(substring(max(pdata$date),1,4))
```

```
pdata$avg_star <- as.numeric(pdata$avg_star)
stars.ts = ts(pdata$avg_star, frequency=12, start=c(date1), end=c(date2+1))
plot(stars.ts, ylab="Star rating",xlab="Year")
x <- c()
j<- c(0.01,0.05,0.1,0.5,1,1.5,2,2.5,3,3.5,4,4.5,5,5.5,6)
q=length(stars.ts)/2+1
for (i in j)
{

  mstar = cpt.mean(stars.ts,penalty="Manual",Q=q,
      pen.value=i,method="BinSeg",test.stat="Normal",class=TRUE,
      param.estimates=TRUE)
  x <- c(x, length(cpts(mstar)))
  print (x)

}
y <- elbowpoints(x)
m <-match(c(1),y)
mstar = cpt.mean(stars.ts,penalty="Manual",Q=q,
    pen.value=2,method="BinSeg",test.stat="Normal",class=TRUE,
    param.estimates=TRUE)
cpts(mstar)
pdata$date[cpts(mstar)]
x  <- pdata$date
y2<- read.csv(file="# read file where sentiment score is saved”)
score.ts = ts(y2$senti_value, frequency=12, start=c(date1), end=c(date2+1))
```

# Plotting change point and sentimental sore in time seriesof star rating

```
par(mar=c(5,4,4,5)+.1)
plot(mstar,type="l",ylab="",xlab="Year")
mtext("Star rating",side=2,line=3)
par(new=TRUE)
plot(score.ts,type="l", col="blue",,xaxt="n",yaxt="n",xlab="",ylab="")
axis(4)
mtext("Sentiment score",side=4,line=3)
legend("topleft",col=c("black","red","blue"),lty=1,legend=c("Star
rating","Change point","Sentiment score"),cex =.5)
```