YELP DATASET CHALLENGE

ANALYZING SEASONAL TRENDS

TEAM:
AKSHAY SRIRANGAM(axs156630)
NEHA NIRMALA SRINIVAS(nxn150930)
VENKATA SAI RAGHURAM G. (vxg150030)

Table of Contents

1. DESIGN

a) Problem Definition

Seasonal Trends is a characteristic of a time series in which the data experiences regular and predictable changes which recur every calendar year. Any predictable change or pattern in a time series that recurs or repeats over a one-year period can be said to be seasonal.

For example, in a climate with cold winters and warm summers, a home's heating costs probably rises in the winter and fall in the summer. This seasonality of heating costs would recur every year. Similarly, a company that sells sunscreen and tanning products would see sales jump up in the summer, but drop in the winter.

Seasonal Trend for the provided dataset is to be determined. In doing so, we can predict which businesses fare well during which seasons. From an investors point of view we can predict which business it would be profitable to invest in. Also, considering the number of people coming to an establishment at different periods of times, this would help in addressing staffing concerns. Overall by analyzing the trend we can increases the maximum profitability of a business.

b) Description of Input Data

The input data was obtained from https://www.yelp.com/dataset_challenge. This gave rise to 2GB JSON content with 5 JSON format files with individual files relevant to businesses, reviews, check-ins and tip
The individual files have contents corresponding to key value pairs (JSON script).
Number of instances in corresponding 4 datasets in use turned out to be above 70500 each with 6-7 features distributed across the files.
Size of each data set happens to be above 100MB with an average of 120MB and the maximum is 1.8GB.

The files that are of relevance are business, reviews and check-in. The following is the structure of these files:

Business:

```
{
   'type': 'business',
   'business_id': (encrypted business id),
   'name': (business name),
   'neighborhoods': [(hood names)],
   'full_address': (localized address),
   'city': (city),
   'state': (state),
   'latitude': latitude,
   'longitude': longitude,
   'stars': (star rating, rounded to half-stars),
   'review_count': review count,
   'categories': [(localized category names)]
   'open': True / False (corresponds to closed, not business hours),
   'hours': {
       (day_of_week): {
           'open': (HH:MM),
           'close': (HH:MM)
       },
       ...
   },
   'attributes': {
       (attribute_name): (attribute_value),
       ...
   },
}
```

Review:

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

Checkin:

```
{
    'type': 'checkin',
    'business_id': (encrypted business id),
    'checkin_info': {
        '0-0': (number of checkins from 00:00 to 01:00 on all Sundays),
        '1-0': (number of checkins from 01:00 to 02:00 on all Sundays),
        ...
        '14-4': (number of checkins from 14:00 to 15:00 on all Thursdays),
        ...
        '23-6': (number of checkins from 23:00 to 00:00 on all Saturdays)
    }, # if there was no checkin for a hour-day block it will not be in the dict
}
```

The fields that are of major interest would be the business id, city, stars and categories from the business table, the date and stars from the review table and all the info in the check-in table.

c) Algorithm for Analysis of Data

In order to determine the seasonal trends for various business or for business types we categorize the seasons as follows:

- Spring : March-May
- Summer : June - August
- Autumn : September - November
- Winter : December - February

Load the business, review and checkin json files into spark and convert them into spark sql tables using spark sql.

➢ To calculate future rating of business:

Select the business_id, review stars, year of review date for each season spring, summer, autumn and winter from the business and review tables.

for each season
select the business_id and the average star rating for each year from 2009 to 2015.

Normalize the tables to have a yearly average rating of 2.5 if there is no average rating information available.

These tables will be used in future as input to do predictive analysis using ARIMA forecasting technique.

➢ Seasonal trends based on count of number of reviews for business:

Select the business_id , review stars , year of review date for each season spring ,summer ,autumn and winter from the business and review tables.

for each season
select the business_id and the count of number of star ratings received for each year from 2009 to 2015.

This table will be used to find how different businesses fare in different seasons.

➢ Seasonal trends based on count of number of reviews per city:

Select the business_id, city, review stars , year of review date for each season spring ,summer ,autumn and winter from the business and review tables.

for each season
Select the business_id,city and the count of number of star ratings received for each year from 2009 to 2015.

This table will be used to find how different businesses in a particular city fare in different seasons.

➢ Analysis of checkin information:

In order to analyze the trend people following while visiting different types of establishments during a week, we categorize the week as follows:
weekdays : Monday - Thursday
weekends : Friday - Sunday
morning : 8am - 11am
noon : 12 pm -3pm
eve : 4pm- 7pm
night : 8pm -11pm

From the checkin table segregate information based on the above classification.
This table will be used to analyze most profitable hours for an establishment or type of establishment.

d) Big Data Strategy

Traditionally for implementing big data solutions few of these resources which are handy are:

- UTD cluster
- VM's like Cloudera, Horton sandbox, Map.

But these are limited in terms of resource and computational power. In fact, these are pseudo distributed system where in there is only one master and one slave. So it truly doesn't replicated the full features of a distributed system for improved computational and storage performance.
So we decided on creating our own cluster. This consists of connecting multiples systems as nodes and then creating an architecture so as to deploy a disturbed system.

We created a cluster with 1 master and 3 nodes, all running on Linux machines. We built these machines from starch, from installing Linux into them to installing spark.

So basically we created a dedicated distributed system around spark which when tested gives at most performance when compared to the UTD cluster or Cloudera VM image.

So we used our own cluster for data analysis.

In order to analyze the dataset the following techniques were considered :
PIG, HIVE and Spark.

The input to be analyzed are JSON files of different sizes. JSON files are considered as structured input. PIG is a tool used to take highly unstructured data and then convert it into a meaningful form. It is an abstraction layer on top of map reduce and is workflow driven. As the data was structured there was no strong intent to use PIG.
HIVE is also an abstraction layer on top of map reduce but requires the data to be structured.

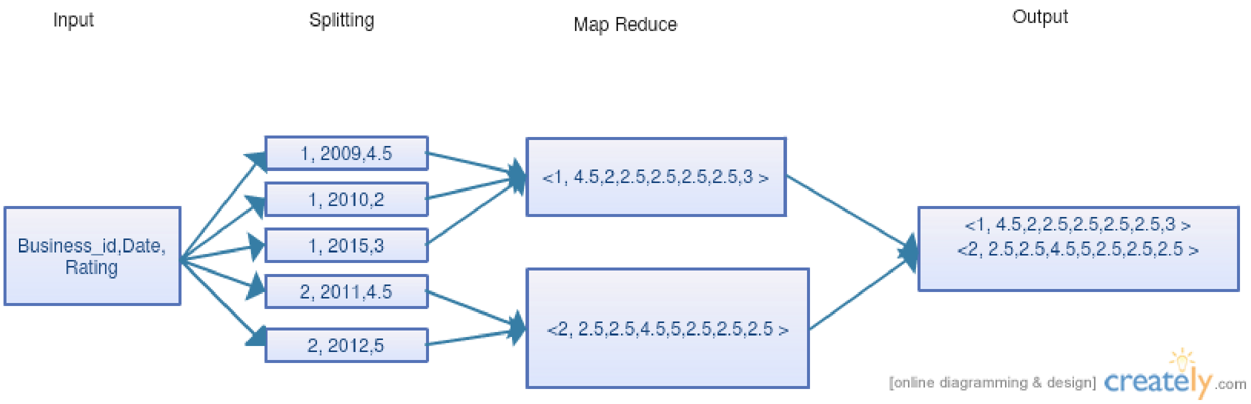
SPARK is a successor to map reduce and performs in memory computing. Spark also allows parsing of data. Spark SQL allows SQL actions to be performed with 100x faster performance than map reduce.

Considering the above analysis of spark hive and pig, in order to analyze the dataset spark was chosen considering its computational ability and its ability to parse data.

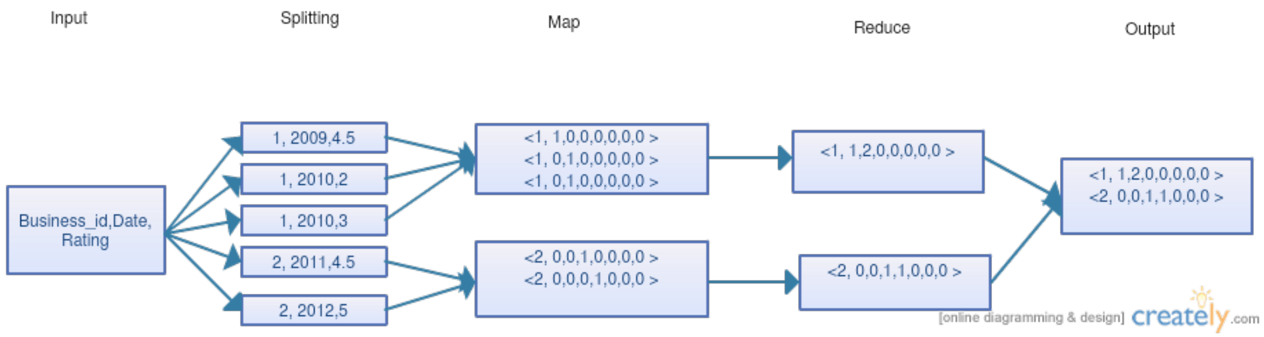
In addition to using SPARK and SPARK SQL to analyze the data, R was used to do perform predictive analysis on the data to predict future star ratings. R was also used to graphically interpret the data and to plot seasonal trends.

e) Data Flow Diagram

The data flow diagram for the generation of input for the predictive analysis is as follows :



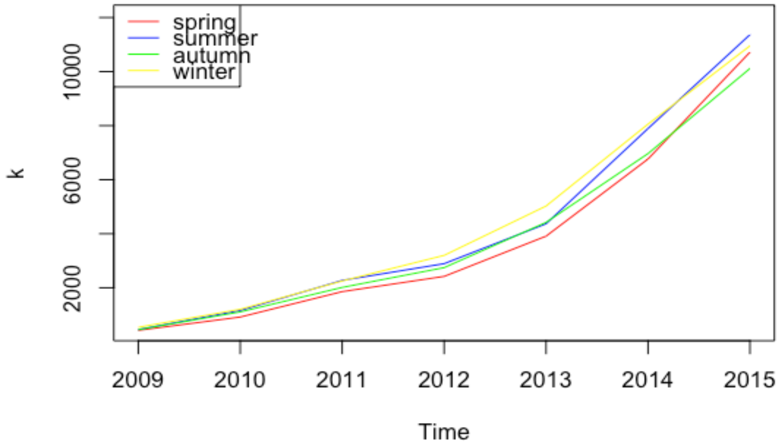
The data flow diagram for the generation of count of number of reviews per business throughout the years is as follows:



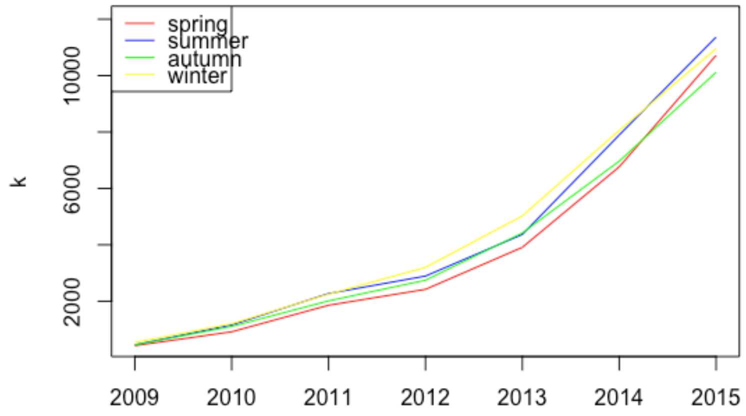
2. ANALYSIS OF RESULTS

From the analysis of the dataset some of the major categories present we as follows: Restaurants, shopping, food, beauty and spa, nightlife, health and medical, home services:

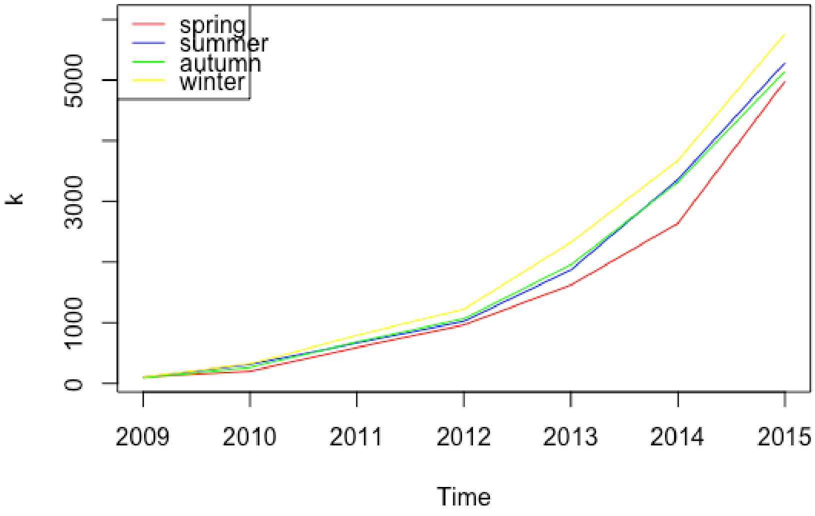
Analysis of which business trend during which season was attempted based on the input data:



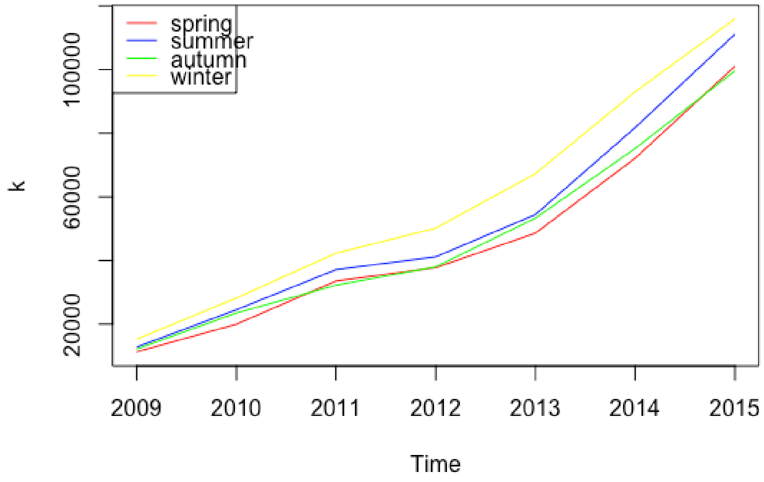
The plot indicates the number of reviews received for establishments related to beauty and spa in different seasons from 2009 to 2015. From the plot we see that beauty and spa establishments see on an average lesser number of reviews are received during autumn when compared to the other seasons.



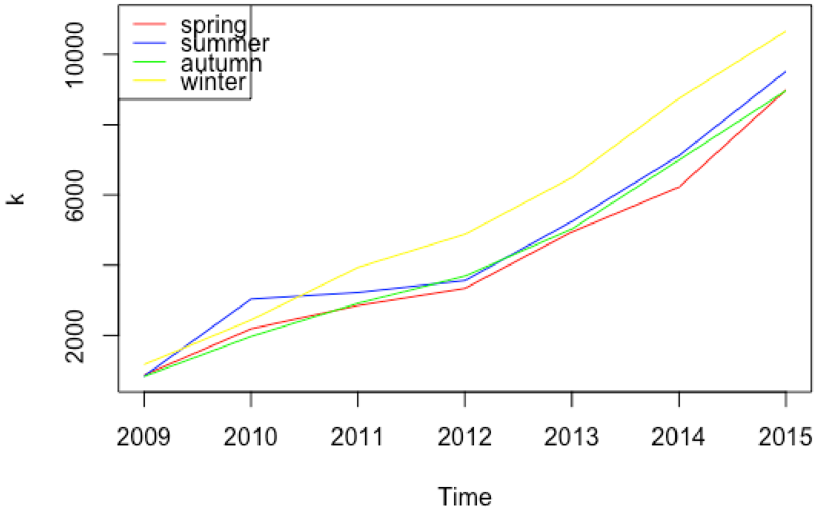
The plot indicates the number of reviews received for establishments related to food in different seasons from 2009 to 2015. From the plot we see that food establishments see on an average greater number of reviews are received during winter



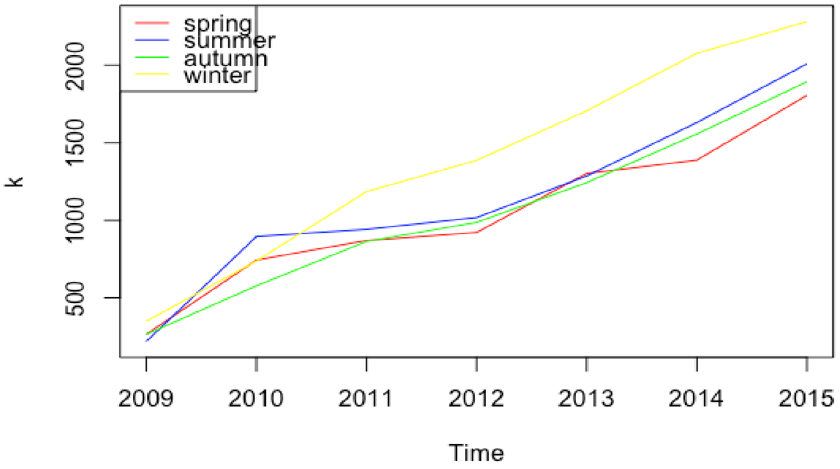
The plot indicates the number of reviews received for health and medical establishments in different seasons from 2009-2015. From the plot it is seen that there are less number of reviews in spring and greater number of reviews in winter. This could be due to good weather conditions in spring and cold climate in winter.



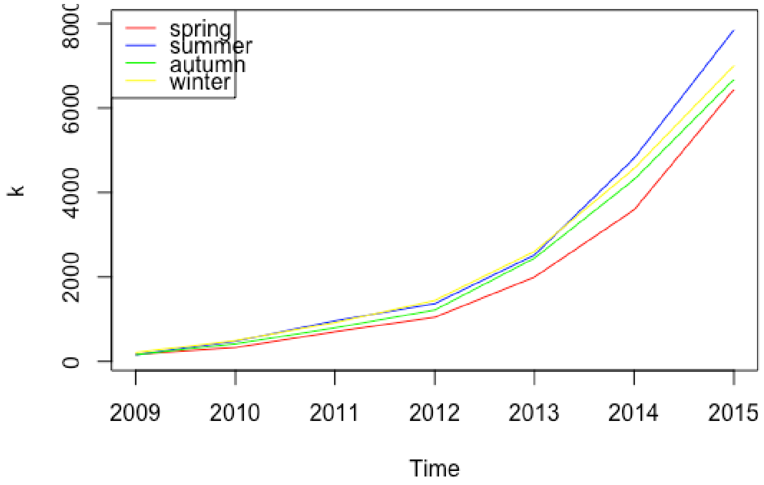
The plot indicates the number of reviews received for restaurants in different seasons from 2009-2015. From the plot it is clear that a greater number of restaurants are reviewed during winter. We could draw a conclusion that it could be related to the holiday season.



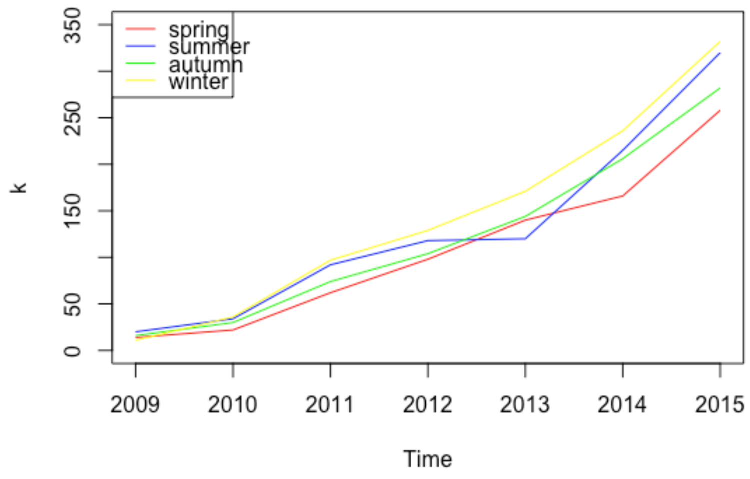
The plot indicates the number of reviews received for shopping establishments in different seasons from 2009-2015. From the plot it is clear that a greater number of customers reviewed shops during winter. We could draw a conclusion that it could be related to the holiday season.



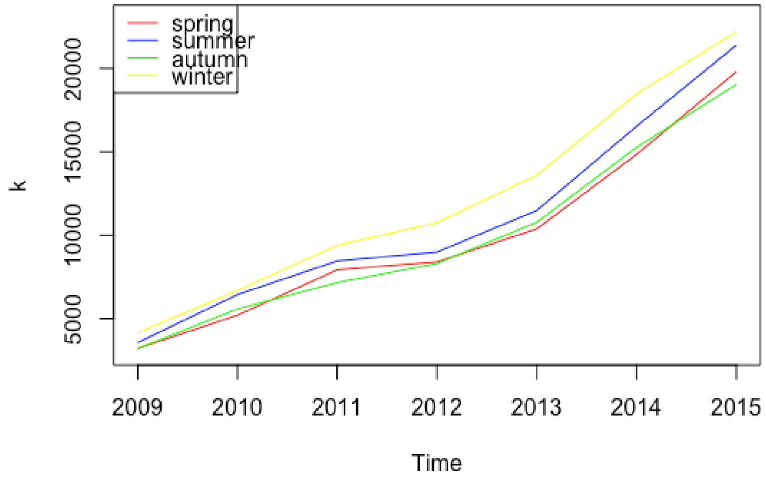
The plot indicates the number of reviews received for fashion establishments in different seasons from 2009-2015. From the plot it is clear that a greater number of fashion establishments are reviewed during winter. This could be due to new trends for the next year



The plot indicates the number of reviews received for home services in different seasons from 2009-2015. From the plot it is clear that a greater number of home services are reviewed during summer. This could be due to the weather conditions favorable to renovations



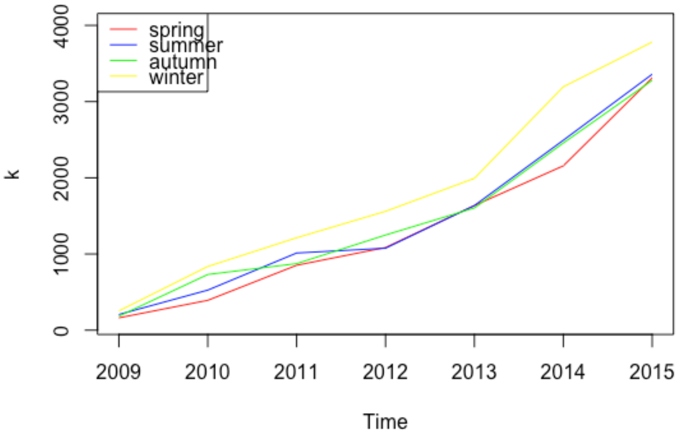
The plot indicates the number of reviews received for nightlife in different seasons from 2009-2015. From the plot it is clear that a greater number of nightlife establishments are reviewed during winter. This could be related to the holidays.

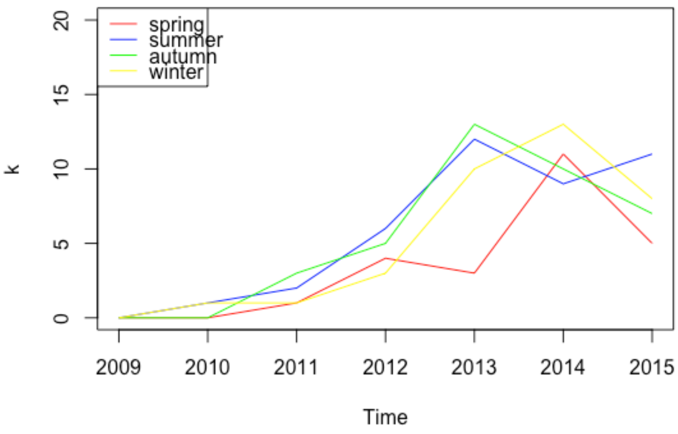


The plot indicates the number of reviews received for schools in different seasons from 2009-2015. Schools receive more reviews during winter followed by summer. This could be in relation to the opening and closing of the school year.

Shopping Trends in different cities in different seasons:

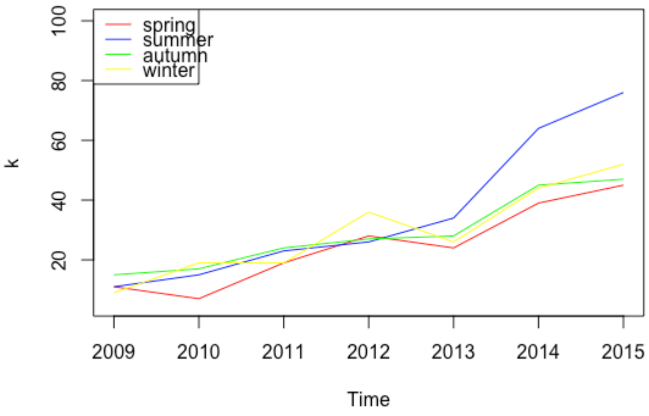
The below plots show the shopping trend in two different cities Las Vegas and Montreal. From the plots it is observed that in Las Vegas people tend to shop more during the winter, where as in Montreal on an average people tend to shop more in the autumn.

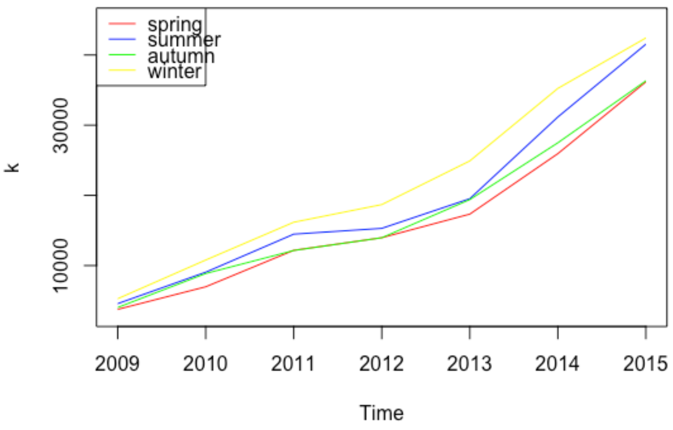


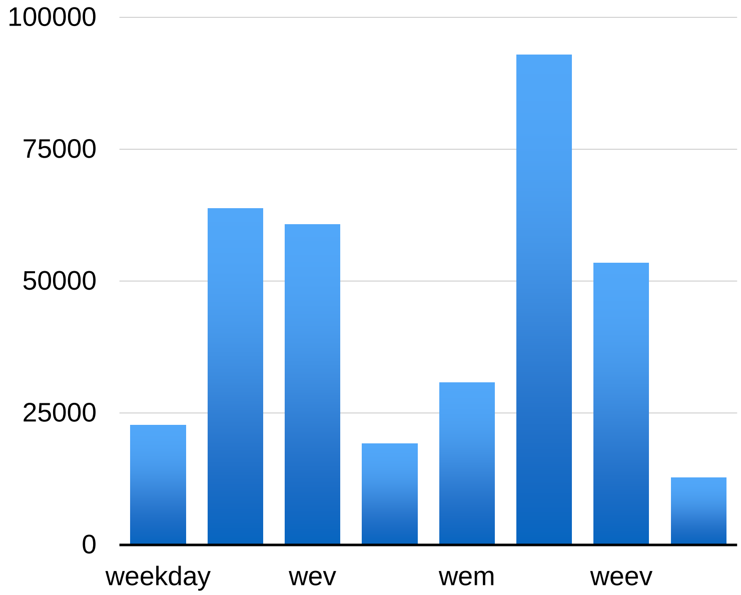


Trends in different cities in different seasons:

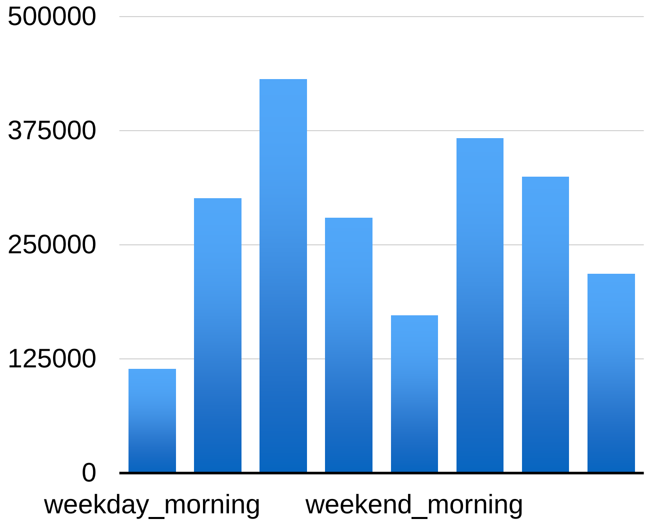
The below plots show the eating out trends in two different cities Las Vegas and Montreal. From the plots it is observed that in Las Vegas people tend to eat out more during the winter, where as in Montreal on an average people tend to shop more in the autumn.



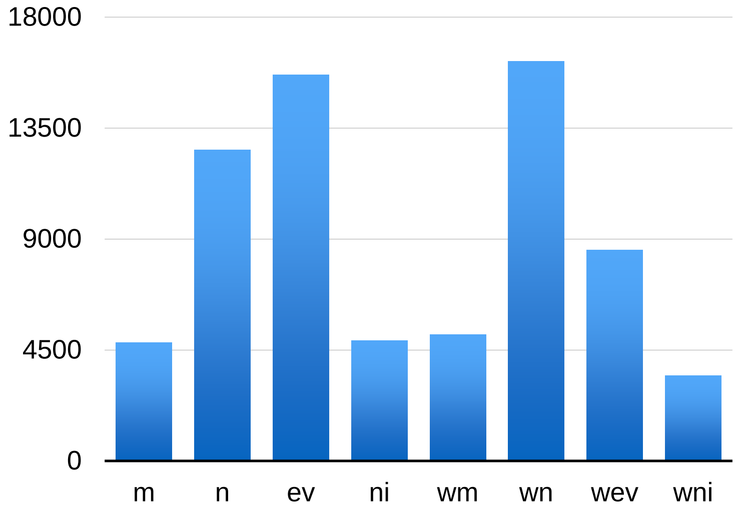




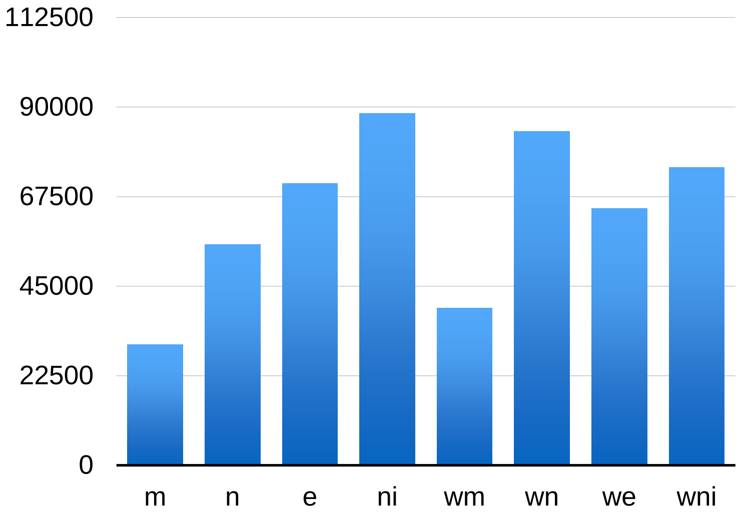
This histogram provides information with respect to the number of people checking into shopping establishments during different times in a week. It is seen that greater number of people tend to shop during weekend afternoons. The morning and night times are times people prefer least to shop.



This histogram provides information with respect to the number of people checking into restaurants during different times in a week. It is seen that greater number of people tend to eat out during weekday evenings followed by weekend afternoons. The number of people checkins into restaurants on weekend mornings is more when compared to weekdays.



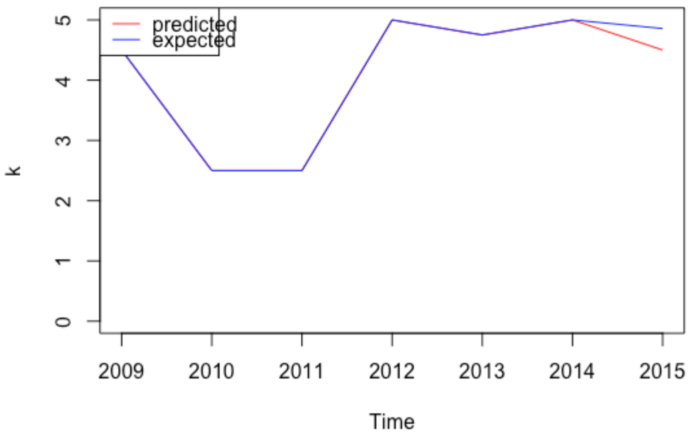
This histogram provides information with respect to the number of people checking into beauty and spa related establishments during different times in a week. It is seen that greater number of people visit these establishments on weekday evening and weekend afternoons.

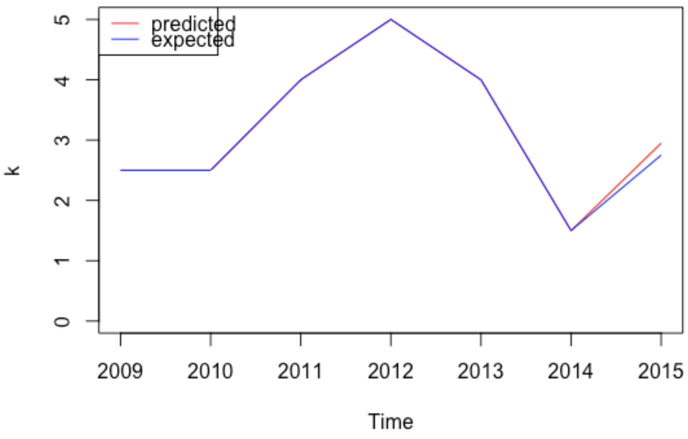


This histogram provides information with respect to the number of people checking into entertainment related establishments during different times in a week. It is seen that greater number of people visit these establishments on weekday night which could be following work. This is followed by weekend afternoons and weekend nights.

Output of Predictive Analysis Using ARIMA:

Here are some sample outputs of the predictive analysis where we see the predicted value slightly varying from the expected value:





3. Conclusion

Considering the amount of data that is required to be processed it would not have been possible without Big Data. Big Data provided an efficient and quick way to process and analyze the data as we required.

During the course of the project some of the key learning made are the various Big data strategies available. By analyzing each one of the strategies in order to determine which would be most suitable for this project we gained insight into PIG, HIVE and SPARK. Familiarized with SPARK SQL by using it in the project. Became familiar with different time series algorithms. Gained knowledge of R to do predictive analysis and to visualize the results obtained from the data analysis.

4. Role of Team Members

Akshay: Setting up of cluster and predictive analysis using R.

Neha: Analysis of data using spark and generation of trend results.

Raghu: Analysis of data using spark and generation of trend results.

5. References

https://rstudio-pubs-static.s3.amazonaws.com/127992_a060e7d374d549998df02fc11ac8c334.html

http://spark.apache.org/docs/latest/sql-programming-guide.html