Final Project

ALY 6015 - 70864

Members : Akash Raj, Harshit Gaur, Mirav Parekh

Professor : Dr. Amin Karimpour

Date : 03/06/2022

Table of Contents

# Introduction

Data analytics is the process of analyzing raw data and generating actionable insights. It comprises of the processes, tools and techniques of data analysis and management, including the collection, organization, and storage of data. Organizations use data analytics to gain competitive advantage by improving their performance and operational efficiency. Data analytics is performed on a variety of big data sets, like transactions, click streams, server logs, electronic health records, insurance claims, etc. Different analytical techniques and algorithms can be applied on these data sets to accomplish different objectives. These different types of analytical techniques are colloquially called:-

1. **Descriptive analytics** - Summarizing the data to understand past events and performance
2. **Diagnostic analytics** - Investigating the root cause of certain events
3. **Predictive analytics** - Predicting the future for planning
4. **Prescriptive analytics** - Recommending the optimal outcomes

Irrespective of the type of analytics being performed, the basis of every method or algorithm in data analytics is descriptive/inferential statistics and machine learning. In this analysis report, we will leverage descriptive statistics to generate insights from the data.

**Problems Statement :**

CarDekho.com is India’s leading car search venture that helps users buy cars that are right for them. It’s website and app carry rich automotive content such as expert reviews, detailed specs and prices, comparisons as well as videos and pictures of all car brands and models available in India.

We have the sales data of all the cars sold during the time frame of 1983 to 2020. We are going to analyse this data set in order to help them expand their business, gain and retain customers, and stand out the competitions they face.

The data set has 8128 data points with 13 features in it related to :

**Car Details -** *Car name, transmission, fuel type, number of seats, year of manufacturing.*

**Engine Details -** *Mileage, Engine type, Torque, Maximum power in BHP.*

**Sale Details -** *Selling price, kilometers driven by the car.*

Later, we will implement learning algorithms and modelling techniques to understand the patterns and achieve high quality, consistent results targeting the following points :

Identify the right prospects at the right time

Build customer loyalty.

Promote efficiency across the departments.

Marketing expenditures and supply chain management.

# Data Analysis

**Importing the libraries required for the analysis**

# Declaring the names of packages to be imported  
packageList <- c("tidyverse", "vtable", "RColorBrewer", "corrplot", "car", "psych", "stargazer", "scales", "DT")  
  
for (package in packageList) {  
 if (!package %in% rownames(installed.packages()))   
 { install.packages(package) }  
   
 # Import the package  
 library(package, character.only = TRUE)  
}

**Importing the data set for analysis**

## Descriptive Statistics

**Descriptive Statistics of ‘Total Votes’ in 2008, 2012, 2016**

totalVotesStats <- ElectionData %>%   
 select(v2008, v2012, v2016)  
  
# Kable Classic Method  
totalVotesStats <- totalVotesStats %>%   
 describe(quant = c(.25, .75), IQR = TRUE) %>%   
 mutate(year = c(2008, 2012, 2016)) %>%   
 relocate(year)  
round(totalVotesStats, 2) %>%   
 kbl(caption = "Table 1: Descriptive Statistics for Total Votes") %>%   
 kable\_classic(html\_font = "Cambria")

Table 1: Descriptive Statistics for Total Votes

year

vars

n

mean

sd

median

trimmed

mad

min

max

range

skew

kurtosis

se

IQR

Q0.25

Q0.75

v2008

2008

1

3112

42153.95

119995.7

11040.5

18272.39

11473.84

79

3318248

3318169

11.40

227.29

2151.03

24091.75

4973.50

29065.25

v2012

2012

2

3112

39505.56

105629.7

10732.5

17507.81

11050.56

64

2427869

2427805

9.14

139.88

1893.50

22862.50

4765.75

27628.25

v2016

2016

3

3112

41745.37

113404.8

10947.0

18177.94

11456.79

64

2652072

2652008

9.25

143.74

2032.88

23976.00

4820.50

28796.50

**Descriptive Statistics of ‘Total Democratic Votes’ in 2008, 2012, 2016**

totalDemocraticVotesStats <- ElectionData %>%   
 select(vd2008, vd2012, vd2016) %>%   
 describe(quant = c(.25, .75), IQR = TRUE) %>%   
 mutate(year = c(2008, 2012, 2016)) %>%   
 relocate(year)  
# Kable Classic Method  
round(totalDemocraticVotesStats, 2) %>%   
 kbl(caption = "Table 2: Descriptive Statistics for Total Democratic Votes") %>%   
 kable\_classic(html\_font = "Cambria")

Table 2: Descriptive Statistics for Total Democratic Votes

year

vars

n

mean

sd

median

trimmed

mad

min

max

range

skew

kurtosis

se

IQR

Q0.25

Q0.75

vd2008

2008

1

3112

22311.25

77155.32

4499.0

7759.49

4958.56

8

2295853

2295845

13.74

315.52

1383.08

10579.75

1806.25

12386.00

vd2012

2012

2

3112

19990.04

65991.90

3948.5

6925.40

4442.61

5

1672164

1672159

11.31

211.30

1182.96

9545.25

1555.00

11100.25

vd2016

2016

3

3112

20060.65

71998.07

3153.0

6057.85

3701.31

4

1893770

1893766

11.78

226.83

1290.63

8442.50

1166.00

9608.50

**Descriptive Statistics of ‘Total Republican Votes’ in 2008, 2012, 2016**

totalRepublicanVotes <- ElectionData %>%   
 select(vg2008, vg2012, vg2016) %>%   
 describe(quant = c(.25, .75), IQR = TRUE) %>%   
 mutate(year = c(2008, 2012, 2016)) %>%   
 relocate(year)  
# Kable Classic Method  
round(totalRepublicanVotes, 2) %>%   
 kbl(caption = "Table 3: Descriptive Statistics for Total Republican Votes") %>%   
 kable\_classic(html\_font = "Cambria")

Table 3: Descriptive Statistics for Total Republican Votes

year

vars

n

mean

sd

median

trimmed

mad

min

max

range

skew

kurtosis

se

IQR

Q0.25

Q0.75

vg2008

2008

1

3112

19207.53

44866.80

6312.0

9881.96

6607.95

67

956425

956358

8.37

114.86

804.28

13154.75

2881.25

16036.00

vg2012

2012

2

3112

18890.39

41731.45

6393.5

9909.65

6657.62

54

699600

699546

6.98

75.11

748.07

13059.00

2889.75

15948.75

vg2016

2016

3

3112

19622.38

40442.74

7164.5

10825.88

7525.68

57

620285

620228

6.25

59.14

724.97

14242.25

3206.00

17448.25

**Descriptive Statistics of ‘Total Unemployment Rate’ in 2011, 2012, 2013, 2014, 2015**

unemploymentRate <- ElectionData %>%   
 select(unemp11, unemp12, unemp13, unemp14, unemp15) %>%   
 describe(quant = c(.25, .75), IQR = TRUE) %>%   
 mutate(year = c(2011, 2012, 2013, 2014, 2015)) %>%   
 relocate(year)  
# Kable Classic Method  
round(unemploymentRate, 2) %>%   
 kbl(caption = "Table 4: Descriptive Statistics for Total Unemployment Rate") %>%   
 kable\_classic(html\_font = "Cambria")

Table 4: Descriptive Statistics for Total Unemployment Rate

year

vars

n

mean

sd

median

trimmed

mad

min

max

range

skew

kurtosis

se

IQR

Q0.25

Q0.75

unemp11

2011

1

3140

4420.26

16317.13

1109.5

1750.46

1243.90

4

600541

600537

19.43

605.90

291.19

2476.25

435.00

2911.25

unemp12

2012

2

3140

4001.55

14725.93

985.0

1567.57

1103.80

4

536500

536496

19.06

585.56

262.80

2193.25

385.00

2578.25

unemp13

2013

3

3140

3666.48

13446.77

909.0

1453.96

1020.03

4

484628

484624

19.02

574.57

239.97

2015.50

361.75

2377.25

unemp14

2014

4

3140

3079.51

11316.39

756.0

1215.57

842.12

4

414341

414337

19.29

597.22

201.95

1661.25

307.00

1968.25

unemp15

2015

5

3140

2644.24

9419.03

663.5

1064.86

731.66

4

336860

336856

18.34

546.73

168.09

1467.50

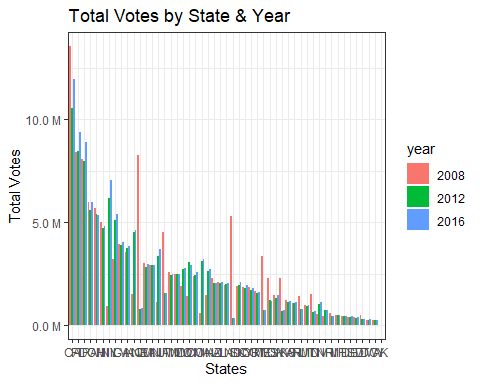
274.50

1742.00

## Exploratory Data Analysis

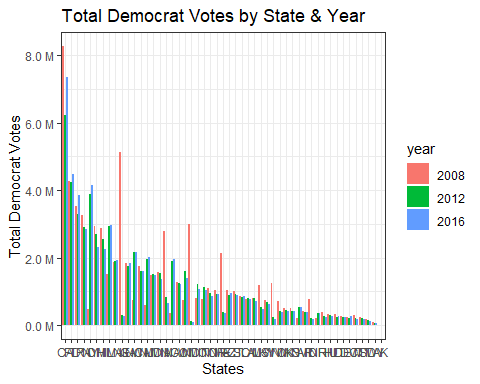
**Plot of outcome variable ‘Total Votes’ by state and year**

totalVotesL <- ElectionData %>%   
 select(state, v2008, v2012, v2016) %>% group\_by(state) %>% summarise('2008' = sum(v2008, na.rm = TRUE),  
 '2012' = sum(v2012, na.rm = TRUE),  
 '2016' = sum(v2016, na.rm = TRUE)) %>%  
 gather(year, tVotes, c('2008', '2012', '2016'))  
  
ggplot(data = totalVotesL, mapping = aes(x = reorder(factor(state), tVotes, function(x) -1\*sum(x)), y = tVotes, fill = year)) +  
 geom\_bar(position = "dodge", stat = "identity") +  
 labs(title = "Total Votes by State & Year") +   
 scale\_x\_discrete(name ="States") +   
 scale\_y\_continuous(name = "Total Votes", labels = label\_number(suffix = " M", scale = 1e-6)) +  
 theme\_bw()



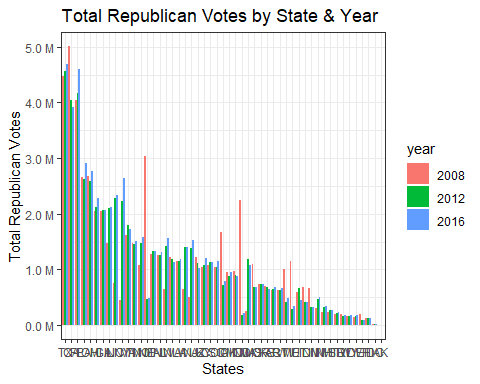
**Plot of outcome variable ‘Total Democratic Votes’ by state and year**

totalDVotesL <- ElectionData %>%   
 select(state, vd2008, vd2012, vd2016) %>% group\_by(state) %>% summarise('2008' = sum(vd2008, na.rm = TRUE),  
 '2012' = sum(vd2012, na.rm = TRUE),  
 '2016' = sum(vd2016, na.rm = TRUE)) %>%  
 gather(year, tdVotes, c('2008', '2012', '2016'))  
  
ggplot(data = totalDVotesL, mapping = aes(x = reorder(factor(state), tdVotes, function(x) -1\*sum(x)), y = tdVotes, fill = year)) +  
 geom\_bar(position = "dodge", stat = "identity") +  
 labs(title = "Total Democrat Votes by State & Year") +   
 scale\_x\_discrete(name ="States") +   
 scale\_y\_continuous(name = "Total Democrat Votes", labels = label\_number(suffix = " M", scale = 1e-6)) +  
 theme\_bw()



**Plot of outcome variable ‘Total Republican Votes’ by state and year**

totalRVotesL <- ElectionData %>%   
 select(state, vg2008, vg2012, vg2016) %>% group\_by(state) %>% summarise('2008' = sum(vg2008, na.rm = TRUE),  
 '2012' = sum(vg2012, na.rm = TRUE),  
 '2016' = sum(vg2016, na.rm = TRUE)) %>%  
 gather(year, tgVotes, c('2008', '2012', '2016'))  
  
ggplot(data = totalRVotesL, mapping = aes(x = reorder(factor(state), tgVotes, function(x) -1\*sum(x)), y = tgVotes, fill = year)) +  
 geom\_bar(position = "dodge", stat = "identity") +  
 labs(title = "Total Republican Votes by State & Year") +   
 scale\_x\_discrete(name ="States") +   
 scale\_y\_continuous(name = "Total Republican Votes", labels = label\_number(suffix = " M", scale = 1e-6)) +  
 theme\_bw()



# Conclusion

The data set of car sales by CarDekho.com has provided various insights about the types of cars sold in the car industry and the patterns between them. The data set contains 8128 data points along with 13 features related to car details, engine details, and sale details.

Data cleaning and data pre-processing are important to prepare the data in the correct format before building the regression model. We extracted numerical values from text fields like engine, mileage, torque, along with brand information from car name.

Unnecessary data that skew the results were also filtered out.

To ensure that the numerical variables are not skewed, it’s crucial to remove or impute the missing values and outliers. We imputed the missing values using kNN imputation method.

Feature engineering and exploratory data analysis were performed to gather more meaningful information from the data.

Derived variables were created using the existing features and skewed variables were scaled.

Apart from this, various data visualization, like box plot, frequency plot, histogram, pair plot, correlation matrix and scatter plot were created to understand the uni-variate distribution and multi-variate relationship of the data.

Once the data cleaning and exploratory analysis is performed, hypothesis testing is performed to validate certain assumptions on the sample data.

We used one sample t-Test and two sample t-Test to compare the variables ‘km\_driven’ and ‘selling\_price’ with the overall sample average and compare across two groups respectively. Based on the exploratory analysis performed earlier, we wanted to validate the following hypotheis using t-test:

1. True Mean of kilometers driven by ‘Individual’ seller-type is greater than the overall average
2. True Mean of selling price of cars sold by ‘Dealer’ seller-type is greater than the overall average
3. True Mean of kilometers driven by ‘First Owner’ owner-type is less than the overall average
4. True Mean of selling price of cars sold by ‘First Owner’ owner-type is greater than the overall average
5. Mean kilometers driven for small cars is not equal to the kilometers driven for medium cars
6. Mean kilometers driven for dealers is not equal to the kilometers driven for trustmark dealer

In the first five test, based on the t-statistic and p-value we obtained sufficient evidences to reject the NULL hypothesis, however in the last test, we did not obtain enough evidence to reject the NULL hypothesis.

We also performed Chi-Square Test and One-wav ANOVA test on categorical and continuous variables in our data set. The p-value of all the tests except the one where groups of Owner and Engine were taken came out to be very less than our assumed alpha = 0.05. Therefore, we were able to reject the Null Hypotheses in all but one tests successfully.

Lastly, we performed Regression Analysis on our data set to model the prices of cars listed on CarDekho.com. We log transformed the Selling price variable to convert it into normal distribution and used lm() function to build the linear regression model.

# Bibliography

Vehicle dataset. (2020, October 24). Kaggle. <https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho>

Xie, Y. C. D. (2021, October 7). R Markdown Cookbook. R Markdown. Retrieved October 30, 2021, from <https://bookdown.org/yihui/rmarkdown-cookbook/>

Bluman, A. (2017). Elementary Statistics: A Step By Step Approach (10th ed.). McGraw-Hill Education.

Kabacoff, R., I. (2022). R in Action, Third Edition. Manning.

CarDekho. (2021). About Us | CarDekho.com. <https://www.cardekho.com/info/about_us>

F. (2021, April 2). tidyverse in r – Complete Tutorial. R-Bloggers. Retrieved October 30, 2021, from <https://www.r-bloggers.com/2021/04/tidyverse-in-r-complete-tutorial/>

D. (2021a, March 22). Data Analytics for Car Dealers. Automated Metrics. <https://www.automatedmetrics.io/data-analytics-for-car-dealers/>

# Appendix

The RMD file of the analysis is included with the analysis report.