National College of Ireland

Data Warehousing and Business Intelligence Project

on

Passenger Car Analysis for New and Old Cars

Vaibhav Aher
X18104215

MSc/PGDip Data Analytics – 2018/9

Submitted to: Dr. Simon Caton

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Vaibhav Aher
Student ID:	X18104215
Programme:	MSc Data Analytics
Year:	2018/9
Module:	Data Warehousing and Business Intelligence
Lecturer:	Dr. Simon Caton
Submission Due Date:	26/11/2018
Project Title:	Passenger Car Analysis on New and Old Cars

I hereby certify that the information contained in this (my submission) is information pertaining to my own individual work that I conducted for this project. All information other than my own contribution is fully and appropriately referenced and listed in the relevant bibliography section. I assert that I have not referred to any work(s) other than those listed. I also include my TurnItIn report with this submission.

ALL materials used must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is an act of plagiarism and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	November 26, 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table 1: Mark sheet – do not edit

Criteria	Mark Awarded	Comment(s)
Objectives	of 5	
Related Work	of 10	
Data	of 25	
ETL	of 20	
Application	of 30	
Video	of 10	
Presentation	of 10	
Total	of 100	

Project Check List

This section capture the core requirements that the project entails represented as a check list for convenience.

- ☒ Used LaTeX template
- ☐ Three Business Requirements listed in introduction
- ☐ At least one structured data source
- ☐ At least one unstructured data source
- ☐ At least three sources of data
- ☐ Described all sources of data
- ☐ All sources of data are less than one year old, i.e. released after 17/09/2017
- ☐ Inserted and discussed star schema
- ☐ Completed logical data map
- ☐ Discussed the high level ETL strategy
- ☐ Provided 3 BI queries
- ☐ Detailed the sources of data used in each query
- ☐ Discussed the implications of results in each query
- ☐ Reviewed at least 5-10 appropriate papers on topic of your DWBI project

Passenger Car Analysis on New and Old Cars

Vaibhav Aher
X18104215

November 26, 2018

Abstract

This Project involves the analysis of new car prices with respect to old car prices for USA region. The goal is to show that which car gives better return price after the usage. The main parameter for the used car is the total run by the car (Mileage). Comparison of the new price with the old price of a car, it is helpful to judge the used price of the car after the specific run. This has been done by collecting the data of cars with the latest prices and for an old car, data has taken from the portal where users put their car for sale. Also, there are some specific states in the USA where the stolen rate is higher. This analysis is trying to put a focus on the total average loss of stolen cars with respect to old prices. In this project with the help of five different dataset analysis is conducted against prices, Number of stolen vehicle and stolen vehicles types.

1 Introduction

Everyone has its dream to have his own car but due to insufficient fund, most of the people opt for used vehicles. In this era of internet and social media customer and seller got an innovative platform to sell their vehicle and the one who is looking for cars also can find the best suitable deals. There are many websites where users can post their cars for sale. The important factor in the used car industry is the year of the vehicle and its mileage. One can negotiate the price of the car but what customer thinks about its budget and condition of vehicle(Singer, A.(2016)) Reviewer can see the car details and most of the times they use filter or sort as per their requirements. The aim of my project is to analyze the prices of new and old cars as per year. It is a symptom of a wise customer to first analyze the market of the car, it's yearly fallen prices and return on investment.

Another concern for the car market is the vehicle stolen rate. It is one of the biggest challenges in front of authorities. The common man has thought that higher is the price of car greater will be the risk from thefts. However, the data of stolen vehicle according to type and the state from where it got stolen is easily available on the government of USA website. It is better to study these things in a statistical way. Researchers working on the model which can be used in a traffic control system. If stolen vehicle is detected or if it passes through the system then the red light will be burn. (Swarup Suresh Kulkarni and Dr Roshani Ade (2017))But it is better to implement this system as a trial basis where car stolen rate is higher or which specific types of cars are targeted by theft. For this, the data with the number of cars stolen and the type of cars stolen has taken for analysis. On this topic, this project will cover the following requirements.

(Req-1) My first requirement is based on the percent of price changes per year for an old car with respect to the new car for same car model.

(Req-2) The second requirement is based on the state with the highest number of cars have any co-relation with a number of a vehicle stolen?

(Req-3) The third requirement covers the question Does car with higher cost has been rapidly targeted by the thieves?

2 Data Sources

For this project, total of five data sets are used from which one is unstructured and rest four are structured.

2.1 Source 1: Kaggle

Released Date :27/09/2017

This is the data set of the used car with its prices and year and which is downloaded from:`https://www.kaggle.com/jpayne/852k-used-car-listings`

Dataset provides 8 columns of information on 800 thousand used cars posted for resale on truecar.com.Each row represent one used car listing. Data is collected on 24/09/2017. Columns present in this dataset are as follow Make, Model, Mileage, price, Year, VIN, City, State However for this project I have taken Make, Model, Mileage, Price, Year, City and state into consideration.

This dataset provides information about the used car prices and the year of the used car. In the first requirement I have compared the prices of the old car with respect to the new car. For this on the basis of the car make and car model the average price is taken into consideration. Also, it provides information about which city in which state that car is available for sale. So for first query the attributes from this dataset used are the year, make, model and price.

2.2 Source 2: Cars.com

`https://www.cars.com/research/` This is one of the leading websites in the USA used for new car specifications and its prices. This is unstructured data obtained by web scrapping using R code. As per requirement for first BI query and third BI query The data scrapped by considering car make with different car models as unique combination parameter. Columns of this data set after web scrapping are New Car Make, Model, Year and Price. In first BI query to compare how the car prices dropped year by year according to the various model I have used New Car Model, Make, year and price with respect to Old car Make, Model and prices. Moreover in third BI query for Car price comparison whether the car with highest prices are targeted by thieves or not their price and car Model are used.

2.3 Source 3: www.iii.org

Release Date : September 2018

This is structured data obtained by web scrapping from `https://www.iii.org/table-archive/20900` This dataset contains column as vehicle stole, state, Rank, Year

Source	Type	Brief Summary
Kaggle	Structured	This dataset contains used cars details
Cars.com	Unstructured	This dataset provides the specifications and prices of the new car as per their models.
www.iii.org	Structured	This dataset provides yearly state wise stolen vehicle rate.
Statista.com	Structured	This dataset provides top 10 frequently stolen car models in the USA.
Federal Highway Administration	Structured	This data provides the information regarding state wise vehicle owned by the public.

Table 2: Summary of sources of data used in the project

from which year, vehicle stolen and State are useful for analysis. Yearly data was given on the website. For the current analysis, I have selected 2015,2016 and 2017 data only. This data set addresses the business requirements for the second query. In which does the state which owned a maximum number of cars have the greater rate of the car stolen? In this State and Stolen cars, these two parameters are used.

2.4 Source 4: www.statista.com

Release Date : September 2018

This is structured data obtained from `https://www.statista.com/statistics/424163/us-top-ten-frequently-stolen-passenger-vehicles-by-model/` This dataset provides information about the top 10 most frequently stolen car models for the year 2017. In this dataset, some cars are only given with Model like a pickup, So as per requirement I have done changes in the model using r code. This dataset has Car type, stolen numbers. For BI query 3 Does thieves target only high price vehicle? By analysis of those factors, we can conclude that bigger price cars are stolen mostly or not. For this query, I have used the stolen numbers column.

2.5 Source 5: Federal highway Administration

Release Date : November 2017

This is structured dataset obtained from `https://www.fhwa.dot.gov/policyinformation/statistics/2016/mv7.cfm` It Contains total fifteen columns like Buses, Automobiles, Trucks, Total, Total vehicles, States, from which the data like the number of vehicles per state is taken into consideration for BI query in which the number of total cars and number of stolen cars are compared.

3 Related Work

Used Car Industry is one of the growing Industry in the world market. Changing prices of the car created the unmanageable task for dealers. Without proper analysis, it will be economically risky to maintain an inventory of the used car. As per customer demand, the variety should be available to the dealer. In (AutomotiveNews(2018)) mentioned the reasons why used car Industry changes seasonally. Most of the customer prefer small cars

in city so we can say that size of car also the factor in used car. However (Peterson(2014)) mentioned in his research topic that there are four important factors in used car which are Engine,Air conditioning ,vehicle body and transmission. Common man thinking is thieves theft car for sailing the parts and making the money but as per (Ragavan(1999)) research car theft is big black market. Many crimes including drugs supply,Human trafficking. Car theft is challenging task in front of police.

4 Data Model

For my project I have build Star schema which consist of One Fact Table named Fact_Cars and three dimension tables Dim_Car,Dim_Location,Dim_Year. In which fact table consist of measures Old_Price,New_Price,Old_mileage,Type_St_number,St_Total_number,

Total_Cars_owned,Car_Id,Location_Id,Year_Id . In which Old_Price is from Kaggle data source which indicates the price of used cars, New_Price is from cars.com which indicates the price of new cars.Old_mileage which is mileage of used cars and it comes kaggle dataset.Type_St_number is shows the total number of stolen vehicle as per car model and this data set belongs to Statista.St_Total_number give us information of statewise stolen cars the data source is www.iii.org . Total_Cars_owned shows us the total number of publicly owned vehicles per state and source for this data is Federal Highway Administration. Primary key of all dimension table are present in Fact table. Car_Id, Location_Id, Year_Id is present in Fact table. As I have inner joined New_car,Theft_Car,Stolen_Car,Vehicle_Owned_car all with Old_car on Model, Year,States respectively.

For First Dimension Dim_Car, there are three attributes Car_Id, Old_model, Old_Make from which Old_Model and Old_make are basically from an Old_cars table but as I have mentioned Earlier that due to inner joined only matching pairs from the new and Old car only will come.

Do not do this:

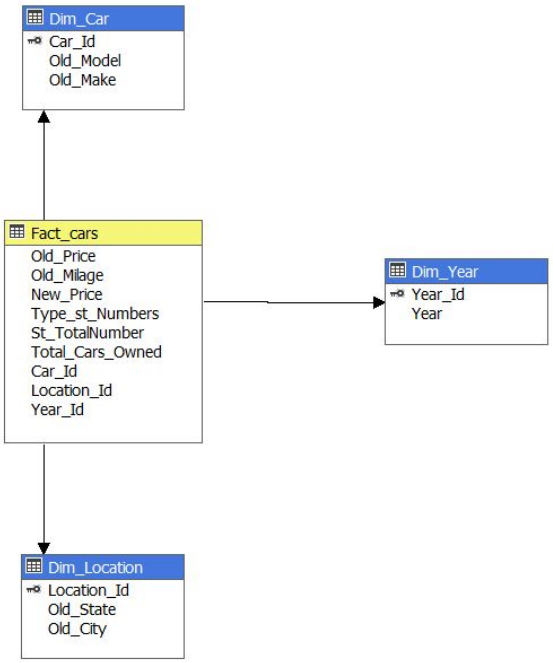

Figure 1: Car Price Analysis Star Schema

So i have 4 dimensions, as seen in Figure 1. The second dimension is Dim_Location in which there are three attributes which are Location _Id, Old_State, Old_City. Here The states from the Old_Cars table is inner joined with Theft_Cars Table So only matching pairs are reflected in this dimension.

The third dimension is Dim_Year in which The year from old car is inner join with Stolen_Cars year for joining purpose. And the dimension comes from kaggle dataset. We

will need the year in first BI query to show percentage price changed yearly of a new car with respective to old car

5 Logical Data Map

Table 3: Logical Data Map describing all transformations, sources and destinations for all components of the data model illustrated in Figure 1

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
Kaggle	Make	Dim_Car	Dim_Make	Dimension	Filtered above car Make GMC,Honda,Chevrolet,Toyota,Nissan,Dodge.
Kaggle	Model	Dim_Car	Dim_Model	Dimension	For Some car Models Special sub categories were attached eg. ChargerSE changed to Charger
Kaggle	Price	Fact_Cars	Old_price	Fact	Removed ´ ,´ and changed data type from string to number by $extract_numeric$ function
Kaggle	State	Dim_Location	Old_State	Dimension	State names with abbreviation were given , by using in build library converted to full name of state
Kaggle	City	Dim_Location	Old_City	Dimension	Column name changed to Old_City
Kaggle	Mileage	Fact_Cars	Old_Mileage	Fact	Changed type from String to integer
Cars.com	New_Price	Fact_Cars	New_Price	Fact	Data is crapped from website only type is changed from string to integer
www.iii.org	Year	Dim_Year	Year	Dimension	As data scraped from website where it was yearly given. For that year column has been created
www.iii.org	Stolen Vehicle	Fact_Cars	St_Total _Numbers	Fact	During web scrapping clean the data to get integer number
Statista	Number of stolen vehicle	Fact_Cars	Type_st _Numbers	Fact	Changed the column name as Type_st _Numbers
Federal Highway Administration	Total Motor Vehicles	Fact_Cars	Total_Cars_Owned	Fact	From number ',' is removed and column name changed to Total_Cars_Owned

6 ETL Process

In this project, My ETL process is divided into six parts. It starts with R script execution. In single R script it consists Data loading from the website then data cleaning and finally, data is exported to Excel file. Various packages are used as per the requirement. In some datasets, values are manipulated as per requirement this part is explained in commenting section of R code. Major areas in cleaning are removing null values, changes in column name for identification, Slicing of unwanted string, Changes for the car models for specific car model manually using for loop, Extraction of the number from the string, splitting and slicing of the matrix.

After cleaning the data frames are stored in different excel sheet. In SSMS the database with name carsdb is created. After project creation in SSIS, the first and main step is attached Execute SQL task in the data flow. As during testing, the project was run several times, it is observed that duplicate data was repeatedly stored into different tables. For fresh execution, tables should be truncated. This step was done after observing the staging table duplicate values. After that, the cleaned data is extracted to SSIS by using Flate file source. All five data sets are imported into flat file source and are connected to different OLEDB destinations where after successful execution the tables are created with the flat file source data.The Table Names Are New_Car,Old_Car,Theft_Car,Stolen_Car,Owned_Car Create table query in OLEDB destination creates the staging table in SSMS for the selected database.

Now the flow comes to the dimension table creation stage. the Data from OLEDB destination is imported to OLEDB source in new data flow task. In OLEDB source only that table is selected from which dimension comes, In my case different dimension old_city, Old_state, Old_Make, Old_Model, Old_Year are from Old_car table. Sort is used before the dimension table creation step. In sort it only allows the unique combination of selected entities. Like I have applied sort for car model, year, states and cities. the output of different sort is attached respective OLEDB destinations where three dimension tables are created. While cresting dimension table primary key is assigned during query by using Identity in SQL. In my project, I have created three dimensions namely Dim_car, Dim_Location, Dim_year with unique primary key as ID like Cat_ID, Location_ID, Year_ID.

After creating dimension tables, In new data flow task new OLEDB source is taken where the different tables are joined using SQL join query. In my case, I have selected all the columns that I want from table NewCars and OldCars, where Model is inner, joined with NewModel from New_Car table whereas Car Make from OldCars inner joined with Make of StolenCar table. State and year are joined on Theft_Car table.This will selects all the matching rows between two different tables which creates relation between the two different tables.

After Joining all the tables by inner join with respect to different fields the output channel of data is joined to Lookup1 which is applied on Make and Model which will transfer only relative data to next lookup2. In lookup 2 it is applied on State and city In which the matched output data from lookup 1 will only retrieve and transfer as output in same manner lookup3 will work for the year after passing data through all three lookups it will come to OLEDB destination where Fact table will be created. Now In fact table only measures like Old_Price,Old_Mileage,New_Price,type_st_number,st_total_numbers,

total_cars_owned,Car_Id,Location_ID, Year_Id will come.

In the last stage, I found some Car Null Id and Location ID in fact tables which are

removed by attaching execute SQL task in the last step. Now Fact table can be used for further deployment process.

7 Application

As per requirements mentioned in 1,Following are the BI queries with their solutions are represented in graphical manner.

7.1 BI Query 1: How much percentage of Price difference occurs for New Car with respect to used car as per year

For this query, the contributing sources of data are kaggle and www.Cars.com This visualization obtained as illustrated in Figure 4This demonstrates the percentage change in New car price with respect to old car price this is represented yearly in three slides.

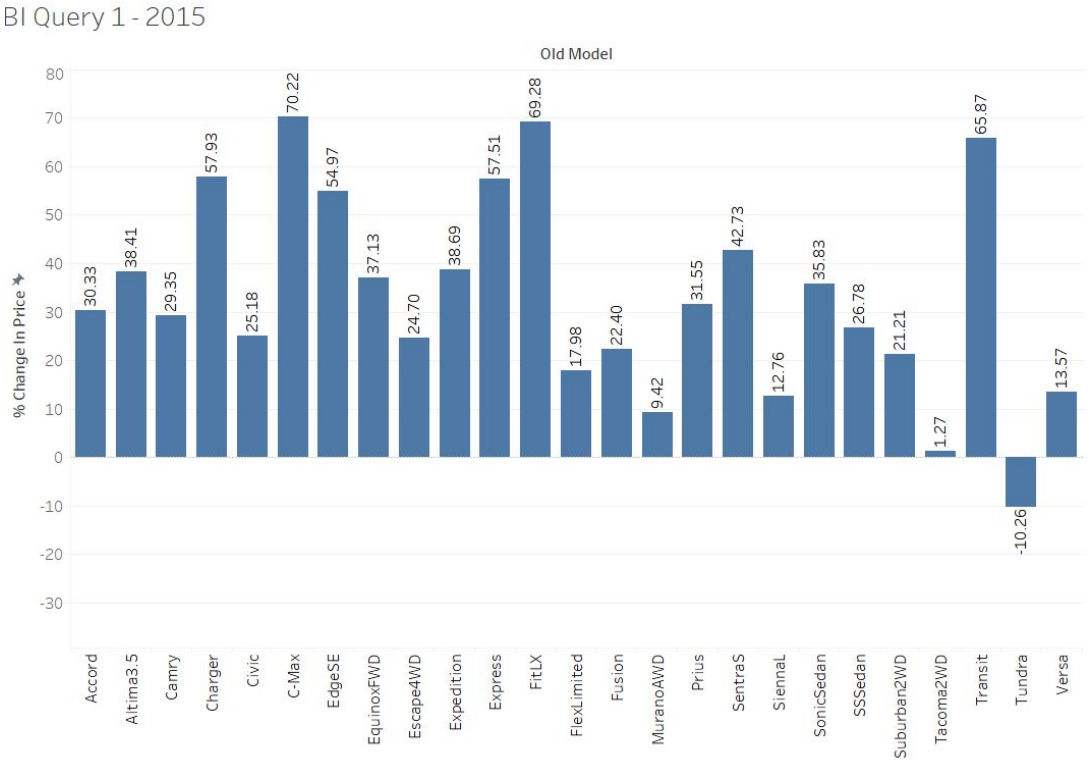

Figure 2: Results for BI Query 1 Year 2015

In 2015 It can be seen that almost all listed car models prices are above 20 per cent except Tundra, The new Car model of Tundra is 10 per cent cheap as compare to the Old price of used car And as we increase the year the percentage price changes appear to be lower and lower. However in the case of FITLX model in 2017 appears to be better because the new car from the same model is by fifty per cent extra price. From this, we can conclude that resale value would be good for the car whose new model price is increasing rapidly.

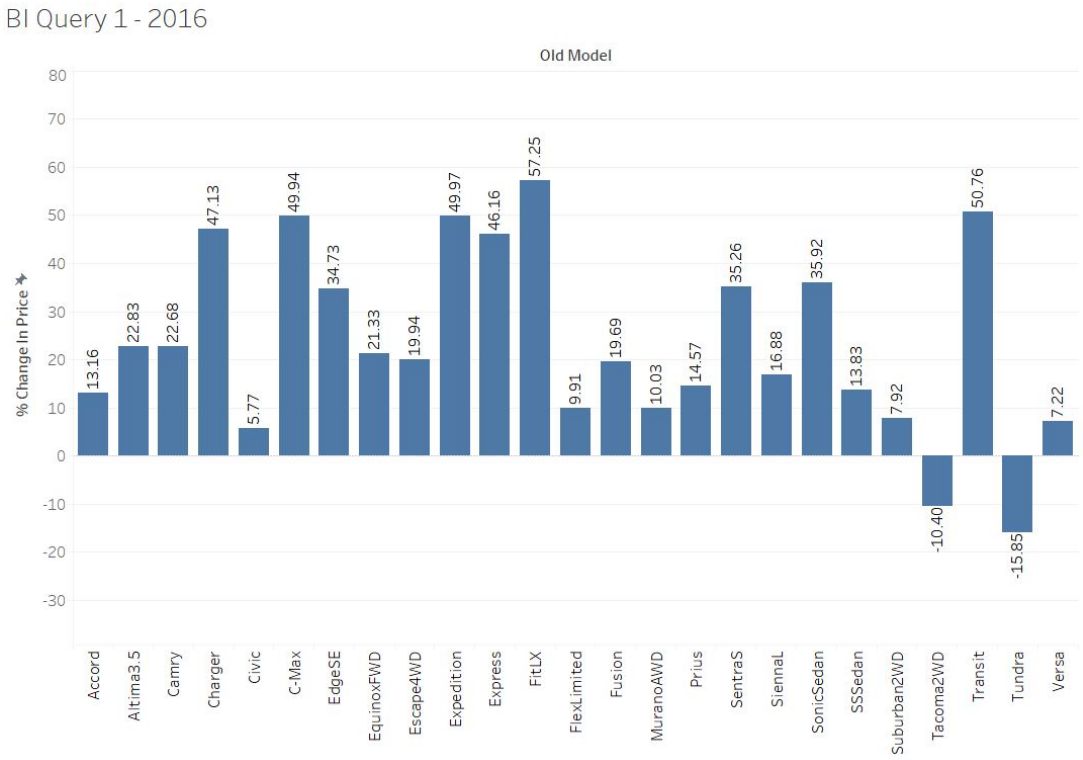

Figure 3: Results for BI Query 1 Year 2016

7.2 BI Query 2: Is there any relationship between number of cars in state and car theft ?

www.iii.org,Federal Highway Administration are two data source used for this second BI query. Federal Highway Administration data source is used for finding out the state wise vehicles in the USA whereas www.iii.org gives the average number of vehicle stolen as per state for three years.From figure Figure 5

It can be seen that state with higher number of cars has higher rank in car stolen. California has highest number of cars at the same time number of car stolen is also high.If we compare Total cars in state and Stolen cars then Ohio state have large number of cars but less stolen cars as compare to Illinois. So for some states its true that state with larger number of vehicle have higher number of car stolen rate.

7.3 BI Query 3: Does thieves intensively target the car which have higher price ?

For this query, the contributing sources of data are kaggle, Statista. Major parameters are Number of Car Stolen, Car Model and Car price. For this query, I am considering old price as car price. From Figure 6 It can be seen that Chevrolet Suburbarn2WD has a comparatively higher price and the number of Car same car Model stolen are also higher. But inverse results are seen from the Toyota Camry and Toyota Corolla model. For these two models though the price is different the stolen number rate is in the same range. So we can conclude that other than price there might be other factors that affect the Car stolen rate as per car model. Figure 6.

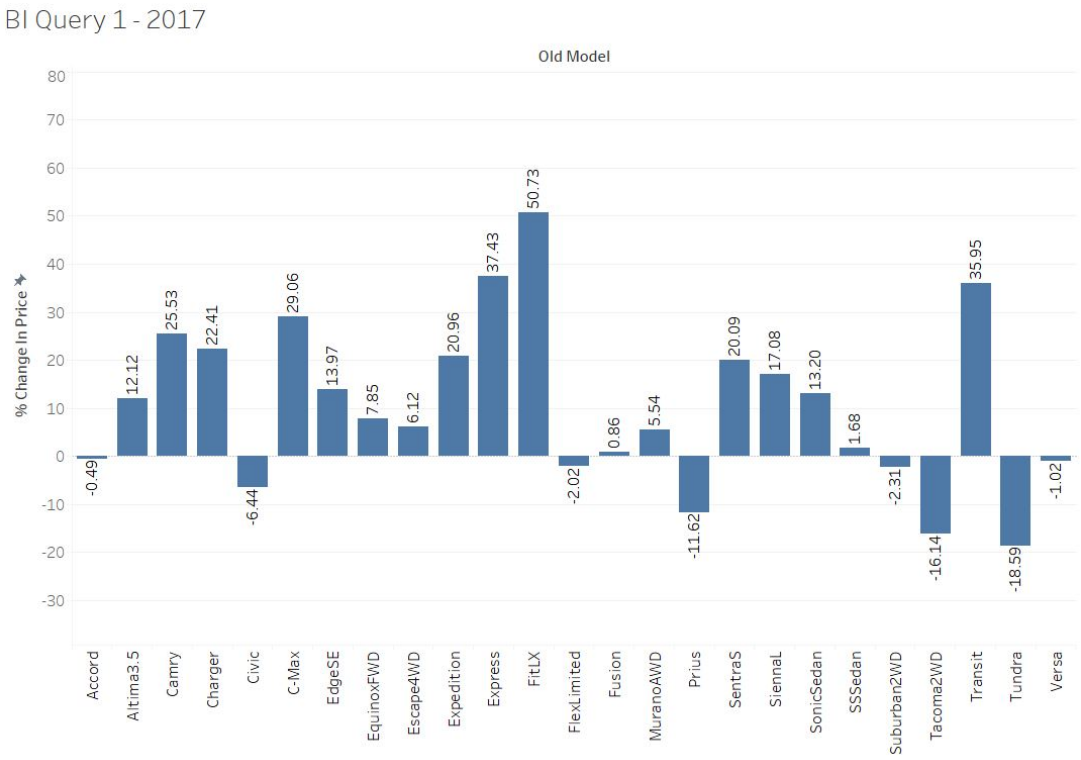

Figure 4: Results for BI Query 1 Year 2017

7.4 Discussion

This reports give an idea about better deals of the car, After comparing the graph of three years it will be helpful to find which car value will give better return if one want to sell it after few years.As mentioned in Introduction many people have low budget because of that customer may opt for used car. But some times it can be seen that there is no large difference in used car prices and new car prices. Moreover there are some specific car models with large price which are intensively targeted by thieves. Car company or Customer can implement extra security system for car. There are specific states in USA where Car stolen rates are higher.

8 Conclusion and Future Work

The implementation of data warehouse helps in analyzing the the degradation of price for old car with respect to new one for same model. Also with the help of state wise survey of stolen car and cost of the car, tracking devices may be used for the state where cr stolen rate is high. For further study If more car details like width, condition, Engine type, fuel type will help the project for more accurate result.

References

Appendix

R code

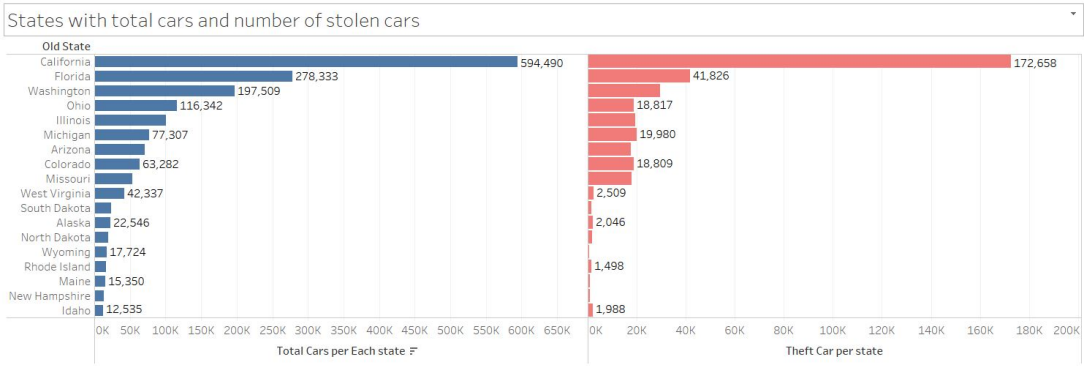

Figure 5: Results for BI Query 2

```
library(rvest)
library(purrr)
library('gmp')
library(xml2)
library(tidyr)
library(readr)
library(readxl)
library(stringi)
library(Hmisc)


#*************** Unstructured Data of New Cars and Its prices **************

#Iterating through webpages
Vehicle <- c("ford/","gmc/","honda/","chevrolet/","toyota/","nissan/",
"dodge/","audi/","mitsubishi/")
url_base <- "https://www.cars.com/research"

#Addind Extracted data in to the Data frame
map_df(Vehicle[1:9], function(i) {
  pg1 <- paste(url_base,i,sep = "/")
  pg <-  WS1 <- html_session(pg1)
  model <- capitalize(gsub("/","",i))

  data.frame(Model=stri_trim(gsub(model,"",html_text(html_nodes
  (pg, "#make-model-list␣.cui-heading-4"))),side =
  c("both","left","right")),
 Price=extract_numeric(html_text(html_nodes
 (pg, ".make-model-card--msrp-range"))),
              Year = "2018",
              Make= model)
}) -> Vehicle_df

#Removing rows with NA Values
Vehicle_df <- Vehicle_df[!is.na(Vehicle_df$Price),]
names(Vehicle_df)[1] <- paste("New_Model")
names(Vehicle_df)[2] <- paste("New_Price")
```

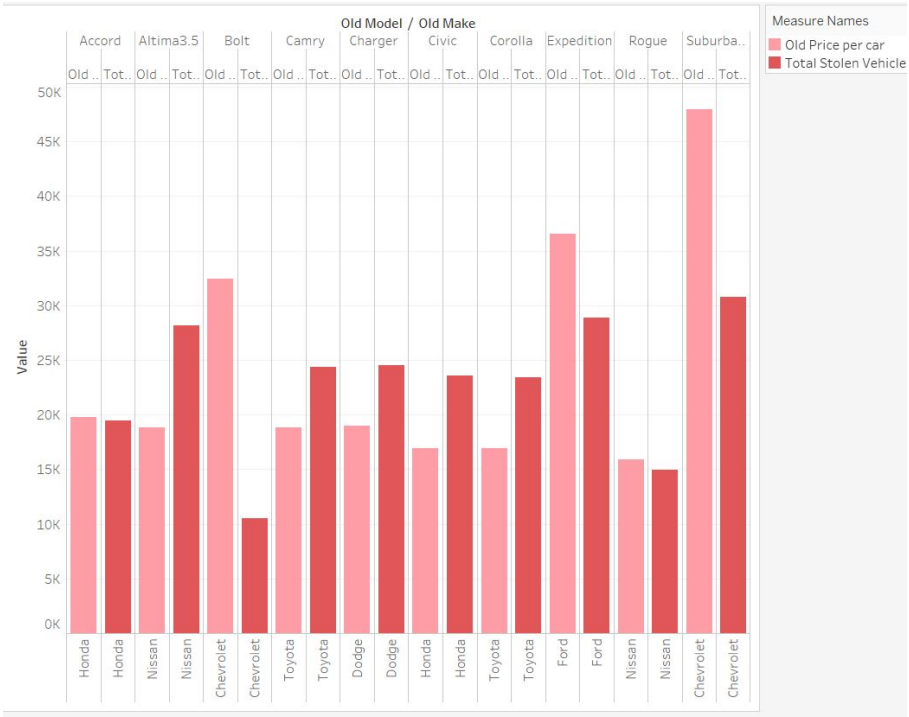

Figure 6: Results for BI Query 3

```
names(Vehicle_df)[3] <- paste("New_Year")
names(Vehicle_df)[4] <- paste("New_Make")

#For Some car Model Car with extra model
name was given like GMC Acadia
1500 for that we are considering model
as Acadia only
#Because in Used old car data model is given
only as Acadia engine capacity not mentioned so removed
it while cleaning
i = 31
while(i <41){
  Vehicle_df$New_Model[i]
  <- strsplit(Vehicle_df$New_Model[i], "\\s+")[[1]][2]
  i = i+1
}
#While web scrapping some models are comes
with extra Values or with features seperated
from during scrapping
#To it has been replaced with original one

Basic_Car <- c("C-Max","FiestaSedan","Express","TaurusSE","EdgeSE","Escape4W
"Explorer4WD","FlexLimited","Transit","TerrainAWD","CanyonCrew","FitLX",
"Pilot2WD","RidgelineRTS","OdysseyEX","Bolt","CamaroCoupe","SSSedan",
"SonicSedan","VoltLT","EquinoxFWD","Suburban2WD","TahoeLS","TraverseFWD",
"ColoradoCrew","86Manual","Tacoma2WD","SiennaL","Altima3.5","LEAFS",
"MaximaS","SentraS","JUKES","MuranoAWD","PathfinderS")
N <- c( 1,3,76,13,15,16,20,21,30,33,38,48,51,52,53,54,56,61,62,
```

```
64,66,67,68,69,71,77,101,103,105,107,108,109,113,115,116)
Vehicle_df$New_Model[N] <- Basic_Car


#Export cleaned New Car Prices Data to csv
write.csv(Vehicle_df, file = "C:\\Users\\Vaibhav\\Desktop\\DWBI␣Project\\Cle

#********************** Structured Data set of Second Hand Cars and its Pric
print("Importing␣Sencond␣Hand␣cars␣data␣will␣take␣time␣as␣it
contains␣more␣than␣8␣lacs␣rows")
Used_Cars <- read.csv("C:\\Users\\Vaibhav\\Desktop\\Car␣DWBI␣Datasets\\true_

# Removed VIN column
Used_Cars <- Used_Cars[-6]

#Remove all data older than 2015
Used_Cars <- Used_Cars[!Used_Cars$Year< 2015,]

# Cleaning with respect to selected brands like "Gmc","Honda","Chevrolet","T
Used_Cars <- Used_Cars[Used_Cars$Make == "Ford" | Used_Cars$Make == "GMC"
 | Used_Cars$Make == "Honda"    | Used_Cars$Make == "Chevrolet"
  | Used_Cars$Make == "Toyota"  | Used_Cars$Make == "Nissan"
  | Used_Cars$Make == "Dodge"| Used_Cars$Make == "Audi"| Used_Cars$Make == "

# Give full names to states from Abbrevation here state.Name is inbuild data
names(state.name) <-state.abb
Used_Cars$State <-state.name[Used_Cars$State]

#Removed the states with Unkwon Abbrevations
Used_Cars <- Used_Cars[!is.na(Used_Cars$State),]

#Remove All NA columns from Dataset
Used_Cars <- Used_Cars[!is.na(Used_Cars$Price) |!is.na(Used_Cars$Year)
|
                          !is.na(Used_Cars$City)    |!is.na(Used_Cars$Mileage)
                          !is.na(Used_Cars$Make)    |!is.na(Used_Cars$Model),]

#Rearrange the row sequence number and cast list to data frame
rownames(Used_Cars) <- NULL
df_Used_Cars <- as.data.frame(Used_Cars)

names(df_Used_Cars)[1] <- paste("Old_Price")
names(df_Used_Cars)[2] <- paste("Old_Year")
names(df_Used_Cars)[3] <- paste("Old_Milage")
names(df_Used_Cars)[4] <- paste("Old_City")
names(df_Used_Cars)[5] <- paste("Old_State")
names(df_Used_Cars)[6] <- paste("Old_Make")
names(df_Used_Cars)[7] <- paste("Old_Model")

#Removed Sub category of Charger SE car to Charger
k = 0
```

```
for (val in df_Used_Cars$Old_Model) {
  k = k+1
  if(val =="ChargerSE"){
    df_Used_Cars$Old_Model[k] <- "Charger"
  }
}

#Export cleaned Used Data to csv
write.csv(df_Used_Cars, file = "C:\\Users\\Vaibhav\\Desktop\\DWBI Project\\C

# Structured Dataset for USA statewise Number of stolen Cars by Model from h

Theft_Car_Data <- html_text(html_nodes(read_html
("https://www.iii.org/table-archive/20900"),'td'))

#Removing string that contains brief description of Dataset
value <- "Most"
value2 <- "Highest"
Theft_Car_Data<-Theft_Car_Data[-(grep(value2,Theft_Car_Data))]
Theft_Car_Data<-Theft_Car_Data[-(grep(value,Theft_Car_Data))]

#Merge All values in to matrix and split matrix in to sub Matrices
1st Matrix is for 2017, 8th = 2015, 9th = 2016
for(i in 2:length(Theft_Car_Data)) output=rbind(output,matrix(Theft_Car_Data
output <- matrix(unlist(Theft_Car_Data), ncol = 3, byrow = TRUE)
My_Matrix <- split(as.data.frame(output), rep(1:9, each = 20))

Theft_Car_2017 <- My_Matrix$`1`[-1]
Theft_Car_2016 <- My_Matrix$`9`[-1]
Theft_Car_2015 <- My_Matrix$`8`[-1]
Theft_Car_2017$Year <- 2017
Theft_Car_2016$Year <- 2016
Theft_Car_2015$Year <- 2015

names(Theft_Car_2017)[1] <- paste("State")
names(Theft_Car_2017)[2] <- paste("Car Stolen")
names(Theft_Car_2016)[1] <- paste("State")
names(Theft_Car_2016)[2] <- paste("Car Stolen")
names(Theft_Car_2015)[1] <- paste("State")
names(Theft_Car_2015)[2] <- paste("Car Stolen")

Theft_Car_2017$`Car Stolen` <- extract_numeric(Theft_Car_2017$`Car Stolen`)
Theft_Car_2016$`Car Stolen` <- extract_numeric(Theft_Car_2016$`Car Stolen`)
Theft_Car_2015$`Car Stolen`<- extract_numeric(Theft_Car_2015$`Car Stolen`)

Theft_Car_2017 <- Theft_Car_2017[order(Theft_Car_2017$`Car Stolen
`,decreasing = TRUE),]
Theft_Car_2016 <- Theft_Car_2016[order(Theft_Car_2016$`Car Stolen
`,decreasing = TRUE),]
Theft_Car_2015 <- Theft_Car_2015[order(Theft_Car_2015$`Car Stolen`
,decreasing = TRUE),]
```

```
Theft_Car <- rbind(c,Theft_Car_2015,Theft_Car_2016,Theft_Car_2017)
Theft_Car <- Theft_Car[!is.na(Theft_Car$State),]
rownames(Theft_Car) <- NULL
df_Theft_Cars <- data.frame(Theft_Car)

names(df_Theft_Cars)[1] <- paste("St_State")
names(df_Theft_Cars)[2] <- paste("St_TotalNumber")
names(df_Theft_Cars)[3] <- paste("St_Year")

#Export cleaned Used Data to csv
write.csv(df_Theft_Cars, file = "C:\\Users\\Vaibhav\\Desktop\\DWBI␣Project\\


#********************** Structured data from Statista ***************
*****(Which Type of vehicle stolen) ****************************
url2 <- "C:\\Users\\Vaibhav\\Desktop\\Car␣DWBI␣Datasets\\Statista␣Data\\stat
-2017-by-model␣(1).xlsx"

Type_Of_Vehicle_Stolen <- read_excel(url2,col_names = FALSE,col_types = NULL
,sheet = 2,na="",skip = 3)
names(Type_Of_Vehicle_Stolen)[1] <- paste("Model")
names(Type_Of_Vehicle_Stolen)[2] <- paste("Stolen_Numbers")
Type_Of_Vehicle_Stolen <- Type_Of_Vehicle_Stolen[!is.na(Type_Of_Vehicle_Stol

#Removed Values in Bracket
Type_Of_Vehicle_Stolen$Model <- sapply(strsplit(Type_Of_Vehicle_Stolen$Model
,split = "␣(",fixed = TRUE),function(x) (x[1]))
Type_Of_Vehicle_Stolen$Make <- sapply(strsplit(Type_Of_Vehicle_Stolen$Model,
split = "␣",fixed = TRUE),function(x) (x[1]))
Type_Of_Vehicle_Stolen$Model <- sapply(strsplit(Type_Of_Vehicle_Stolen$Model
split = "␣",fixed = TRUE),function(x) (x[2]))

# Here Only car model for few were given as 'Pick up' so I have replace pick
Type_Of_Vehicle_Stolen$Model[6] <- "Rogue"
Type_Of_Vehicle_Stolen$Model[3] <- "Expedition"
Type_Of_Vehicle_Stolen$Model[10] <- "Bolt"
Type_Of_Vehicle_Stolen$Model[4] <- "Suburban2WD"
Type_Of_Vehicle_Stolen$Model[8] <- "Charger"

df_Type_Of_Vehicle_Stolen <- data.frame(Type_Of_Vehicle_Stolen)
names(df_Type_Of_Vehicle_Stolen)[1] <- paste("Type_Model")
names(df_Type_Of_Vehicle_Stolen)[2] <- paste("Type_st_Numbers")
names(df_Type_Of_Vehicle_Stolen)[3] <- paste("Type_Make")

df_Type_Of_Vehicle_Stolen$Type_Year <- 2017

write.csv(df_Type_Of_Vehicle_Stolen, file = "C:\\Users\\Vaibhav\\Desktop\\DW

#*************** Structured data for Number of vehicles woned statewise*****
Path <- "C:\\Users\\Vaibhav\\Desktop\\Car␣DWBI␣Datasets\\Total␣STates␣Vehicl
```

```
data <- read_excel(path = Path,sheet = 1,skip = 6,na = "",col_names = FALSE,
Number_Vehicle_Owned <- data[14]
States_Name <- data[1]
Number_Vehicle_Owned <- data.frame(States_Name,Number_Vehicle_Owned)
Number_Vehicle_Owned <- Number_Vehicle_Owned[c(2:52),]
row.names(Number_Vehicle_Owned) <- NULL
names(Number_Vehicle_Owned)[1] <- paste("States")
names(Number_Vehicle_Owned)[2] <- paste("Total_Cars_Owned")
write.csv(Number_Vehicle_Owned, file = "C:\\Users\\Vaibhav\\Desktop\\DWBI␣Pr
```


[Automotive News (2018)]

Small used-car prices rise unexpectedly; Retailers are struggling to stock adequate supply. Available at: http://search.ebscohost.com/login.aspx?direct=trueAuthType=ip,cookie,shibdb=eds livescope=sitecustid=ncirlib (Accessed: 26 November 2018).

[Swarup Suresh Kulkarni and Dr. Roshani Ade (2017)]

Swarup Suresh Kulkarni and Dr. Roshani Ade (2017) Intelligent Traffic Control System Implementation for Traffic Violation Control, Congestion Control and Stolen Vehicle Detection, International Journal of Recent Contributions from Engineering, Science IT, Vol 5, Iss 2, Pp 57-71 (2017), (2), p. 57. doi: 10.3991/ijes.v5i2.7230.

[Singer, A. (2016)]

Singer, A. (2016) Purchasing a Used Vehicle, Salem Press Encyclopedia. Available at: http://search.ebscohost.com/login.aspx?direct=trueAuthType=ip,cookie,shibdb=ersAN=113928203sit livescope=sitecustid=ncirlib (Accessed: 24 November 2018)

[Peterson, J. R. and Schneider, H. S. (2014)]

Peterson, J. R. and Schneider, H. S. (2014) Adverse selection in the used-car market: evidence from purchase and repair patterns in the Consumer Expenditure Survey, RAND Journal of Economics (Wiley-Blackwell), 45(1), pp. 140154. doi: 10.1111/1756-2171.12045.