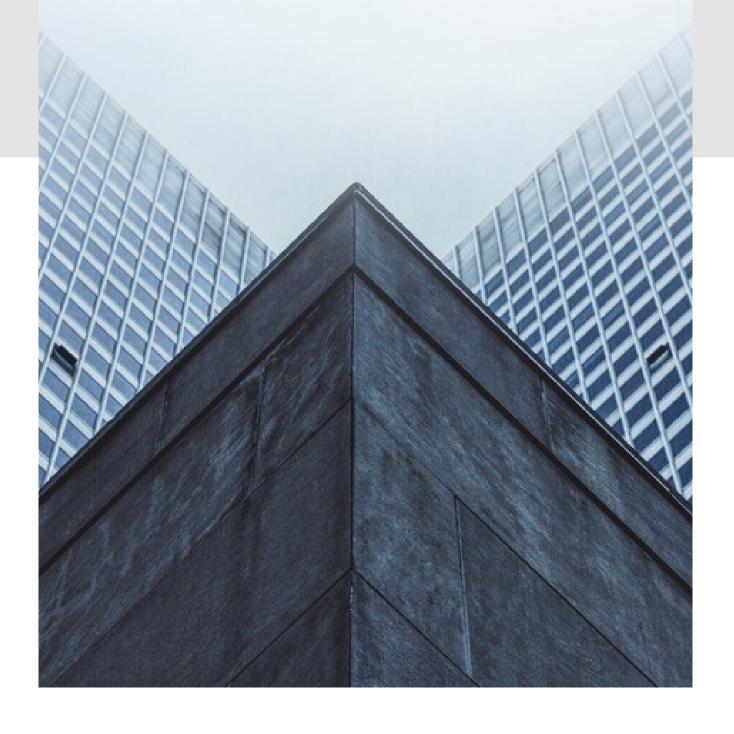
## GROUP K

## R PROJECT

CASE STUDY - USED CARS AGENCIES



Data description: The data that already stored in the excel file is data set contains information about used cars that are already sold or not by the agency.

#### The data set are as follows:

- 1.Name, (Model)
- 2.Year
- 3. Selling price
- 4.Kmdriven
- 5.Fueltype
- 6.Sellertype
- 7. Transmissiontype
- 8.Owner(First or second)
- 9.Mileage(Fuel consume per km)
- **10.Engine Capacity**
- 11.Maximum horse power
- 12.Engine torque
- 13. Number of seats
- 14.Name of dealer

#### The Agency has three dealers working full time:

- 1-Mr. Henry Spelman
- 2-Ms. Anny Grouper
- 3-Mr. David Soliman

# TASK DISTRIBUTION

## **VARUN NANDA**

DESCRIPTIVE ANALYSIS ( DATA READING and QUESTIONS )

# HIMANSHU DHAHANA

HIMANSHU DHAHANA – PREDICTIVE ANALYSIS ( PLUS GRAPHS )

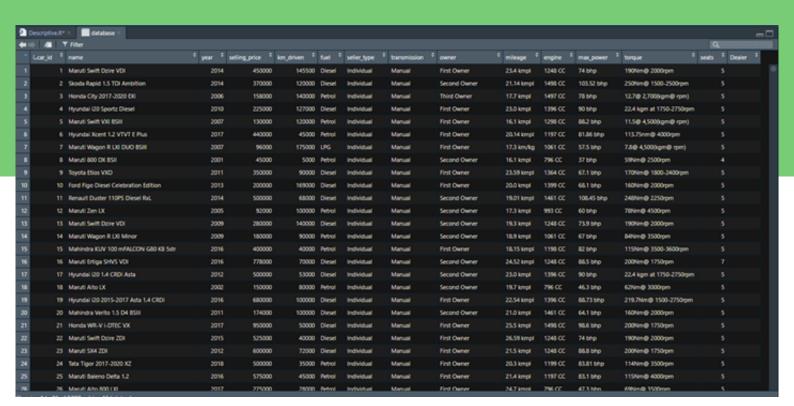
# SUKHMAN AND BHUMANYU SINGH

REPORTS AND PRESENTATION

## IMPORTING DATA

setwd("D:/r programs/big data")

database<- read.csv("carbook.csv") view(database)



## DESCRIPTIVE ANALYSIS

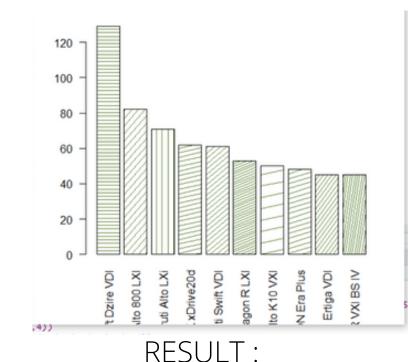
#### Q.1 WHICH CAR MODEL SOLD THE MOST?

#### CODE:

#### **GRAPH:**

#### TABLE:





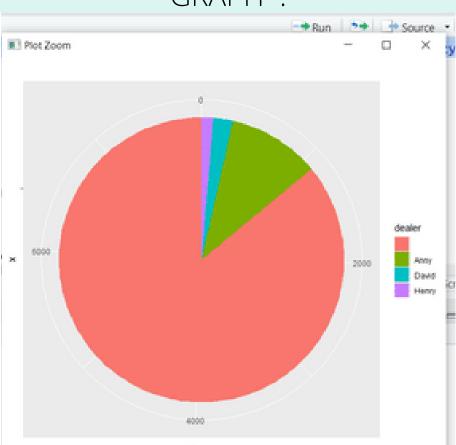
# Q.2 Which dealer sold more cars, and his total sales?

```
CODE:
```

#### **RESULT:**

```
> print(main_dealer)
dealer
Anny David Henry
6800 828 175 104
>
```

#### GRAPH:



# Q.3 WHAT IS THE AVERAGE PRICE FOR EACH CAR MODEL?

#### CODE:

```
# average price of each car model

car_model <- database[ ,c(2,4)]

View(car_model)

choose_model <- filter(car_model, name=='Maruti Swift Dzire VDI')

print(choose_model)

avg_price <- mean(choose_model$selling_price)

print(avg_price)</pre>
```

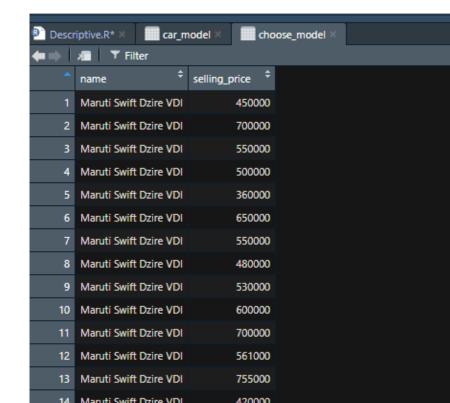
#### FOR ALL



#### RESULT: AVERAGE PRICE OF CAR MODELS

```
>
> print(avg_price)
[1] 578658.9
> |
```

#### For a choosen model



#### Q.4 WHAT IS THE NEWEST CAR AND THE OLDEST CAR?

CODE:

```
#newest and oldest car
year_name <- database[,c(2,3)]

view(year_name)

all_year<- (table(years))

print(all_year , decreasing=TRUE)

#newest car
newest_cars=filter(year_name, year==2020)

View(newest_cars)
#oldest car
oldest_cars=filter(year_name, year==1994)

View(oldest_cars)</pre>
```

#### YEAR:

#### Descriptive.R\* × year\_name × all\_year ▶ 🔏 🍸 Filter 10 2003

#### YEAR NAME:



#### Oldest cars

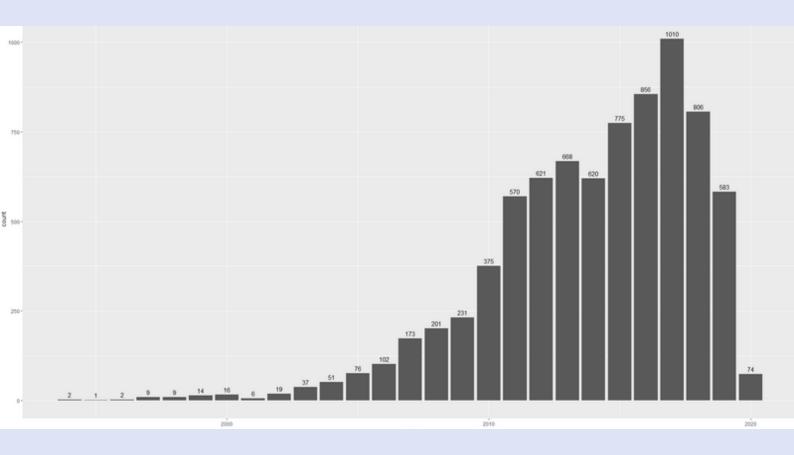
#### **NEWEST CARS**

] [	Descriptive.R* × year_name ×	all_year ×
į	→ In The Filter  →	
	name	year ‡
3	Toyota Innova Crysta 2.7 GX AT 8 STR	2020
4	Hyundai Grand i10 Sportz	2020
5	Maruti Eeco 5 Seater AC BSIV	2020
5	Maruti Dzire LXI	2020
7	BMW X7 xDrive 30d DPE	2020
3	Maruti Swift AMT VXI	2020
9	Tata Zest Revotron 1.2 XT	2020
)	Tata Zest Revotron 1.2 XT	2020
1	Maruti Alto 800 LXI	2020
2	Hyundai Xcent 1.2 VTVT E	2020
3	Maruti Eeco CNG 5 Seater AC BSIV	2020
4	Datsun GO T Option	2020
5	Maruti Alto 800 VXI	2020
5	Mahindra Scorpio S11 4WD BSIV	2020
7	Tata Zest Revotron 1.2 XT	2020
3	Kia Seltos HTX Plus AT D	2020
9	Volkswagen Polo 1.0 TSI Highline Plus	2020
)	Tata Harrier XZ Plus	2020

oldest cars

newest\_cars

# GRAPH:



# Q.5 Which is the best car related to price, and low mileage? CODE:

```
#best car related to price, low mileage
mileage_table <- database[ ,c(2,4,10)]

view(mileage_table)

min_price <- min(mileage_table$selling_price)

filtered_car<- filter(mileage_table, selling_price==min_price)

view(filtered_car)</pre>
```

#### **OUTPUT:**

<u> </u>	Descriptive.R* × mileage_table × fil	tered_car ×	
<b>(#</b> 1	→ Image: Im		
*	name	selling_price ‡	mileage ‡
1	Maruti Swift Dzire VDI	450000	23.4 kmpl
2	Skoda Rapid 1.5 TDI Ambition	370000	21.14 kmpl
3	Honda City 2017-2020 EXi	158000	17.7 kmpl
4	Hyundai i20 Sportz Diesel	225000	23.0 kmpl
5	Maruti Swift VXI BSIII	130000	16.1 kmpl
6	Hyundai Xcent 1.2 VTVT E Plus	440000	20.14 kmpl
7	Maruti Wagon R LXI DUO BSIII	96000	17.3 km/kg
8	Maruti 800 DX BSII	45000	16.1 kmpl
9	Toyota Etios VXD	350000	23.59 kmpl
10	Ford Figo Diesel Celebration Edition	200000	20.0 kmpl
11	Renault Duster 110PS Diesel RxL	500000	19.01 kmpl
12	Maruti Zen LX	92000	17.3 kmpl
13	Maruti Swift Dzire VDi	280000	19.3 kmpl
14	Maruti Wagon R LXI Minor	180000	18.9 kmpl
15	Mahindra KUV 100 mFALCON G80 K8 5str	400000	18.15 kmpl
16	Maruti Ertiga SHVS VDI	778000	24.52 kmpl
17	Hyundai i20 1.4 CRDi Asta	500000	23.0 kmpl
18	Maruti Alto LX	150000	19.7 kmpl
19	Hyundai i20 2015-2017 Asta 1.4 CRDi	680000	22.54 kmpl
20	Mahindra Verito 1.5 D4 BSIII	174000	21.0 kmpl
21	Honda WR-V i-DTEC VX	950000	25.5 kmpl

#### **RESULT:**

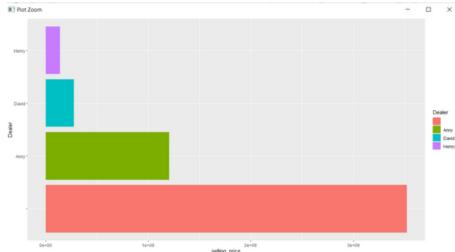
name <sup>‡</sup> sellir	ng_price † mileage †
1 Maruti 800 AC 2999	99 16.1 kmpl

# Q.6 What is the total revenue from Individual selling?

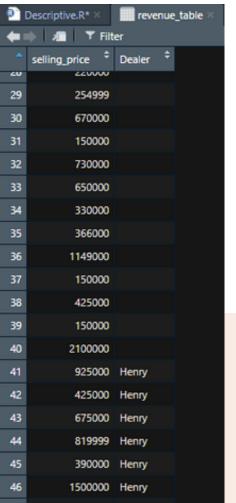
CODE:

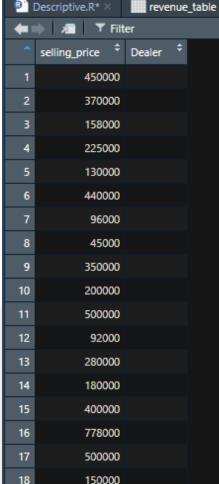


GRAPH:



TABLES:





#### **RESULT**

> print(revenue)
[1] 3521387320
>



# PREDICTIVE ANALYSIS

BY GROUP K

# 1.Installation of packages needed for analysis and running the code

```
1 install.packages("dplyr")
2 install.packages("ggplot2")
3 install.packages("readr")
4 install.packages("tidyverse")
```

# 2. Loading the packages into R by running the code

```
6 library(dplyr)
7 library(ggplot2)
8 library(readr)
9 library(tidyverse)
```

- 3. i) Loading data into R
  - ii) Choose required columns from data frame
  - iii) Check the data for null values

```
carbook <- read_csv("carbook.csv")
|
cardf <- carbook %>% select(c(name, selling_price,mileage))
view(cardf)

str(cardf)

#Changing mileage format
cardf$mileage<- gsub("km/kg","",as.character(cardf$mileage))
view(cardf$mileage)

# checking for null values
sum(is.na(cardf))</pre>
```

## **RESULTING DATA TABLE:**

^	name \$	selling_price ‡	mileage ‡
1	Maruti Swift Dzire VDI	450000	23.4
2	Skoda Rapid 1.5 TDI Ambition	370000	21.14
3	Honda City 2017-2020 EXi	158000	17.7
4	Hyundai i20 Sportz Diesel	225000	23.0
5	Maruti Swift VXI BSIII	130000	16.1
6	Hyundai Xcent 1.2 VTVT E Plus	440000	20.14
7	Maruti Wagon R LXI DUO BSIII	96000	17.3
8	Maruti 800 DX BSII	45000	16.1
9	Toyota Etios VXD	350000	23.59
0	Ford Figo Diesel Celebration Edition	200000	20.0
1	Renault Duster 110PS Diesel RxL	500000	19.01
2	Maruti Zen LX	92000	17.3
3	Maruti Swift Dzire VDi	280000	19.3
4	Maruti Wagon R LXI Minor	180000	18.9
5	Mahindra KUV 100 mFALCON G80 K8 5str	400000	18.15
6	Maruti Ertiga SHVS VDI	778000	24.52
7	Hyundai i20 1.4 CRDi Asta	500000	23.0
8	Maruti Alto I X	150000	19.7

4. Performing simple regression using summary (car\_ln) .we want to estimate the most profitable car related to selling price and mileage.

```
#Linear regression model
car_lm = lm(mileage~name, data = cardf)
summary(car_lm)

car_lm1 = lm(mileage~name +selling_price, data = cardf)
summary(car_lm1)

#checking for coefficients
coefs <- coefficients(car_lm1)

#sorting the coefficients
coefs_sorted <- coefs[order(coefs)]

tail(coefs_sorted)</pre>
```

```
car_lm1 = lm(mileage~name +selling_price, data = cardf)
> summary(car_lm1)
lm(formula = mileage ~ name + selling_price, data = cardf)
Residuals:
Min 1Q Median 3Q Max
-4.3804 -0.0333 0.0000 0.0382 4.3945
Coefficients:
                                                                        Estimate Std. Error t value Pr(>|t|)
                                                                      1.273e+01 5.494e-01 23.173 < 2e-16 ***
-2.167e-02 7.770e-01 -0.028 0.977751
-4.244e-02 7.770e-01 -0.055 0.956445
(Intercept)
nameAmbassador Classic 2000 DSZ AC PS
nameAmbassador Grand 1500 DSZ BSIII
nameAmbassador Grand 2000 DSZ PW CL
                                                                       5.871e-01 7.770e-01 0.756 0.449912
                                                                                                  9.094
nameAshok Levland Stile LE
                                                                       7.067e+00 7.771e-01
```

# **Residuals:**

Min 1Q Median 3Q Max -4.3804 -0.333 0 0.0382 4.3945

## THE ERROR RANGE

 MEDIAN = ERROR MEDIAN in this case, 0

## The coefficients section shows:

- 1. The estimates (Estimate) for model parameters the value of y-intercept, in this case 1.273e+01 and the estimate effect is -2.167e-02
- 2. The standard error of the estimated values (std error)i.e. precision of estimate
- 3.. The test statistic i.e. T value.
- 4. The p-value (Pr(>|t|)) i.e. the probability of finding the given t-statistic if the null hypothesis of no relationship were true.

The most important thing to note is the p-value, 2.2e-16 or almost 0 which indicate whether the model fits the data well

```
signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.5494 on 5923 degrees of freedom
Multiple R-squared: 0.9861, Adjusted R-squared: 0.9815
F-statistic: 212.3 on 1983 and 5923 DF, p-value: < 2.2e-16
```

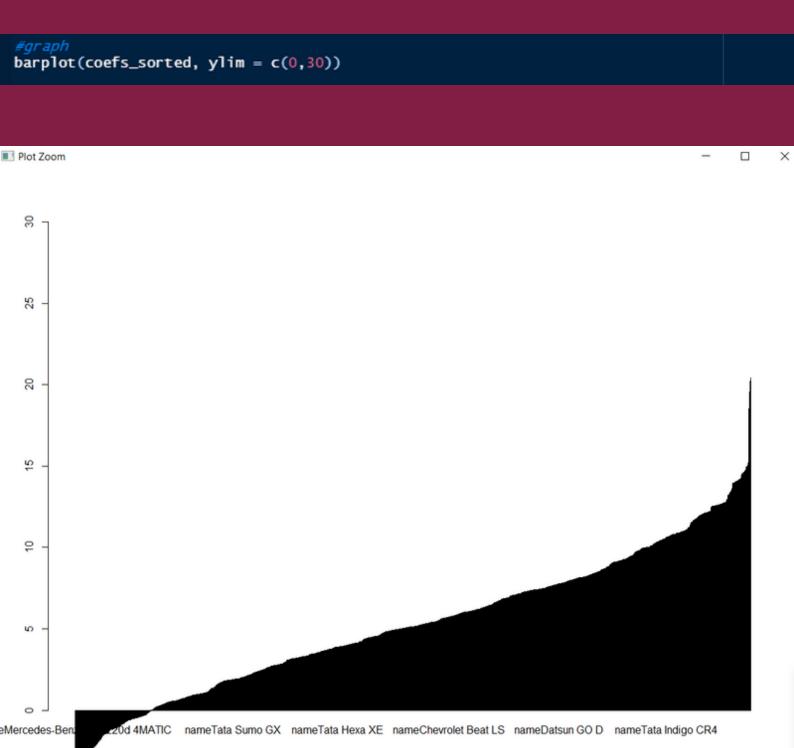
- Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 variable possible values
  - Residual standard error: 0.5494 on 5923 degrees of freedom

The standard deviation of the error

- Multiple R-squared: 0.9861, Adjusted R-squared: 0.9815
   The amount of variance explained by the model
- F-statistic: 212.3 on 1983 and 5923 DF, p-value: < 2.2e-16 checking if at least the variable is significantly different from zero.
- 5. We arrange the coefficient in descending order and observe the highest value.

# RESULT :- Maruti Alto 800 CNG LXI Optional

## **GRAPH OF LINEAR REGRESSION**



y- intercept = selling price + mileage x- intercept = car models

