PSG Institute of Management
PSG College of Technology

BLITZ JOBS

Internship Report: Summer 2020

Name: Akash M

Roll No:19PGM03

EXECUTIVE SUMMARY

Blitz jobs has been established with a vision to provide best-in class workforce solutions to organizations, colleges and individuals alike. Blitz jobs is the ultimate solution for all the HR problems organizations face. We are 100% committed to providing our clients with top notch solutions that leave them satisfied and accelerate their growth.

We have built our credibility by consistently performing well and have emerged as a service partner of choice with various organizations of repute like Peacock Solar and Eureka Forbes. We have a brilliant team of HR professionals and analysts. Our aim is to provide a significant competitive edge to our clients in a fast-growing market like India. We offer the right talent at the earliest possible and enable our clients to do better business every day.

We build enduring relationships with our clients and candidates and this approach empowers organizations to meet and exceed their business goals. We always commit to our motto, **"Employment for Everyone".** We have exemplary business practices and robust work ethics when it comes to our clients. In order to have our team work as seamlessly as possible, we also apply the same ethics to ourselves. Our team comprises of brilliant, talented people who work in a highly nurturing and productive environment where they get ample opportunities to grow as a professional and as an individual.

We are committed to providing superlative and satisfactory services to our clients and we need a highly motivated team for the job. Well, our team is more than up to the task and delivers consistently thanks to the wonderful guidance and support available to everyone.

CERTIFICATE OF COMPLETION

ACKNOWLEDGEMENTS

I acknowledge my sincere gratitude to Dr. Vijaya T G, Professor, Director, PSG Institute of Management for her kind patronage.

I also express my sincere gratitude to Mr. Harish V, Assistant Professor (Senior Grade), Department of Operations Management, PSG Institute of Management for his valuable guidance and long hours of inspirational suggestions, timely lessons and patience. His constant support and motivating guidance has helped me to stay focused and complete my report on time.

I am extremely happy to express my thankfulness to all my friends, my family members and the respectable faculty whose kindness and support throughout has helped me in completed the report work with finesse.

TABLE OF CONTENTS

Contents

5.INTRODUCTION

Blitz jobs is a Human resources solution to the startup companies, who faces difficulty in hiring and also seeking for reduction in costing of HR processes. Generally, companies take 15 to 20 days to screening a resume, call them and arrange an interview.

Pays 2 months' salary to the employee which is wastage of money due to the delay in hiring process. After investing this effort, time and money, still unable to get right talent with appropriate skills. Blitz Jobs introducing **INSTANT HIRING** concept, where committing to hiring within 6 days. Thus, reduces difficulty in hiring process.

Along with it, it also committing an zero notice period for an employer, who don't want to waste money for 2 months of poor quality work.

Employers can also get extra benefits like free on boarding formalities, exit formalities, payroll services etc. Through the portal name KREDILY Generally, consultancies take 8 to 10 cent fee of employee CTC from employer for every hiring. Where blitz jobs charges only 5 cents of employee CTC.

MISSION AND VISION

Massive outreach

We have a massive network of connections around the world. We are connected with the top talents and organizations, in fact, we have a comprehensive database of over 200+ candidates (from premier institutions like IIT, IIM) and 1000+ organizations contacts.

Invest in your success

We help you prepare for the selection process in many ways. Our dedicated team not only offers training programs, but we also guide you through every step of the process to ensure that you get the best.

Speed is our NorthStar

Gone are the days of anxious waiting that were a part of the hiring process. We ensure that it takes minimum time for the organizations to hire the perfect candidate without any fuss or loss of time and money.

PRODUCTS

FOR STARTUP AND MNC

- Talent Acquisition
- Talent Retention
- On-boarding Formalities
- Exit Formalities
- Third Party Payroll Services
- Performance Appraisal
- Attendance Management

- Training and Development

FOR COLLEGES

- Career Counselling
- Placement Services
- Resume Making
- Mock Interviews
- Skill Development Trainings
- HR Sessions

SKILLS DEVELOPMENT PROGRAMME

- Digital Marketing
- SEO
- Excel
- MySQL
- Python

Company believe in complete transparency and have a policy of open communication. Work life balance is essential and we at Blitz jobs are very aware. We ensure that the balance is maintained an workload is manageable. We also reward hard work and superior performance, that is, we believe in meritocracy. We also have training and development programs so that while working with us, not only do you gain valuable experience, but you also learn and grow.

6.BACKGROUND OF THE STUDY

SOURCE OF THE DATASET

Given from the company, so analysis should be done based on the given dataset.

Variable	Description
date Crawled	this ad was first crawled, all field-values are taken from this date
Name	name of the car
Seller	private or dealer
offer Type	offer type which car seller provides
Price	the price on the ad to sell the car
Abtest	abtest
vehicle Type	vehicle type
yearOfRegistration	at which year the car was first registered
Gearbox	gearbox of the car
powerPS	power of the car in PS
Model	model of the car
Kilometer	how many kilometers the car has driven
monthOfRegistration	at which month the car was first registered
fuelType	fuel type of the car
Brand	car brand
notRepairedDamage	if the car has a damage which is not repaired yet
dateCreated	the date for which the ad at ebay was created
nrOfPictures	number of pictures in the ad
postalCode	postal code
lastSeen	the crawler saw this ad last online

OBJECTIVE OF THE PROJECT

- This dataset has information for about 3,70,000 used cars.
- The objective of this project is to understand the dataset, analyze the variables and predict the price of the used car from various

factors which determines the price of the used car.

- The result of this project will project whether the price is nominal or high based on few influencing factors. This project will help the seller to fix the price of the used cars, based on various factors which influence the price

DATA PREPROCESSING

Data preprocessing involves transforming data into a basic form that makes it easy to work with. One characteristics of a tidy dataset is that: one observation per row and one variable per column.

LITERATURE REVIEW

COMPANIES ADOPTING ANALYTICS WIDELY.

Arunachalam (2018), has done a detailed study on the previous business analytics papers and given an idea that many companies have not started to set the basic infrastructure for sensing a data. It should be done first. Systems like RFID and sensors should be installed in logistics, so that it pays way for performing analytics. He also asked companies to establish policies regarding sharing of information i.e. What should be shared and what should not be. The paper also tells that industries should be in ready mode by initiating the first level of analytics which is descriptive analytics which tells us what are all the things happening inside the company.so that it can transform to diagnostic analytics in next stage which helps to understand why it is happening in the company. **Ray.Y. Zhong (2016),** mentioned that it takes 635 years to process 1K petabytes, he also showed his worries about processing of the bigdata. He tells that handling such a huge data requires very huge processors and technologies which is still creates

more challenges additionally. Paper also talks about the future technologies like smart cloud-based infrastructure to store a data in huge sum and also Intelligent processor to process the data to get valuable insights. Self-learning models have the capability of learning by themselves from the massive bigdata input. Deep machine learning (DML) will be embedded in the decision-making models, so that it can able to make decisions by own. These self-learning models will also act as a smart learning models by collaborating with parallel models and inputs.

HOW MACHINE LEARNING CHANGES NEW WORLD.

Bongsug chae (2014), paper tells about how twitter data can be extracted as a data and converted into insightful information. Generally, it classified into three methodologies, first one is descriptive analytics, here, study focus descriptive statistics, generally helpful in finding types of tweet, hashtag, number of tweets which can be useful in surveys. Secondly, content analytics (ca), which tells about Natural Language Processing (NLP), where unstructured data are cleaned using intelligence and data mining, which is further used in sentiment analysis. Thirdly, Network Analytics (NA) which extracts network information using network theory, and helps in building friendship network (using followers and following) and interpersonal relationship among twitter users. **Gang Wang (2016),** In this paper, he had drafted an frame work for a supply chain in that he classified logistics and management into two major categories which is operations and strategy. In operations it talks about the coordination between the functional area of an organization to control the cost effectively and look after the processes to increase the operational efficiency. In strategy, firstly, paper talks about how the firm is collaborative to share and protect the information and also conduct exploratory research. Second is Agile stagey which how quick the firm is responding to the changes. It tells all about monitoring and usage of analytics in efficient manner. Thirdly, sustainable strategy is telling about how the data in the

organization is gathered, analyzed data sustainably to bring out effective decisions in supply chain management.

7.PROJECT METHODOLOGY AND STEPS OF EXECUTION

UNDERSTANDING THE DATA

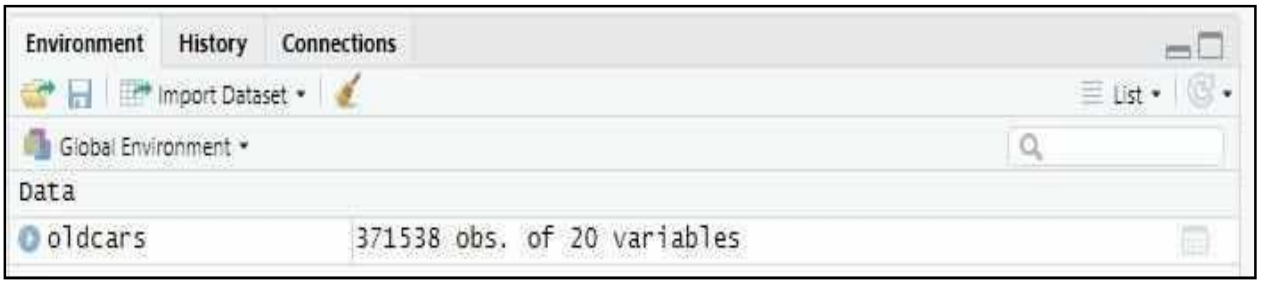
Any machine learning models that you build are only as good as the data that you provide them. The first step in understanding your data is to actually look at some raw values and calculate some basic statistics.

LOAD THE REQUIRED PACKAGES

No need for any packages for understanding the data.

LOAD THE DATASET

Code: oldcars<- read.csv(file.choose(),header=T)



VIEW THE CLASS

Code: class(oldcars)

DIMENSION OF DATASET

Code: dim(oldcars)

NO OF ROWS AND COLUMNS

Code: nrow(oldcars)

ncol(oldcars)

```
> class(oldcars)
[1] "data.frame"
> dim(oldcars)
[1] 371538     20
> nrow(oldcars)
[1] 371538
> ncol(oldcars)
[1] 20
```

VIEW THE COLUMN NAME

Code: names(oldcars)

```
> names(oldcars)
 [1] "dateCrawled"        "name"             "seller"           "offerType"        "price"
 [6] "abtest"             "vehicleType"      "yearOfRegistration" "gearbox"        "powerPS"
[11] "model"              "kilometer"        "monthOfRegistration" "fuelType"      "brand"
[16] "notRepairedDamage"  "dateCreated"      "nrOfPictures"     "postalCode"       "lastSeen"
```

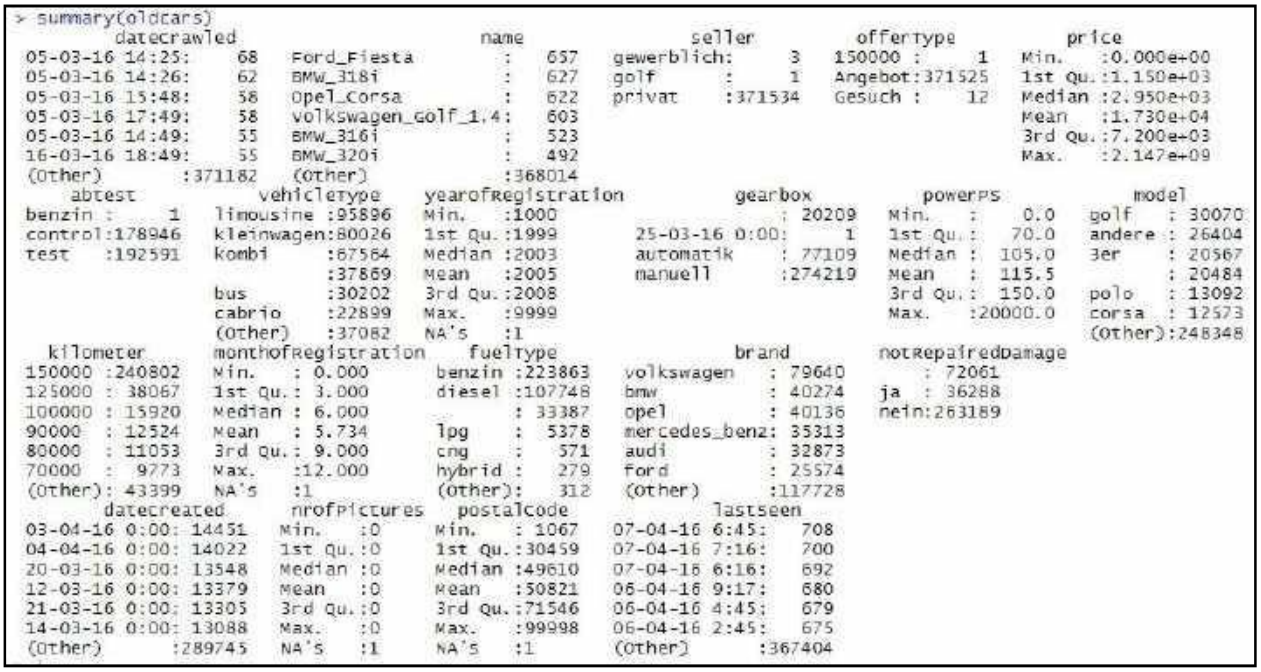
VIEW THE STRUCTURE OF THE DATASET

Code: str(oldcars)

```
> str(oldcars)
'data.frame':   371538 obs. of  20 variables:
 $ dateCrawled        : Factor w/ 15623 levels "01-04-16 0:06",..: 11709 11687 6744 8386 15351 1790 347 10384 1985 8233
 ...
 $ name               : Factor w/ 233525 levels "'Showcar_und_Messefahrzeug'_Opel_Astra_G_Cabrio",..: 80264 4563 92307 8
2386 172980 28930 147768 215521 64563 217950 ...
 $ seller             : Factor w/ 3 levels "gewerblich","golf",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ offerType          : Factor w/ 3 levels "150000","Angebot",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ price              : int  480 18300 9800 1500 3600 650 2200 0 14500 999 ...
 $ abtest             : Factor w/ 3 levels "benzin","control",..: 3 3 3 3 3 3 3 3 2 3 ...
 $ vehicleType        : Factor w/ 10 levels "","andere","bus",..: 1 5 9 6 6 8 4 8 3 6 ...
 $ yearOfRegistration : int  1993 2011 2004 2001 2008 1995 2004 1980 2014 1998 ...
 $ gearbox            : Factor w/ 4 levels "","25-03-16 0:00",..: 4 4 3 4 4 4 4 4 4 4 ...
 $ powerPS            : int  0 190 163 75 69 102 109 50 125 101 ...
 $ model              : Factor w/ 253 levels "","1_reihe","100",..: 121 1 122 121 106 13 9 43 59 121 ...
 $ kilometer          : Factor w/ 14 levels "10000","100000",..: 4 3 3 4 14 4 4 8 7 4 ...
 $ monthOfRegistration: int  0 5 8 6 7 10 8 7 8 0 ...
 $ fuelType           : Factor w/ 8 levels "","andere","benzin",..: 3 5 5 3 5 3 3 3 3 1 ...
 $ brand              : Factor w/ 41 levels "","alfa_romeo",..: 40 3 16 40 33 4 27 40 12 40 ...
 $ notRepairedDamage  : Factor w/ 3 levels "","ja","nein": 1 2 1 3 3 2 3 3 1 1 ...
 $ dateCreated        : Factor w/ 115 levels "","01-02-16 0:00",..: 92 92 57 67 115 16 4 82 16 67 ...
 $ nrOfPictures       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ postalCode         : int  70435 66954 90480 91074 60437 33775 67112 19348 94505 27472 ...
 $ lastSeen           : Factor w/ 18706 levels "","01-04-16 0:15",..: 4476 4404 3000 9883 3697 3796 3075 14670 2520 1835
1 ...
```

VIEW THE SUMMARY OF THE DATASET

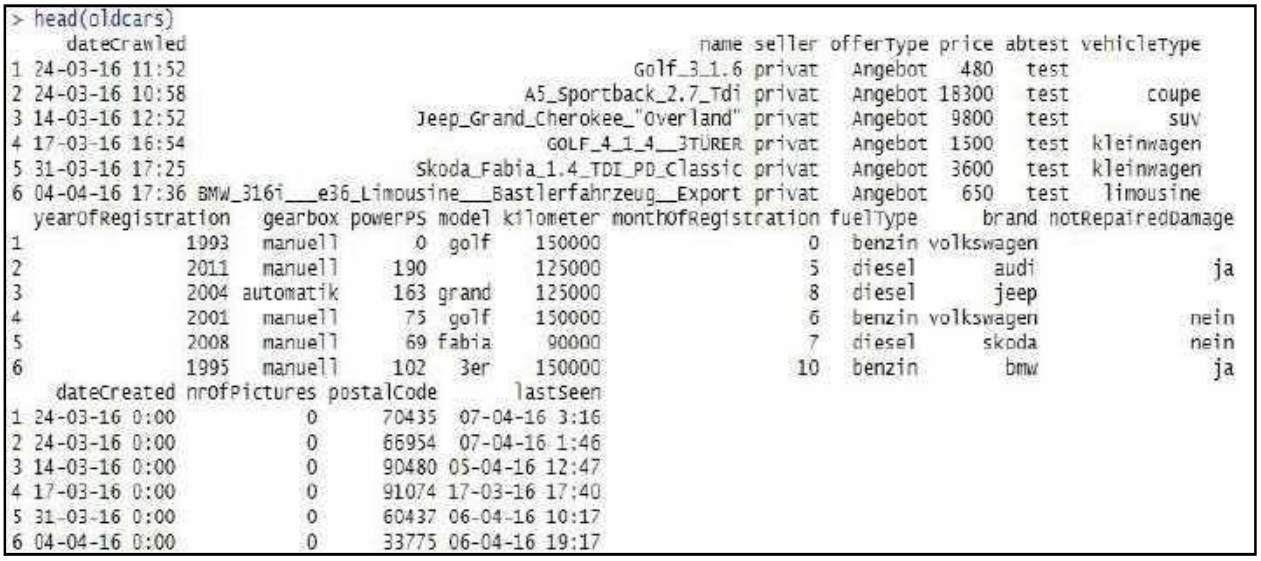
Code: summary(oldcars)



```
> summary(oldcars)
       dateCrawled                 name                seller          offerType          price
 05-03-16 14:25:    68   Ford_Fiesta        :   657   gewerblich:     3   150000 :     1   Min.   :0.000e+00
 05-03-16 14:26:    62   BMW_318i           :   627   golf      :     1   Angebot:371525   1st Qu.:1.150e+03
 05-03-16 15:48:    58   Opel_Corsa         :   622   privat    :371534   Gesuch :    12   Median :2.950e+03
 05-03-16 17:49:    58   volkswagen_Golf_1.4:   603                                        Mean   :1.730e+04
 05-03-16 14:49:    55   BMW_316i           :   523                                        3rd Qu.:7.200e+03
 16-03-16 18:49:    55   BMW_320i           :   492                                        Max.   :2.147e+09
 (Other)       :371182   (Other)            :368014
    abtest           vehicleType    yearOfRegistration          gearbox          powerPS            model
 benzin :     1   limousine :95896   Min.   :1000                  : 20209   Min.   :    0.0   golf    : 30070
 control:178946   kleinwagen:80026   1st Qu.:1999       25-03-16 0:00:     1   1st Qu.:   70.0   andere  : 26404
 test   :192591   kombi     :67564   Median :2003       automatik     : 77109   Median :  105.0   3er     : 20567
                              :37869   Mean   :2005       manuell       :274219   Mean   :  115.5             : 20484
                  bus       :30202   3rd Qu.:2008                                 3rd Qu.:  150.0   polo    : 13092
                  cabrio    :22899   Max.   :9999                                 Max.   :20000.0   corsa   : 12573
                  (Other)   :37082   NA's   :1                                                      (Other):248348
  kilometer      monthOfRegistration    fuelType                brand          notRepairedDamage
 150000 :240802   Min.   : 0.000      benzin :223863   volkswagen    : 79640          : 72061
 125000 : 38067   1st Qu.: 3.000      diesel :107748   bmw           : 40274   ja     : 36288
 100000 : 15920   Median : 6.000             : 33387   opel          : 40136   nein   :263189
 90000  : 12524   Mean   : 5.734      lpg    :  5378   mercedes_benz: 35313
 80000  : 11053   3rd Qu.: 9.000      cng    :   571   audi          : 32873
 70000  :  9773   Max.   :12.000      hybrid :   279   ford          : 25574
 (Other): 43399   NA's   :1           (Other):   312   (Other)       :117728
      dateCreated      nrOfPictures   postalCode          lastSeen
 03-04-16 0:00: 14451   Min.   :0   Min.   : 1067   07-04-16 6:45:   708
 04-04-16 0:00: 14022   1st Qu.:0   1st Qu.:30459   07-04-16 7:16:   700
 20-03-16 0:00: 13548   Median :0   Median :49610   07-04-16 6:16:   692
 12-03-16 0:00: 13379   Mean   :0   Mean   :50821   06-04-16 9:17:   680
 21-03-16 0:00: 13305   3rd Qu.:0   3rd Qu.:71546   06-04-16 4:45:   679
 14-03-16 0:00: 13088   Max.   :0   Max.   :99998   06-04-16 2:45:   675
 (Other)      :289745   NA's   :1   NA's   :1       (Other)       :367404
```

VIEW THE HEAD OF THE DATASET

Code: head(oldcars)



```
> head(oldcars)
     dateCrawled                                                name seller offerType price abtest vehicleType
1 24-03-16 11:52                                      Golf_3_1.6 privat   Angebot   480   test
2 24-03-16 10:58                                A5_Sportback_2.7_Tdi privat   Angebot 18300   test       coupe
3 14-03-16 12:52                       Jeep_Grand_Cherokee_"Overland" privat   Angebot  9800   test         suv
4 17-03-16 16:54                                GOLF_4_1_4__3TÜRER privat   Angebot  1500   test  kleinwagen
5 31-03-16 17:25                       Skoda_Fabia_1.4_TDI_PD_Classic privat   Angebot  3600   test  kleinwagen
6 04-04-16 17:36 BMW_316i___e36_Limousine___Bastlerfahrzeug__Export privat   Angebot   650   test   limousine
  yearOfRegistration   gearbox powerPS model kilometer monthOfRegistration fuelType      brand notRepairedDamage
1               1993   manuell       0  golf    150000                  0   benzin volkswagen
2               2011   manuell     190           125000                  5   diesel       audi              ja
3               2004 automatik     163 grand     125000                  8   diesel       jeep
4               2001   manuell      75  golf    150000                  6   benzin volkswagen            nein
5               2008   manuell      69 fabia      90000                  7   diesel      skoda            nein
6               1995   manuell     102   3er    150000                 10   benzin        bmw              ja
    dateCreated nrOfPictures postalCode      lastSeen
1 24-03-16 0:00           0      70435  07-04-16 3:16
2 24-03-16 0:00           0      66954  07-04-16 1:46
3 14-03-16 0:00           0      90480 05-04-16 12:47
4 17-03-16 0:00           0      91074 17-03-16 17:40
5 31-03-16 0:00           0      60437 06-04-16 10:17
6 04-04-16 0:00           0      33775 06-04-16 19:17
```

VIEW THE TAIL OF THE DATASET

Code: tail(oldcars)

```
> tail(oldcars)
        dateCrawled                                                name seller offerType price
371533  21-03-16 9:50                                    Mitsubishi_Cold privat    Angebot     0
371534 14-03-16 17:48                          Suche_t4___vito_ab_6_sitze privat    Angebot  2200
371535 05-03-16 19:56        smart_smart_leistungssteigerung_100ps privat    Angebot  1199
371536 19-03-16 18:57          volkswagen_Multivan_T4_TDI_7DC_UY2 privat    Angebot  9200
371537 20-03-16 19:41                       VW_Golf_Kombi_1_9l_TDI privat    Angebot  3400
371538 07-03-16 19:39 BMW_M135i_vollausgestattet_NP_52.720_____Euro privat    Angebot 28990
        abtest vehicleType yearOfRegistration   gearbox powerPS       model kilometer
371533 control                           2005    manuell       0        colt    150000
371534    test                           2005                  0                20000
371535    test       cabrio              2000 automatik     101       fortwo    125000
371536    test          bus              1996    manuell     102 transporter    150000
371537    test        kombi              2002    manuell     100        golf    150000
371538 control    limousine              2013    manuell     320    m_reihe     50000
        monthOfRegistration fuelType          brand notRepairedDamage    dateCreated
371533                   7   benzin     mitsubishi                ja 21-03-16 0:00
371534                   1           sonstige_autos                   14-03-16 0:00
371535                   3   benzin          smart               nein 05-03-16 0:00
371536                   3   diesel     volkswagen               nein 19-03-16 0:00
371537                   6   diesel     volkswagen                    20-03-16 0:00
371538                   8   benzin            bmw               nein 07-03-16 0:00
        nrofPictures postalCode     lastseen
371533            0       2694 21-03-16 10:42
371534            0      39576  06-04-16 0:46
371535            0      26135 11-03-16 18:17
371536            0      87439  07-04-16 7:15
371537            0      40764 24-03-16 12:45
371538            0      73326  22-03-16 3:17
```

8.ALTERNATIVES AND SOLUTIONS

CHI SQUARE ANALYSIS

To find weather there is any relationship between the categorical variables.

TEST OF INDEPENDENCE.

Null hypothesis: Variables are independent in nature.

Chi-square summary data

	0	1	
0	644	374	1018
1	772	400	1172
	1416	774	2190

Expected Values

	0	1	Total
0	658.214	359.786	1018
1	757.786	414.214	1172
Total	1416	774	2190

Chi-Square Test

SUMMARY		Alpha	0.05
Count	*Rows*	*Cols*	*df*
2190	2	2	1

CHI-SQUARE

	chi-sq	*p-value*	*x-crit*	*sig*	*Cramer V*	*Odds Ratio*
Pearson's	1.62281	0.2027	3.84146	no	0.02722	0.89219
Max likelil	1.62181	0.20284	3.84146	no	0.02721	0.89219

INFERENCE

- Variables taken are Ab test and Fuel type.
- After conducting test, found that test is failed to reject null hypothesis and p value is greater 0.05.
- So we concluded there is no significant relation between the independent variables and it shown no sign of multicollinearity.

ONE WAY ANOVA TEST.

To find weather there is any relationship between the categorical and continuous variables.

Null hypothesis: there is no significant relation between variables.

ANOVA: Single Factor

DESCRIPTION					Alpha	0.05		
Group	*Count*	*Sum*	*Mean*	*Variance*	*SS*	*Std Err*	*Lower*	*Upper*
price	2190	1.11E+08	50687.93	4.57E+12	1E+16	26377.95	-1021.42	102397.3
kilometer	2190	2.8E+08	127840.2	1.74E+09	3.8E+12	26377.95	76130.83	179549.5

ANOVA

Sources	*SS*	*df*	*MS*	*F*	*P value*	*F crit*	*RMSSE*	*Omega Sq*
Between G	1.81E+13	1	1.81E+13	11.904	0.000564	3.842875	0.044195	0.001657
Within Grc	1E+16	6568	1.52E+12					
Total	1E+16	6569	1.53E+12					

INFERENCE

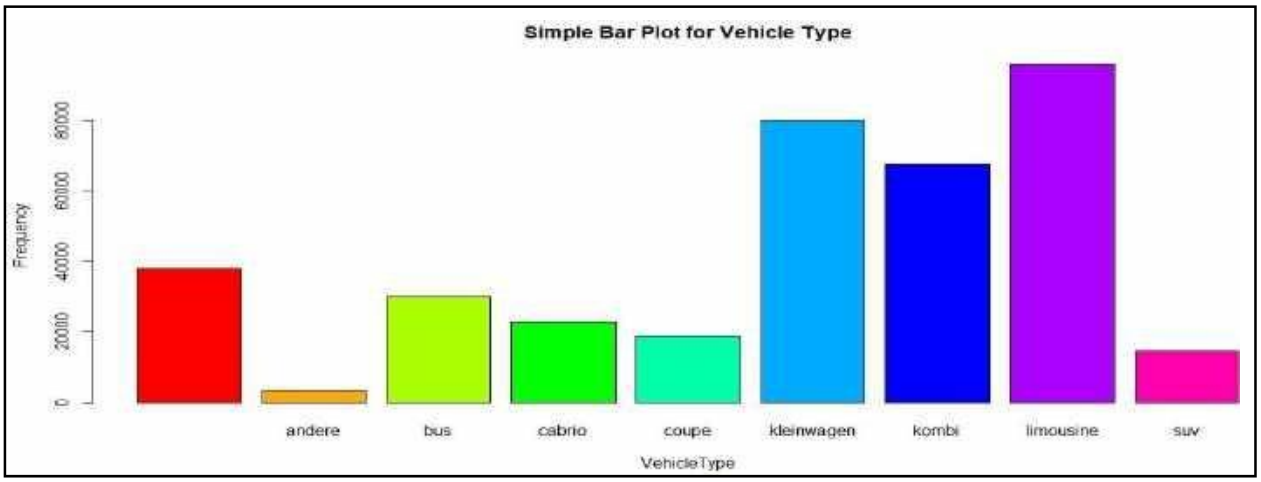
- Variables taken are kilometers (independent variable) and price (dependent variable)
- Result shows there f value is greater than f critical and P value is less than 0.05.
- So, we reject the null hypothesis and conclude that kilometer is having significant impact on price.

BARPLOT

Code: type<-table(oldcars$vehicleType)

barplot(type, xlab = "VehicleType", main = "Simple Bar Plot for Vehicle Type",

ylab ="Frequency",col = rainbow(9))



Inference:

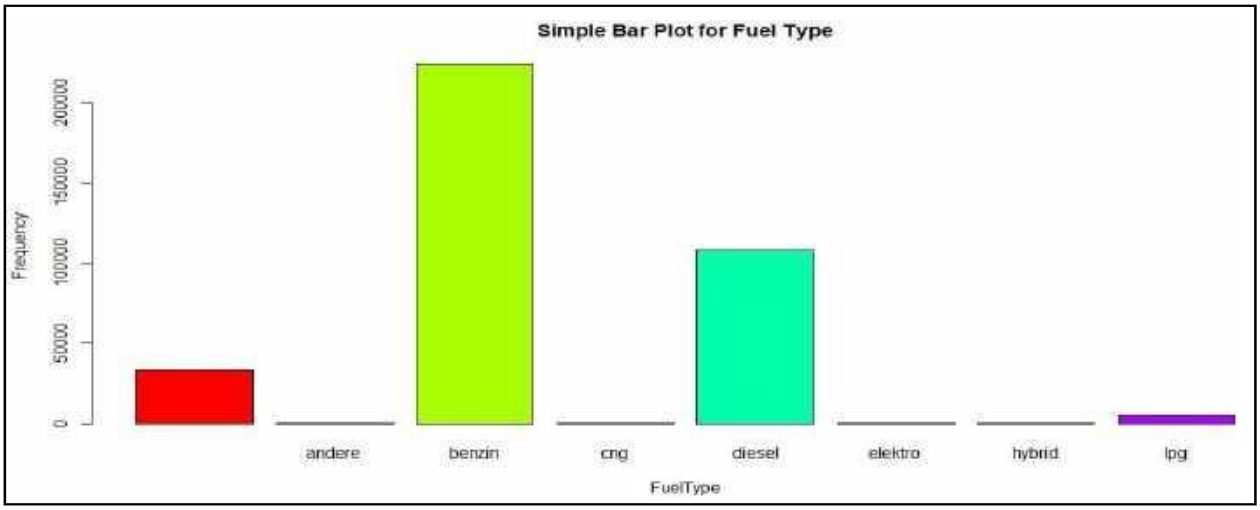
- Limousine vehicle type is the greatest number of vehicles in this dataset.
- SUV, bus and cabrio has vehicles between 20000 and 40000.
- Very few vehicle type in this dataset is andere type.

Code:

type<-table(oldcars$fuelType)

barplot(type, xlab = "FuelType", main = "Simple Bar Plot for Fuel Type", ylab

="Frequency",col = rainbow(9))



Inference:

- The most number of vehicles has benzon fuel type.
- Around 1 lakh vehicles have diesel fuel type.
- Very few cars has lpg, cng, electric and hybrid fuel types.

Code: counts<-table(oldcars$gearbox,oldcars$abtest)

barplot(counts, xlab = "abtest", main = "Car Distribution by gearbox and abtest",

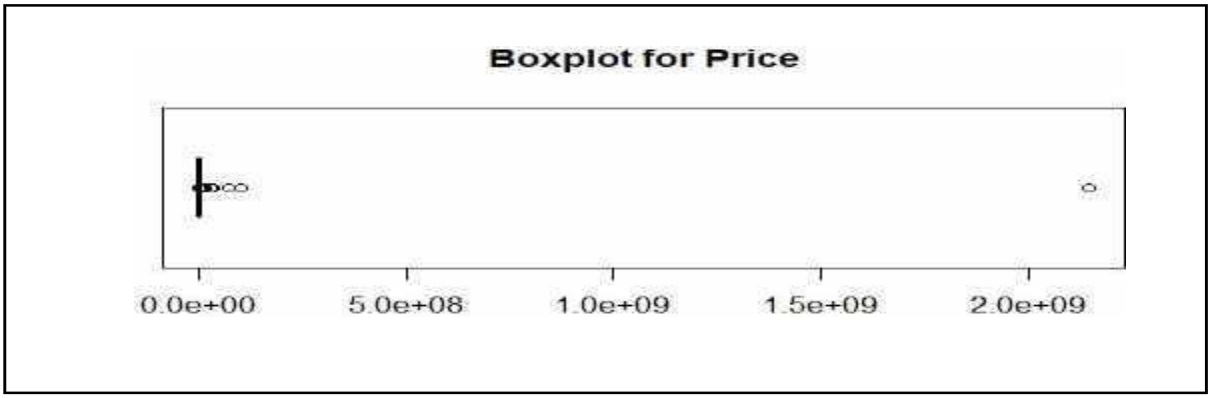
ylab ="Frequency",legend=rownames(counts),col = rainbow(3), beside = TRUE)

Inference:

- No of manual gearbox cars is very high when compared to automatic type in both the abtest types.
- Very few cars haven't disclosed their gearbox type.
- No of automatic cars which has control abtest is quiet higher than the automatic cars which has test abtest

BOXPLOT

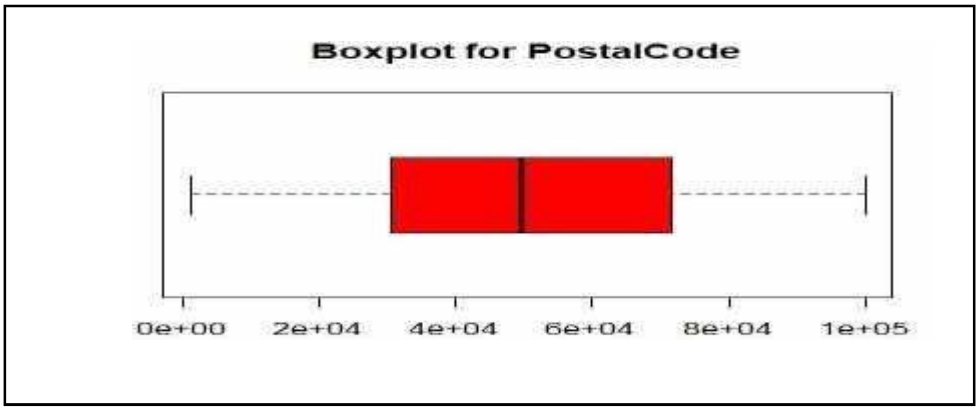
Code: boxplot(oldcars$price,col="red", main ="Boxplot for Price",horizontal = TRUE)



Inference:

- Outlier exists.

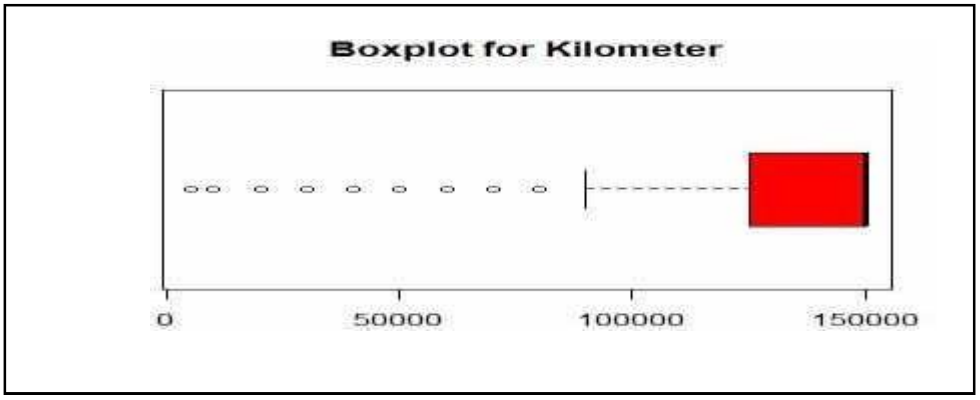
Code: boxplot(oldcars$postalCode,col="red", main ="Boxplot for PostalCode",horizontal = TRUE)



Inference:

- Outlier not exist.

Code: boxplot(oldcars$kilometer,col="red", main ="Boxplot for Kilometer",horizontal = TRUE)



Inference:

- Outlier exist.

PIE PLOT

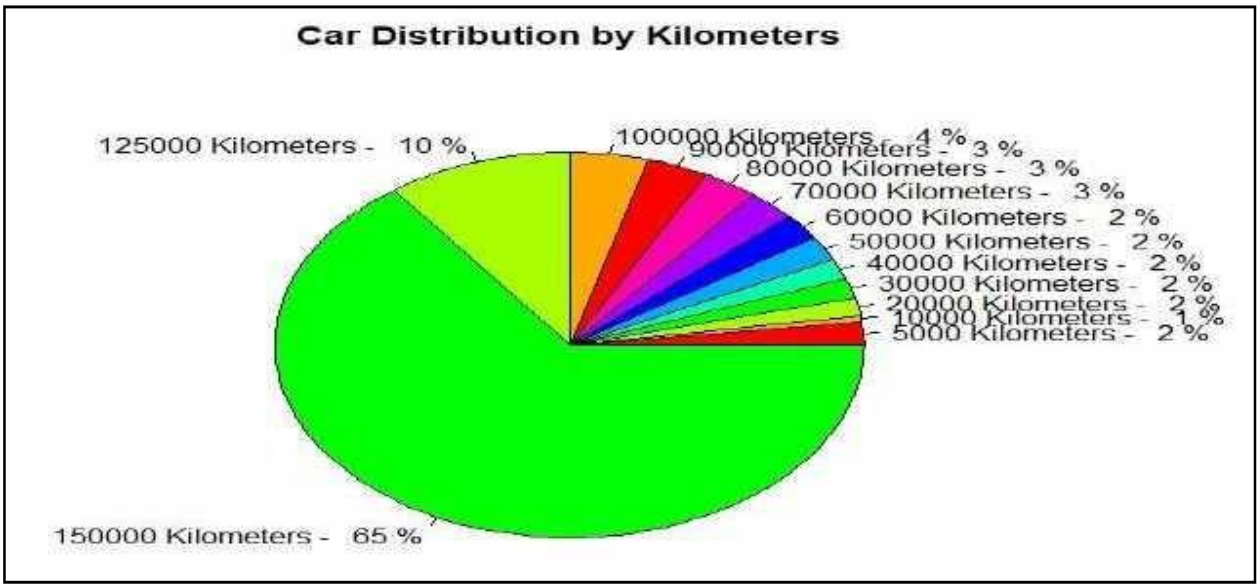
gear_count<- table(oldcars$kilometer)

Percentage_calc<- round(gear_count/sum(gear_count)*100)

samp_label<- paste(rownames(gear_count),"Kilometers -"," ",Percentage_calc,"%")

pie(gear_count,samp_label,main="Car Distribution by Kilometers", col= rainbow(9))

Output:

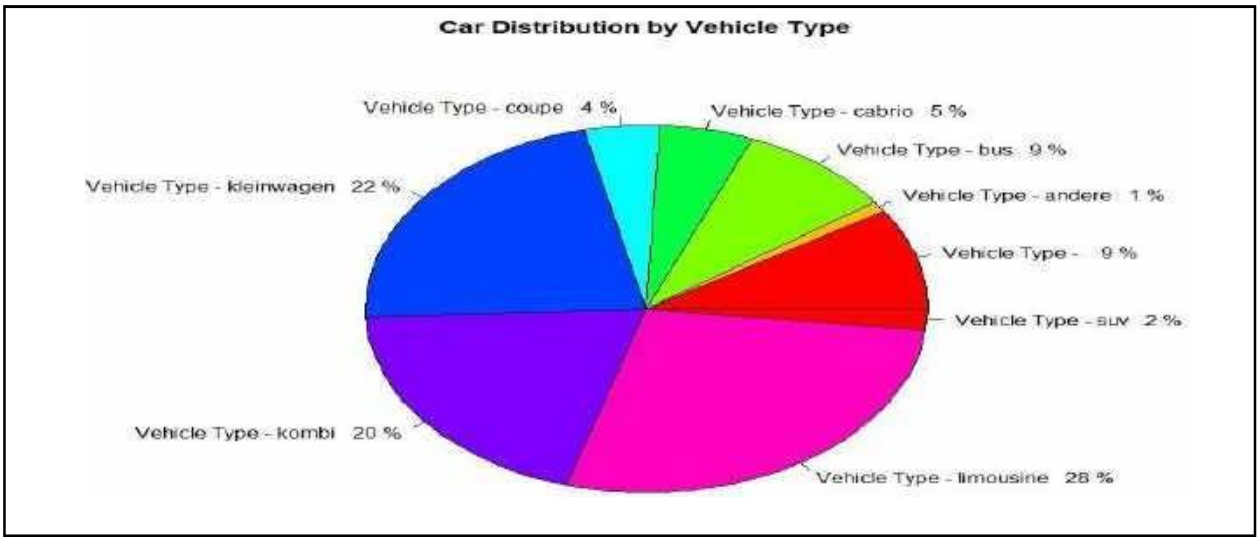


Inference:

- 65% of the cars in this dataset has 150000 kilometers.
- 10% of the cars in this dataset has 125000 kilometers.
- 4% of the cars in this dataset has 100000 kilometers.
- Rest all the cars in this dataset has kilometers less than 90000

Code:

```
library(plotrix)

gear_count<- table(oldcars$vehicleType)

Percentage_calc<- round(gear_count/sum(gear_count)*100)

samp_label<- paste("Vehicle Type -",rownames(gear_count)," ",Percentage_calc,"%")

pie(gear_count,samp_label,

   main="Car Distribution by Vehicle Type", col= rainbow
```



Inference:

- 28% of the cars in this dataset has limousine vehicle type.
- SUV Type cars are very few in this dataset.
- Around 20% of the cars are Kombi and kleinwagen type cars each.
- 5% of the cars are couple and cabrio type cars in this dataset.

DEALING WITH OUTLIERS

Code to treat outliers:

```
boxplot(oldcars$price, horizontal = TRUE)

x<-oldcars$price

qnt<-quantile(x, probs = c(.25,.75), na.rm = T)

caps<-quantile(x, probs = c(.05,.95), na.rm = T)

H<-1.5*IQR(x,na.rm = T)

x[x<(qnt[1]-H)]<-caps[1]

x[x>(qnt[2]+H)]<-caps[2]

price<-x

boxplot(price,main="Boxplot for Price (Treated)", col="green", horizontal = TRUE)
```

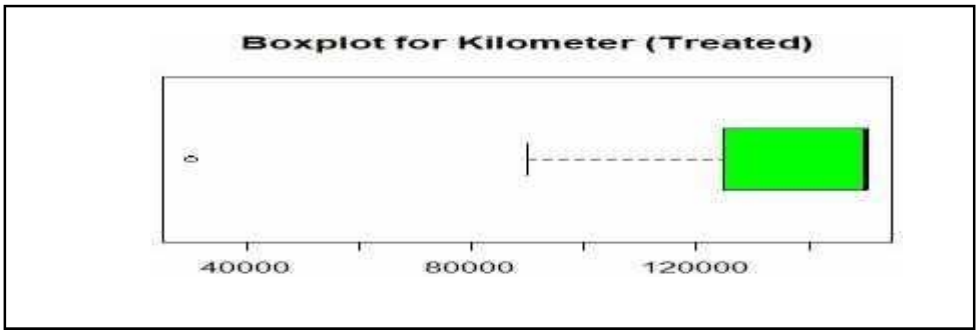
Inference:

- Outliers has been

treated. Code to treat outliers:

```
boxplot(oldcars$kilometer, horizontal = TRUE)
x<-oldcars$kilometer
qnt<-quantile(x, probs = c(.25,.75), na.rm = T)
caps<-quantile(x, probs = c(.05,.95), na.rm = T)
H<-1.5*IQR(x,na.rm = T)
x[x<(qnt[1]-H)]<-caps[1]
x[x>(qnt[2]+H)]<-caps[2]
kilo<-x
boxplot(kilo,main="Boxplot for Kilometer (Treated)", col="green", horizontal =TRUE)
```

Output:



Inference:

- Outlier has been treated.

DEALING WITH MISSING VALUES

Code:

```
library(DataExplorer)

any(is.na(mtcars[]))
```

Output:

```
[1] FALSE
```

Code:

```
sum(is.na(oldcars[]))

colSums(is.na(oldcars))
```

Output:

```
> sum(is.na(oldcars[]))
[1] 0
> colSums(is.na(oldcars))
       dateCrawled               name             seller          offerType
                 0                  0                  0                  0
             price             abtest        vehicleType  yearOfRegistration
                 0                  0                  0                  0
           gearbox            powerPS              model          kilometer
                 0                  0                  0                  0
monthOfRegistration           fuelType              brand   notRepairedDamage
                 0                  0                  0                  0
       dateCreated        nrOfPictures         postalCode           lastSeen
                 0                  0                  0                  0
```

Inference:

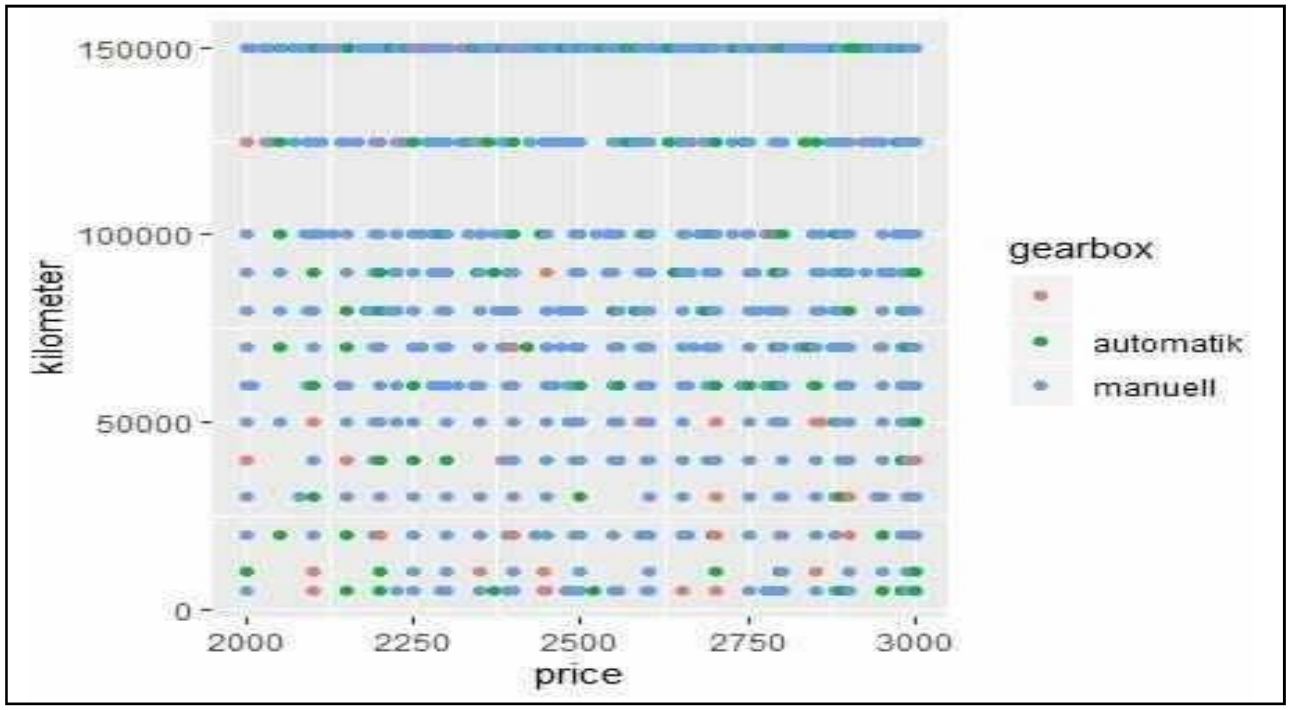
- The above result clearly shows that there are no NA values in the dataset.

DATA VISUALIZATION USING GGPLOTS

Code:

```
ggplot(data = oldcars) + geom_point(mapping = aes(x=price, y=kilometer, colour=gearbox))
```

Output:

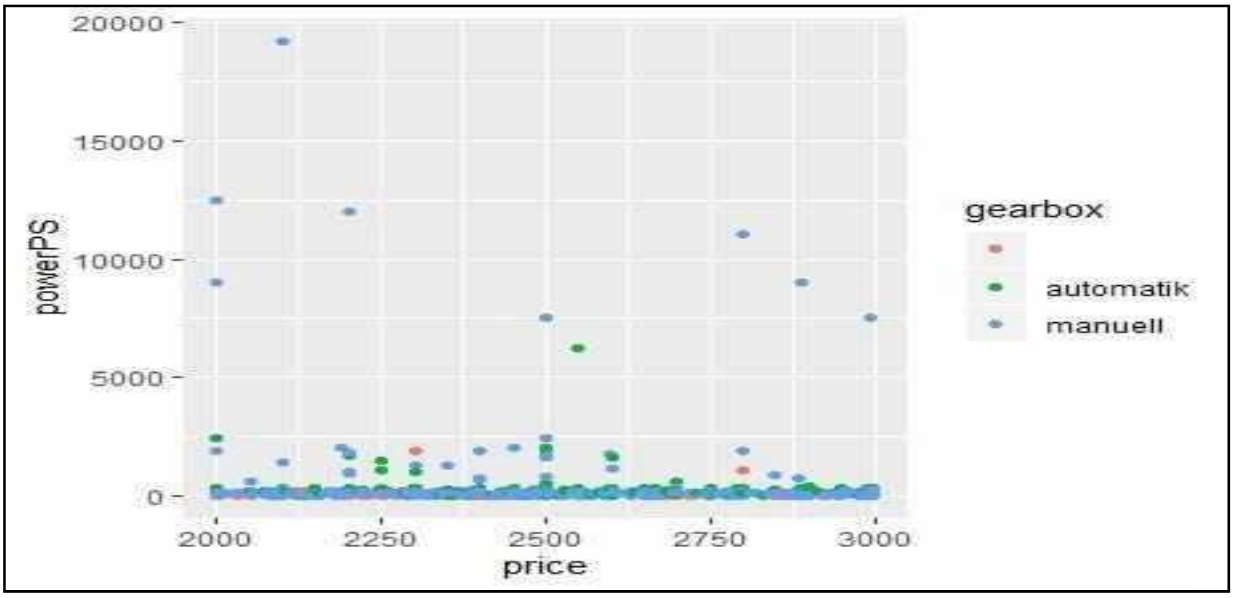


Inference:

- Price of automatic cars are higher than that of manual cars
- There is huge number of manual cars for kilometer greater than 100000.
- Price is very high for automatic cars irrespective of the kilometers of the used car.
- Number of used cars which has kilometer less than 100000 is very high.

Code:

```
ggplot(data = oldcars) + geom_point(mapping = aes(x=price, y=powerPS,
colour=gearbox)) Output:
```



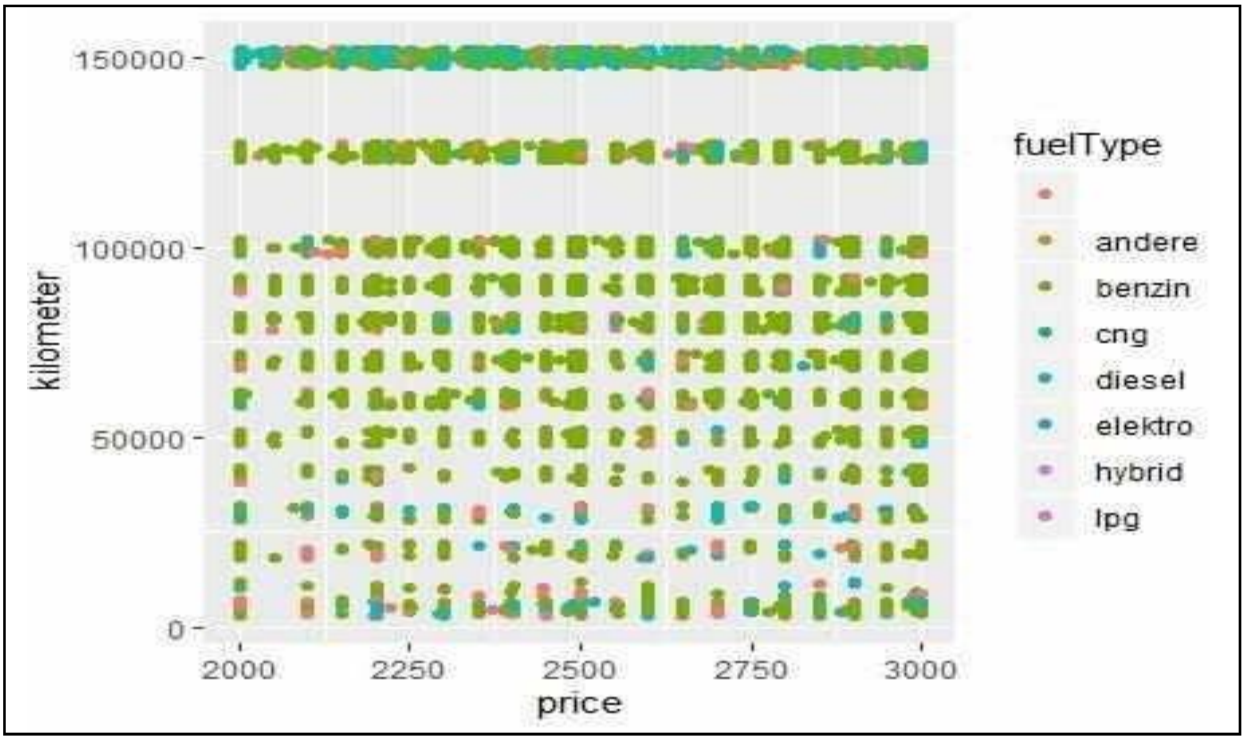
Inference:

- Most of the cars has PowerPS less than 2500
- Few cars which has higher PowerPS and manual gearbox, the price is high.
- None of the automatic gearbox cars has PowerPS greater than 7000
- The car which has highest PowerPS having manual gearbox, the price is lower.

Code:

```
ggplot(data = oldcars) + geom_point(mapping=aes(x=price ,

y=kilometer,colour=fuelType),position = "jitter") + labs(x="price", y="kilometer")
```

Output:



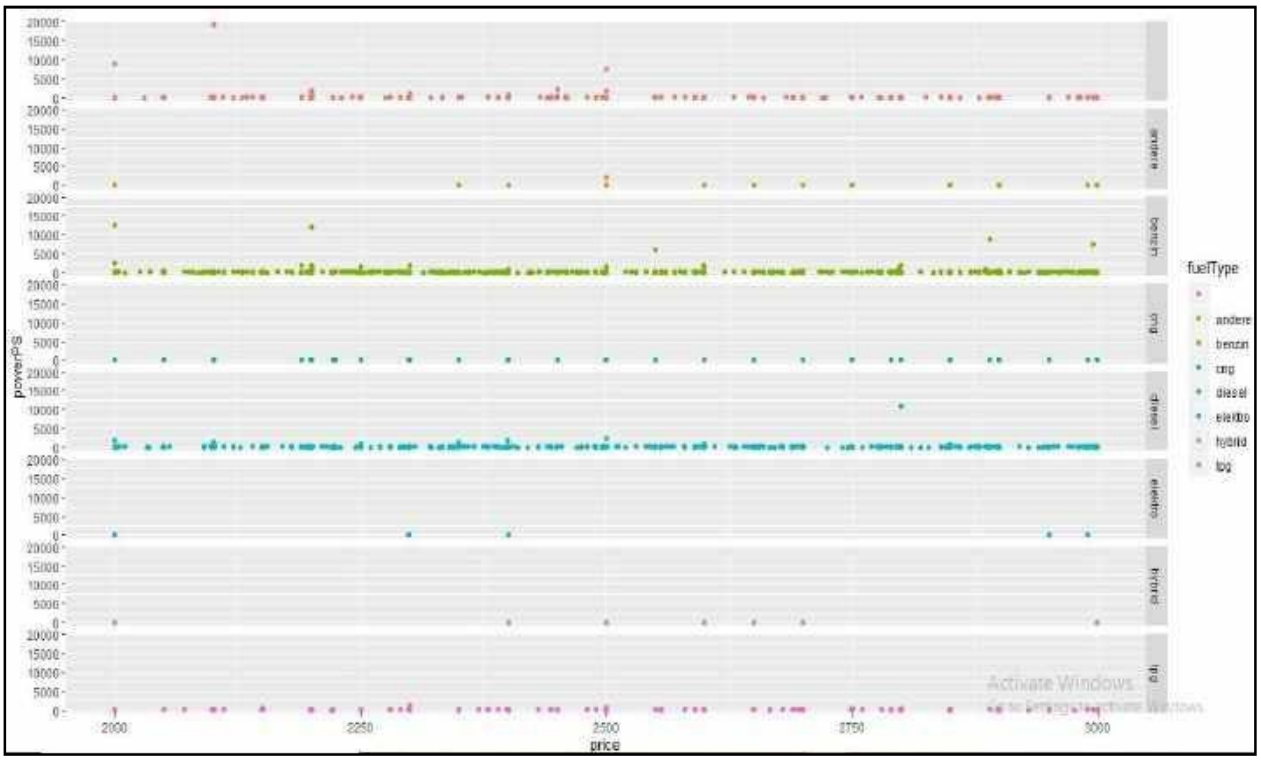
Inference:

- The price of the lpg fuel type cars price is not dependant on the kilometers.
- Most number of cars has benzin fuel type and when kilometers increases, price gradually increases.
- Most of the diesel fuel type cars have more kilometers when compared to the rest all fuel types.
- Electric fuel type cars price is quiet high when compared to benzin fuel type cars with same kilometers.

Code:

```
ggplot(data = oldcars) + geom_point(mapping=aes(x=price ,
    y=powerPS, colour=fuelType)) + facet_grid(oldcars$fuelType) +
    labs(x="price", y="powerPS")
```

Output:



Inferences:

- None of the hybrid and electric type cars has PowerPS greater than 2500.
- Benzin fuel type cars has the highest PowerPS.
- Price of the diesel type cars are equally distributed.
- The number of cng type cars is low when compared to benzin and diesel but price of cng type cars is higher than the price of those counterpart in diesel type cars.
- Few cars they haven't mentioned the type of the fuel type and those cars price are equally distributed.
- Andere fuel type cars have PowerPS less than 2000

MODELING

3.1 LINEAR REGRESSION

Code:

```
#linear regression

oldcars<- read.csv(file.choose(),header=T)

names(oldcars)

str(oldcars)

oldcars$fuelType<-as.numeric(oldcars$fuelType)

oldcars$abtest<-as.numeric(oldcars$abtest)

oldcars$yearOfRegistration<-as.numeric(oldcars$yearOfRegistration)

oldcars$model<-as.numeric(oldcars$model)

input<-
oldcars[c("price","powerPS","kilometer","fuelType","yearOfRegistration","abtest","model")]

print(head(input))

model<-lm(price~powerPS+kilometer+fuelType+yearOfRegistration+abtest+model,data=input)

print(model)

summary(model)
```

Output:

```
> print(head(input))
  price powerPS kilometer fuelType yearOfRegistration abtest model
1    20       2     20000        3                2000      2     2
2  4970       2      5000        6                2012      3     2
3  4970       2     10000        6                2012      3     2
4  1300       2     40000        3                2000      3    18
5  1300       2     50000        3                2005      3    18
6  1300       2     10000        3                2002      2    18
```

```
> print(model)

Call:
lm(formula = price ~ powerPS + kilometer + fuelType + yearOfRegistration +
    abtest + model, data = input)

Coefficients:
       (Intercept)             powerPS           kilometer            fuelType
         -1.448e+06           6.673e+01           8.073e-03          -9.309e+00
yearOfRegistration             abtest               model
          7.210e+02           1.090e+02          -1.461e+01
```

```
> summary(model)

Call:
lm(formula = price ~ powerPS + kilometer + fuelType + yearOfRegistration +
    abtest + model, data = input)

Residuals:
     Min       1Q   Median       3Q      Max
-17111.4  -1728.9    -90.4   1956.9  13019.6

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.448e+06  1.905e+04 -75.996  < 2e-16 ***
powerPS            6.673e+01  1.043e+00  63.985  < 2e-16 ***
kilometer          8.073e-03  2.414e-03   3.344 0.000829 ***
fuelType          -9.309e+00  4.146e+01  -0.225 0.822348
yearOfRegistration 7.210e+02  9.478e+00  76.069  < 2e-16 ***
abtest             1.090e+02  6.764e+01   1.611 0.107120
model             -1.461e+01  3.982e+00  -3.669 0.000245 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3019 on 7992 degrees of freedom
Multiple R-squared:  0.5714,    Adjusted R-squared:  0.5711
F-statistic:  1776 on 6 and 7992 DF,  p-value: < 2.2e-16
```

Analysis:

- More the starts more the significant.
- Based on the above intercept and coefficient values, we create the mathematicalequation

Y=a+XpowerPS*x1+Xkilometer*x2+XyearofRegistration*x3+Xmodel*x4

9.RECOMMENDATIONS

- In this project we have built Linear regression model, using Used car dataset
- Accuracy of Linear regression model is 57.14 %
- As we can see powerPS, kilometer, year of registration and model impacts more on the dependant variable and they are more significant variables.
- R- square value is 0.5714 so this model is less robust.
- p value is less than 0.05.
- ab test and fueltype are less significant as per the model.
- So seller need to mainly focus on the cars on the significant variables mentioned above.
- Main focus should be on manual type cars rather than automatic because as we can clearly see that manual cars sales is much better than automatic cars.
- Seller should also focus on the yearofRegistration of the car, because it directly impacts the price of the car.

10. CONCLUSION

There are agreeably numerous constraints to overcome before analytics in present world. But in future, data is a future asset of every firm. Since tech companies like apple, google, Microsoft are already leading in the market, the big other firms are also beginning to enter analytics market with the entrepreneur spirits.

Before that we have plenty of queries like how we are going to connect all the consignment with internet connectivity? How entire industries going to be data driven?? But after seeing the trends shown in the paper, we confirm that INFORMATION is already become the fourth production factors in the industry.

Till date, we did not have complete access to most of the data in the market. And there is lot of shortage in skilled scientists due to the various combination of subjects. Some government regulations to access deep learning of a data is also a major constraint at present. Though we have now scarcity in talents, investments and availability of data, the values and insights we get from analytics will overthrow all this limitation and make huge footprints in the development and advancements of technology in future.

11. BIBLIOGRAPHY

1. Arunachalam, Deepak, Niraj Kumar, and John Paul Kawalek. "Understanding big data analytics capabilities: Unravelling the issues, challenges and implications for practice." *Transportation Research Part E: Logistics and Transportation Review* 114 (2018): 416-436.
2. Zhong, Ray Y., et al. "Big Data for management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives." *Computers & Industrial Engineering* 101 (2016): 572-591.
3. Wang, Gang, et al. "Big data analytics in future world: Certain investigations for research and applications." *International Journal of Production Economics* 176 (2016): 98-110.
4. Chae, Bongsug Kevin. "Insights from hashtag# supply chain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research." *International Journal of Production Economics* 165 (2015): 247-259.
5. Schoenherr, Tobias, and Cheri Speier-Pero. "Data science, predictive analytics, and big data in management: Current state and future potential." *Journal of Business Logistics* 36.1 (2015): 120-132.
6. Chen, Daniel Q., David S. Preston, and Morgan Swink. "How the use of big data analytics affects value creation in management." *Journal of Management Information Systems* 32.4 (2015): 4-39.

12.APPENDICES