R Notebook

Code

DAB501-002 - Project#1

Please find the below HTML Documentation of Project_1

Group Members

Name	Student ID
Nishi Shrivastava	0770047
Diksha Diksha	0794012
Shashank Tripathi	0789803
Preet Patel	0795875

Academic Integrity

The group has acknowledged St Clair's College's Academic Integrity policies in completing the project. The project is the result of our own findings and research.

Versions and Package Name

Package Name	Version
R	Version 4.1.2
RStudio	Version 2021.09.2
dplyr	version 2.1.1
gganimate	version 1.0.7
ggplot2	version 3.3.5
tidyverse	version 1.3.1
tibble	version 3.1.6
readr	version 2.1.2

purr	version 0.3.4
stringr	version 1.4.0
forcats	version 0.5.1

Dataset Summary

The data set has been picked from https://vincentarelbundock.github.io/Rdatasets/doc/MASS/Cars93.html (https://vincentarelbundock.github.io/Rdatasets/doc/MASS/Cars93.html)

- TitleName: **Data from 93 Cars on Sale in the USA in 1993**
- The data set represents the data of cars from the year 1993.Cars were selected at random from among 1993 passenger car models that were listed in both the Consumer Reports issue and the PACE Buying Guide.Pickup trucks and Sport/Utility vehicles were eliminated due to incomplete information in the Consumer Reports source. Duplicate models were listed at most once.
- Data were obtained from two sources, The 1993 Cars - Annual Auto Issue from Consumer Reports and PACE New Car & Truck 1993 Buying Guide.
- Cars93 dataset has total of 93 Observations and 27 Variables. Below is the detailed information about the variables

Car Sale dataframe containing :

1.Manufacturer : Information about the Car Manufacturer.

2.Model : Car Model.

3.Type: A factor with levels “Small”, “Sporty”, “Compact”, “Midsize”, “Large” and “Van”.

4.Min.Price : Minimum Price (in $1,000): price for a basic version.

5.Price : Midrange Price (in $1,000): average of Min.Price and Max.Price.

6.Max.Price: Maximum Price (in $1,000): price for “a premium version”.

7.MPG.city : City MPG (miles per US gallon by EPA rating).

8.MPG.highway : Highway MPG.

9.AirBags : Air Bags standard. Factor: none, driver only, or driver & passenger.

10.DriveTrain : Drive train type: rear wheel, front wheel or 4WD; (factor).

11.Cylinders : Number of cylinders (missing for Mazda RX-7, which has a rotary engine).

12.EngineSize : Engine size (litres).

13.Horsepower : Horsepower (maximum).

14.RPM : RPM (revs per minute at maximum horsepower).

15.Rev.per.mile : Engine revolutions per mile (in highest gear).

16.Man.trans.avail: Is a manual transmission version available? (yes or no, Factor).

17.Fuel.tank.capacity : Fuel tank capacity (US gallons).

18.Passengers : Passenger capacity (persons)

19.Length : Length (inches).

20.Wheelbase : Wheelbase (inches).

21.Width : Width (inches).

22.Turn.circle : U-turn space (feet).

23.Rear.seat.room: Rear seat room (inches) (missing for 2-seater vehicles).

24.Luggage.room : Luggage capacity (cubic feet) (missing for vans).

25.Weight : Weight (pounds).

26.Origin : Of non-USA or USA company origins? (factor).

27.Make : Combination of Manufacturer and Model (character).

1: Two plots displaying distribution of a single continuous variable.

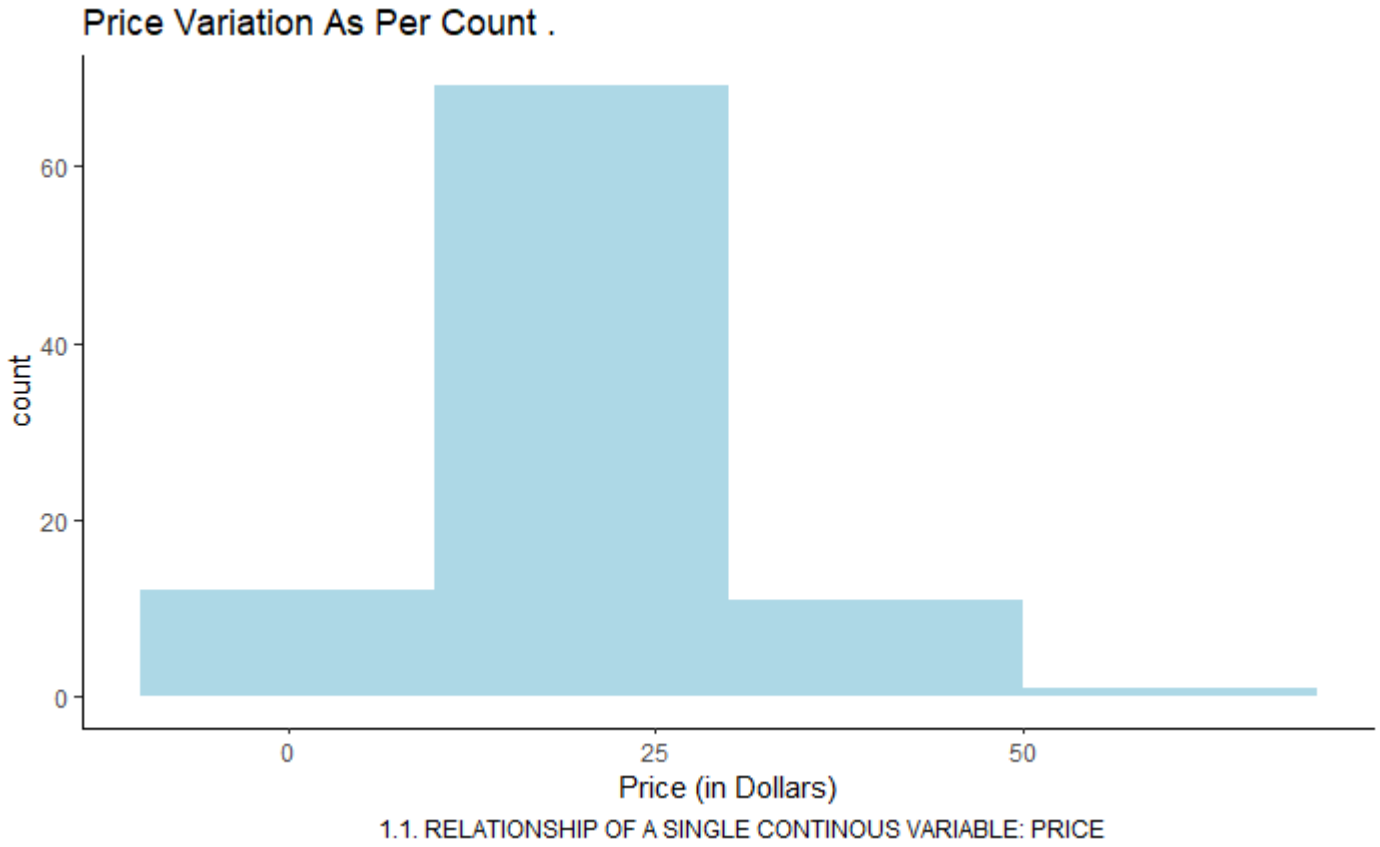
Hide

```
library(tidyverse)
summary(Cars93)
```

```
      ...1    Manufacturer          Model               Type             Min.Price         Price
 Min.   : 1   Length:93          Length:93          Length:93          Min.   : 6.70   Min.   :
7.40
 1st Qu.:24   Class :character   Class :character   Class :character   1st Qu.:10.80   1st Qu.:1
2.20
 Median :47   Mode  :character   Mode  :character   Mode  :character   Median :14.70   Median :1
7.70
 Mean   :47                                                           Mean   :17.13   Mean   :1
9.51
 3rd Qu.:70                                                           3rd Qu.:20.30   3rd Qu.:2
3.30
 Max.   :93                                                           Max.   :45.40   Max.   :6
1.90

   Max.Price       MPG.city      MPG.highway      AirBags           DriveTrain         Cylinders
 Min.   : 7.9   Min.   :15.00   Min.   :20.00   Length:93          Length:93          Length:93
 1st Qu.:14.7   1st Qu.:18.00   1st Qu.:26.00   Class :character   Class :character   Class :cha
racter
 Median :19.6   Median :21.00   Median :28.00   Mode  :character   Mode  :character   Mode  :cha
racter
 Mean   :21.9   Mean   :22.37   Mean   :29.09
 3rd Qu.:25.3   3rd Qu.:25.00   3rd Qu.:31.00
 Max.   :80.0   Max.   :46.00   Max.   :50.00

   EngineSize      Horsepower         RPM        Rev.per.mile  Man.trans.avail    Fuel.tank.capa
city
 Min.   :1.000   Min.   : 55.0   Min.   :3800   Min.   :1320   Length:93          Min.   : 9.20
 1st Qu.:1.800   1st Qu.:103.0   1st Qu.:4800   1st Qu.:1985   Class :character   1st Qu.:14.50
 Median :2.400   Median :140.0   Median :5200   Median :2340   Mode  :character   Median :16.40
 Mean   :2.668   Mean   :143.8   Mean   :5281   Mean   :2332                      Mean   :16.66
 3rd Qu.:3.300   3rd Qu.:170.0   3rd Qu.:5750   3rd Qu.:2565                      3rd Qu.:18.80
 Max.   :5.700   Max.   :300.0   Max.   :6500   Max.   :3755                      Max.   :27.00

   Passengers        Length        Wheelbase         Width        Turn.circle    Rear.seat.room
 Min.   :2.000   Min.   :141.0   Min.   : 90.0   Min.   :60.00   Min.   :32.00   Min.   :19.00
 1st Qu.:4.000   1st Qu.:174.0   1st Qu.: 98.0   1st Qu.:67.00   1st Qu.:37.00   1st Qu.:26.00
 Median :5.000   Median :183.0   Median :103.0   Median :69.00   Median :39.00   Median :27.50
 Mean   :5.086   Mean   :183.2   Mean   :103.9   Mean   :69.38   Mean   :38.96   Mean   :27.83
 3rd Qu.:6.000   3rd Qu.:192.0   3rd Qu.:110.0   3rd Qu.:72.00   3rd Qu.:41.00   3rd Qu.:30.00
 Max.   :8.000   Max.   :219.0   Max.   :119.0   Max.   :78.00   Max.   :45.00   Max.   :36.00
                                                                                 NA's   :2

  Luggage.room       Weight        Origin              Make
 Min.   : 6.00   Min.   :1695   Length:93          Length:93
 1st Qu.:12.00   1st Qu.:2620   Class :character   Class :character
 Median :14.00   Median :3040   Mode  :character   Mode  :character
 Mean   :13.89   Mean   :3073
 3rd Qu.:15.00   3rd Qu.:3525
 Max.   :22.00   Max.   :4105
 NA's   :11
```

Hide

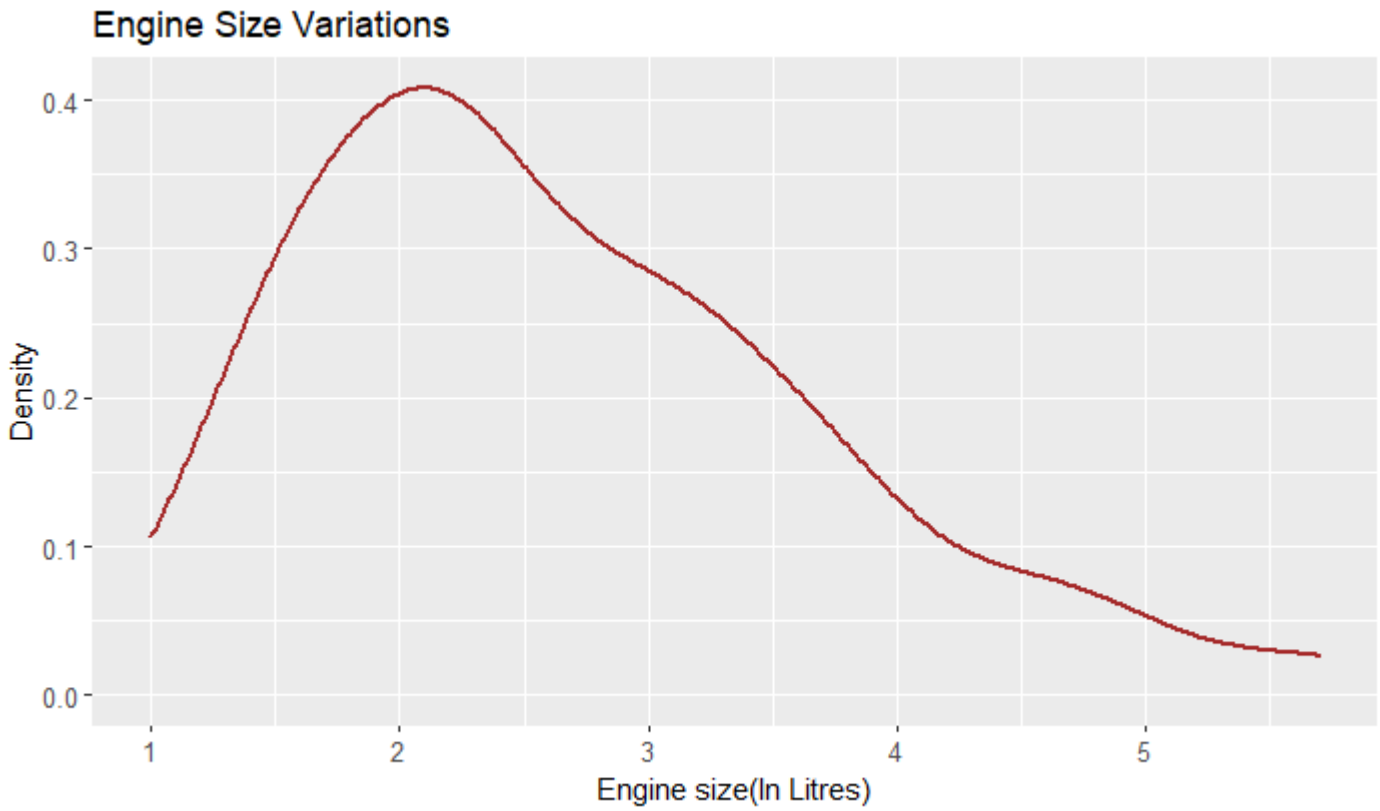
```
ggplot(Cars93, aes(Price)) +
        geom_histogram(binwidth= 20, fill = "light blue")+
  theme_classic()+
  labs(x="Price (in Dollars)", title = "Price Variation As Per Count .", caption = "1.1. RELATIO
NSHIP OF A SINGLE CONTINOUS VARIABLE: PRICE                                                    ")
```



The Histogram shows the relationship between Price and Count. The height of the bar represents the count of the price in dollars.

Hide

```
ggplot(Cars93,aes(x=EngineSize))+
  geom_density(size=1,color="Brown",)+
  theme(axis.text.x = element_text(size=10)) +
  theme(axis.text.y = element_text(size=10 ))+
  labs( x = "Engine size(In Litres)",y="Density", title = "Engine Size Variations", caption =
"1.2. RELATIONSHIP OF A SINGLE CONTINOUS VARIABLE: Engine Size                                 ")
```

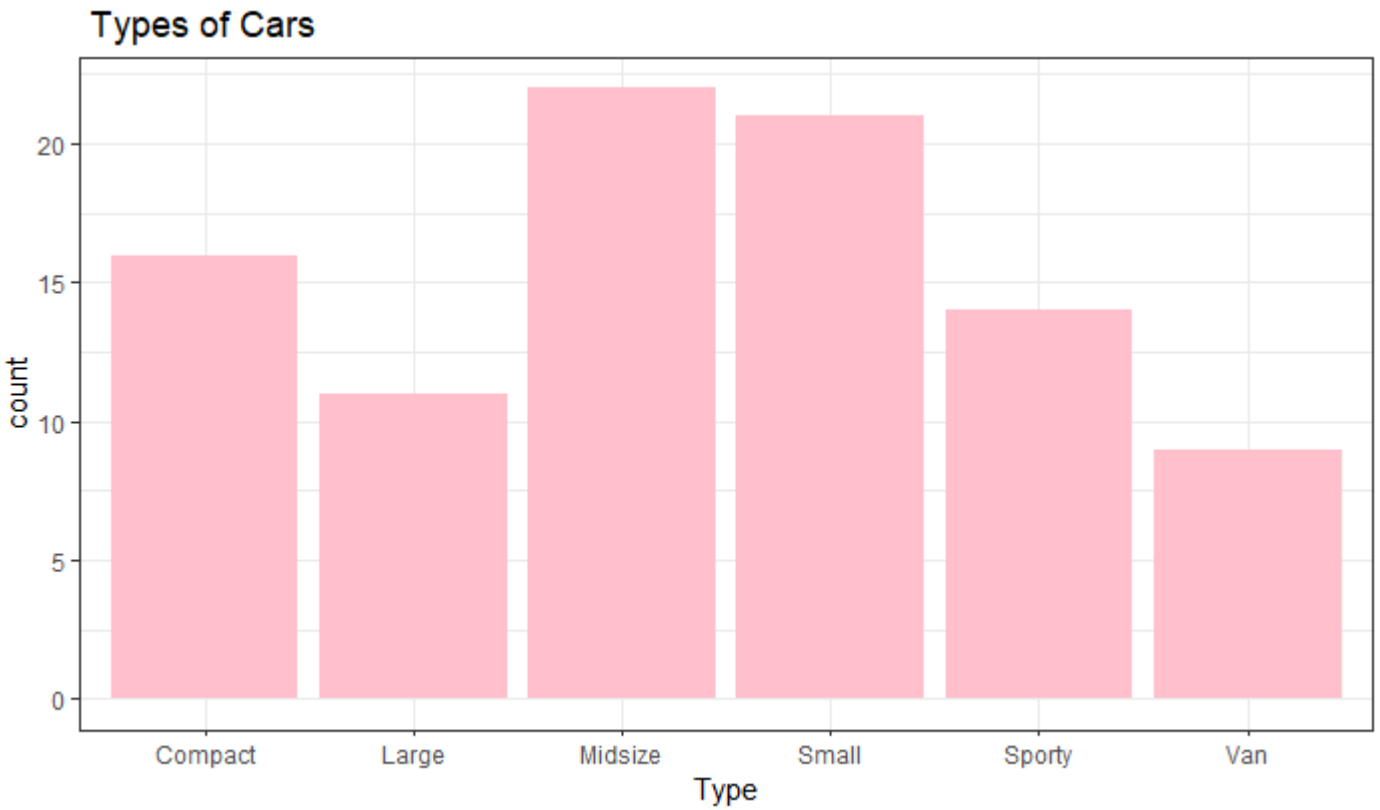


1.2. RELATIONSHIP OF A SINGLE CONTINOUS VARIABLE: Engine Size

2: Two plots displaying distribution about a single categorical variable.

Hide

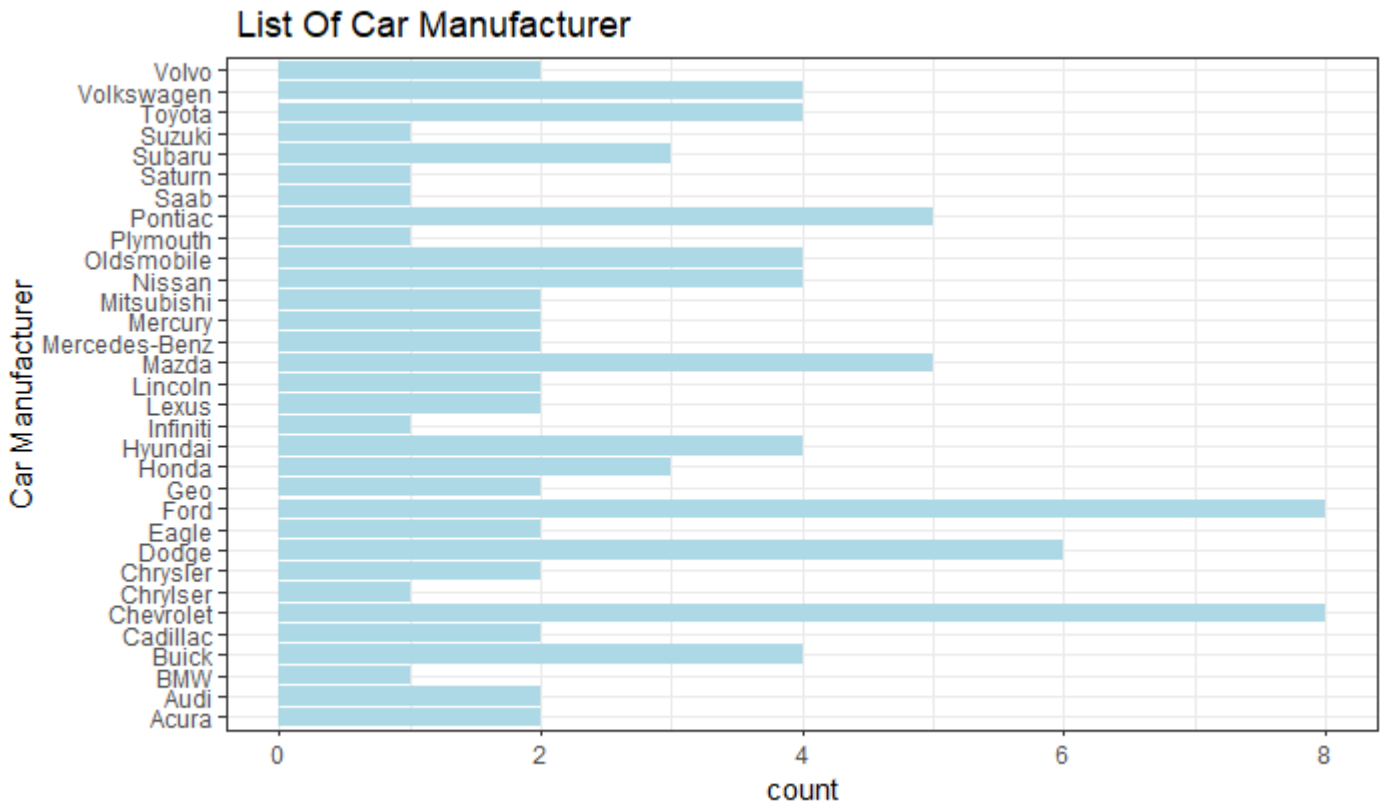
```
ggplot(data=Cars93)+

geom_bar(mapping = aes(x= Type ), fill= 'pink')+
  labs(x= " Type ", title =  " Types of Cars ", caption = "2.1. RELATIONSHIP OF A SINGLE CATEGOR
ICAL VARIABLE: Types of cars                                  ")+

  theme_bw()
```



2.1. RELATIONSHIP OF A SINGLE CATEGORICAL VARIABLE: Types of cars

Hide

```
ggplot(data=Cars93)+

geom_bar(mapping = aes(x= Manufacturer ), fill= ' light blue')+
  labs(x= " Car Manufacturer ", title =  " List Of Car Manufacturer ",caption = "2.2. RELATIONSH
IP OF A SINGLE CATEGORICAL VARIABLE: Manufacturer                                  ")+

  coord_flip()+


  theme_bw()
```

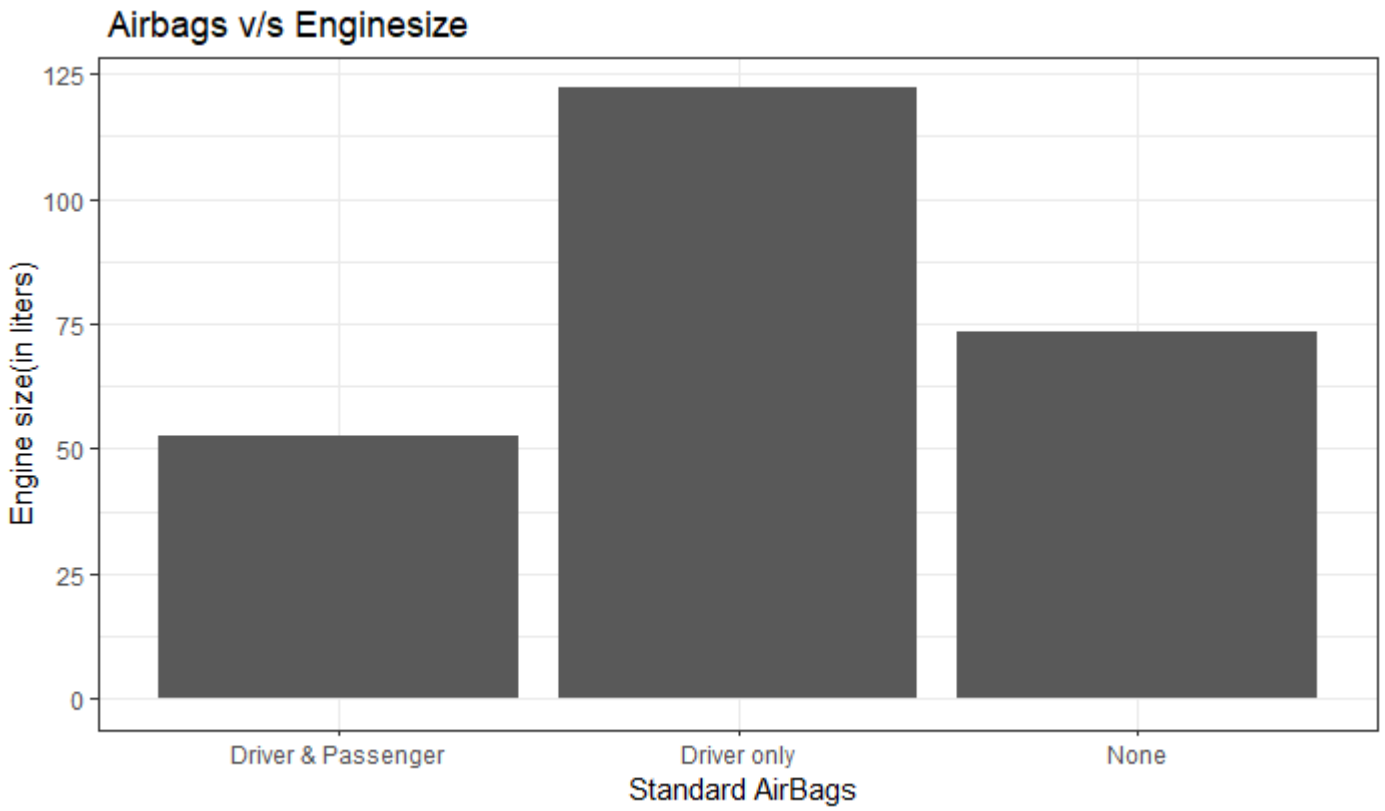


2.2. RELATIONSHIP OF A SINGLE CATEGORICAL VARIABLE: Manufacturer

3: One plot displaying information about both a continuous variable and a categorical variable.

Hide

```
ggplot(data=Cars93)+

geom_col(aes(AirBags,EngineSize))+
  labs(x= " Standard AirBags", y= "Engine size(in liters)",  title =  " Airbags v/s Enginesize
", caption = "3. RELATIONSHIP OF A CONTINOUS AND CATEGORICAL VARIABLE: Air Bags v/s Engine Size
")+
  theme_bw()
```



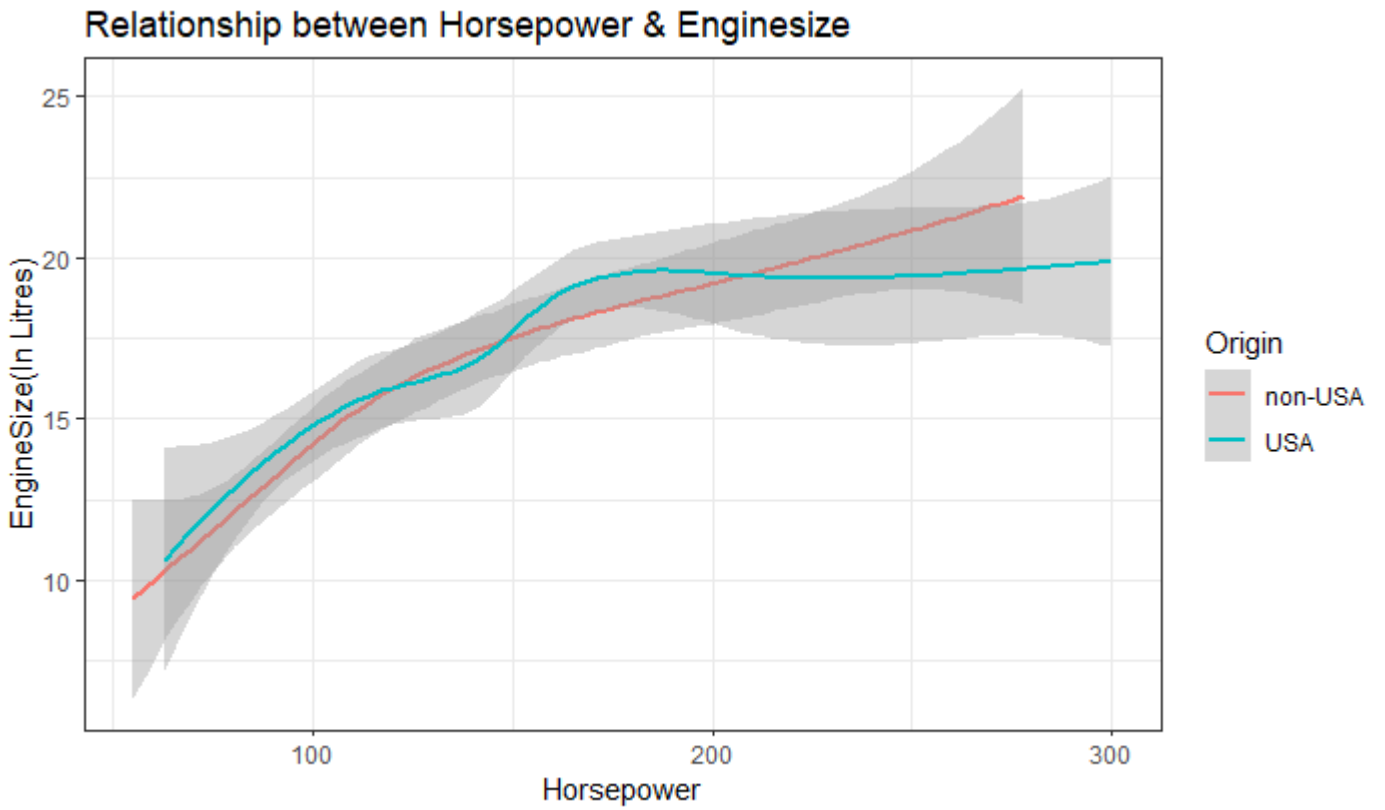
3. RELATIONSHIP OF A CONTINOUS AND CATEGORICAL VARIABLE: Air Bags v/s Engine Size

4: two plots should display information that shows a relationship between two variables.

Hide

```
ggplot(data=Cars93)+
geom_smooth(mapping=aes(Horsepower,Fuel.tank.capacity,colour=Origin))+
labs(x="Horsepower", y="EngineSize(In Litres)" ,
title="Relationship between Horsepower & Enginesize", caption = "4.1. RELATIONSHIP BETWEEN TWO V
ARIABLE: Air Bags v/s Engine Size           ")+
theme_bw()
```

```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

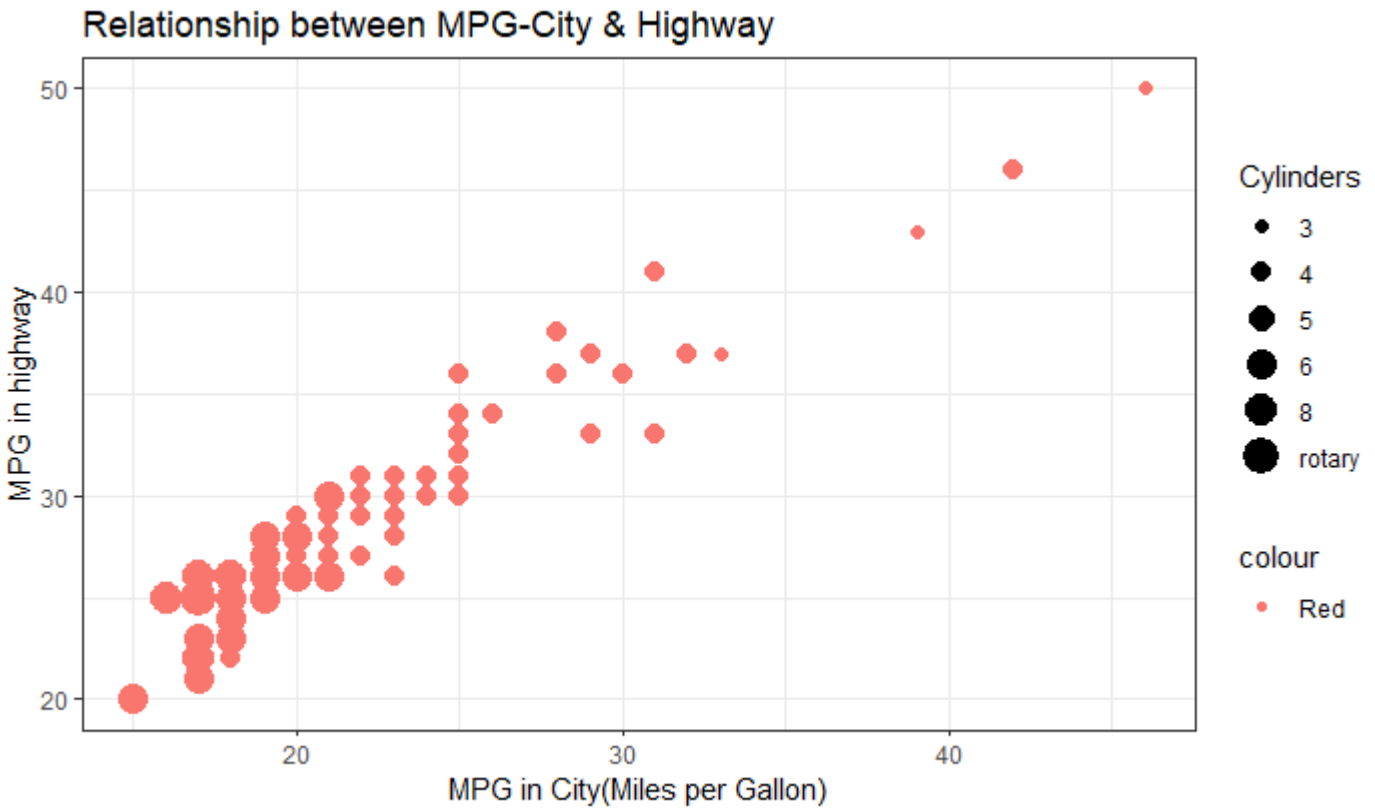


4.1. RELATIONSHIP BETWEEN TWO VARIABLE: Air Bags v/s Engine Size

Hide

```
ggplot(data=Cars93)+
geom_point(mapping = aes(x=MPG.city, y=MPG.highway, size= Cylinders, colour = 'Red'))+
  labs(x="MPG in City(Miles per Gallon)", y="MPG in highway" ,
title="Relationship between MPG-City & Highway ", caption = "4.2. RELATIONSHIP BETWEEN TWO VARIA
BLE: MPG City v/s Highway           ")+
theme_bw()
```

```
Warning: Using size for a discrete variable is not advised.
```

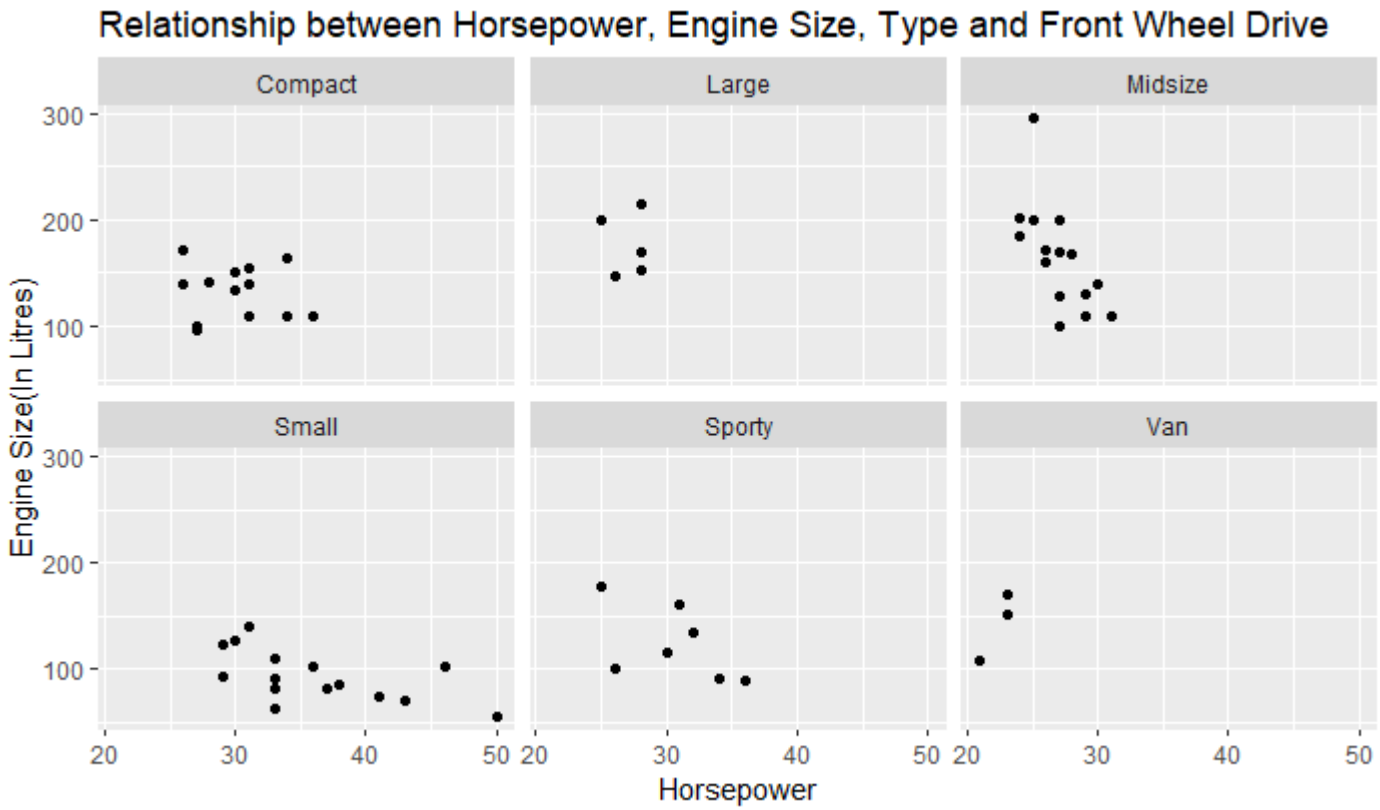


4.2. RELATIONSHIP BETWEEN TWO VARIABLE: MPG City v/s Highway

5: One plot should use faceting and display information about 4 variables.

Hide

```
DriveT <- Cars93 %>% filter(DriveTrain == "Front")
ggplot(data = DriveT) +
geom_point(mapping = aes(x = MPG.highway, y = Horsepower)) +
facet_wrap(~ Type, nrow = 2)+
labs(x="Horsepower", y="Engine Size(In Litres)" ,
title="Relationship between Horsepower, Engine Size, Type and Front Wheel Drive", caption="5. Re
lationship Between Four Variables (Horsepower,Engine Size,Type, Front Wheel Drive)
")+
theme_replace()
```



5. Relationship Between Four Variables (Horsepower,Engine Size,Type, Front Wheel Drive)

Question 1: In what ways do you think data visualization is important to understanding a dataset?

Answer1: Understanding the data set is important to see the trends and pattern more clearly by presenting visual context through maps or graphs. Data sets is help us to solve our business problem. In order to build a data visualization maps and graphs, we need to explore and understand our data sets before plotting anything. We should spend most of our time to explore, clean and preparing data for modelling. It really helps us to build the accurate model.

Question 2: In what ways do you think data visualization is important to communicating important aspects of a data set?

In today's business world, so much data is being collected and analyzed, and we need a way to interpret it. By providing visual context to the information through maps or graphs, data visualization helps us understand the data. In this way, data can be analyzed by a human mind in a way that is more natural. This makes trends, patterns, and outliers easier to discern in large data sets. Businesses are able to better predict sales volumes and future growth with the help of visualized data.

Question 3: What role does your integrity as an analyst play when creating a data visualization for communicating results to others?

Answer3: While communicating the result to others as analyst, first of all he is responsible for making backups to company files in a secure location that protects the data from all the storage devices.Analyst responsibility is to identify improvements areas that needs immediate attention with respect to the company, he has to predict the sales volumes of the company by plotting the data on graph and charts.

Question 4: How many variables do you think you can successfully represent in a visualization? What happens when you exceed this number?

To depict a good visualization, it is important to represent the infographic clearly and precisely. The count of variables can't be determined, however data variables till 4 variables will be a reasonable number on a two-axis chart.Exceeding the variables beyond the reasonable count will lead to clutter and confusion and may result in failure of the design attributes of information.

Contribution to each team member in Project

Name	Contribution
Nishi, Diksha, Shashank and Preet	Did Research on Finalizing the data set and documentation of Dataset
Shashank	Created Plots 2.2,4.2 and answer 1
Diksha	Created Plots 3,4.1 and answer 4
Nishi	Created plots 1.1, 2.1 and answer 3
Preet	Created plots 1.2, 5 and answer 2

References

- Lock, R. H. (1993) 1993 New Car Data. Journal of Statistics Education (https://https://www.tandfonline.com/doi/full/10.1080/10691898.1993.11910459)
- Venables, W. N. and Ripley, B. D. (1999) Modern Applied Statistics with S-PLUS. Third Edition. Springer.