COMP 8157 Advanced Database Topics

University of Windsor, School of Computer Science

Lab 1

CONFIDENTIALITY AGREEMENT & STATEMENT OF HONESTY

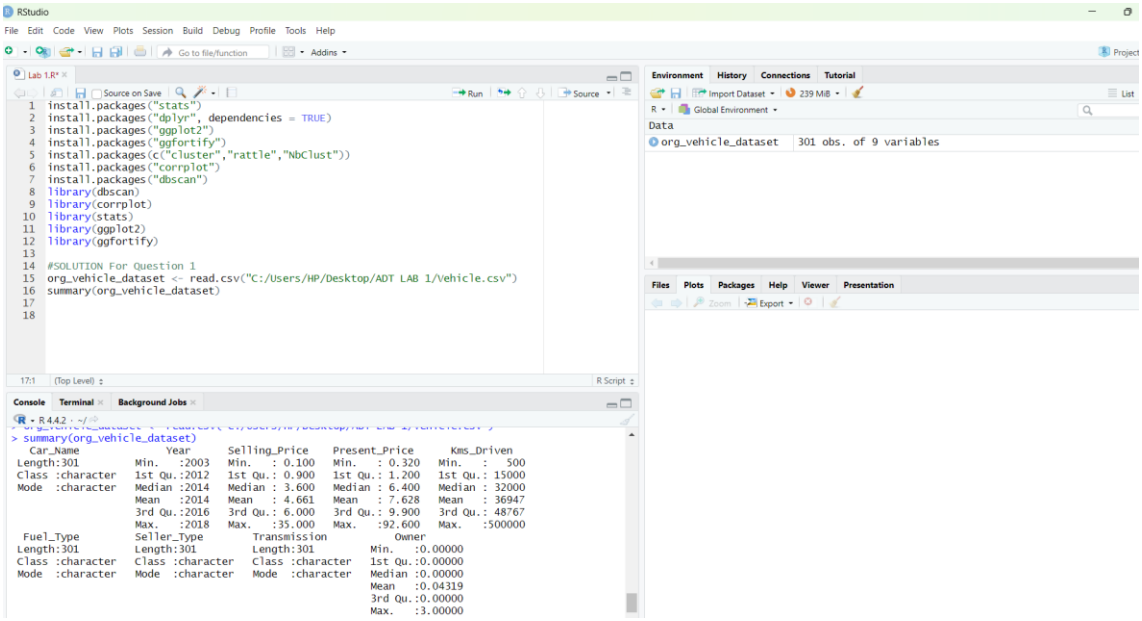
I, **Firas Hussain Mohammed** verify that the submitted work is my own, original work, and that I did not use Generative AI tools (e.g., ChatGPT, Bard) to produce this lab report. I confirm knowing that a mark of 0 may be assigned for sharing or copying this work.

Firas Hussain Mohammed	Firas Hussain Mohammed	110173878
Student Signature	**Student Name**	**Student I.D.**

Student Signature	**Student Name**	**Student I.D.**

Part 1: Data Exploration

1. Import the Vehicle dataset, summarize it and explain the output?

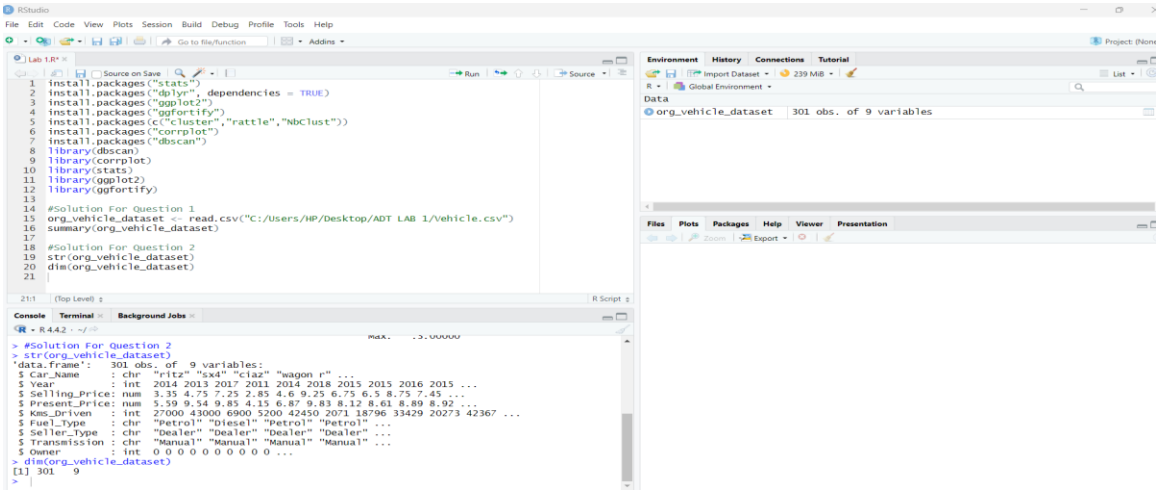


```
install.packages("stats")
install.packages("dplyr", dependencies = TRUE)
install.packages("ggplot2")
install.packages("ggfortify")
install.packages(c("cluster","rattle","NbClust"))
install.packages("corrplot")
install.packages("dbscan")
library(dbscan)
library(corrplot)
library(stats)
library(ggplot2)
library(ggfortify)

#SOLUTION For Question 1
org_vehicle_dataset <- read.csv("C:/Users/HP/Desktop/ADT LAB 1/Vehicle.csv")
summary(org_vehicle_dataset)
```

```
> summary(org_vehicle_dataset)
   Car_Name              Year      Selling_Price     Present_Price       Kms_Driven
 Length:301         Min.   :2003   Min.   : 0.100   Min.   : 0.320   Min.   :   500
 Class :character   1st Qu.:2012   1st Qu.: 0.900   1st Qu.: 1.200   1st Qu.: 15000
 Mode  :character   Median :2014   Median : 3.600   Median : 6.400   Median : 32000
                    Mean   :2014   Mean   : 4.661   Mean   : 7.628   Mean   : 36947
                    3rd Qu.:2016   3rd Qu.: 6.000   3rd Qu.: 9.900   3rd Qu.: 48767
                    Max.   :2018   Max.   :35.000   Max.   :92.600   Max.   :500000
  Fuel_Type          Seller_Type        Transmission          Owner
 Length:301         Length:301         Length:301         Min.   :0.00000
 Class :character   Class :character   Class :character   1st Qu.:0.00000
 Mode  :character   Mode  :character   Mode  :character   Median :0.00000
                                                          Mean   :0.04319
                                                          3rd Qu.:0.00000
                                                          Max.   :3.00000
```

- The output generated showed us complete details of our dataset which includes all column names type of data such as character or numeric, number of records.
- For numeric records it displayed minimum value, 1st Quartile that is 25% of data, Median which is also 2nd Quartile (50% o data), mean, 3rd Quartile that is 75% of data and maximum value of record in our column.
- Our dataset has 302 records and 9 columns.

2. Show the structure and dimension of the dataset and explain it?

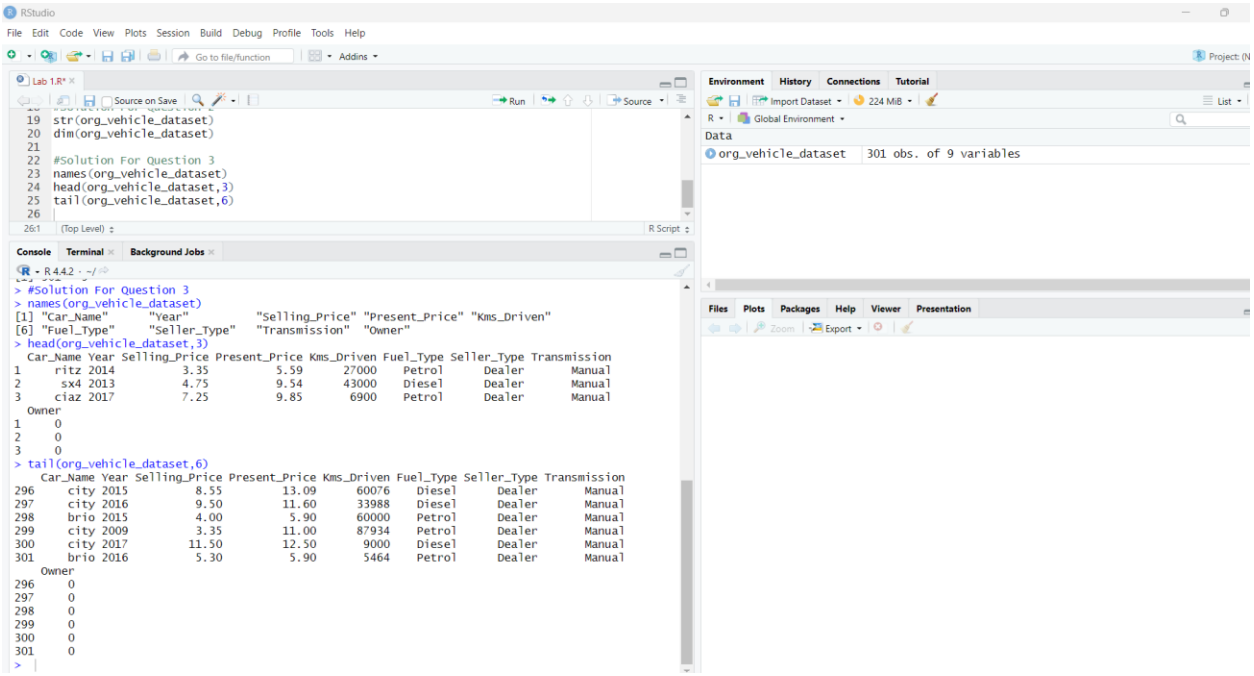


```
install.packages("stats")
install.packages("dplyr", dependencies = TRUE)
install.packages("ggplot2")
install.packages("ggfortify")
install.packages(c("cluster","rattle","NbClust"))
install.packages("corrplot")
install.packages("dbscan")
library(dbscan)
library(corrplot)
library(stats)
library(ggplot2)
library(ggfortify)

#Solution For Question 1
org_vehicle_dataset <- read.csv("C:/Users/HP/Desktop/ADT LAB 1/Vehicle.csv")
summary(org_vehicle_dataset)

#Solution For Question 2
str(org_vehicle_dataset)
dim(org_vehicle_dataset)
```

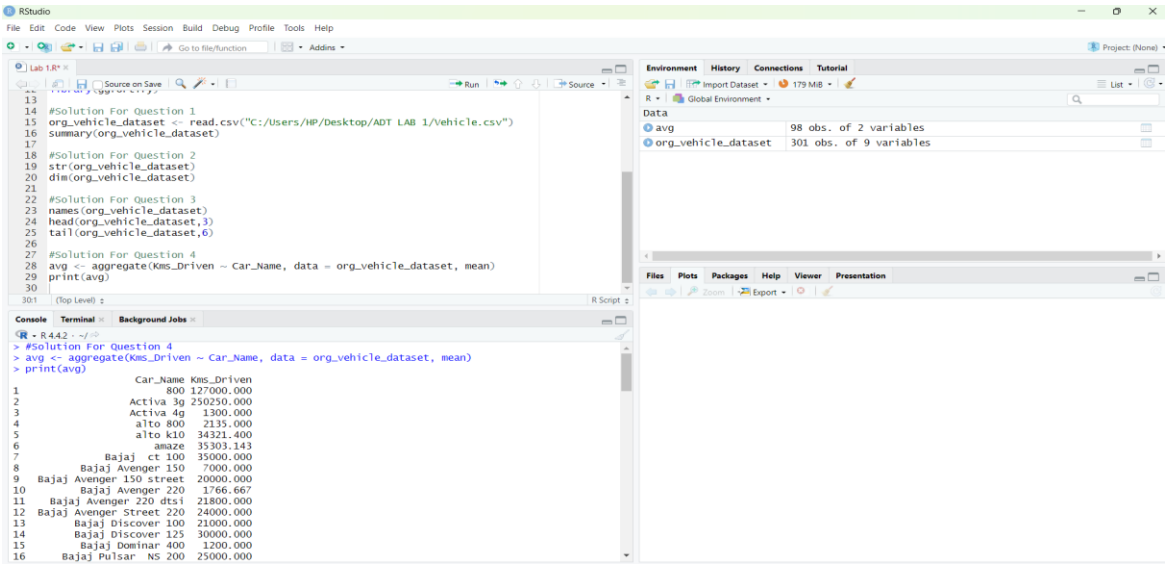
```
> #Solution For Question 2
> str(org_vehicle_dataset)
'data.frame':   301 obs. of  9 variables:
 $ Car_Name     : chr  "ritz" "sx4" "ciaz" "wagon r" ...
 $ Year         : int  2014 2013 2017 2011 2014 2018 2015 2015 2016 2015 ...
 $ Selling_Price: num  3.35 4.75 7.25 2.85 4.6 9.25 6.75 6.5 8.75 7.45 ...
 $ Present_Price: num  5.59 9.54 9.85 4.15 6.87 9.83 8.12 8.61 8.89 8.92 ...
 $ Kms_Driven   : int  27000 43000 6900 5200 42450 2071 18796 33429 20273 42367 ...
 $ Fuel_Type    : chr  "Petrol" "Diesel" "Petrol" "Petrol" ...
 $ Seller_Type  : chr  "Dealer" "Dealer" "Dealer" "Dealer" ...
 $ Transmission : chr  "Manual" "Manual" "Manual" "Manual" ...
 $ Owner        : int  0 0 0 0 0 0 0 0 0 0 ...
> dim(org_vehicle_dataset)
[1] 301   9
>
```

- Through the structure command we got information like type of data each column has and also displayed a snapshot of the records.
- Dimension on the other hand returned only the number of records in our dataset and number o columns.

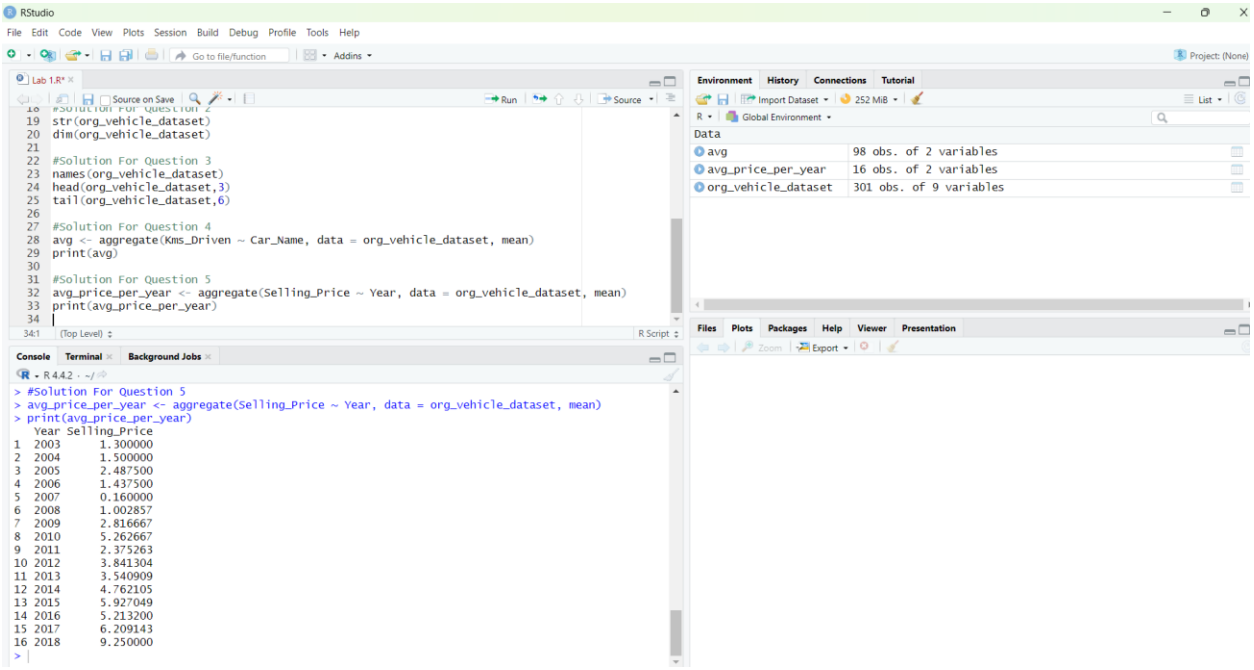
3. Show the column names of the Vehicle dataset and the first 3 rows and the last 6 rows of it?



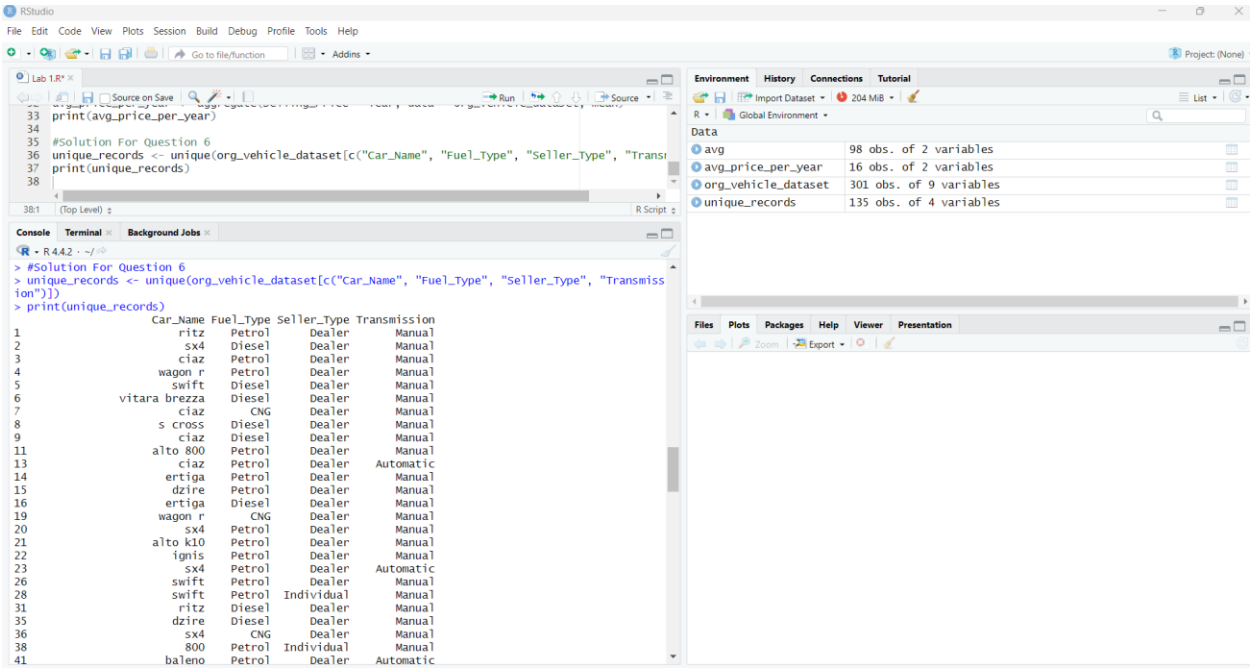
4. Show the average Kms_Driven for each type of car (Car_Name) in the dataset.



5. What is the average Selling_Price of the cars in each year?



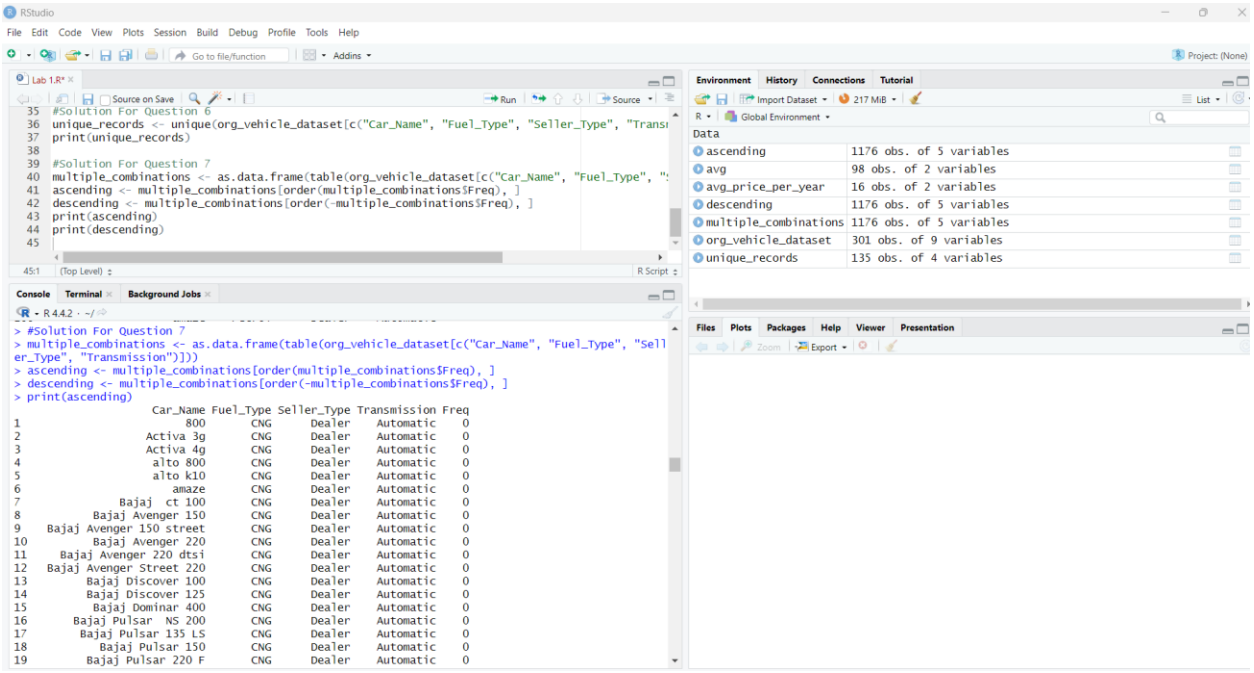
6. Show the unique combinations of Car_Name, Fuel_Type, Seller_Type, and Transmission in the Vehicle dataset.



```
print(avg_price_per_year)

#Solution For Question 6
unique_records <- unique(org_vehicle_dataset[c("Car_Name", "Fuel_Type", "Seller_Type", "Transmission")])
print(unique_records)
```

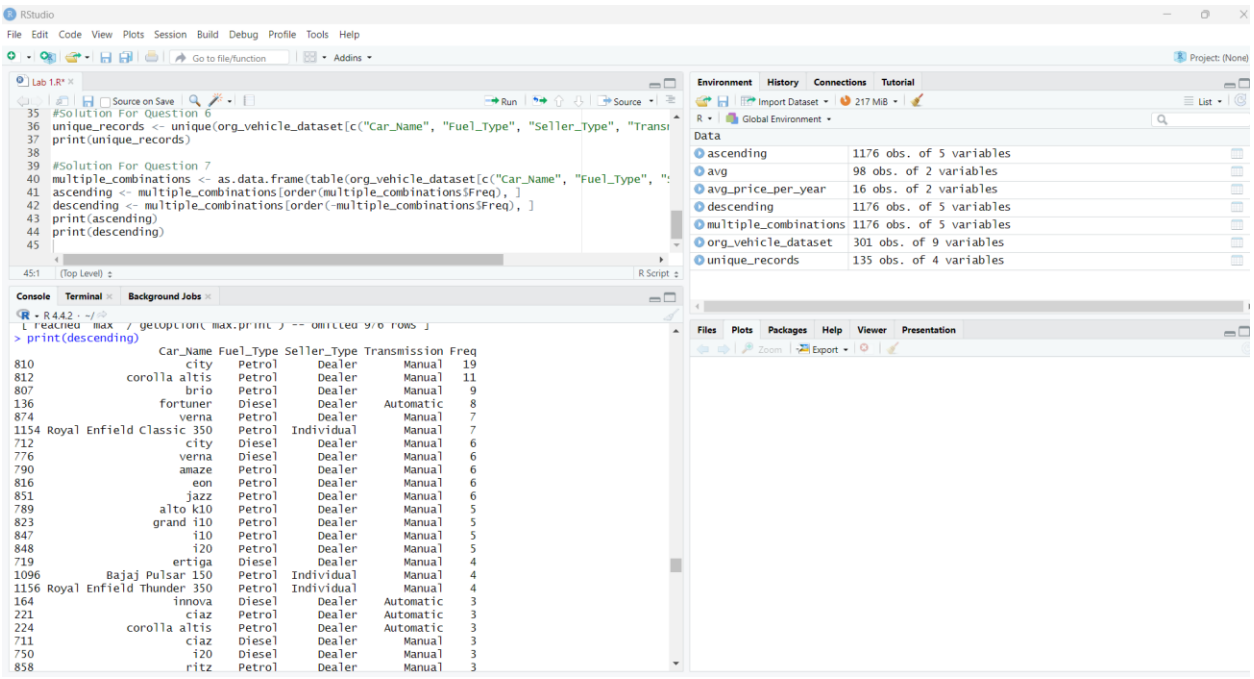
```
> #Solution For Question 6
> unique_records <- unique(org_vehicle_dataset[c("Car_Name", "Fuel_Type", "Seller_Type", "Transmission")])
> print(unique_records)
                  Car_Name Fuel_Type Seller_Type Transmission
1                     ritz    Petrol      Dealer       Manual
2                      sx4    Diesel      Dealer       Manual
3                     ciaz    Petrol      Dealer       Manual
4                  wagon r    Petrol      Dealer       Manual
5                    swift    Diesel      Dealer       Manual
6            vitara brezza    Diesel      Dealer       Manual
7                     ciaz       CNG      Dealer       Manual
8                  s cross    Diesel      Dealer       Manual
9                     ciaz    Diesel      Dealer       Manual
11                alto 800    Petrol      Dealer       Manual
13                    ciaz    Petrol      Dealer    Automatic
14                  ertiga    Petrol      Dealer       Manual
15                   dzire    Petrol      Dealer       Manual
16                  ertiga    Diesel      Dealer       Manual
19                 wagon r       CNG      Dealer       Manual
20                     sx4    Petrol      Dealer       Manual
21                alto k10    Petrol      Dealer       Manual
22                   ignis    Petrol      Dealer       Manual
23                     sx4    Petrol      Dealer    Automatic
26                   swift    Petrol      Dealer       Manual
28                   swift    Petrol  Individual       Manual
31                    ritz    Diesel      Dealer       Manual
35                   dzire    Diesel      Dealer       Manual
36                     sx4       CNG      Dealer       Manual
38                     800    Petrol  Individual       Manual
41                  baleno    Petrol      Dealer    Automatic
```

7. What are the different combinations of Car_Name, Fuel_Type, Seller_Type, and Transmission in the Vehicle dataset, and how many times does it occur? (Display all such in both ascending and descending orders)



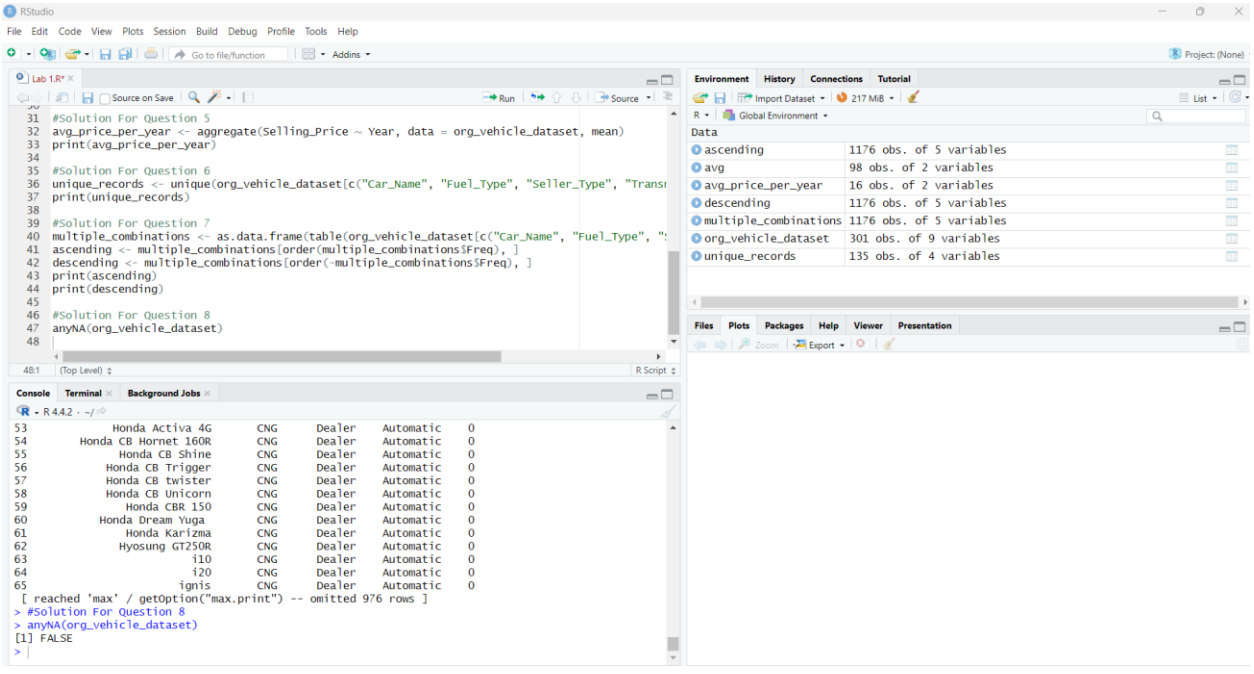
```
#Solution For Question 6
unique_records <- unique(org_vehicle_dataset[c("Car_Name", "Fuel_Type", "Seller_Type", "Transmission")])
print(unique_records)

#Solution For Question 7
multiple_combinations <- as.data.frame(table(org_vehicle_dataset[c("Car_Name", "Fuel_Type", "Seller_Type", "Transmission")]))
ascending <- multiple_combinations[order(multiple_combinations$Freq), ]
descending <- multiple_combinations[order(-multiple_combinations$Freq), ]
print(ascending)
print(descending)
```

```
> #Solution For Question 7
> multiple_combinations <- as.data.frame(table(org_vehicle_dataset[c("Car_Name", "Fuel_Type", "Seller_Type", "Transmission")]))
> ascending <- multiple_combinations[order(multiple_combinations$Freq), ]
> descending <- multiple_combinations[order(-multiple_combinations$Freq), ]
> print(ascending)
                  Car_Name Fuel_Type Seller_Type Transmission Freq
1                      800       CNG      Dealer    Automatic    0
2                Activa 3g       CNG      Dealer    Automatic    0
3                Activa 4g       CNG      Dealer    Automatic    0
4                 alto 800       CNG      Dealer    Automatic    0
5                 alto k10       CNG      Dealer    Automatic    0
6                    amaze       CNG      Dealer    Automatic    0
7            Bajaj  ct 100       CNG      Dealer    Automatic    0
8          Bajaj Avenger 150       CNG      Dealer    Automatic    0
9    Bajaj Avenger 150 street      CNG      Dealer    Automatic    0
10         Bajaj Avenger 220       CNG      Dealer    Automatic    0
11    Bajaj Avenger 220 dtsi       CNG      Dealer    Automatic    0
12   Bajaj Avenger Street 220      CNG      Dealer    Automatic    0
13        Bajaj Discover 100       CNG      Dealer    Automatic    0
14        Bajaj Discover 125       CNG      Dealer    Automatic    0
15         Bajaj Dominar 400       CNG      Dealer    Automatic    0
16      Bajaj Pulsar  NS 200       CNG      Dealer    Automatic    0
17       Bajaj Pulsar 135 LS       CNG      Dealer    Automatic    0
18          Bajaj Pulsar 150       CNG      Dealer    Automatic    0
19        Bajaj Pulsar 220 F       CNG      Dealer    Automatic    0
```

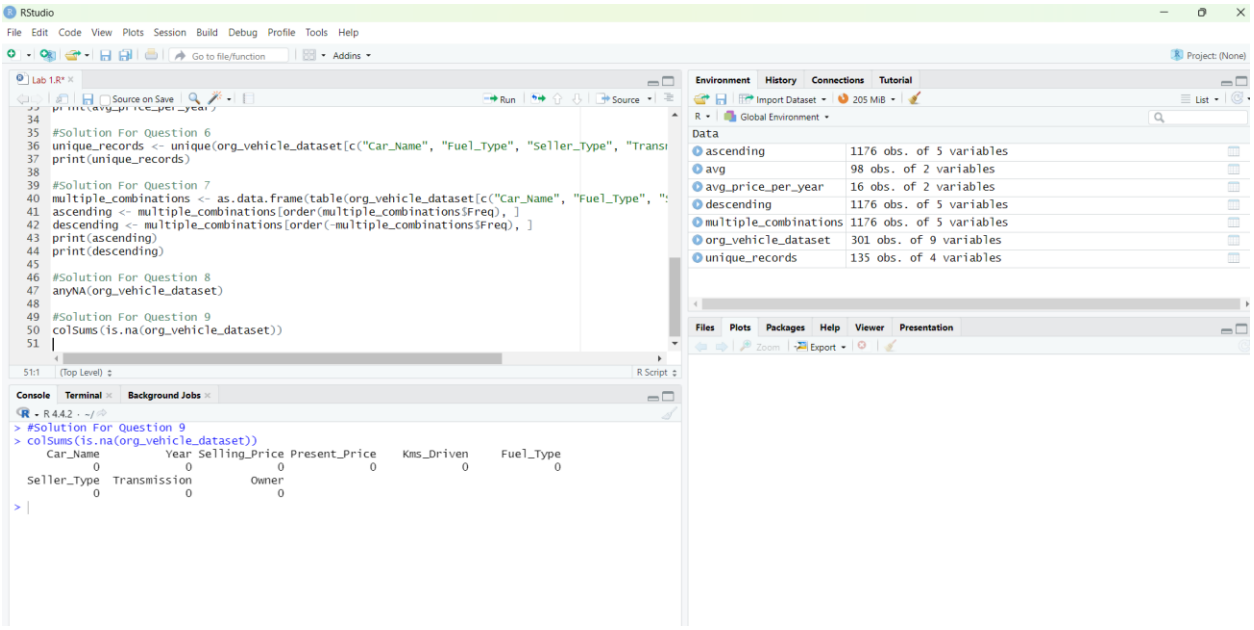


Part 2: Data Pre-Processing

8. Find if there are any missing values in the Vehicle dataset



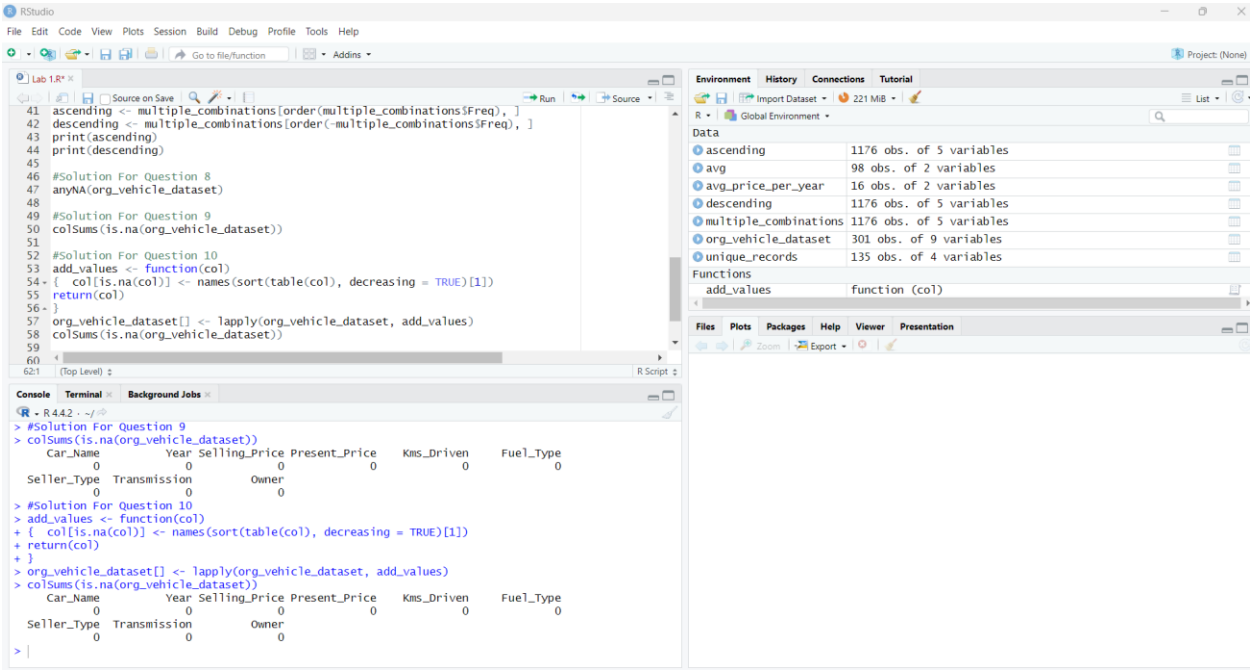
9. Find which columns contain missing values in the vehicles dataset. What are the total missing values for each column?



```
#Solution For Question 6
unique_records <- unique(org_vehicle_dataset[c("Car_Name", "Fuel_Type", "Seller_Type", "Transm
print(unique_records)

#Solution For Question 7
multiple_combinations <- as.data.frame(table(org_vehicle_dataset[c("Car_Name", "Fuel_Type", "
ascending <- multiple_combinations[order(multiple_combinations$Freq), ]
descending <- multiple_combinations[order(-multiple_combinations$Freq), ]
print(ascending)
print(descending)

#Solution For Question 8
anyNA(org_vehicle_dataset)

#Solution For Question 9
colSums(is.na(org_vehicle_dataset))
```

```
> #Solution For Question 9
> colSums(is.na(org_vehicle_dataset))
    Car_Name          Year Selling_Price Present_Price    Kms_Driven     Fuel_Type
           0             0             0             0             0             0
 Seller_Type  Transmission         Owner
           0             0             0
```

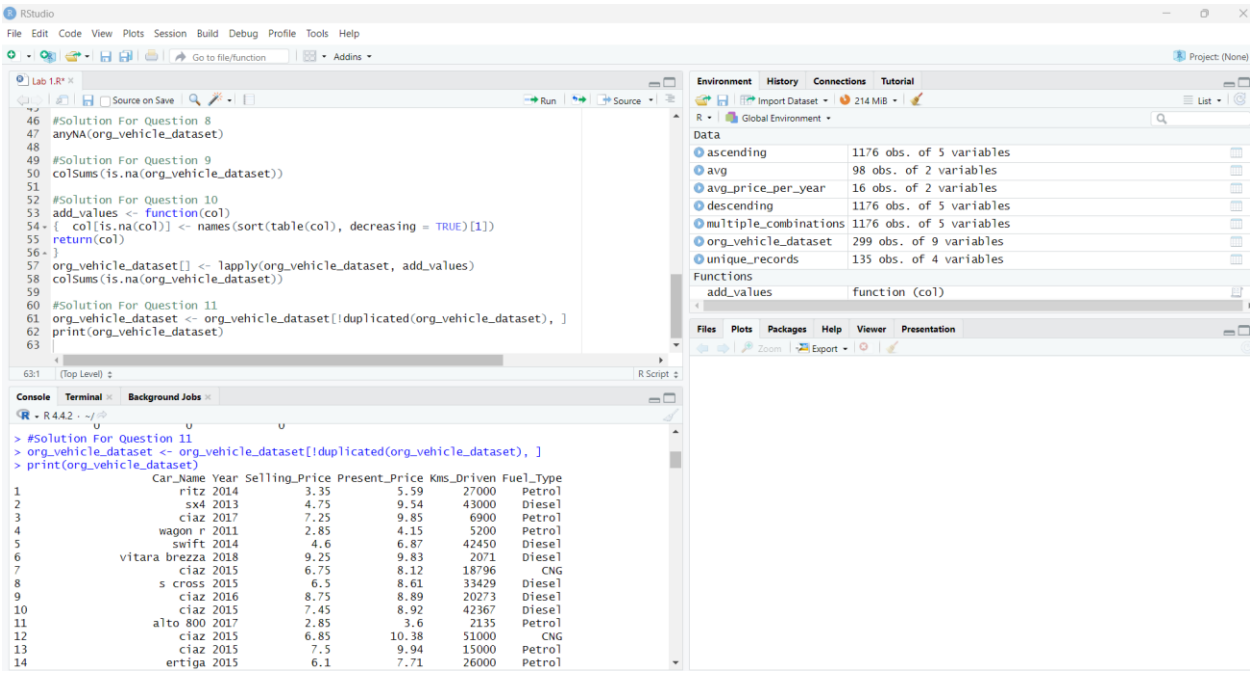
10. Replace the missing values in the dataset with the most repeated value of that field. Check if the missing values were replaced successfully

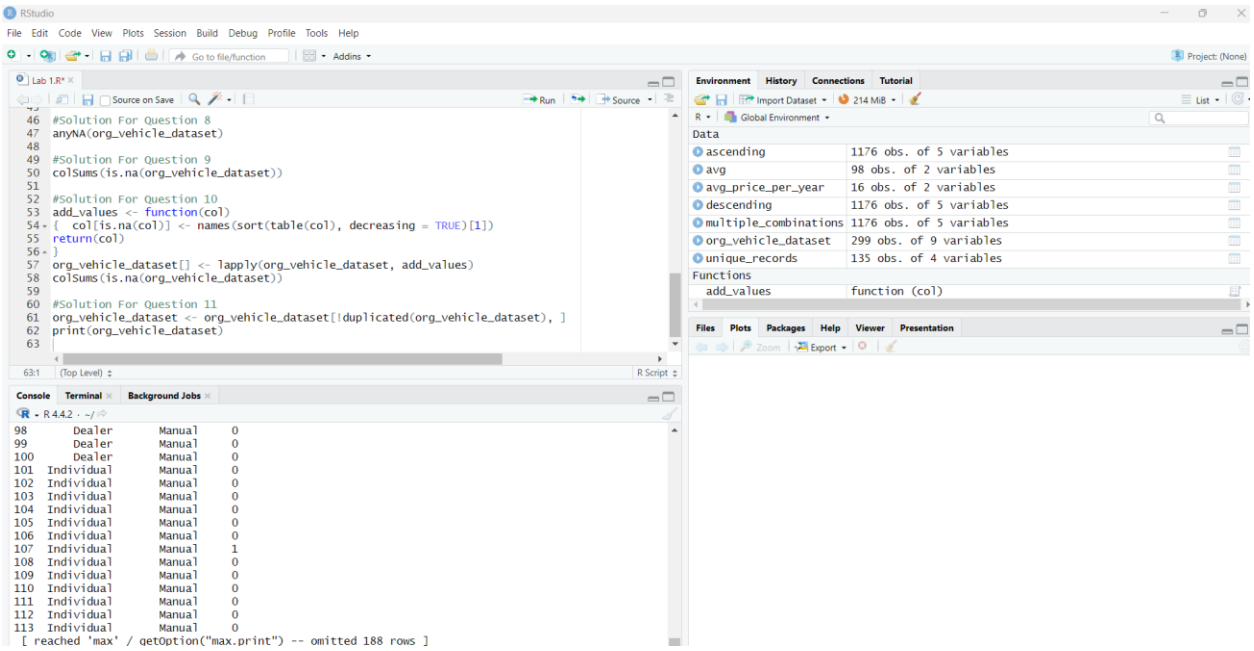


```
ascending <- multiple_combinations[order(multiple_combinations$Freq), ]
descending <- multiple_combinations[order(-multiple_combinations$Freq), ]
print(ascending)
print(descending)

#Solution For Question 8
anyNA(org_vehicle_dataset)

#Solution For Question 9
colSums(is.na(org_vehicle_dataset))

#Solution For Question 10
add_values <- function(col)
{  col[is.na(col)] <- names(sort(table(col), decreasing = TRUE)[1])
return(col)
}
org_vehicle_dataset[] <- lapply(org_vehicle_dataset, add_values)
colSums(is.na(org_vehicle_dataset))
```

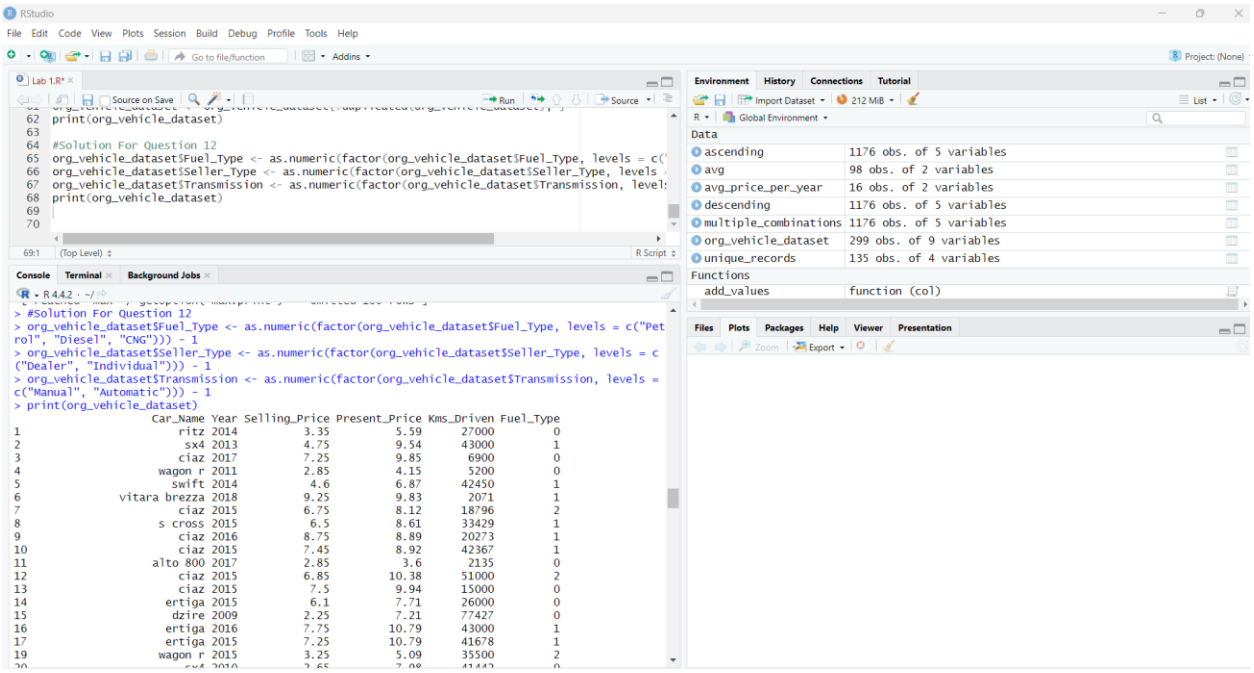
```
> #Solution For Question 9
> colSums(is.na(org_vehicle_dataset))
    Car_Name          Year Selling_Price Present_Price    Kms_Driven     Fuel_Type
           0             0             0             0             0             0
 Seller_Type  Transmission         Owner
           0             0             0
> #Solution For Question 10
> add_values <- function(col)
+ {  col[is.na(col)] <- names(sort(table(col), decreasing = TRUE)[1])
+ return(col)
+ }
> org_vehicle_dataset[] <- lapply(org_vehicle_dataset, add_values)
> colSums(is.na(org_vehicle_dataset))
    Car_Name          Year Selling_Price Present_Price    Kms_Driven     Fuel_Type
           0             0             0             0             0             0
 Seller_Type  Transmission         Owner
           0             0             0
```

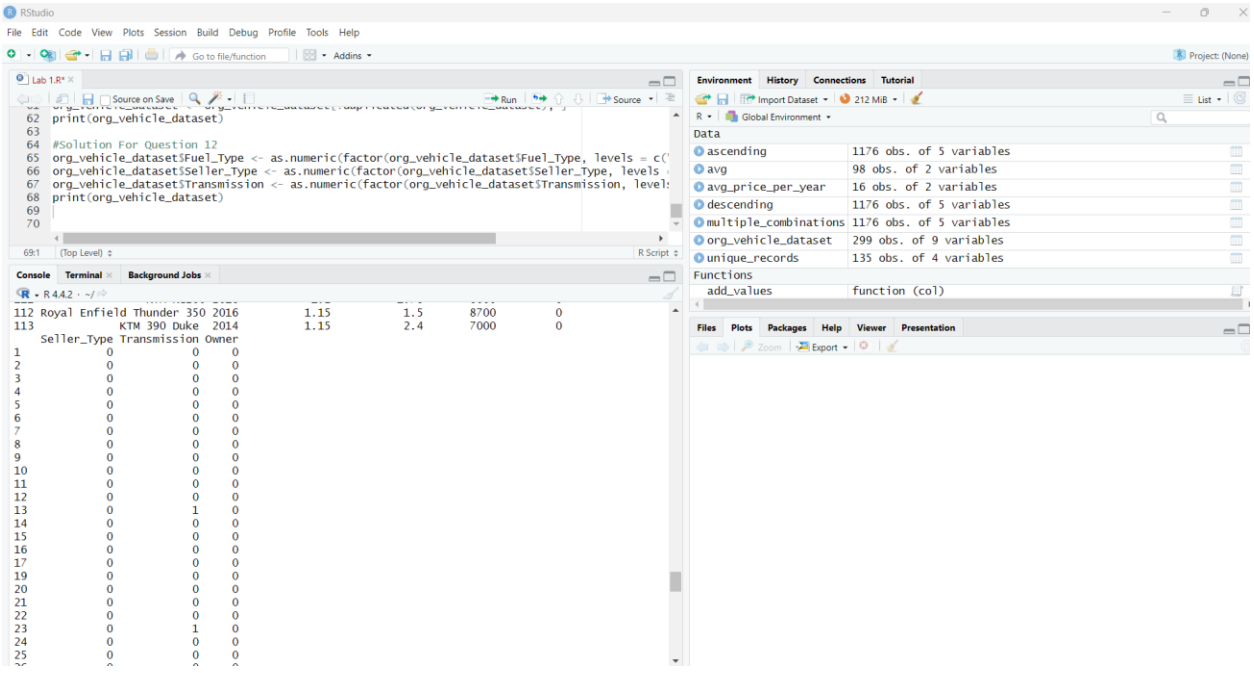
11. Find if the dataset has duplicate rows. Remove them, if exist?



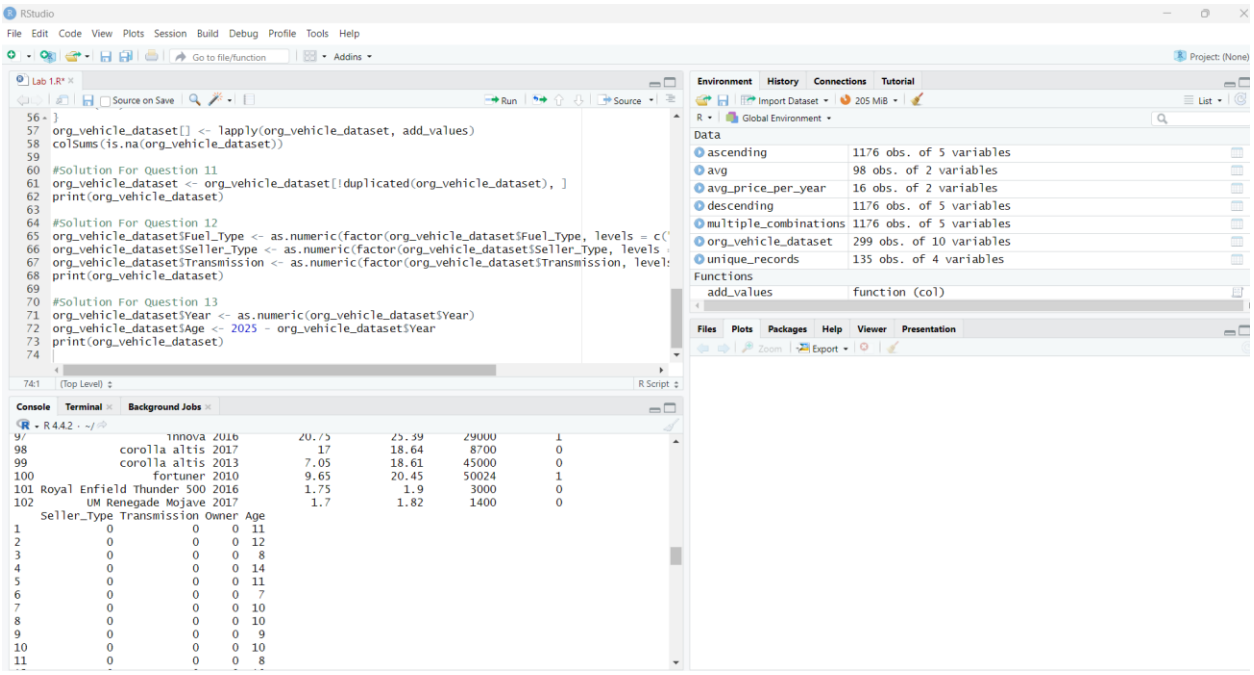


12. Replace the values of the following attributes:
    Fuel_Type: “Petrol”: 0, “Diesel”: 1, “CNG”: 2
    Seller_Type: “Dealer”: 0, “Individual”: 1
    Transmission: “Manual”: 0, “Automatic”: 1
    Show the conversion output of the specific attribute

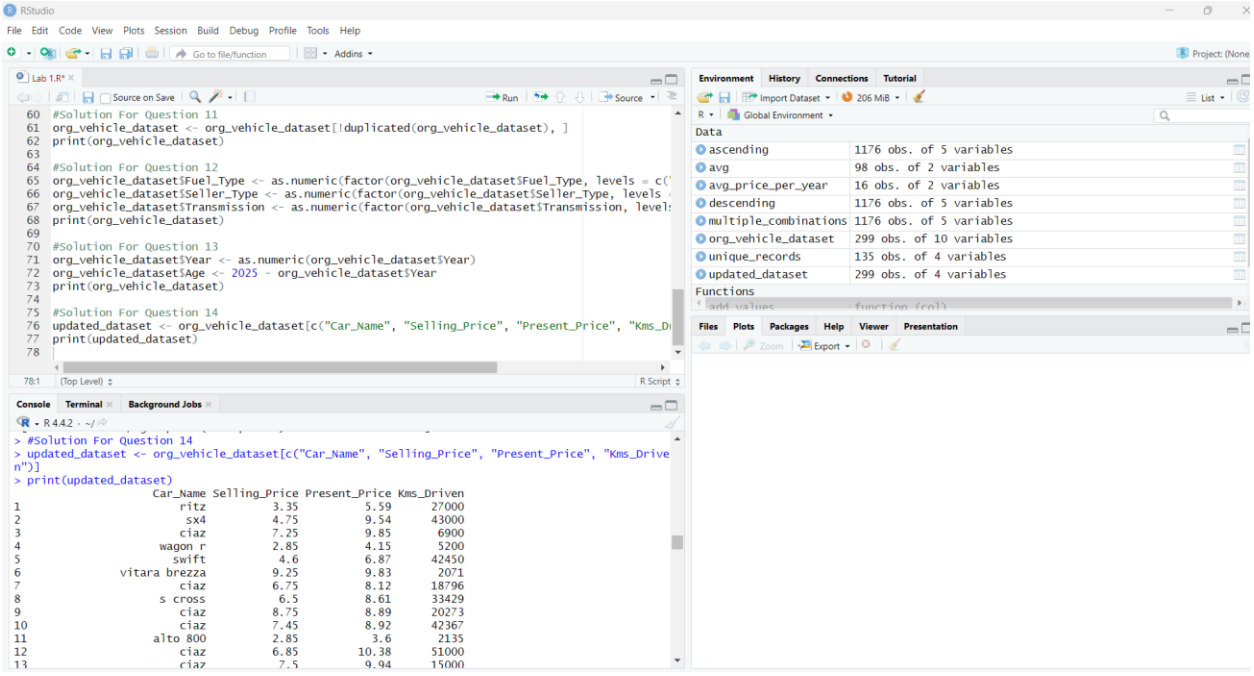




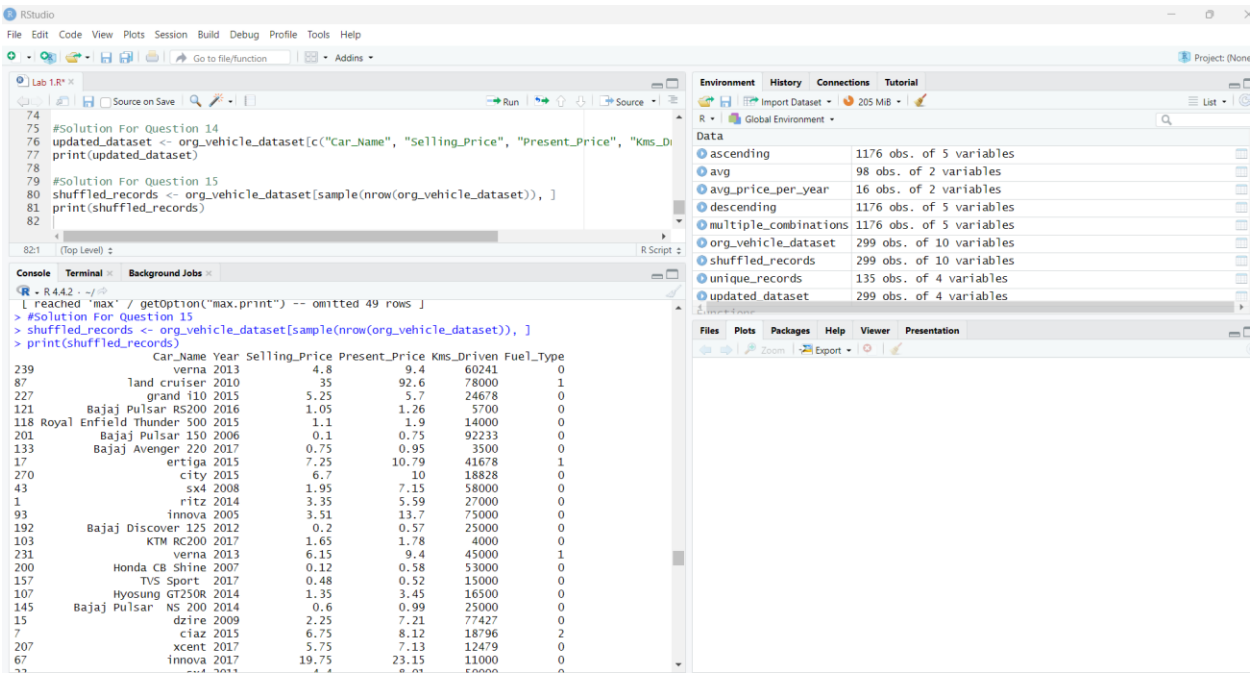
13. Add a new field called ‘Age’, and input the values by using the field Year. Show the output.



14. Create a new dataset by selecting only the columns “Car_name”, “Selling_Price”, “Present_Price”, and “Kms_Drive”. Show the output of the new dataset.

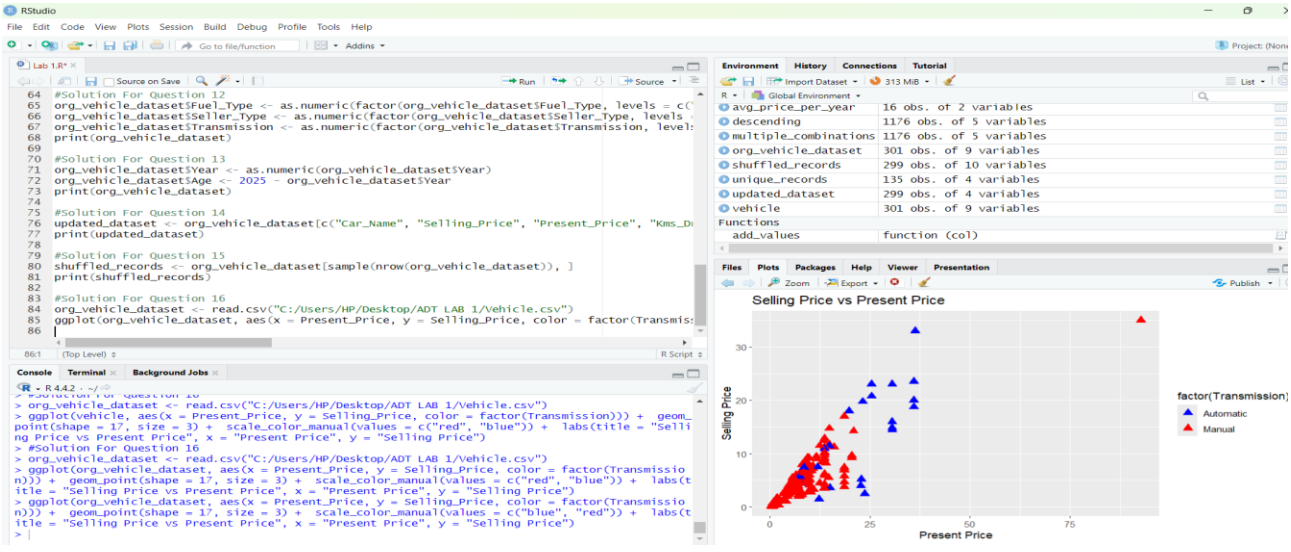


15. Shuffle the rows of the Vehicle dataset randomly and show the output



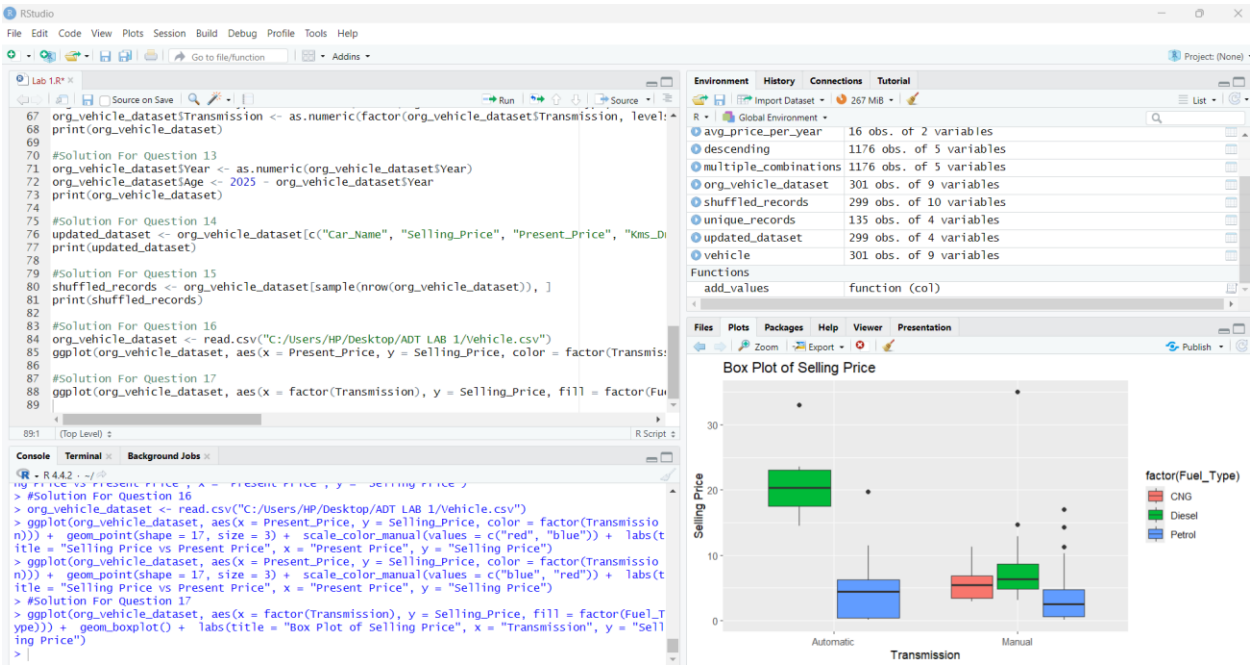
Part 3: Data Visualization

16. Import the Vehicle dataset. Create a scatter plot of the Selling_Price Vs Present_Price. Colour code the points based on the Transmission (5 marks).
    a) Add labels, title and colour to the plot. The colour should be red for Transmission type '0' and blue for '1'.
    b) Add open triangles to the plot.
    c) What do you understand from the output?

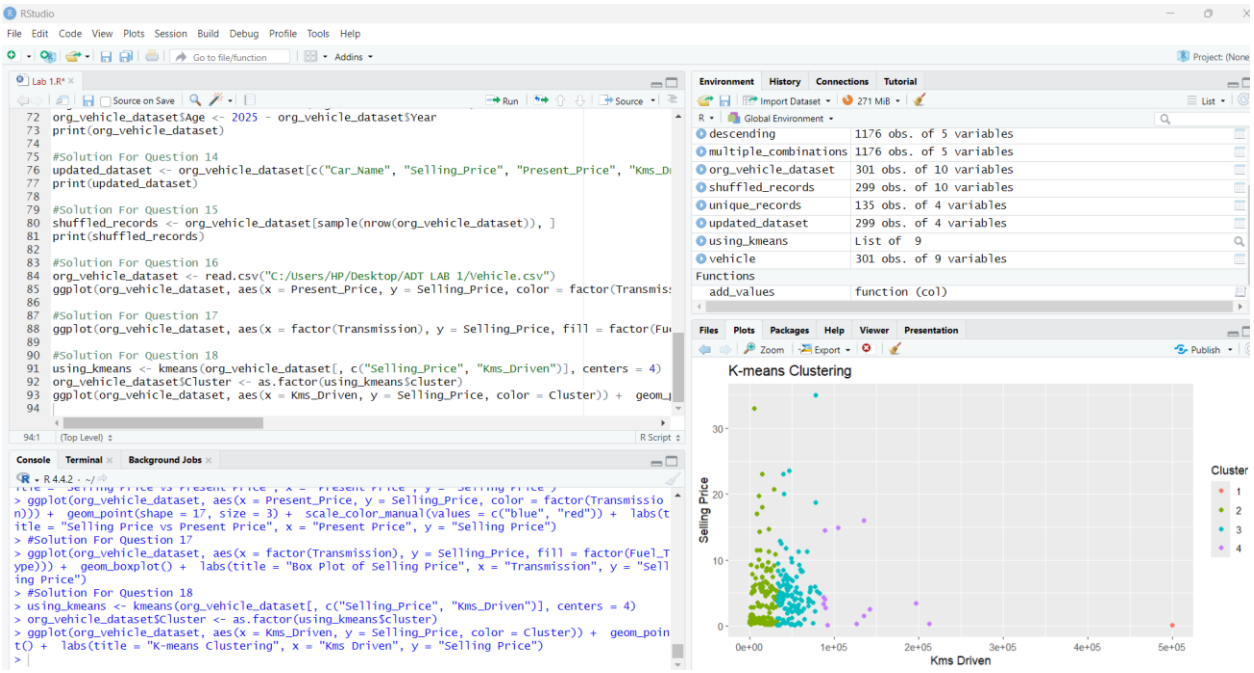


- The scatter plot indicates that manual transmission cars are within a specific range in contrast to automatic transmission.
- The blue clusters formed indicate that mostly everyone prefer manual cars as they are within a range than the automatic transmission.

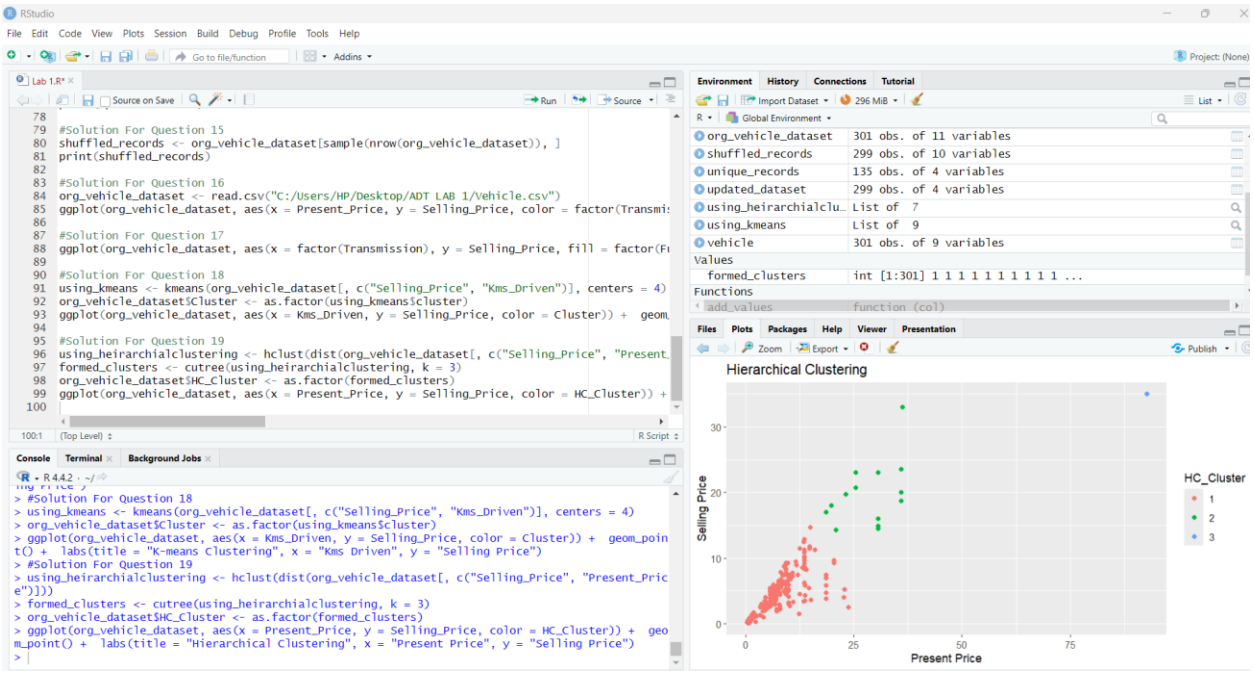
17. Create a box plot of the Selling_Price Vs Transmission and Fuel_Type



18. Create a scatter plot of the Selling_Price Vs Kms_Driven, and use k-means clustering to cluster the points into 4 clusters. Colour-code based on the cluster they belong to.

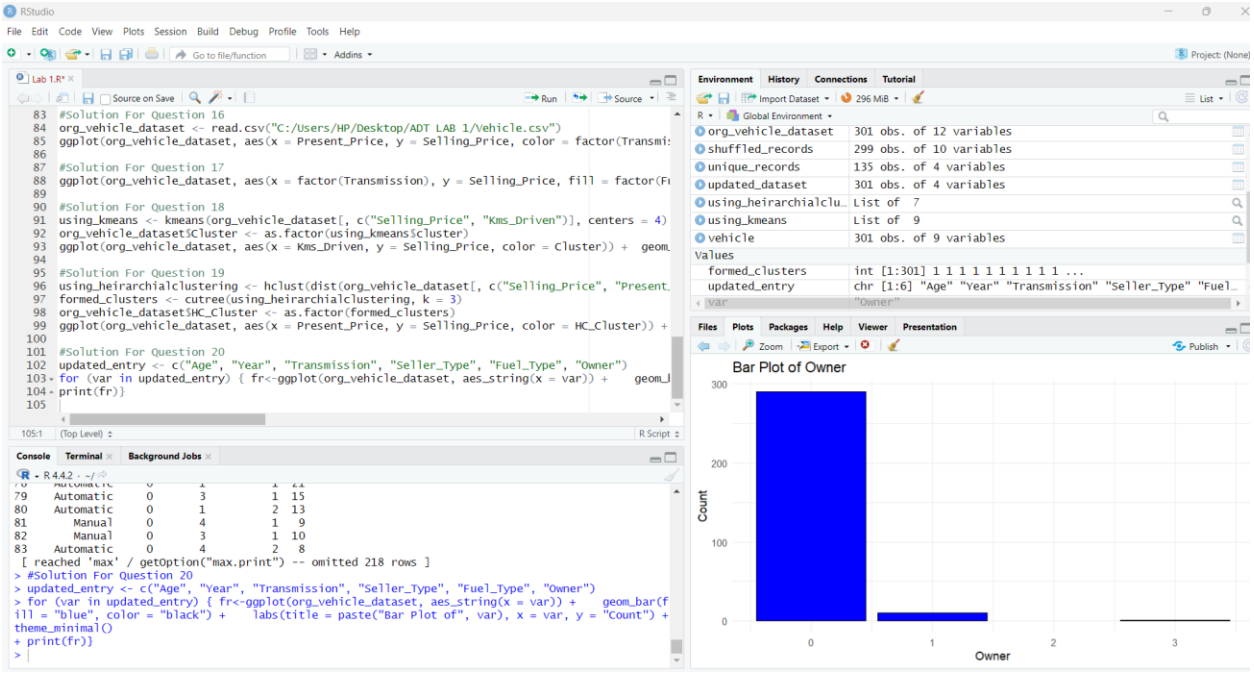


19. Create a scatter plot of the Selling_Price Vs Present_Price, and use hierarchical clustering to cluster the points into 3 clusters? Colour-code the points based on the cluster they belong to.

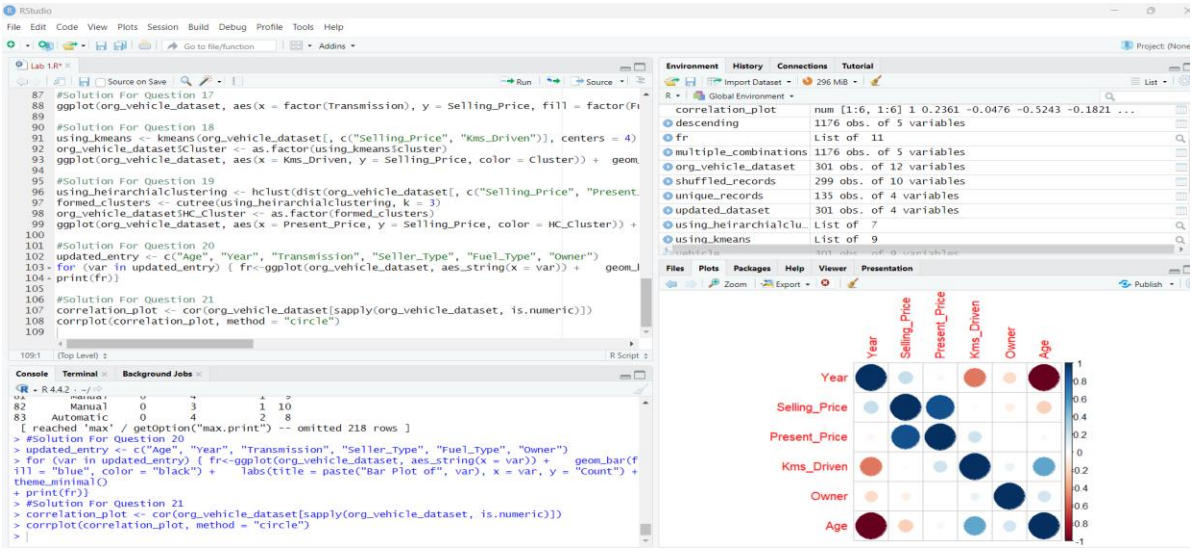


20. Add a new field called ‘Age’, and calculate it using the field ‘Year’. Create a barplot for the following fields of the dataset:

a) ‘Age’, ‘Year’, ‘Transmission’, ‘Seller_Type’, ‘Fuel_Type’ and ‘Owner’
b) Add labels, titles, and colours to the plot.



21. Create a correlation plot of the whole dataset variables and explain the output. Do not forget to convert some of the variable’s datatype if required and possible.



22. Create a scatter plot of the Selling_Price Vs Kms_Driven, and use DBSCAN clustering to cluster the points into 3 clusters. Color-code based on the cluster they belong to. Add a legend to the plot.