Assignment: Introduction to Data Science

Oksana Harapyn, Sergey Romadin

2024-06-24

Introduction

This project involves an exploratory data analysis (EDA) of the mpg dataset, which contains information on various car models and their fuel efficiency. The goal is to investigate key variables such as highway mileage (hwy), engine displacement (displ), and vehicle class to uncover patterns, outliers, and relationships that can inform our understanding of vehicle performance and efficiency. By utilizing statistical summaries, visualizations like histograms and boxplots, and methods for identifying outliers, this analysis aims to provide insights into the factors that influence fuel efficiency, the distribution of vehicle types, and trends over time.

Question 1: Investigation of the Variable hwy (Highway Mileage)

The summary statistics and histograms for highway mileage (hwy) show a mean of 23.44 mpg. The overall density histogram for hwy displays a bimodal distribution, indicating that there are several distinct groups of vehicles with different average highway mileage characteristics. This bimodal pattern suggests underlying differences in vehicle types, likely due to variations in design, weight, and intended use, such as fuel-efficient cars versus larger, less efficient vehicles.

The boxplot identifies a few outliers, which are vehicles with exceptionally high highway mileage. Overall, the data suggest that while there is some variation, most vehicles' highway mileage is clustered around the mean, with certain classes performing better than others.

Defining the measures

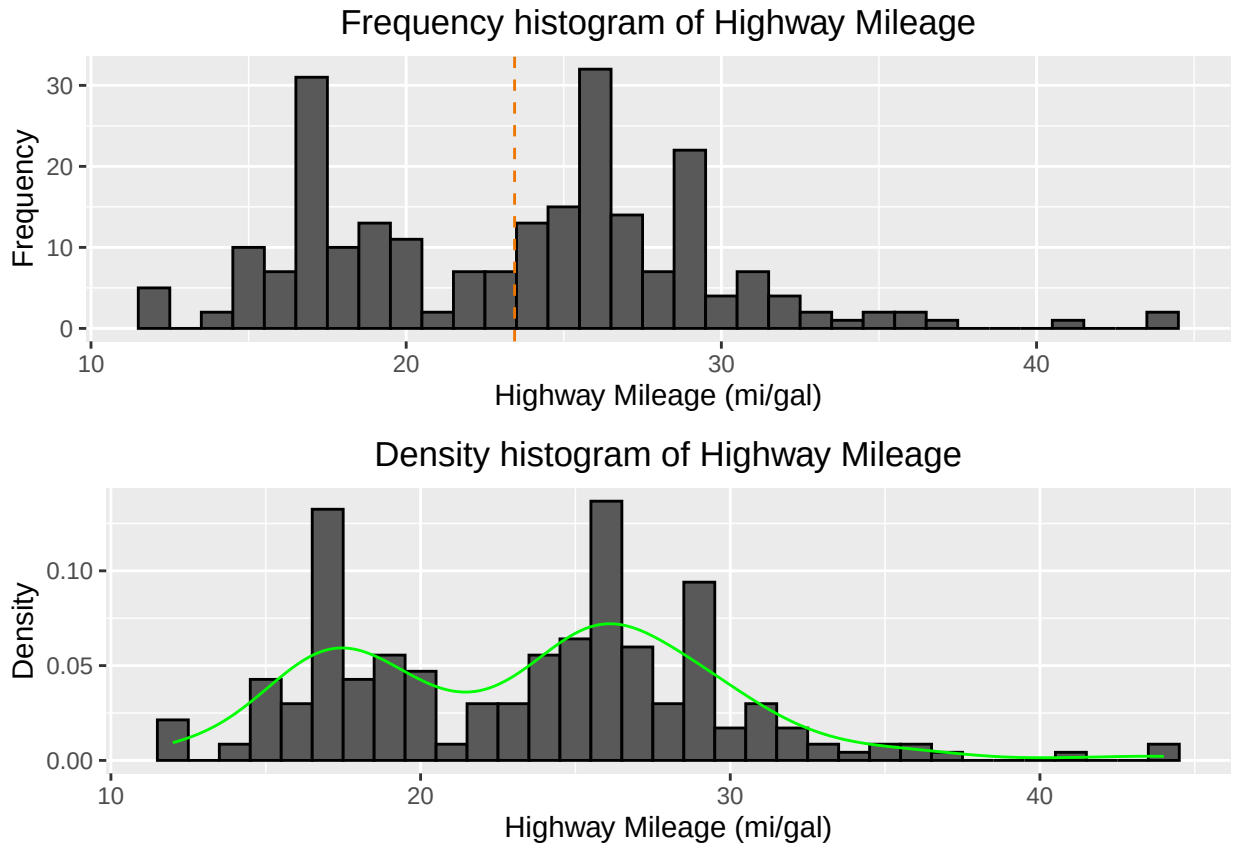
```
summary(mpg$hwy)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   18.00   24.00   23.44   27.00   44.00
```

Highway Mileage distribution

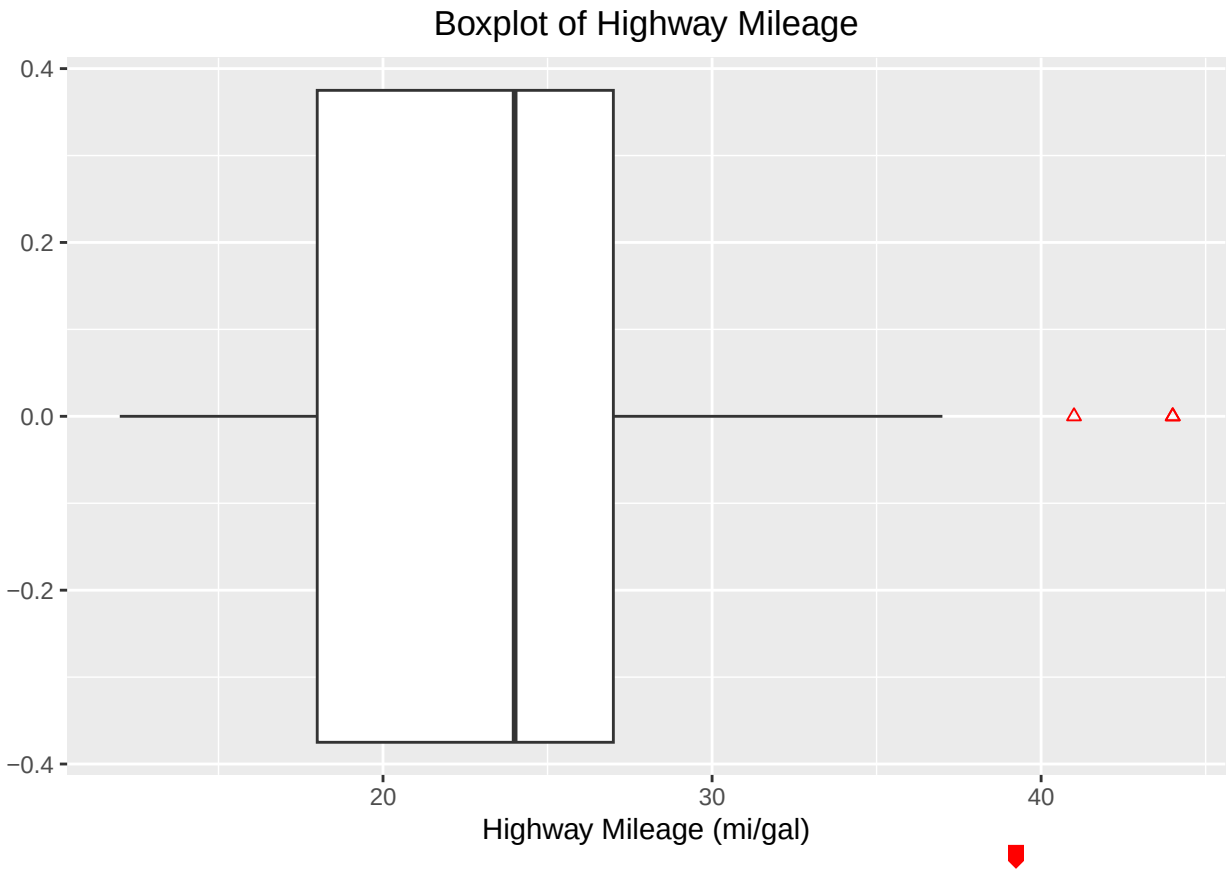
```
p1= ggplot(mpg, aes(x=hwy))+
  geom_histogram(color="black", binwidth = 1)+
  geom_vline(aes(xintercept=mean(hwy)), color="darkorange2", linetype="dashed")+
  ggtitle("Frequency histogram of Highway Mileage")+
  xlab("Highway Mileage (mi/gal)")+
  ylab("Frequency")+
  theme(plot.title = element_text(hjust = 0.5))+
```

```
  geom_text(x=21.5,y=42, label=" =23.44", color="darkorange2",size=3)

p2= ggplot(mpg, aes(x=hwy))+
  geom_histogram(color="black", binwidth = 1, aes(y=after_stat(density)))+
  ggtitle("Density histogram of Highway Mileage")+
  xlab("Highway Mileage (mi/gal)")+
  ylab("Density")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_text(x=21.5,y=42, label=" =23.44", color="darkorange2",size=3)+
  geom_density(color="green")

grid.arrange(p1,p2, nrow=2)
```



Investigation of Outliers using Boxplot

```
ggplot(mpg, mapping=(aes(x=hwy)))+
  geom_boxplot(outlier.shape =2, outlier.colour ="red")+
  ggtitle("Boxplot of Highway Mileage")+
  xlab("Highway Mileage (mi/gal)")+
  theme(plot.title = element_text(hjust = 0.5))
```



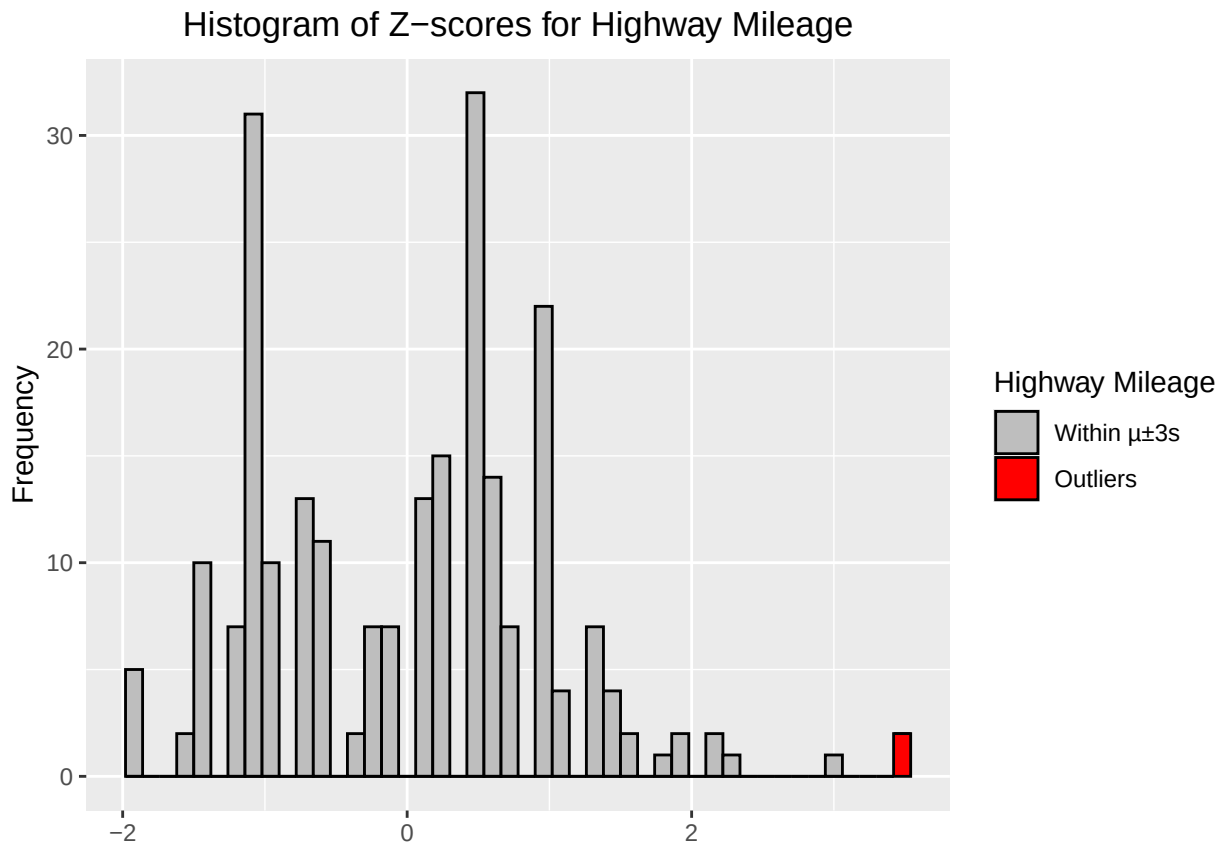
Question 2: Determination of Outliers using different methods

Using the Z-score method, outliers are identified as those points that are more than three standard deviations away from the mean. This method reveals a few outliers with high highway mileage. The IQR method, which looks at the spread of the middle 50% of the data, also identifies outliers but may be more robust against extreme values. Those outliers was detected by the Boxplot of Highway Mileage. Both methods highlight the presence of outliers in the highway mileage data, but they differ slightly in which data points they classify as outliers.

Z-score

```
mpg$hwyz=(mpg$hwy-mean(mpg$hwy))/sd(mpg$hwy)
data.frame(rand=mpg$hwyz,
           cut={
             sd=sd(mpg$hwyz)
             mn=mean(mpg$hwyz)
             cut(mpg$hwyz, c(-Inf, mn -3*sd, mn +3*sd, Inf))
           }) |>
  ggplot(aes(x=mpg$hwyz, fill=cut ))+
  geom_histogram(colour="black",binwidth = 0.12)+
  scale_fill_discrete(name="Highway Mileage", labels=c("Within ±3 ", "Outliers"), type=c("grey","red"))+
  ggtitle("Histogram of Z-scores for Highway Mileage")+
  xlab(NULL)+
```

```
  ylab("Frequency")+
  theme(plot.title = element_text(hjust = 0.5))
```




```
mean_hwy <- mean(mpg$hwy)
median_hwy <- median(mpg$hwy)
sd_hwy <- sd(mpg$hwy)
iqr_hwy <- IQR(mpg$hwy)
z_scores <- (mpg$hwy - mean_hwy) / sd_hwy
outliers_z <- mpg$hwy[abs(z_scores) > 3]
print(paste("Outliers detected by Z_score:", list(outliers_z)))
```

```
## [1] "Outliers detected by Z_score: c(44, 44)"
```

IQR method

```
Q1 <- quantile(mpg$hwy, 0.25)
Q3 <- quantile(mpg$hwy, 0.75)
IQR_hwy <- Q3 - Q1
lower_bound <- Q1 - 1.5 * IQR_hwy
upper_bound <- Q3 + 1.5 * IQR_hwy

outliers_iqr <- mpg$hwy[mpg$hwy < lower_bound | mpg$hwy > upper_bound]
```

```r
print(paste("Lower and upper bounds:", lower_bound, upper_bound))
```

```
## [1] "Lower and upper bounds: 4.5 40.5"
```

```r
print(paste("Outliers detected by IQR method:", list(outliers_iqr)))
```

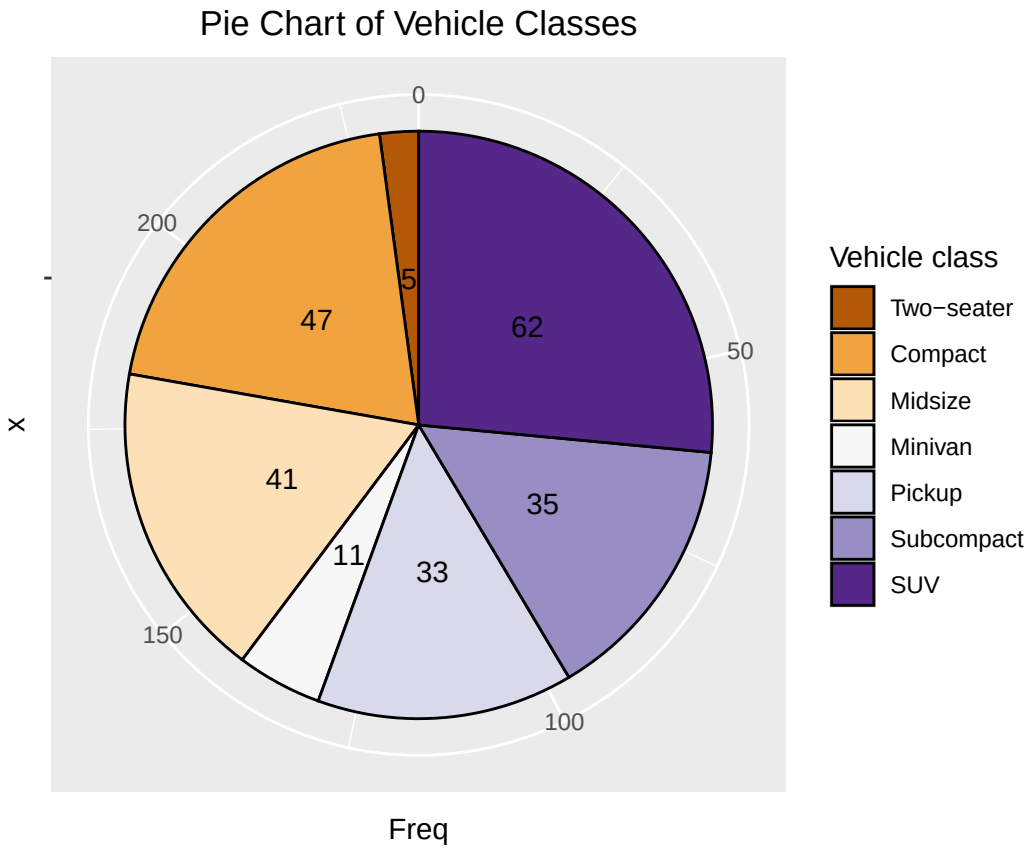
```
## [1] "Outliers detected by IQR method: c(44, 44, 41)"
```

Question 3: Making a Pie Chart of the Variable class (Vehicle Class)

The pie chart indicates a diverse distribution of vehicle types, with a relative equality in spread of most classes. SUVs and compact cars are the most common classes in the dataset. Conversely, the top two rarest classes are two-seaters and minivans. They show a notable scarcity compared to other types of vehicles. This information can be useful for understanding the market composition and targeting specific segments for analysis or marketing.

Pie Chart of Variable class

```r
a=as.data.frame(table(mpg$class))
ggplot(a, aes(x="", y=Freq, fill=Var1))+
  geom_bar(stat="identity", width=1, color="black")+
  coord_polar("y", start=0)+
  scale_fill_brewer(name="Vehicle class",  labels=c("Two-seater", "Compact", "Midsize", "Minivan", "Pic
  geom_text(aes(label=Freq), position=position_stack(vjust = 0.5))+
  ggtitle("Pie Chart of Vehicle Classes")+
  theme(plot.title = element_text(hjust = 0.5))
```



Question 4: Investigating the Change in the Number of Outliers When Separating hwy by Class

Separating highway mileage by vehicle class reveals variations in the number of outliers across different classes. For example, two-seaters and subcompacts tend to have higher highway mileage, whereas pickups have lower mileage. The boxplots and summary statistics provide insights into the distribution of highway mileage across different vehicle classes. More outliers are found in separate vehicle classes because the variation within each class is smaller compared to the entire dataset. This means that individual classes show a more detailed spread of data points, which highlights specific outliers that may be diluted when considering the overall dataset. This detailed breakdown helps in understanding the specific performance and variability within each vehicle class.

Boxplots by Variable class

```
ggplot(mpg, aes(x="", y=hwy)) +
  geom_boxplot(outlier.shape =2, outlier.colour ="red")+
  facet_wrap(~class)+
  ggtitle("Boxplots of Highway Mileage for different Vehicle types")+
  ylab("Highway Mileage (mi/gal)")+
  xlab(NULL)
```

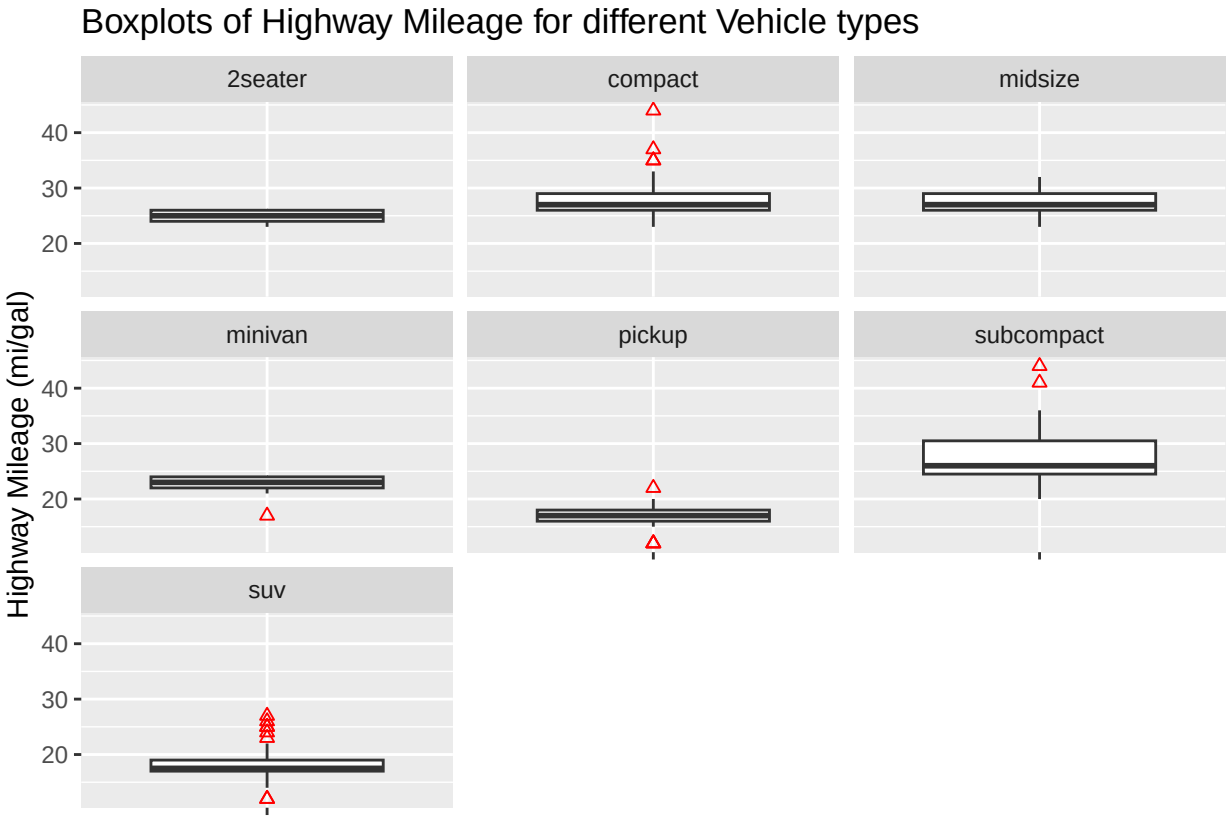


Table of Statisctics

```
f=function(x) c(
  mean = mean(x),
  median = median(x),
  sd = sd(x),
  q0=quantile(x, 0.25),
  q0=quantile(x, 0.75)
  )
table=do.call(data.frame, aggregate(hwy~class, mpg, f))
header=table[,1]
t_table=t(table[, -1])
colnames(t_table)=header
as.data.frame(t_table)
```

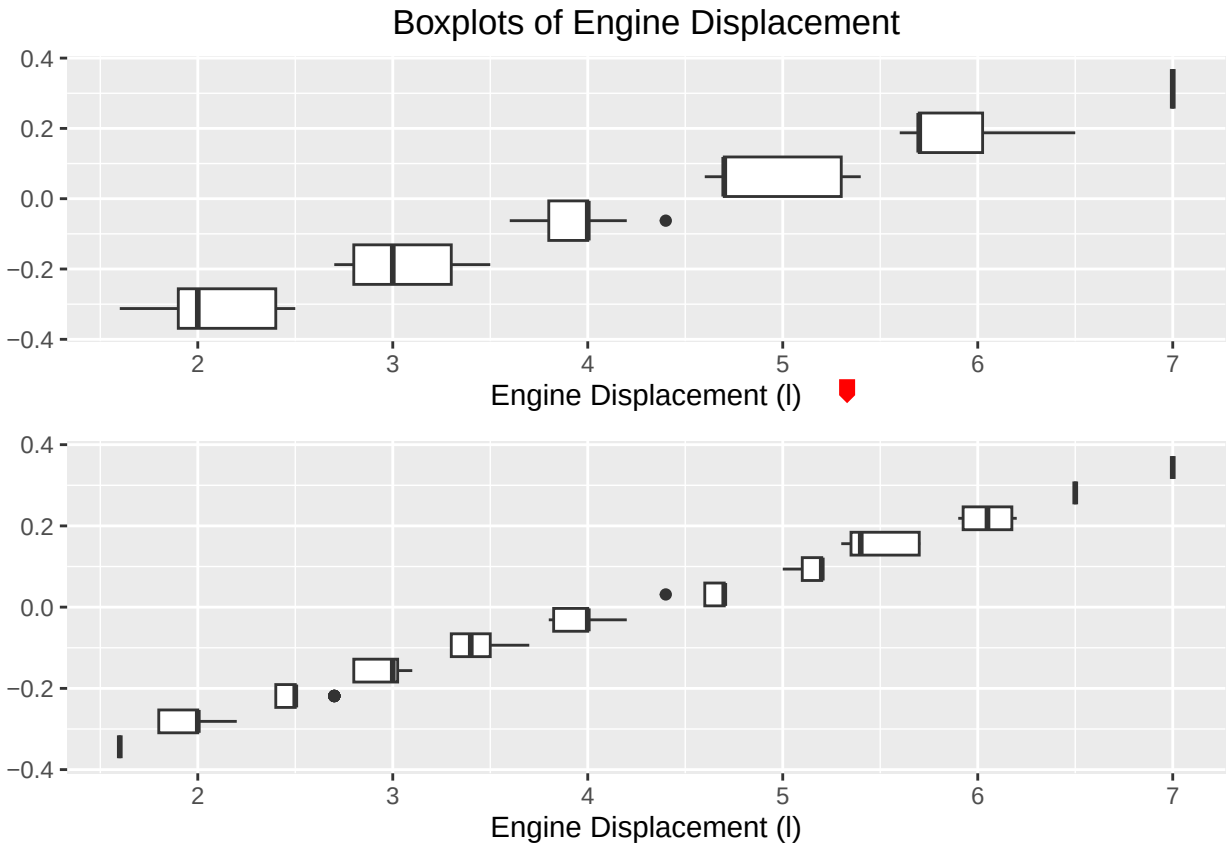
```
##             2seater  compact  midsize   minivan   pickup subcompact       suv
## hwy.mean   24.80000 28.29787 27.29268 22.363636 16.87879  28.142857 18.129032
## hwy.median 25.00000 27.00000 27.00000 23.000000 17.00000  26.000000 17.500000
## hwy.sd      1.30384  3.78162  2.13593  2.062655  2.27428   5.375012  2.977973
## hwy.q0.25. 24.00000 26.00000 26.00000 22.000000 16.00000  24.500000 17.000000
## hwy.q0.75. 26.00000 29.00000 29.00000 24.000000 18.00000  30.500000 19.000000
```

Question 5: Investigation of Boxplots Change

The boxplots of engine displacement with different group widths (1 and 0.5 liters) show how the choice of grouping can affect the interpretation of data - 0.5 liters being more precise grouping. Smaller group widths provide more detailed insights, while larger widths give a broader overview. It is also easier to interpret the boxplot with a larger widths. This demonstrates the importance of selecting appropriate group widths for analyzing continuous data.

Boxplots of Variable displ (engine displacement)

```
p3=ggplot(mpg, mapping=(aes(x=displ))) +
  geom_boxplot(varwidth = FALSE, aes(group = cut_width(displ, 1)))+
  ggtitle("Boxplots of Engine Displacement")+
  xlab("Engine Displacement (l)")+
  theme(plot.title = element_text(hjust = 0.5))
p4=ggplot(mpg, mapping=(aes(x=displ))) +
  geom_boxplot(varwidth = FALSE, aes(group = cut_width(displ, 0.5)))+
  xlab("Engine Displacement (l)")+
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(p3,p4, nrow=2)
```



Question 6: Investigating Influence of Year on hwy

The histograms and mean comparison of highway mileage between the years 1999 and 2008 show little difference in the average highway mileage between these years. The values of both means are almost equal to the computed average highway mileage in the Question 1. The bi-modal distribution also stays according to the histogram. This suggests that different years of construction does not explain bi-modal distribution, which hints that such distribution is supported by other difference in groups of cars.

Histogram by year

```
year.mean=aggregate(mpg$hwy, list(mpg$year), FUN=mean)
colnames(year.mean)=c("year","mean")
ggplot(mpg, aes(x=hwy))+
  geom_histogram(color="black", binwidth = 1)+
  facet_wrap(~year)+
  geom_vline(data=year.mean, aes(xintercept=mean, color="Mean"),
             linetype="dashed")+
  scale_color_manual(name = "Statistics", values = c(Mean = "darkorange2"))+
  ggtitle("Histograms of Highway Mileage for 1999 and 2008")+
  xlab("Highway Mileage (mi/gal)")+
  ylab("Frequency")+
  theme(plot.title = element_text(hjust = 0.5))
```

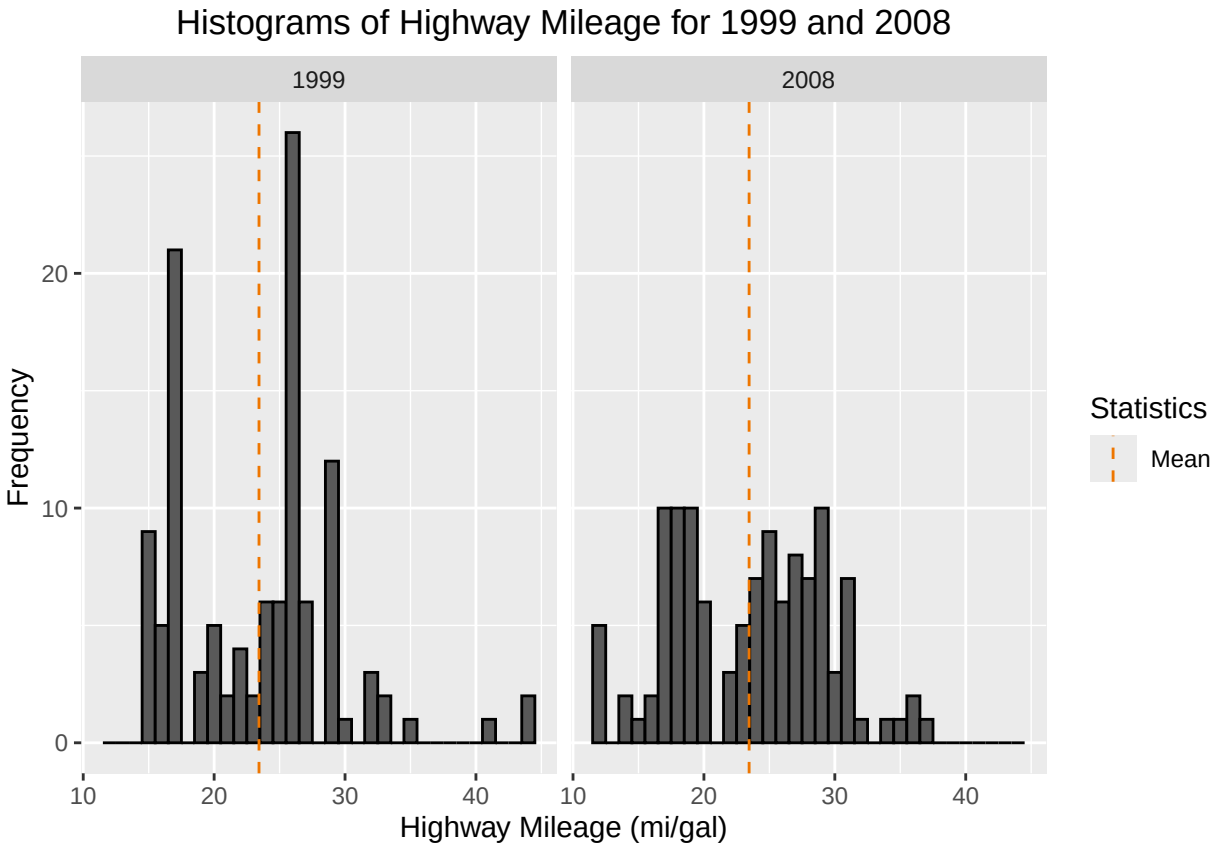


Table of hwy mean by year

```
year.mean
```

```
##   year     mean
## 1 1999 23.42735
## 2 2008 23.45299
```

Question 7: Investigating Influence of Class on hwy

The histograms and means of highway mileage by vehicle class reveal significant differences between classes. The histograms also follow strong unimodal distribution. This indicates that class of the car is the cuase of the initial hwy histogram to follow bimodal distribution. Compact and subcompact cars have the highest average highway mileage, while pickups have the lowest. This highlights the impact of vehicle design and purpose on fuel efficiency, and suggests areas where improvements can be made to enhance fuel economy across different vehicle types.

The scatterplot above shows the relationship between highway mileage and city millage while separating data points by classes. Overall, there is a strong positive relationship between hwy and cty in this sample. SUV and pickups tend to have lower total mileage than other classes, while compact and subcompact classes total mileage may vary greatly from average to high.

Histograms by Variable class

```
class.mean=aggregate(mpg$hwy, list(mpg$class), FUN=mean)
colnames(class.mean)=c("class","mean")
ggplot(mpg, aes(x=hwy))+
  geom_histogram(color="black", binwidth=1)+
  facet_wrap(~class)+
  geom_vline(data=class.mean, aes(xintercept=mean, color="Mean"),
             linetype="dashed")+
  scale_color_manual(name = "Statistics", values = c(Mean = "darkorange2"))+
  ggtitle("Histograms of Highway Mileage for different Vehicle types")+
  xlab("Highway Mileage (mi/gal)")+
  ylab("Frequency")+
  theme(plot.title = element_text(hjust = 0.5))
```

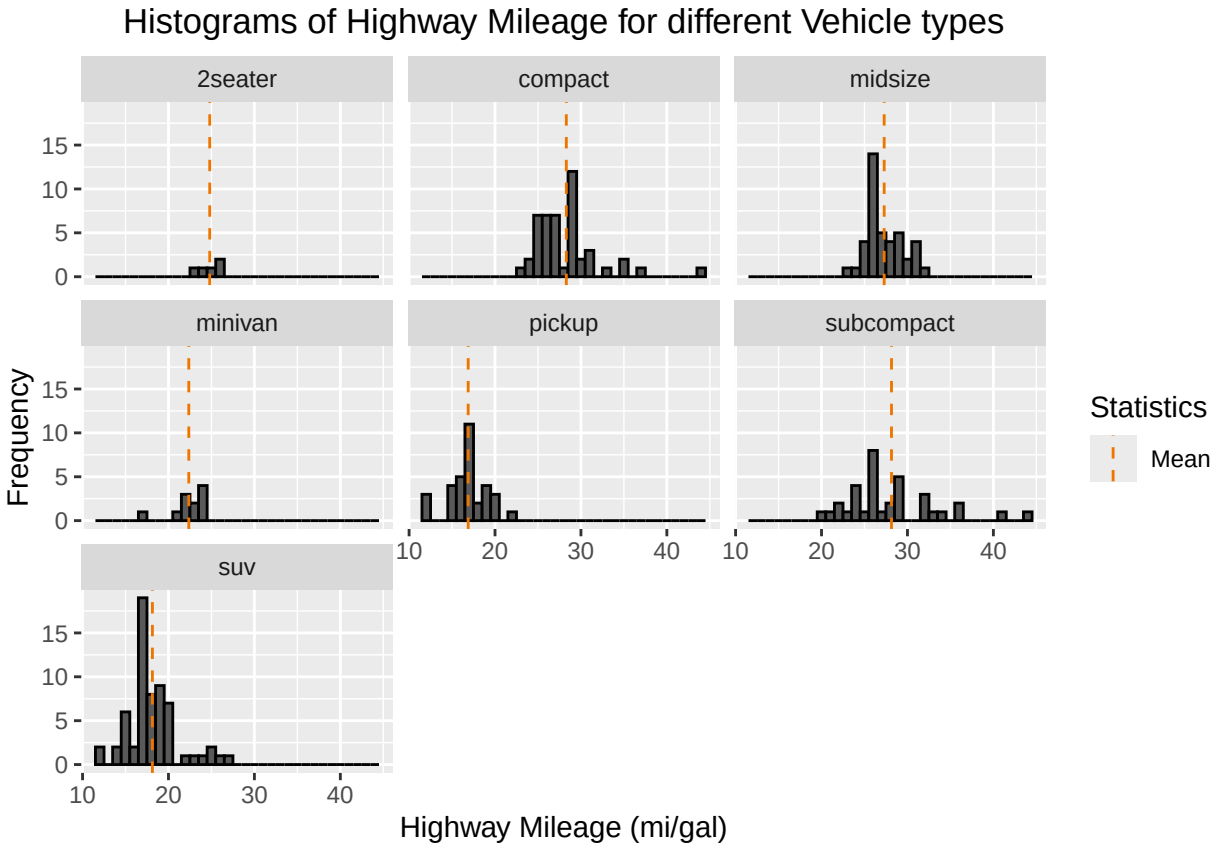


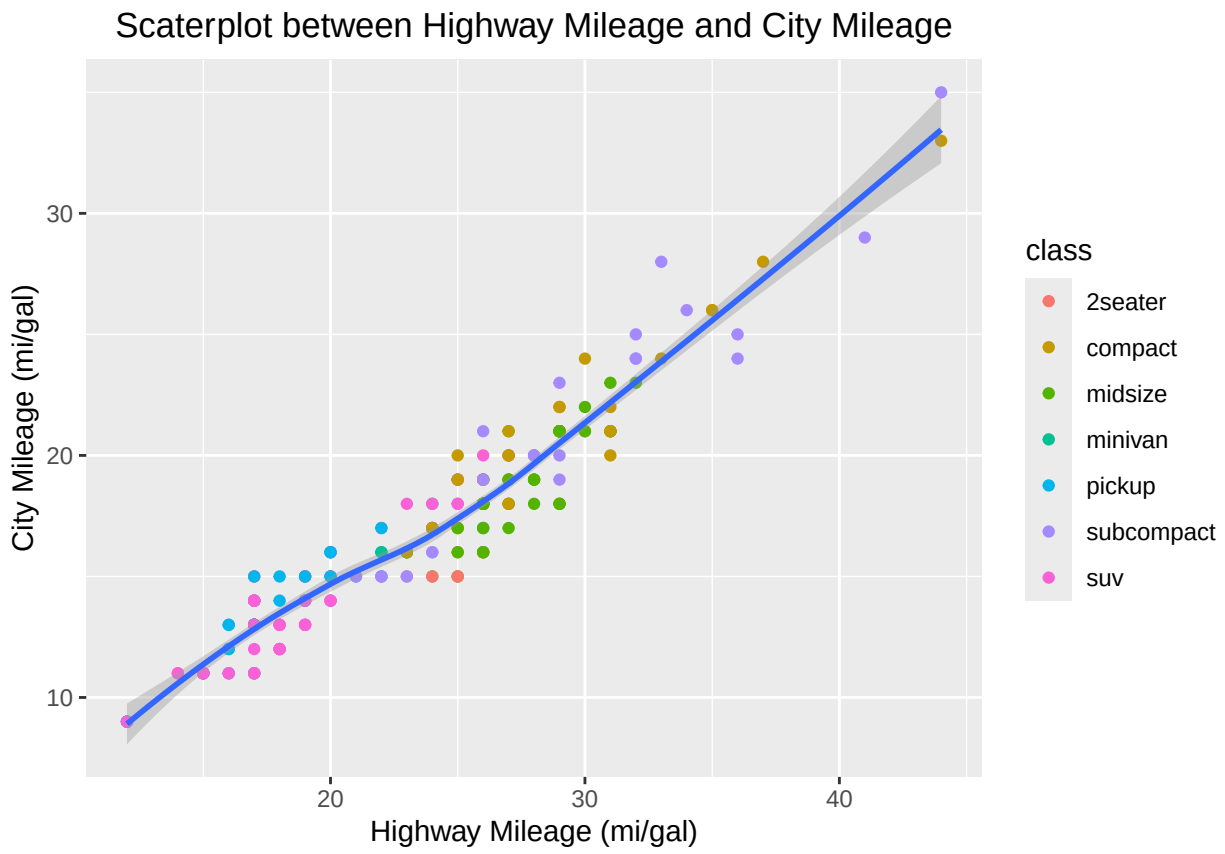
Table of hwy mean by class

```
class.mean
```

```
##        class     mean
## 1    2seater 24.80000
## 2    compact 28.29787
## 3    midsize 27.29268
## 4    minivan 22.36364
## 5     pickup 16.87879
## 6 subcompact 28.14286
## 7        suv 18.12903
```

Scatterplot between the Variables cty and hwy

```
ggplot(mpg, aes(x=hwy, y=cty))+
  geom_point(aes(color=class))+
  ggtitle("Scaterplot between Highway Mileage and City Mileage")+
  xlab("Highway Mileage (mi/gal)")+
  ylab("City Mileage (mi/gal)")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_smooth(formula = y ~ x, method = "loess")
```



Conclusion

This analysis of the mpg dataset highlights several key insights into vehicle fuel efficiency. The highway mileage (hwy) distribution shows distinct groups of vehicles with different fuel efficiencies, emphasizing the importance of vehicle class. Compact and subcompact cars generally achieve higher mileage compared to SUVs and pickups, indicating that vehicle design and purpose significantly impact fuel efficiency.

Comparing highway mileage between 1999 and 2008 indicates no significant improvement in fuel efficiency over time, highlighting a potential area for innovation. The relationship between engine displacement and highway mileage demonstrates that smaller engines generally achieve better mileage, further emphasizing the role of vehicle design.

Outlier analysis revealed more outliers within individual vehicle classes, suggesting greater variability in specific classes. The pie chart analysis shows that SUVs and compact cars are the most common, while two-seaters and minivans are the rarest, providing insights into market composition.

The mean highway mileage differs significantly across vehicle classes (being more to the left/more to the right compared to the "original" mean of hwy distribution in Question 1), which, in combination with relative equality in spread of those classes, explains the bimodal distribution in the overall histogram. Additionally, the scatterplot of city mileage (cty) versus highway mileage (hwy) illustrates a clear correlation between these variables, underscoring the importance of class-specific analysis for a comprehensive understanding of fuel efficiency trends. These findings can guide efforts to improve fuel economy and reduce environmental impact.