Homework_3

Sai Saketh Boyanapalli

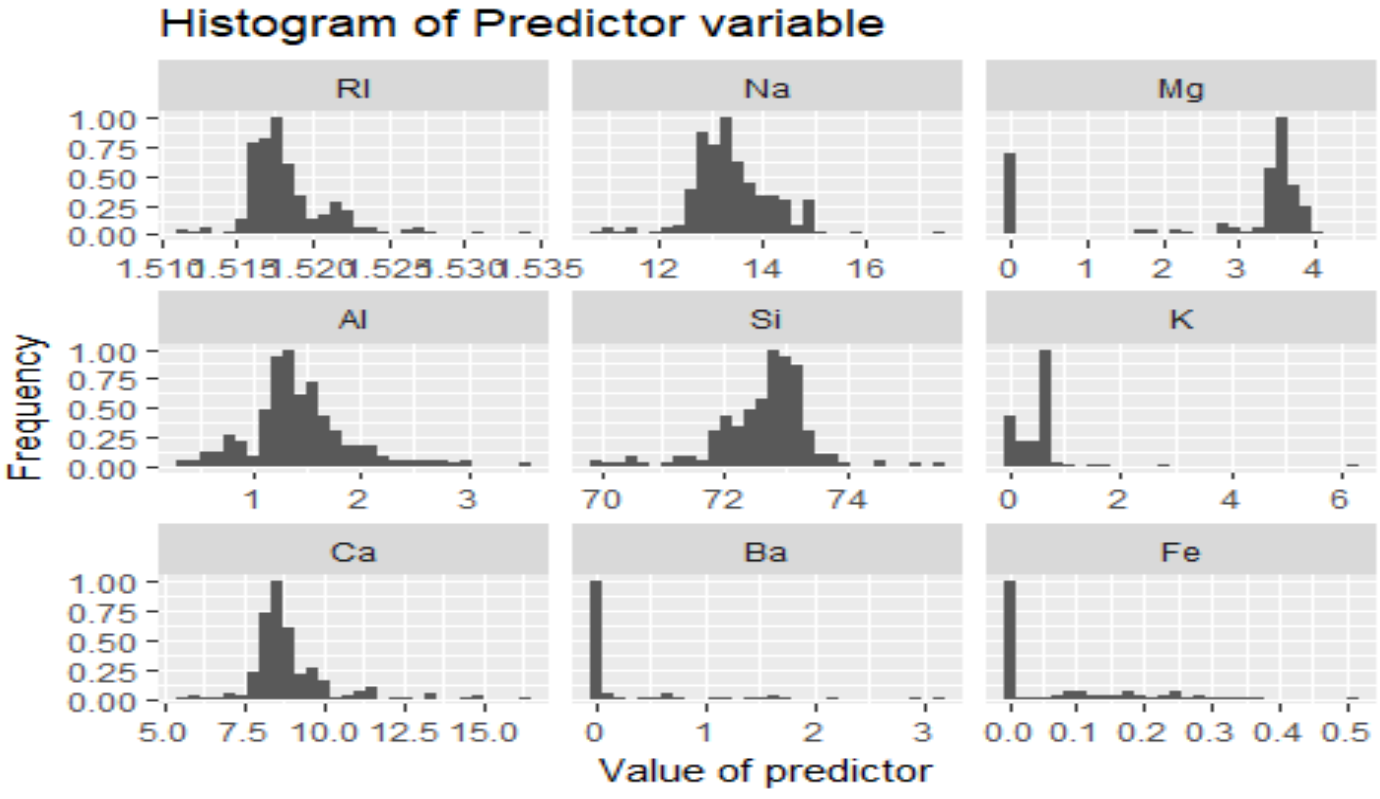
October 4, 2017

```
# library's required
library(EnvStats)
library(mlbench)
library(reshape2)
library(ggplot2)
library(car)
library(scales)
library(gridExtra)
library(Amelia)
library(mice)
library(VIM)
```

1 Glass Identification

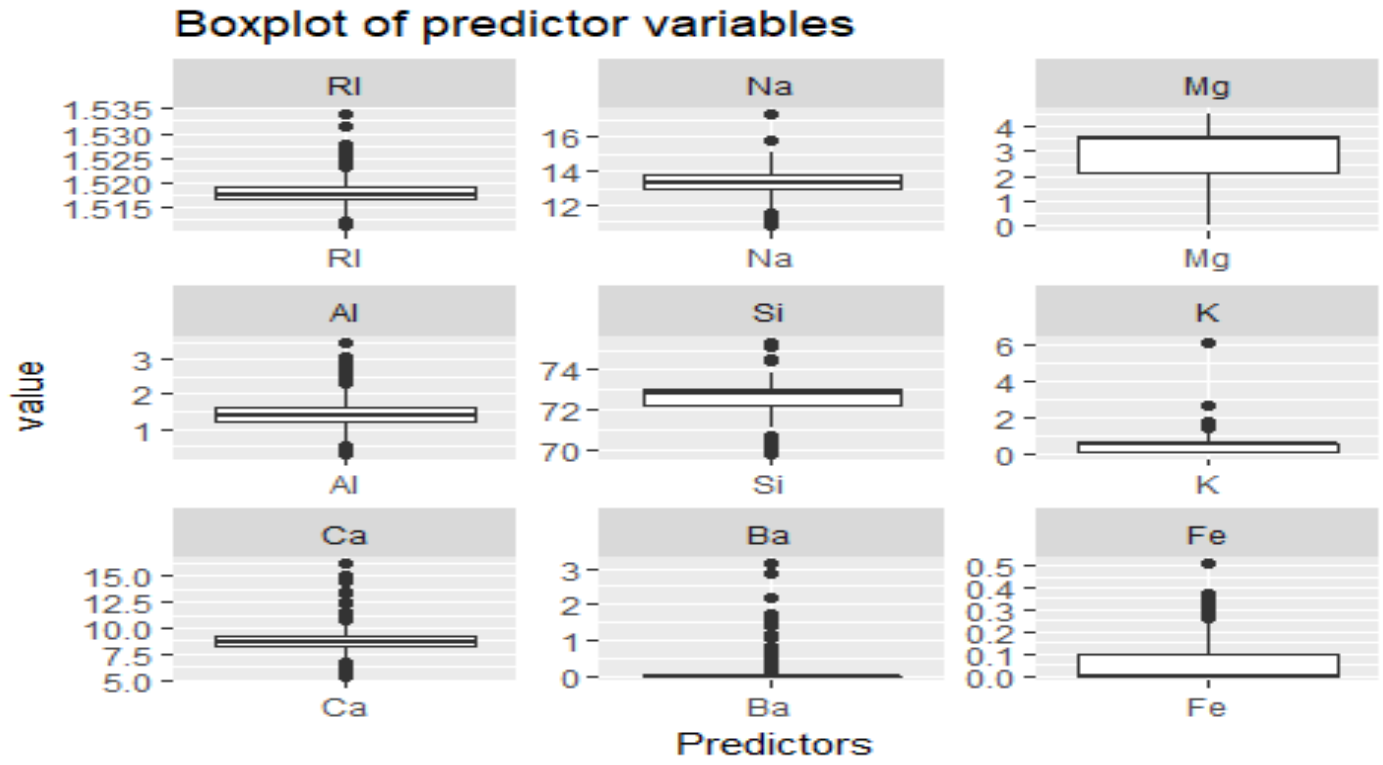
1(a)

```
library(mlbench)
data("Glass") # loading Glass data
names(Glass) # looking at column names for the data

##  [1] "RI"   "Na"   "Mg"   "Al"   "Si"   "K"    "Ca"   "Ba"   "Fe"   "Type"

str(Glass) # looking at structure of the data
GlassMelt <- melt(Glass[,-10]) # melting data into one column
```



We can see that some of the histograms are skewed and some of them are normally distributed.

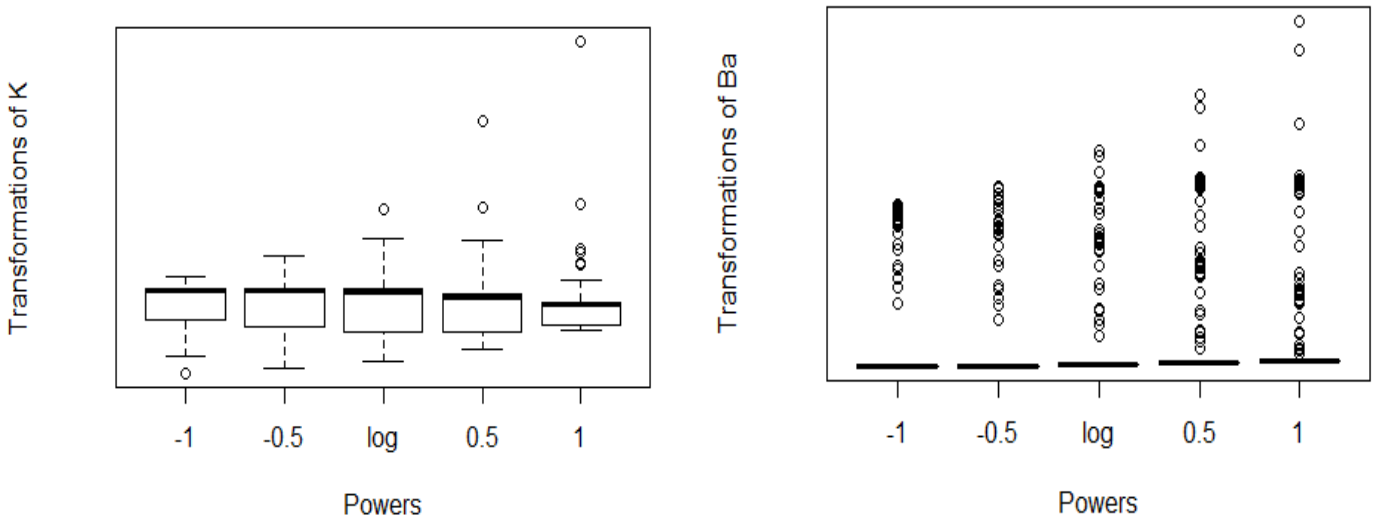
```
ggplot(GlassMelt, aes(factor(variable), value))+
  geom_boxplot() + facet_wrap(~variable, scale="free") +
  xlab("Predictors") +
  ggtitle("Boxplot of predictor variables")
```

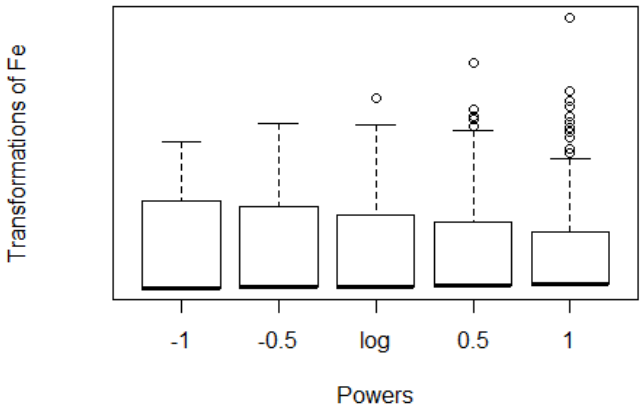


Here we can see that there are lot of outliers in the data using a boxplot. except for Mg there are no outliers in the data.

1(b) I Choose Predictors 'K', 'Ba', 'Fe' as my skewed variables from the histogram

1 (b) i)





In case of symbox transformation of K it fits perfects for power -0.5, and it does not fit to powers for Ba and Fe it is fitting for powers -1, -0.5

1 (b) (ii)

```
library(EnvStats) # to access Box-Cox fuction

boxcox(Glass$K, lambda = c(-3,3), optimize = T)

boxcox(Glass$Ba, lambda = c(-2,2), optimize = T)

boxcox(Glass$Fe, lambda = c(-1,1), optimize = T)
```

The optimal value of Lambda for K, Ba, Fe are 0.088, -0.582, 0.054

1(c)

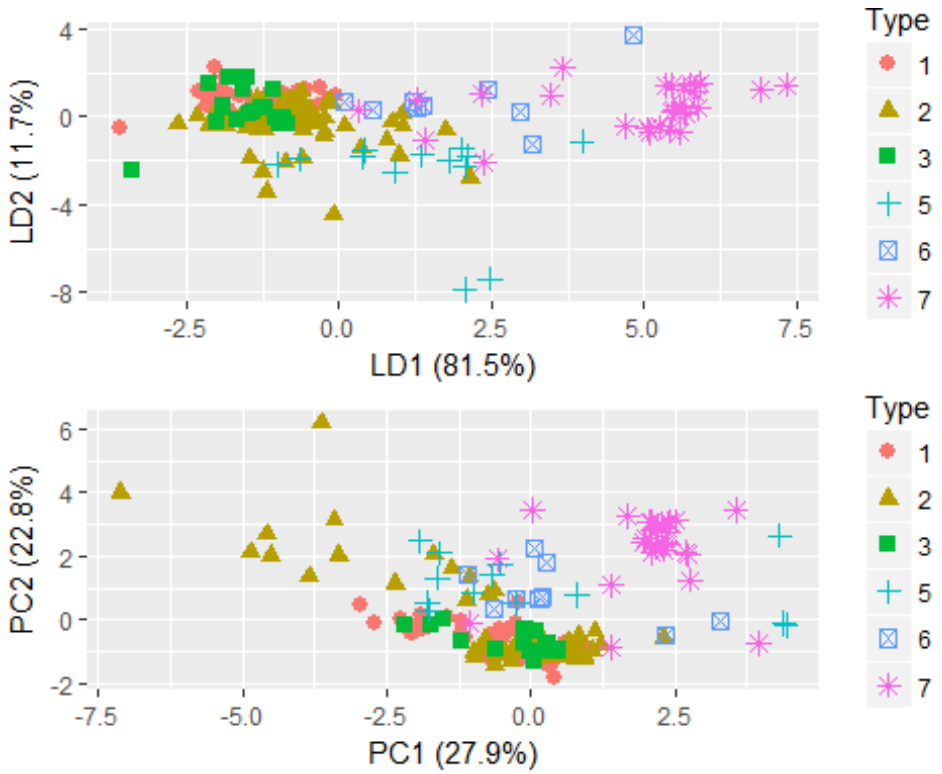
```
## Importance of components%s:
##                           PC1    PC2    PC3    PC4    PC5     PC6    PC7
## Standard deviation     1.585 1.4318 1.1853 1.0760 0.9560 0.72639 0.6074
## Proportion of Variance 0.279 0.2278 0.1561 0.1286 0.1016 0.05863 0.0410
## Cumulative Proportion  0.279 0.5068 0.6629 0.7915 0.8931 0.95173 0.9927
##                           PC8     PC9
## Standard deviation     0.25269 0.04011
## Proportion of Variance 0.00709 0.00018
## Cumulative Proportion  0.99982 1.00000
```

After running Principal component analysis on Glass we can see that PC1 - PC7 holds about 99% of the variance in our data. upto PC5 can provide information about 90% of the data so, we can reduce the dimensions from 9 - 5 using PCA.

1(d)

```
## Proportion of trace:
##    LD1    LD2    LD3    LD4    LD5
## 0.8145 0.1169 0.0413 0.0163 0.0111
```

We can see that Linear discriminants LD1 and LD2 capture around 93% feature separation in the data.



Here PCA classifies data on variance and LDA tries to classify data based on features. In this case LDA does a better job in differentiating predictor variables than PCA.

Question 2 Missing Data

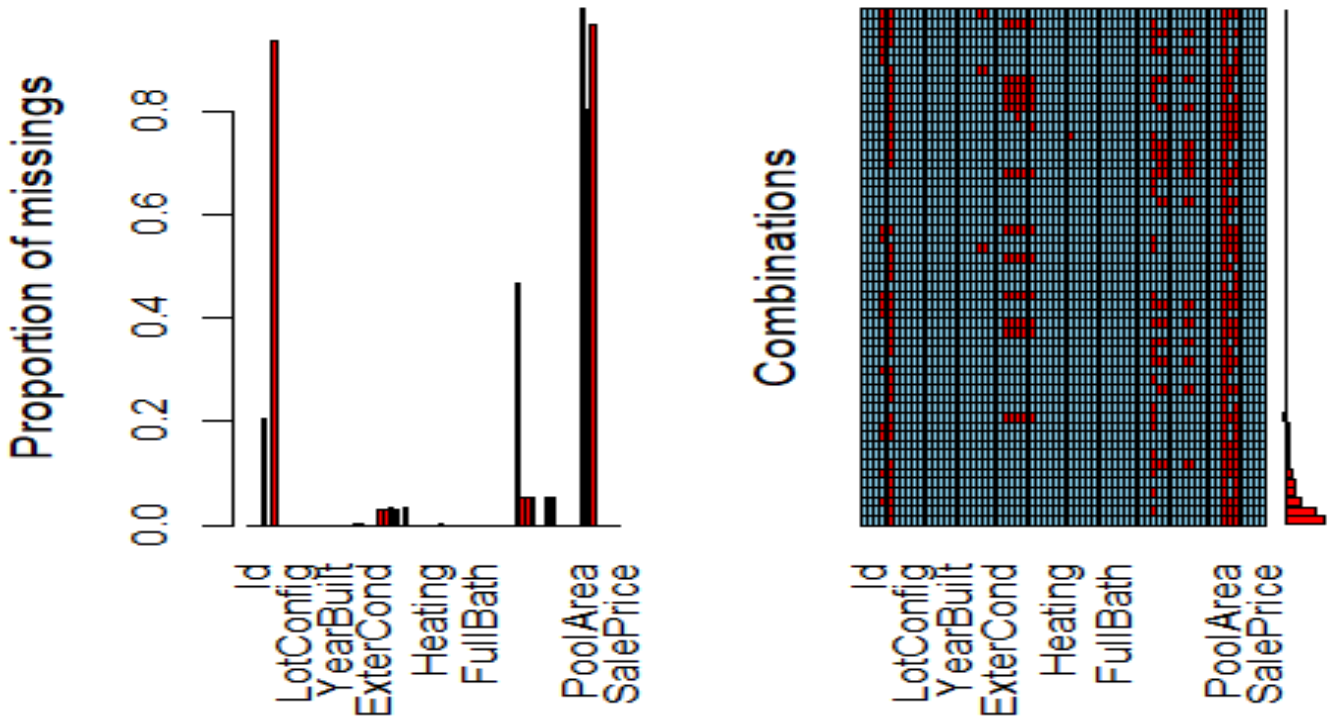
2 (a, b, c) Code Provided.

2(d)

In the above 3 cases the values for the Coefficients are different for each case, in case of listwise deletion we are reducing the scope of our data by removing all the rows without information this will reduce the variability in our data and the adjusted R - squared is found to be 0.9 in case of listwise deletion. Mean imputation tries to preserve the variability of the data by imputing the missing values with mean. but this case the data is skewed to a particular value because the missingness is imputed by mean now the data is skewed to the mean. In the above test some of the coefficients have a major change while some remain constant. the Adjusted R - square is 0.61. In the case of multiple imputation with chained equations it ries to preserve the variablity but but due to multiple iteraations the adjusted R - squared values goes down to 0.8. ideally we want R - squared to be 1. here this was achieved close by listwise deletion. ###Question 3 House prices data

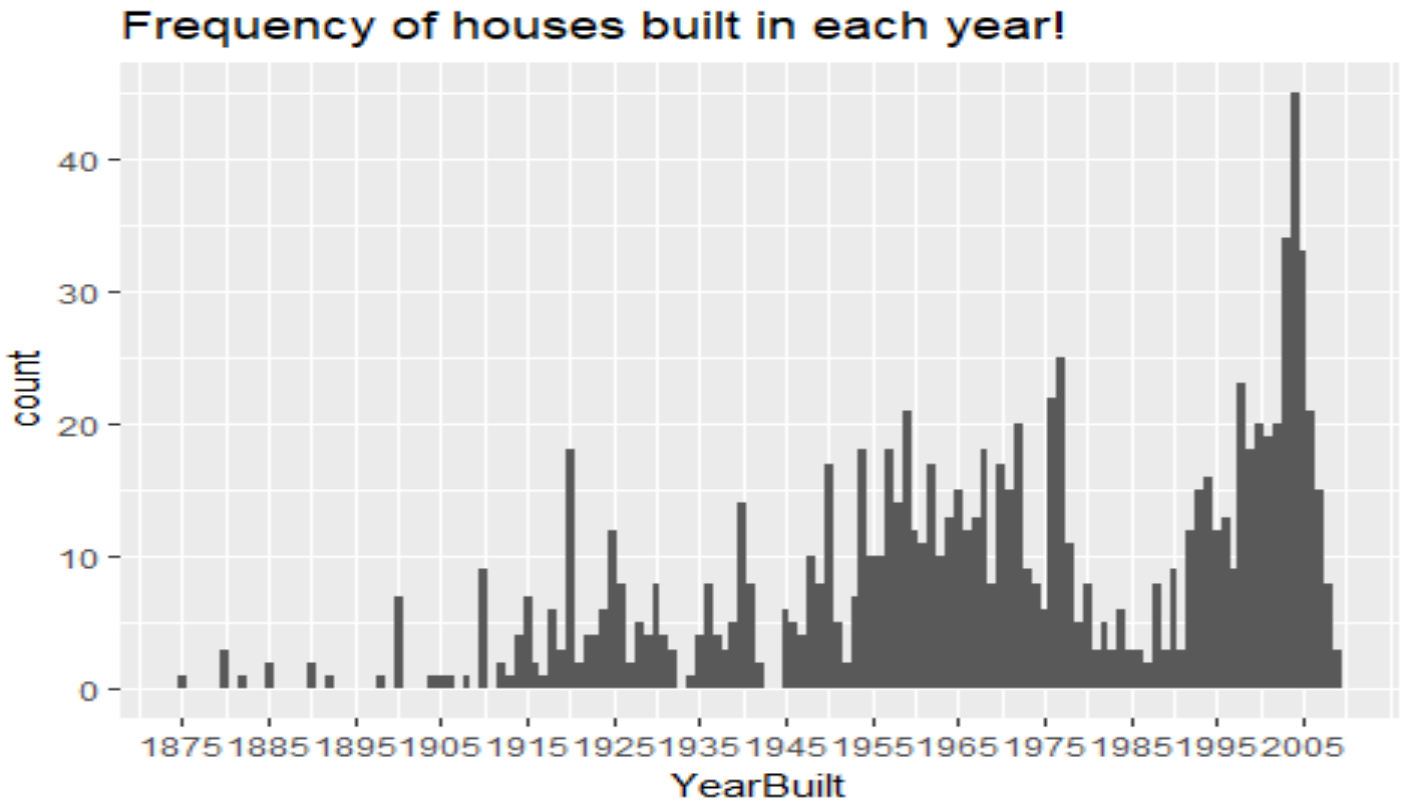
3 (a) Explore and visualize data.

```
housing_data <- read.csv("housingData.csv")
missing <- aggr(housing_data)
```



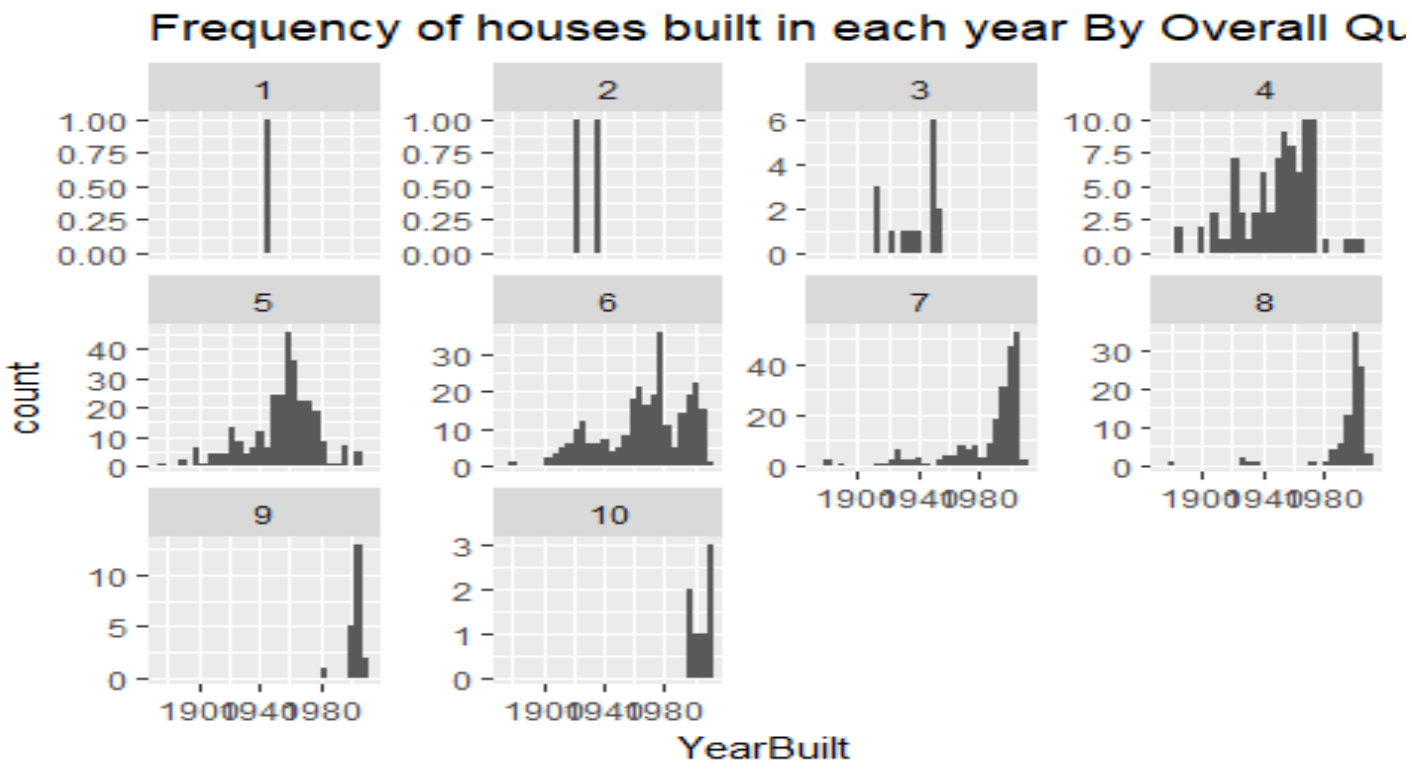
We can see the variables with most missing values. This missingness in the variables can be due to other reasons for example if we take pool area into account not every house has a pool so, that the fact of mssingness there.

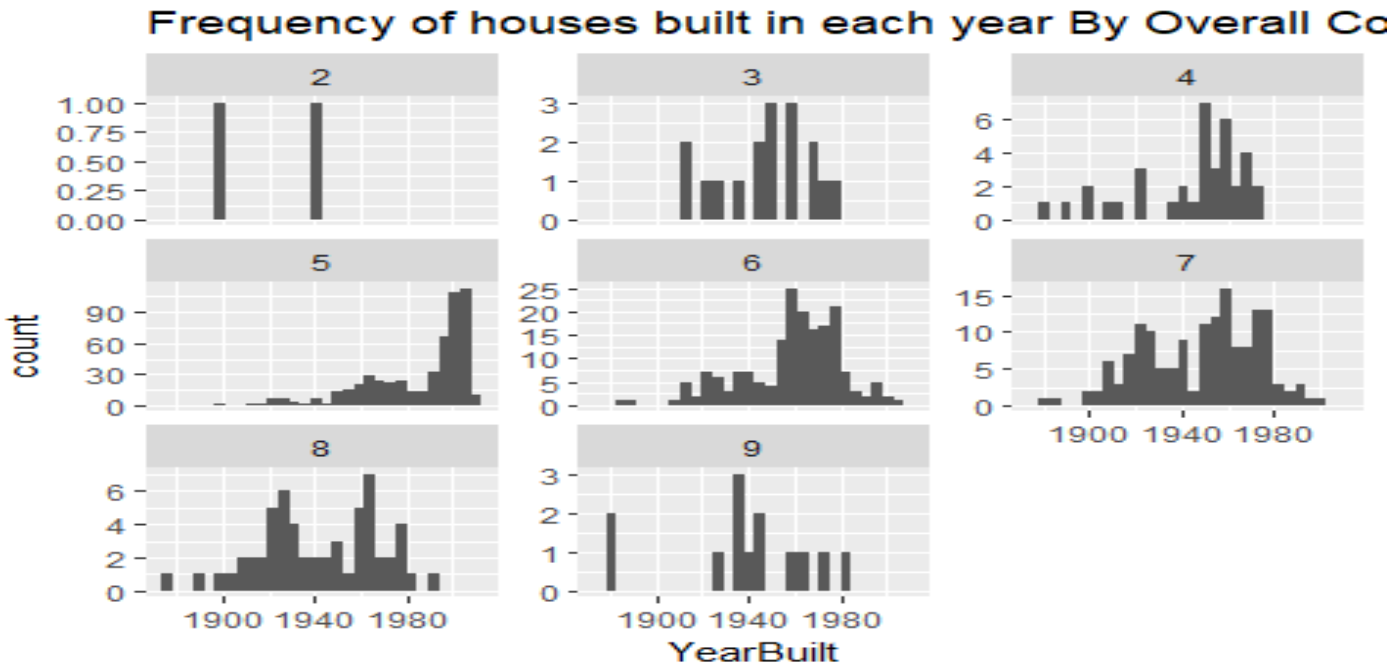
```
ggplot(aes(x = YearBuilt),data = subset(housing_data, !is.na(housing_data$YearBuilt))) +
  geom_histogram(binwidth = 1) +
  scale_x_continuous(breaks = seq(1875, 2009, 10)) +
  ggtitle('Frequency of houses built in each year!')
```



Here we can see that houses built over the year increased and peaked at year 2004. see a downward trend after that. this might be due to recession.

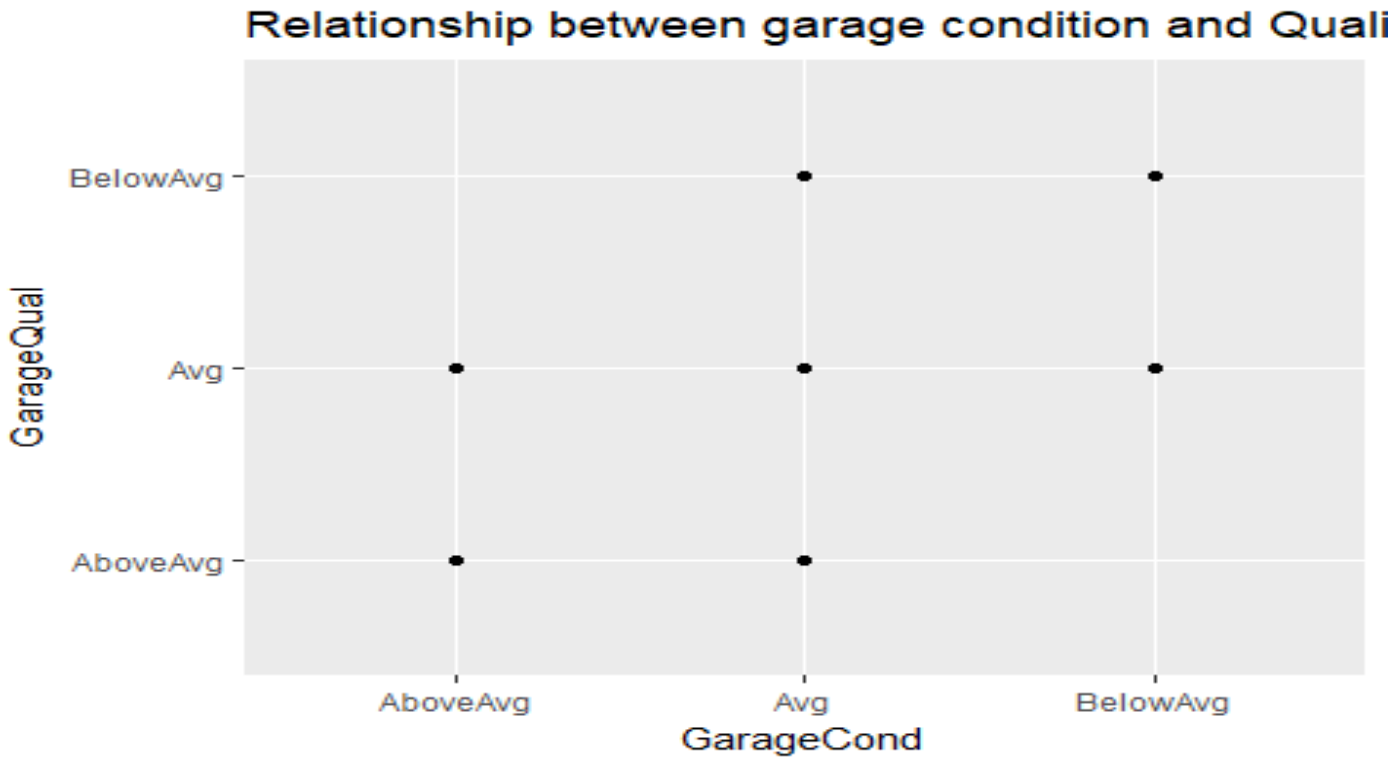
```
range(housing_data$YearBuilt) # Shows yaer built for oldest and newest house in data.

## [1] 1875 2009

ggplot(aes(x = YearBuilt),data = subset(housing_data, !is.na(housing_data$YearBuilt))) +
  geom_histogram() +
  ggtitle('Frequency of houses built in each year By Overall Quality') +
  facet_wrap(~OverallQual, scales = "free_y")
```





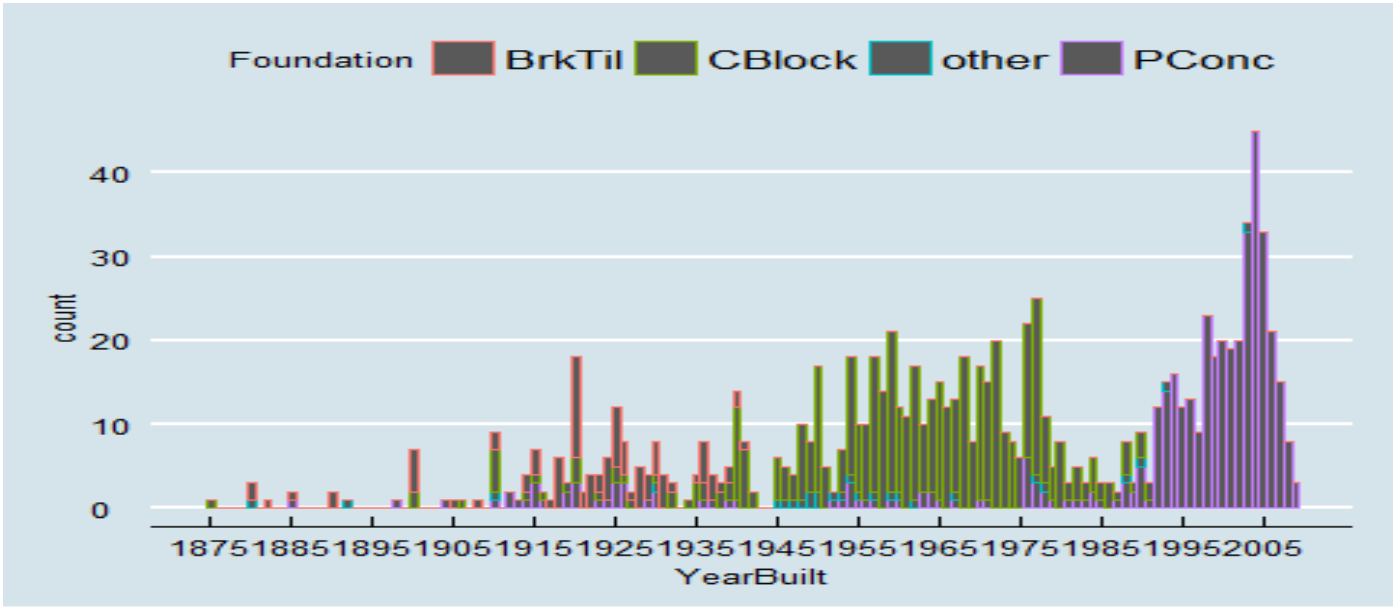
Here we can see that there are just 2 houses in very poor condition and there are lot of houses with Average > Above average > Good Condition. and most of them are built after year 1940. We can see that half of the houses are Above Average in overall quality and condition.

```
ggplot(aes(x = GarageCond, y = GarageQual), data = subset(housing_data,
!is.na(housing_data$GarageQual))) +
  geom_point() + ggtitle('Relationship between garage condition and Quality')
```



Here we can see that AboveAverage Garage quality does not have BelowAvg Garage Condition and viceVersa.

```
ggplot(aes(x = YearBuilt), data = housing_data) +
  geom_histogram(aes(color = Foundation),binwidth = 1) +
  ggthemes::theme_economist() +
  scale_x_continuous(breaks = seq(1875,2009, 10))
```



We can see that Poured Concrete foundation of houses started around year 1990.

3 (b)

```
# some feature construction
housing_data$YearsUsed <- housing_data$YrSold - housing_data$YearBuilt # No of year house
was used.
housing_data$TotalNoFullBath <- housing_data$BsmtFullBath - housing_data$FullBath # Total
number of fullbath's in the house.
housing_data$TotalNohalfBath <- housing_data$BsmtHalfBath - housing_data$HalfBath # Total
number of fullbath's in the house.
housing_data$TotalBathRooms <- housing_data$TotalNoFullBath +
housing_data$TotalNohalfBath # Total no of bathrooms in the house.
housing_data$TotalFloorsqft <- housing_data$X1stFlrSF + housing_data$X2ndFlrSF # Total
Floor area.
# As the overall quality and the condition of the house mostly represents a single thing
lets combine them into a single variable.
housing_data$OverallQualCond = (housing_data$OverallQual + housing_data$OverallCond)/2 #
dividing by 2 to keep a constant scale.
summary(housing_data$OverallQualCond)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   5.500   6.000   5.809   6.500   8.500
```

3 (c)

I have made a feature called years used because this might be an important factor when cosidering buying a used house. I have combined overall Quality and condition beacuse these two try to represent the same feature. Total no of bathrooms in the house, this is an important factor for some when buying a new house. Total square ft is also an important paramenter so, i have created that. I have created feature for total no of full bath and total no of becoz that an important feature when buying an house. I have added all the features to the data frame.

Question 4 Kaggle.com { a little more data understanding

4 (a)

https://www.kaggle.com/c/nyc-taxi-trip-duration

This competition is to build a model that predicts the total ride duration of taxi trips in New York City. primary dataset is one released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables. ####4 (b)

```
taxi <- read.csv("train.csv")
dim(taxi)
```

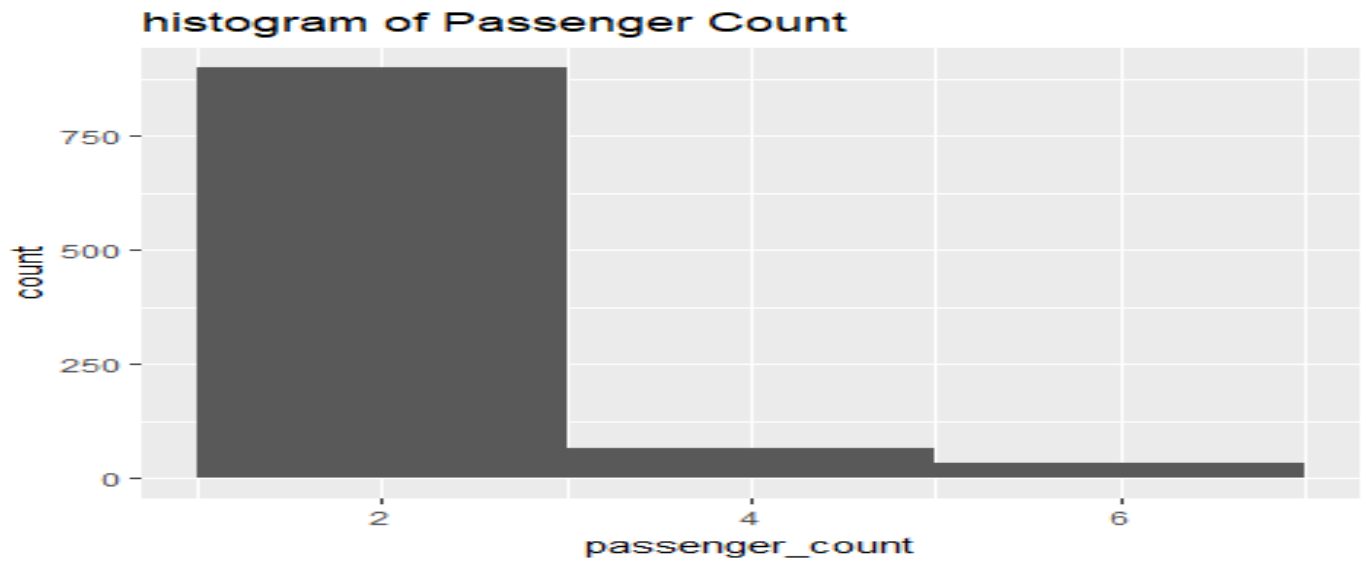
```
## [1] 1458644      11
```

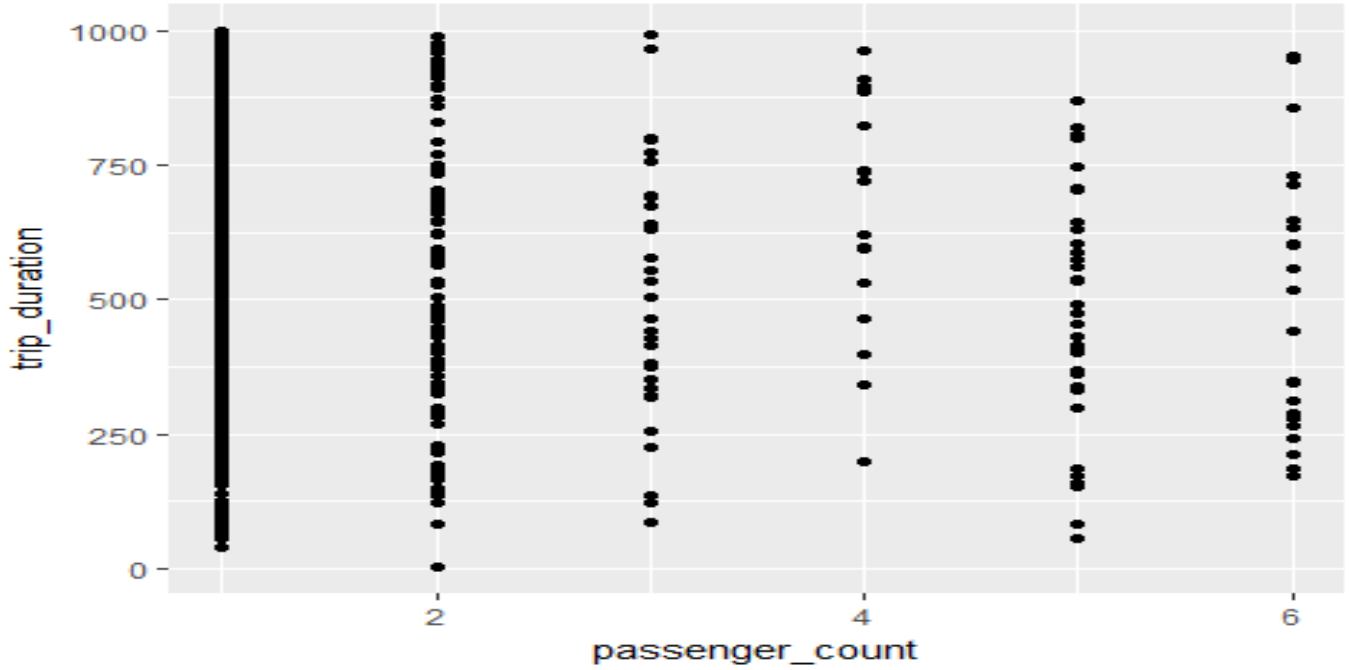
```
# there are 1458644 rows and 11 columns/variables
# this data set is too big for my computer
taxiSubset <- taxi[1:1000,]
# descriptive stats
summary(taxiSubset$passenger_count)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.663   2.000   6.000
```

```
summary(taxiSubset$trip_duration)
```

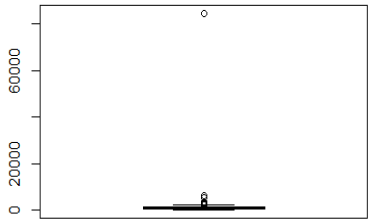
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.0   414.0   672.0   924.1  1074.2 84594.0
```



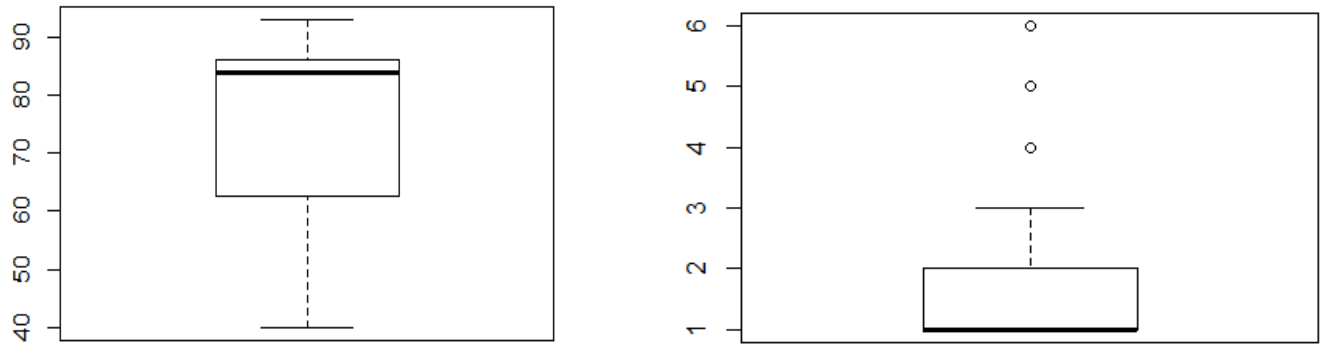


There is no clear relationship between passenger count and trip duaration.

```
boxplot(taxiSubset$trip_duration)
```

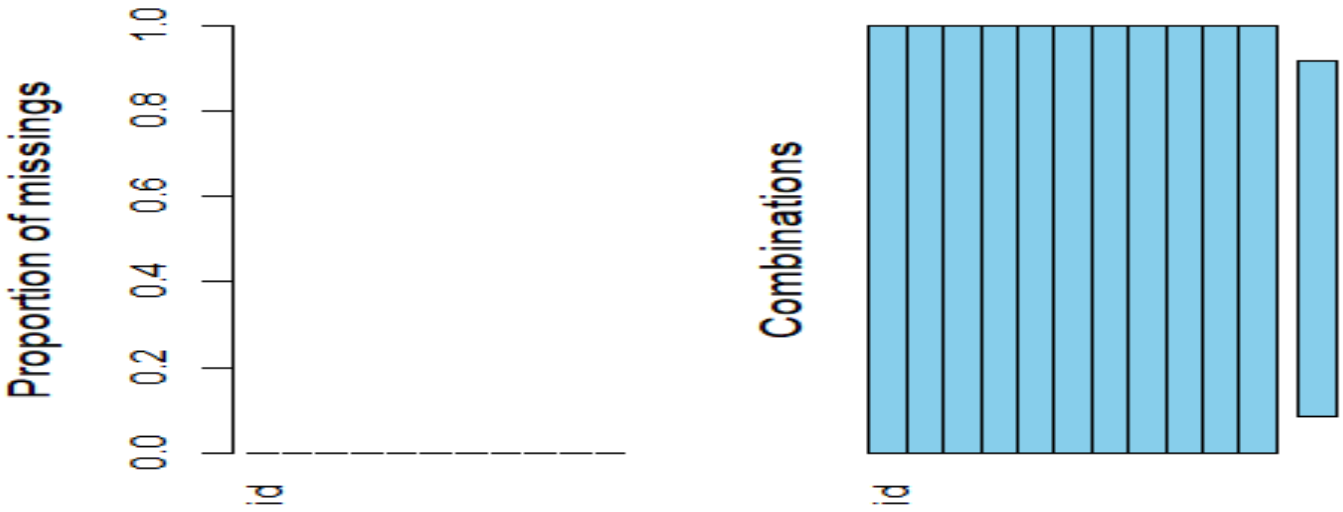


```
taxiNew <- subset(taxiSubset, trip_duration < 100 & trip_duration > 5)
boxplot(taxiNew$trip_duration)
```



So, the extremely long trip duaration and very short trip duration were outliers in the data.

We can see there are few cases with 4, 5 and 6 passengers which is basically not a outlier but less frequent trend in passenger count.



we can see that this dataset is complete and there are no missing variables.