DATA 606 Data Project Proposal

John Mazon

Data Preparation

```
library(RCurl)
library(Hmisc)
library(ggplot2)
```

Research question

Is there a correlation in diamond value with respect to it's carat amount and cut quality?

Cases

This dataset contains the prices and various attributes of almost 54,000 diamonds. Specifically, it is a data frame with 53940 rows and 10 variables. Some characteristics specified for each row are price[in U.S Dollars], carat, cut quality, color, clarity, total depth percentage and table.

Data collection

This classic dataset has been found on the Kaggle site.

Type of study

This is an observational study since we are trying to infer from already collected data.No specific experiment is being conducted data is being observed only.

Data Source

Original data is from Kaggle site: https://www.kaggle.com/shivam2503/diamonds

Dependent Variable

The response variable is the diamond value and it is quantitative.

Independent Variable

Carat amount is the quantitative independent variable Cut quality is the qualitative independent variable

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```
my_data <- getURL("https://raw.githubusercontent.com/johnm1990/DATA607/master/diamonds.csv")
diamond_info <- read.csv(text=my_data)

diamond_df <- data.frame(diamond_info)
```

```
summary(diamond_df$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     326     950    2401    3933    5324   18823
```

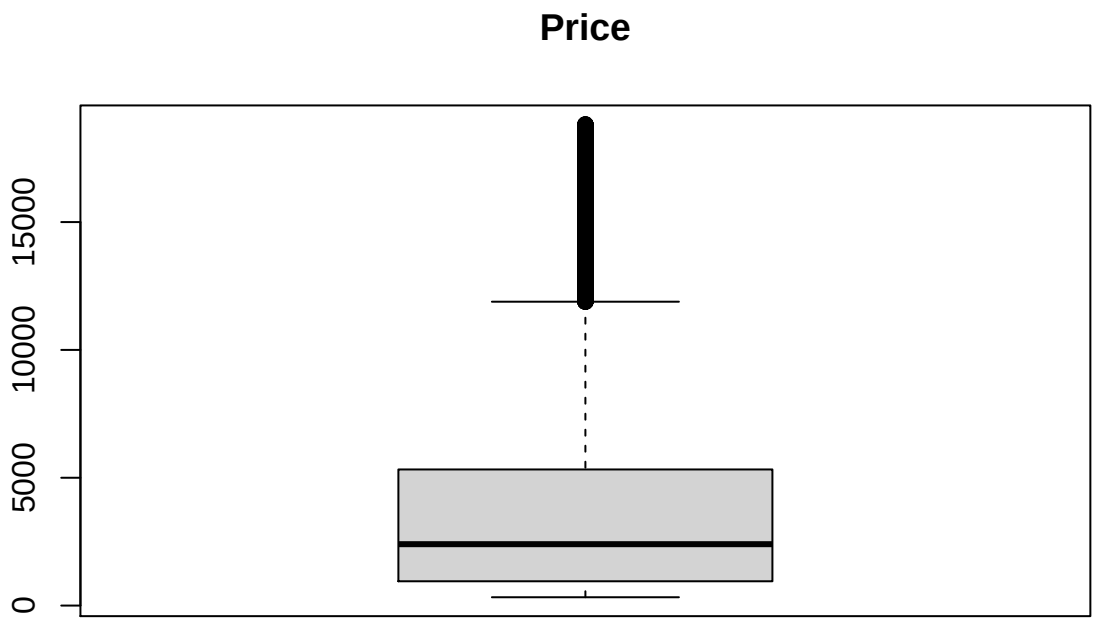
```
summary(diamond_df$carat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2000  0.4000  0.7000  0.7979  1.0400  5.0100
```

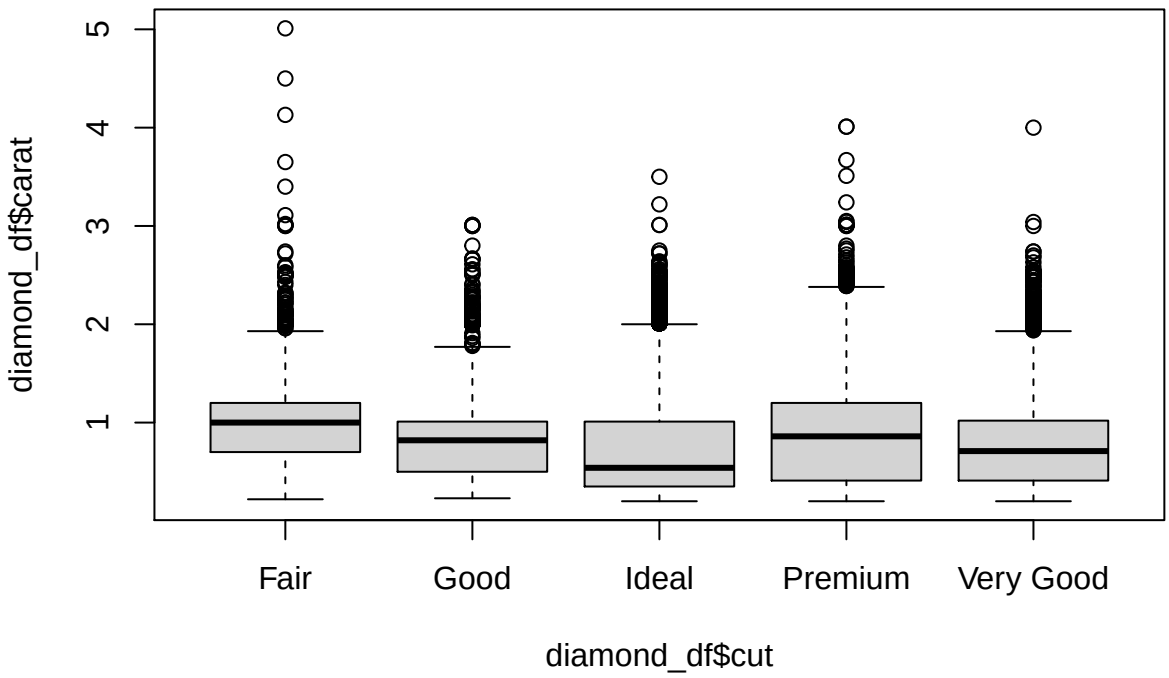
```
summary(diamond_df$cut)
```

```
##    Length     Class      Mode
##     53940 character character
```

```
boxplot(diamond_df$price, main = "Price")
```




```
boxplot(diamond_df$carat ~ diamond_df$cut)
```


diamond_df$carat
1
2
3
4
5
Fair
Good
Ideal
Premium
Very Good
diamond_df$cut