Project1

Tilak Muley

2023-02-10

Background and data chosen

The data that was chosen was diamonds. This data set comes built in-the tidyverse package.The data diamonds has variables: carat, cut, color, clarity, depth, table, price, x, y, and z. The variable carat depends on what weight the diamond is in this data set the carat ranges from .20-5.01. The variable cut identifies what cut quality the diamond is in ranging from ideal, premium, good, very good and fair. The variable clarity identifies how clear the diamond is. The variable depth identifies the depth percentage of the diamond which in this data set it ranges from 43.0-79.0. The variable price identifies the price of the diamonds in U.S. dollars in this data set the prices ranges from $326-$18823. The variable x identifies the length in mm in this data set it ranges from 0mm–10.74mm. The variable y identifies width in mm which in this data set ranges from 0mm–58.9mm. The variable z identifies depth in mm which in this data set ranges from 0mm–31.8mm.

Problem definition

The problem I decided to choose to solve is to find out if there is any difference or if its similar on price and carat depending on what type of cut the diamond has (ideal, premium, good, very good and fair)

Importing Data and Library

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(knitr)
library(ggplot2)
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(tidyr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## v purrr   1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x data.table::between() masks dplyr::between()
## x dplyr::filter()       masks stats::filter()
## x data.table::first()   masks dplyr::first()
## x dplyr::lag()          masks stats::lag()
## x data.table::last()    masks dplyr::last()
## x purrr::transpose()    masks data.table::transpose()
```

```
data("diamonds")
```

EXPLATORY DATA ANALYSIS

Summary of Data

Summary:

```
summary(diamonds)
```

```
##      carat               cut        color        clarity          depth
##  Min.   :0.2000   Fair      : 1610   D: 6775   SI1    :13065   Min.   :43.00
##  1st Qu.:0.4000   Good      : 4906   E: 9797   VS2    :12258   1st Qu.:61.00
##  Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194   Median :61.80
##  Mean   :0.7979   Premium   :13791   G:11292   VS1    : 8171   Mean   :61.75
##  3rd Qu.:1.0400   Ideal     :21551   H: 8304   VVS2   : 5066   3rd Qu.:62.50
##  Max.   :5.0100                     I: 5422   VVS1   : 3655   Max.   :79.00
##                                     J: 2808   (Other): 2531
##      table           price             x                y
##  Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720
##  Median :57.00   Median : 2401   Median : 5.700   Median : 5.710
##  Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735
##  3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540
##  Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900
##
##        z
##  Min.   : 0.000
##  1st Qu.: 2.910
##  Median : 3.530
##  Mean   : 3.539
##  3rd Qu.: 4.040
##  Max.   :31.800
##
```

Data Exploration of diamonds data

First 6 rows of the input data frame:

```
head(diamonds)
```

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
```

```
## 1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

Last 6 rows of the input data frame:

```
tail(diamonds)
```

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.72 Premium   D     SI1      62.7    59  2757  5.69  5.73  3.58
## 2  0.72 Ideal     D     SI1      60.8    57  2757  5.75  5.76  3.5
## 3  0.72 Good      D     SI1      63.1    55  2757  5.69  5.75  3.61
## 4  0.7  Very Good D     SI1      62.8    60  2757  5.66  5.68  3.56
## 5  0.86 Premium   H     SI2      61      58  2757  6.15  6.12  3.74
## 6  0.75 Ideal     D     SI2      62.2    55  2757  5.83  5.87  3.64
```

Dimensions of the data:

```
dim(diamonds)
```

```
## [1] 53940    10
```

DATA MANIPULATION

Data wrangling, munging and cleaning

To Data wrangle and clean up the data I first used the select() command to get only the data of the variables I want for my problem. The variables I wanted was cut, carat and price. So I created a new data set called diamondWanted.

```
diamondWanted = select(diamonds, cut,carat,price)
```

Then I arrange the data set by type of cut so all data for fair together, good is together, very good together, premium together, and ideal together

```
diamondWanted = diamondWanted %>% arrange(cut)
diamondWanted
```

```
## # A tibble: 53,940 x 3
##    cut   carat price
##    <ord> <dbl> <int>
##  1 Fair   0.22   337
##  2 Fair   0.86  2757
##  3 Fair   0.96  2759
##  4 Fair   0.7   2762
##  5 Fair   0.7   2762
##  6 Fair   0.91  2763
##  7 Fair   0.91  2763
##  8 Fair   0.98  2777
##  9 Fair   0.84  2782
## 10 Fair   1.01  2788
## # ... with 53,930 more rows
```

Data Exploration of diamondWanted Data

First 6 rows of the input Data frame with just variables cut, carat, price:

```
head(diamondWanted)
```

```
## # A tibble: 6 x 3
##   cut   carat price
##   <ord> <dbl> <int>
## 1 Fair   0.22   337
## 2 Fair   0.86  2757
## 3 Fair   0.96  2759
## 4 Fair   0.7   2762
## 5 Fair   0.7   2762
## 6 Fair   0.91  2763
```

Last 6 rows of the input Data frame with just variables cut, carat, price:

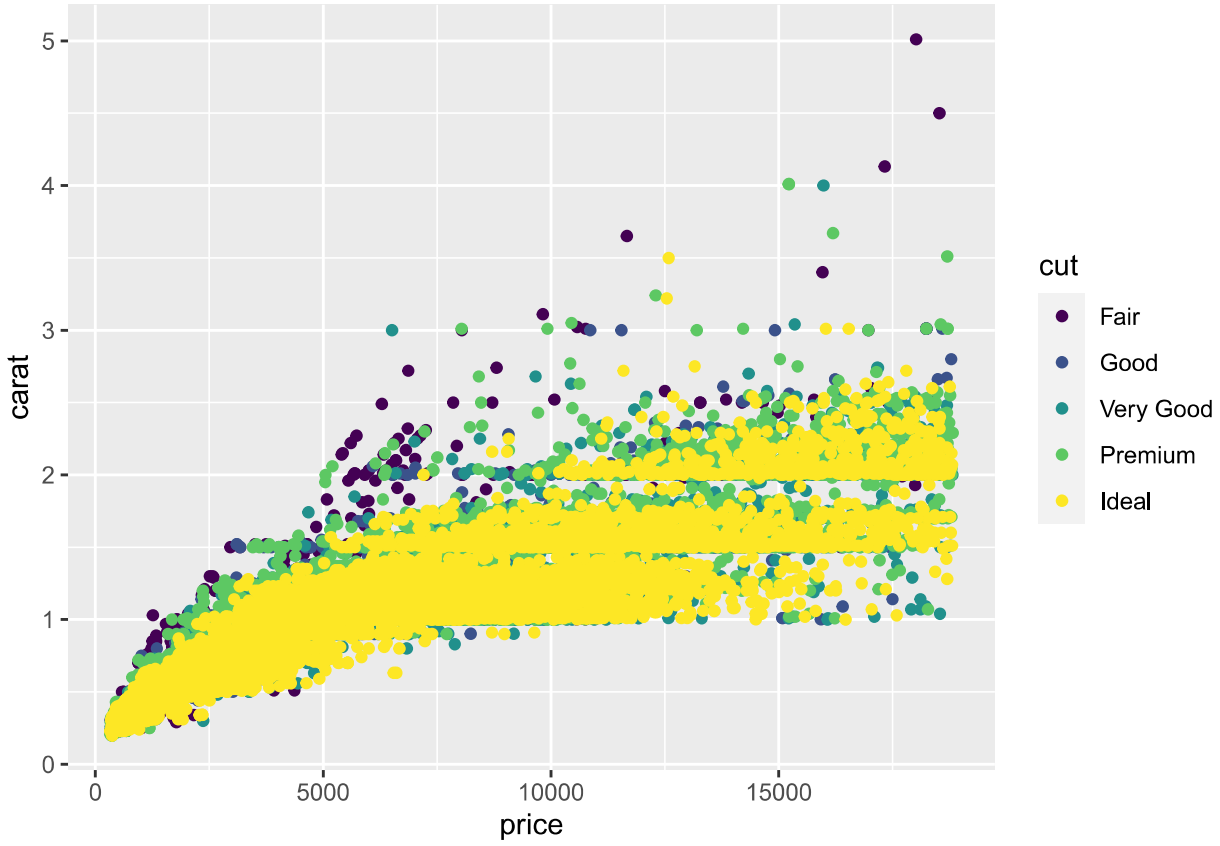
```
tail(diamondWanted)
```

```
## # A tibble: 6 x 3
##   cut   carat price
##   <ord> <dbl> <int>
## 1 Ideal  0.73  2756
## 2 Ideal  0.79  2756
## 3 Ideal  0.71  2756
## 4 Ideal  0.71  2756
## 5 Ideal  0.72  2757
## 6 Ideal  0.75  2757
```

Data Visualization

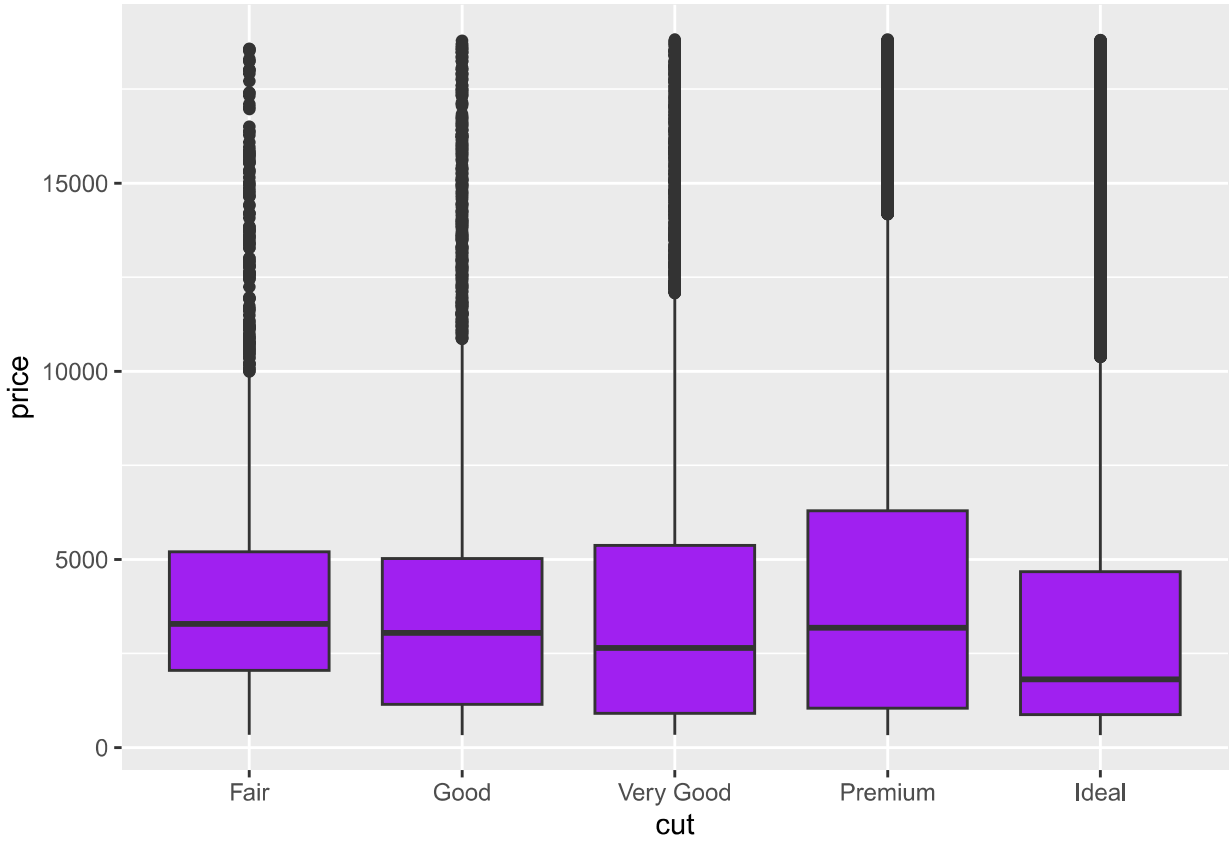
Scatter plot:

```
ggplot(data = diamondWanted, aes(x = price, y = carat, color = cut)) +
  geom_point()
```



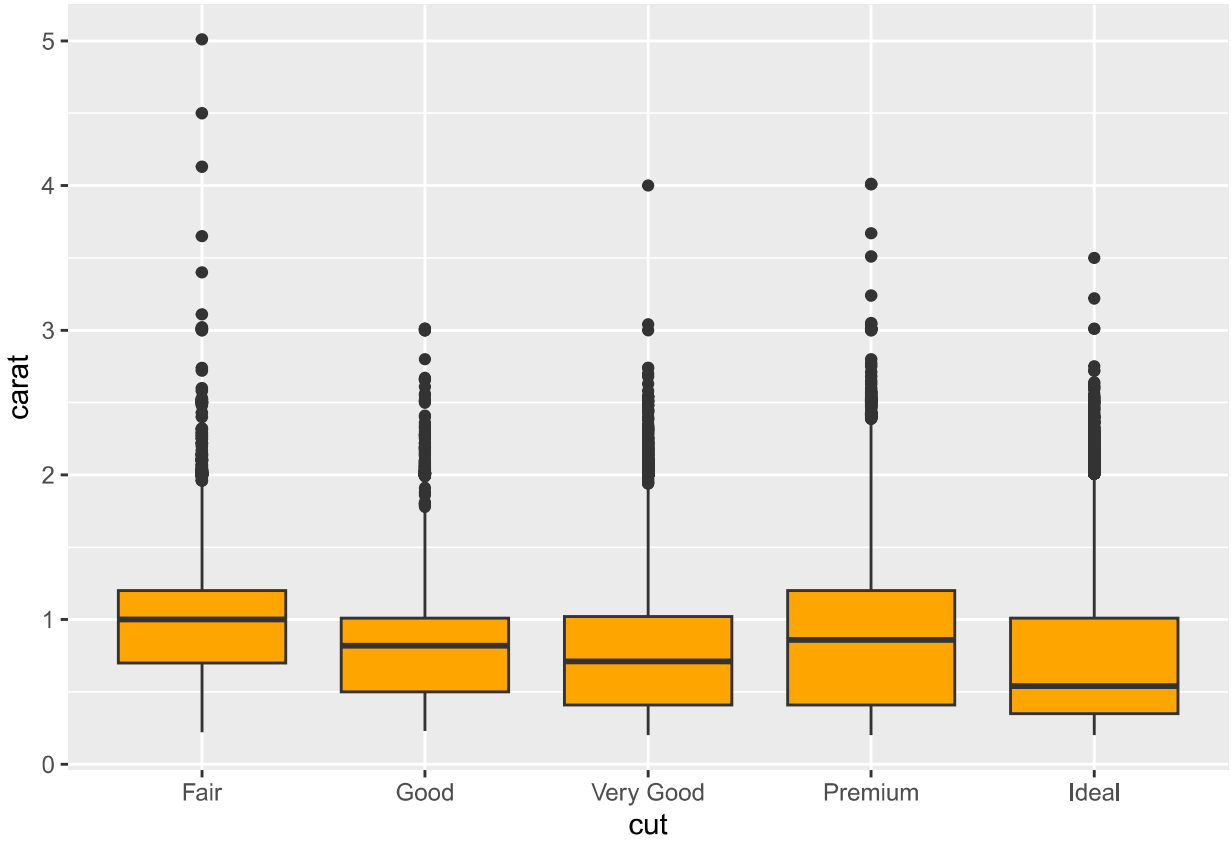
Box plot of cut and price:

```
ggplot(data = diamondWanted,aes(x = cut, y = price))+geom_boxplot(fill="purple")
```



Box plot of cut and carat:

```
ggplot(data = diamondWanted,aes(x = cut, y = carat))+geom_boxplot(fill="orange")
```



Conclusion

In conclusion I found that it isn't that much of difference of carat and price depending on the type of cut. Also that the main thing is the carat of the diamond not the type of cut that affects the price. Overall I found out the cut of the diamond doesn't have much of a impact on the carat or the price of the diamond. But I did find that ideal cut is much more spread out along the price but the carat of any ideal cut diamond does not exist above 3.5. Also learned and found the carat medians of each cut is similar. Also price medians of each cut is similar to each type of cut. While max and min was fairly close in price between all types of cuts big difference was between max of carat between each type of cut with fair cut having a bigger max and good cut having the smaller max. It was also found the smallest carat belongs to a very good cut diamond with a price of $367. While biggest carat belongs to a fair cut diamond with a price of $18018. While lowest price on a diamond is $326 on a ideal cut .23 carat and a premium cut .21 carat. While highest price on diamond is $18823 on a premium cut 2.29 carat diamond.