Lab 04: Analysis of Variance

Isaac Plotkin

2/11/2022

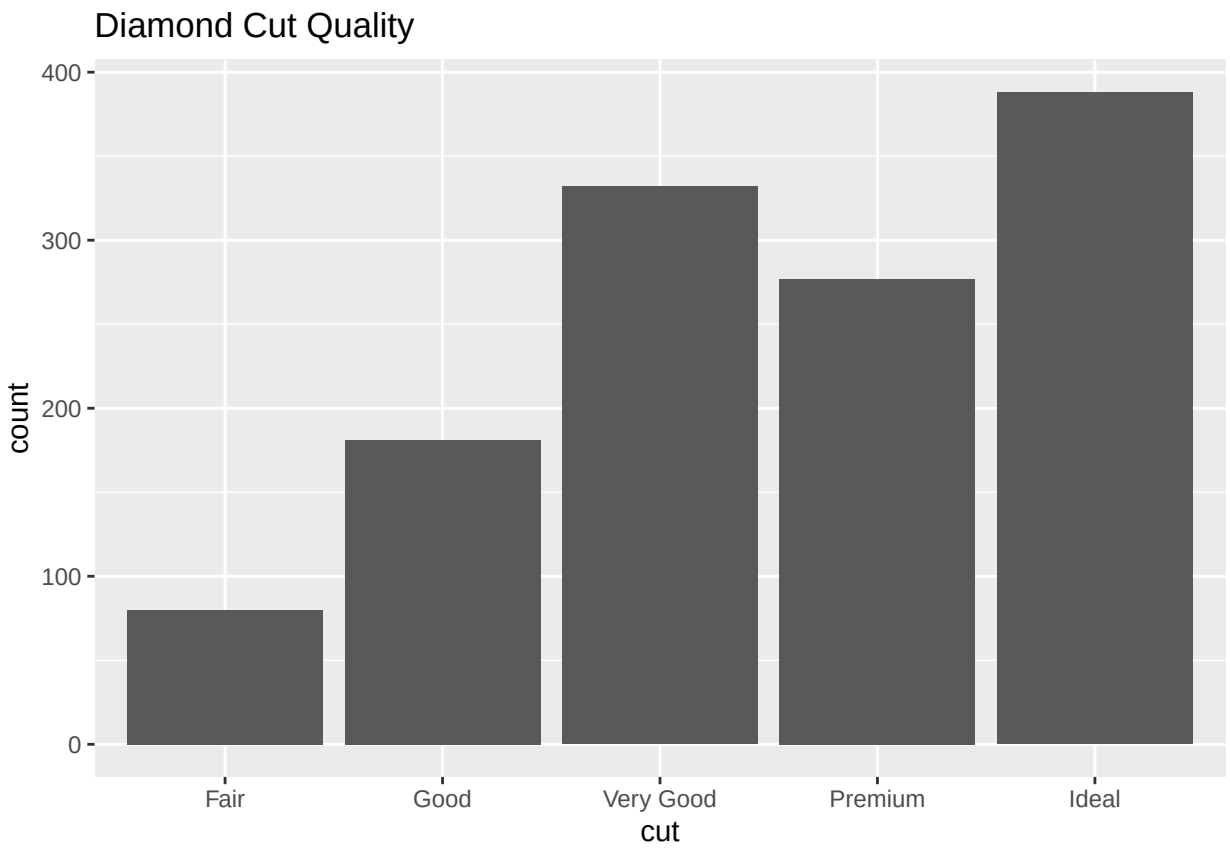
1.

```
data(diamonds)

diamonds_5c <- subset(diamonds, carat == 0.5)
```

There are 1,258 observations in the new dataset.

2.

```
p1 <- ggplot(data = diamonds_5c, aes(x = cut)) +
  geom_bar() +
  labs(title = "Diamond Cut Quality")

p1
```



Fair and Good cuts have the fewest number of observations.

3.

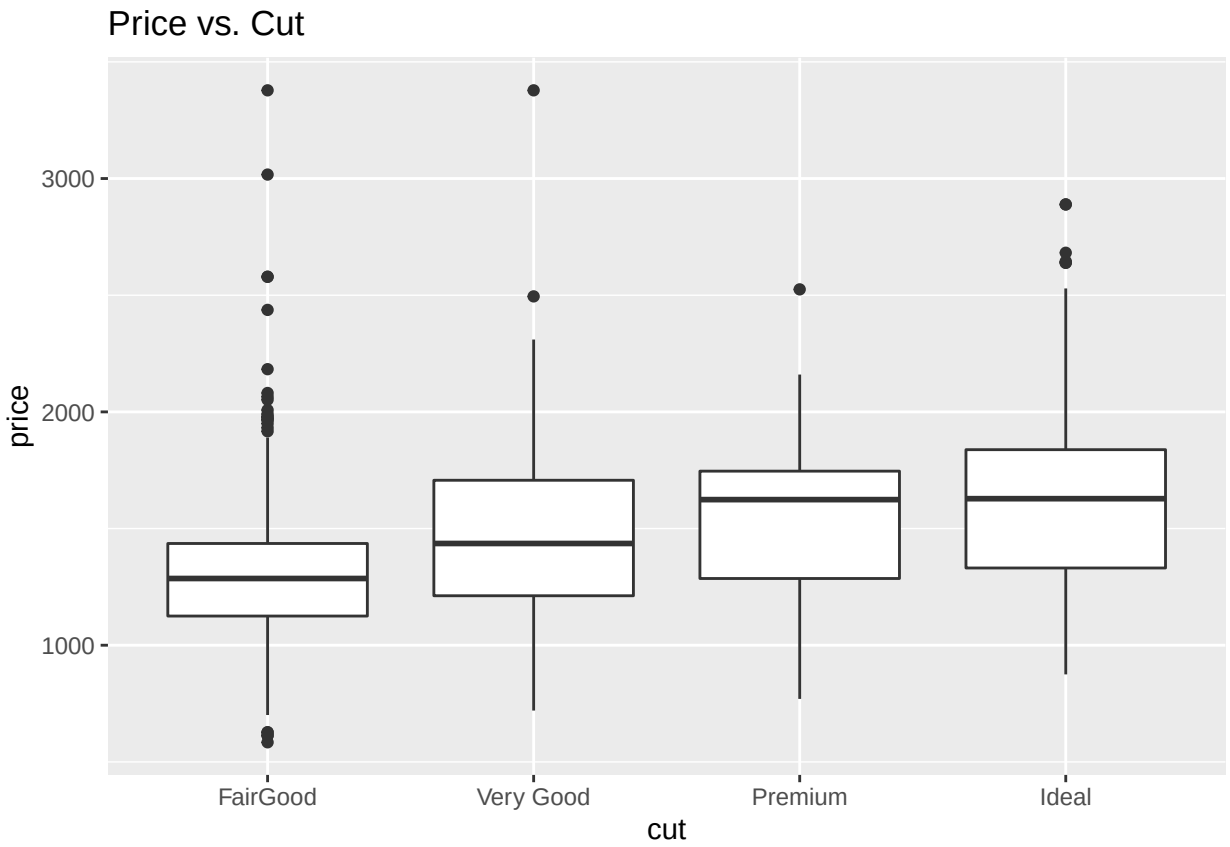
```
diamonds_new <- mutate(diamonds_5c, cut = fct_recode(cut, "FairGood" = "Fair", "FairGood" = "Good"))

diamonds_new
```

```
## # A tibble: 1,258 x 10
##    carat cut       color clarity depth table price     x     y     z
##    <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
##  1   0.5 Ideal     E     VVS2     62.2    54  2889  5.08  5.12  3.17
##  2   0.5 Ideal     E     VVS2     62.2    54  2889  5.09  5.11  3.17
##  3   0.5 FairGood  D     VVS2     62.4    64  3017  5.03  5.06  3.14
##  4   0.5 FairGood  D     IF       63.2    59  3378  4.99  5.04  3.17
##  5   0.5 Very Good D     IF       62.9    59  3378  4.99  5.09  3.17
##  6   0.5 FairGood  F     I1       69.8    55   584  4.89  4.8   3.38
##  7   0.5 FairGood  F     I1       71      57   613  4.87  4.79  3.43
##  8   0.5 FairGood  F     I1       68.4    54   613  4.94  4.82  3.35
##  9   0.5 FairGood  F     I1       67.1    57   627  4.92  4.87  3.28
## 10   0.5 FairGood  F     I1       68.3    58   627  4.91  4.78  3.32
## # ... with 1,248 more rows
```

The variable cut got recoded correctly.

4.

```
pvc <- ggplot(data = diamonds_new, aes(x = cut, y = price)) +
  geom_boxplot() +
  labs(title = "Price vs. Cut")

pvc
```



5.

```
print(summary(diamonds_new$cut))
```

```
##  FairGood Very Good   Premium     Ideal
##       261       332       277       388
```

```
summary <- aggregate(price~cut, diamonds_new, function(x) c(mean = mean(x), sd = sd(x)))

summary[-1][[1]]
```

```
##          mean       sd
## [1,] 1340.644 364.5216
## [2,] 1488.663 339.3630
## [3,] 1531.776 304.1443
## [4,] 1608.668 368.3448
```

6. There appears to be a relationship between the cut and price for diamonds that are 0.5 carats because the box plot and summary statistics show that the mean is increasing as the quality of the cut increases.

ANOVA

7. The assumptions for ANOVA are satisfied because the distribution of price appears normal within each cut category. All the observations are collected separately so independence is satisfied. Lastly constant variance is satisfied because all of the standard deviations are close in size.

8.

```
model <- lm(price ~ cut, data = diamonds_new)
tidy(model, conf.int = TRUE) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1492.438	9.893	150.852	0.000	1473.028	1511.847
cut.L	189.436	19.657	9.637	0.000	150.873	228.000
cut.Q	-35.564	19.787	-1.797	0.073	-74.383	3.255
cut.C	31.011	19.916	1.557	0.120	-8.062	70.083

```
kable(anova(model), format="markdown", digits = 3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cut	3	11507056	3835685.3	31.916	0
Residuals	1254	150706506	120180.6	NA	NA

9. Total variation = SST/DFT

```
total_variation <- 150706506/1257
total_variation
```

```
## [1] 119893.8
```

10. (theta)^2 = SSW/DFW (DFW = 1258-4)

```
11507056/1254
```

```
## [1] 9176.281
```

11. The null hypothesis is that the cut of a diamond does not cause variability in the price of a diamond. The alternative hypothesis is that the cut of a diamond does cause variability in the price of a diamond.

12. The conclusion is that the null hypothesis is rejected because the p value for cut~0.This means that the cut of the diamonds have different means. Therefore the cut of a diamond has a direct affect on the price.

13. Based on the box plot and ANOVA table it is clear that the higher the grade of the cut, the higher the price of the diamond.