HW 8

Arjun Ganesh

4/16/2020

1. [24] Load the ggplot2 package and use the diamonds dataset to answer the following questions. Let's assume that this data on diamonds is representative of all the diamonds in theworld. [Note: D is the best color and Ideal is the best cut of a diamond]

A. [2] Two parts: write a code to isolate and count the number of Ideal cut diamonds in the dataset. Also, write a code to isolate and count the number of Fair cut diamonds in the dataset.[Note: be sure to include the output of your code.]

```
library(ggplot2)
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
```

```
data(diamonds)


fair<-diamonds%>%
  filter(diamonds$cut=="Fair")
count(fair)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  1610
```

```
ideal<-diamonds%>%
  filter(diamonds$cut=="Ideal")
count(ideal)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 21551
```

B. [2] Two parts: write a code to count the number of D colored diamonds among all the Ideal cut diamonds. Also, write a code to count the number of D colored diamonds among all the Fair cut diamonds.[Note: be sure to include the output of your code.]

```
DI<-ideal%>%
  filter(ideal$color=="D")
count(DI)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  2834
```

```
DF<-fair%>%
  filter(fair$color=="D")
count(DF)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   163
```

C. [6] Refer to parts (a) and (b). Write a code to estimate the difference between the two proportions with 95% confidence. (The two proportions are: proportion of D colored diamonds among all the Ideal cut diamonds and the proportion of D colored diamonds among all the Fair cut diamonds). Also, write a sentence interpreting your confidence interval. [Note: be sure to include the output of your code.]

```
prop.test(c(2834,163),c(21551,1610), conf.level = .95)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(2834, 163) out of c(21551, 1610)
## X-squared = 11.909, df = 1, p-value = 0.0005586
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.01451612 0.04600345
## sample estimates:
##    prop 1    prop 2
## 0.1315020 0.1012422
```

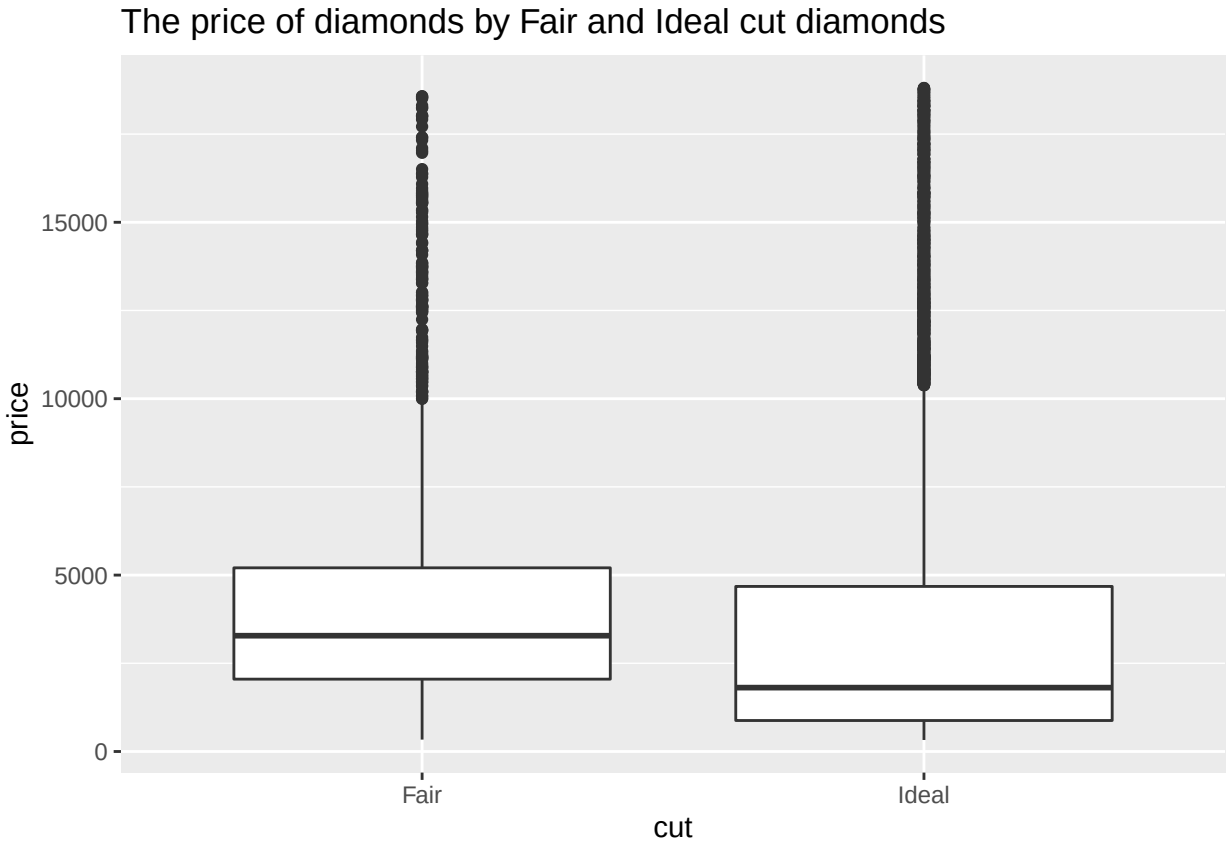
We are 95% confident that the difference between the two proportions is between an interval of 1.451612% to 4.600345%. And 13.15020% are ideal with D 10.12422% are fair with D.

D. [14] Our goal is to conduct a hypothesis test to investigate whether Fair cut diamonds are more expensive than Ideal cut diamonds, on average. Use a 5% level of significance. (a) [2] Data wrangling: filter the data to extract Fair OR Ideal cut diamonds, and save the resulting dataframe.

```
fair_or_ideal<-diamonds%>%
  filter(diamonds$cut == "Ideal" | diamonds$cut =="Fair")
```

(b) [3] Data visualization: make and side-by-side boxplot of the price of diamonds by Fair and Ideal cut diamonds. What does the graph reveal? The price of fair cut is greater than ideal cut.

```
library(ggplot2)
ggplot(fair_or_ideal,aes(cut,price))+geom_boxplot()+ggtitle("The price of diamonds by Fair and Ideal cu
```



(c) [2] Clearly state the null and alternative hypotheses in words.

NH: There is no difference between the price of Fair cut and Ideal cut diamonds. AH: There is a positive difference between the price of Ideal cut diamonds from Fair cut diamonds. Meaning Fair cut diamonds are more worth more than Ideal cut diamonds.

(d) [3] Write a code to output the p-value.

```
t.test(price~ cut, fair_or_ideal, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  price by cut
## t = 9.7484, df = 1894.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  749.079     Inf
## sample estimates:
##  mean in group Fair mean in group Ideal
##            4358.758            3457.542
```

(e) [3] Write your conclusion in context.

The sample estimates of the mean in the Fair group are greater than the Ideal group. Also the p-value calculate is less than 0.05 and so we can reject the null hypothesis and say Fair cut diamonds are more expensive Ideal cut diamonds, on average.

(f) [1] If you end up rejecting the null hypothesis, suggest a potential confounding variable for this decision.

```
fair_clarity<-fair%>%
  filter(fair$clarity=="VVS2"| fair$clarity=="VVS1")
count(fair_clarity)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    86
```

```
ideal_clarity<-ideal%>%
  filter(ideal$clarity=="VVS2"| ideal$clarity=="VVS1")
count(ideal_clarity)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  4653
```

There are far less valuale clarity of fair diamonds than there are ideal so this could be a potential confounding variable as you want the highest level of clarity when purchasing a diamond.

2. [6] Suppose you are a Data Scientist at a credit card bank and you need to forecast how much the average spending will increase or decrease from month to month.

In particular, you are tasked with estimating the actual mean change in spending between the month of December and January.

Import the ConsumerSpending dataset, write a code (3 points) to estimate the mean difference in spending between December and January with 95% confidence. Also, write a sentence interpreting (3 points) your confidence interval. [Hint: every row in the dataset represent a particular customer's amount of expenditure for December,January, February,March and April.]

```
CS<-read.csv("/cloud/project/ConsumerSpending.csv")

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.3.0 --
```

```
## v tibble  2.1.3     v purrr   0.3.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
diff_months<-CS[,1:2]
dif<- pivot_longer(diff_months, c("December","January"), names_to = "Month",values_to="Amount")

t.test(Amount~ Month,dif,conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  Amount by Month
## t = 6.5951, df = 2357.5, p-value = 5.227e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  114.6726 211.7210
```

```
## sample estimates:
## mean in group December  mean in group January
##               728.1820               564.9852
```

We are 95% confident that from December going into January, the mean change in spending between the month of December and January will decrease between an interval of $114.6726 dollars and 211.7210 dollars.