Lab 04

Rohit Kulkarni

02/09/2022

The goal of this lab is to use Analysis of Variance (ANOVA) to understand the variation in price of diamonds that are 0.5 carats. Additionally, you will be introduced to new R function used for wrangling and summarizing data.

Data

In today's lab, we will analyze the diamonds dataset from the ggplot2 package . Type ?diamonds in the console to see a dictionary of the variables in the data set. The primary focus of this analysis will be examining the relationship between a diamond's cut and price.

Before starting the exercises, take a moment to read more about the diamond attributes on the Gemological Institute of America webpage: https://www.gia.edu/diamond-quality-factor.

Exercises

The diamonds dataset contains the price and other characteristics for over 50,000 diamonds. For this analysis, we will only consider diamonds that have a carat weight of 0.5.

Exploratory Data Analysis

1. Create a new data frame that is a subset of diamonds that weigh 0.5 carats. How many observations are in the new dataset?

There are 1,258 observations in the new dataset.

You will use this subset for the remainder of lab.

2. When using Analysis of Variance (ANOVA) to compare group means, it is ideal to have approximately the same number of observations for each group.

   Which two levels of cut have the fewest number of observations? Show the code and output used to support your answer.

```
count(diamonds, cut)
```

The output: Fair 80 Good 181 Very Good 332 Premium 277 Ideal 388

With the count method, I can see that Fair and Good have the fewest number of observations.

See the forcats reference page for ideas on recoding factor variables.

3. Confirm that the variable cut was recoded as expected. Show the code and output used to check the recoding.

Code for the recoding:

```
diamonds$cut <- fct_recode(diamonds$cut, "GoodOrFair" = "Fair", "GoodOrFair" = "Good")
```

Code for checking the recoding:

```
count(diamonds, cut)
```

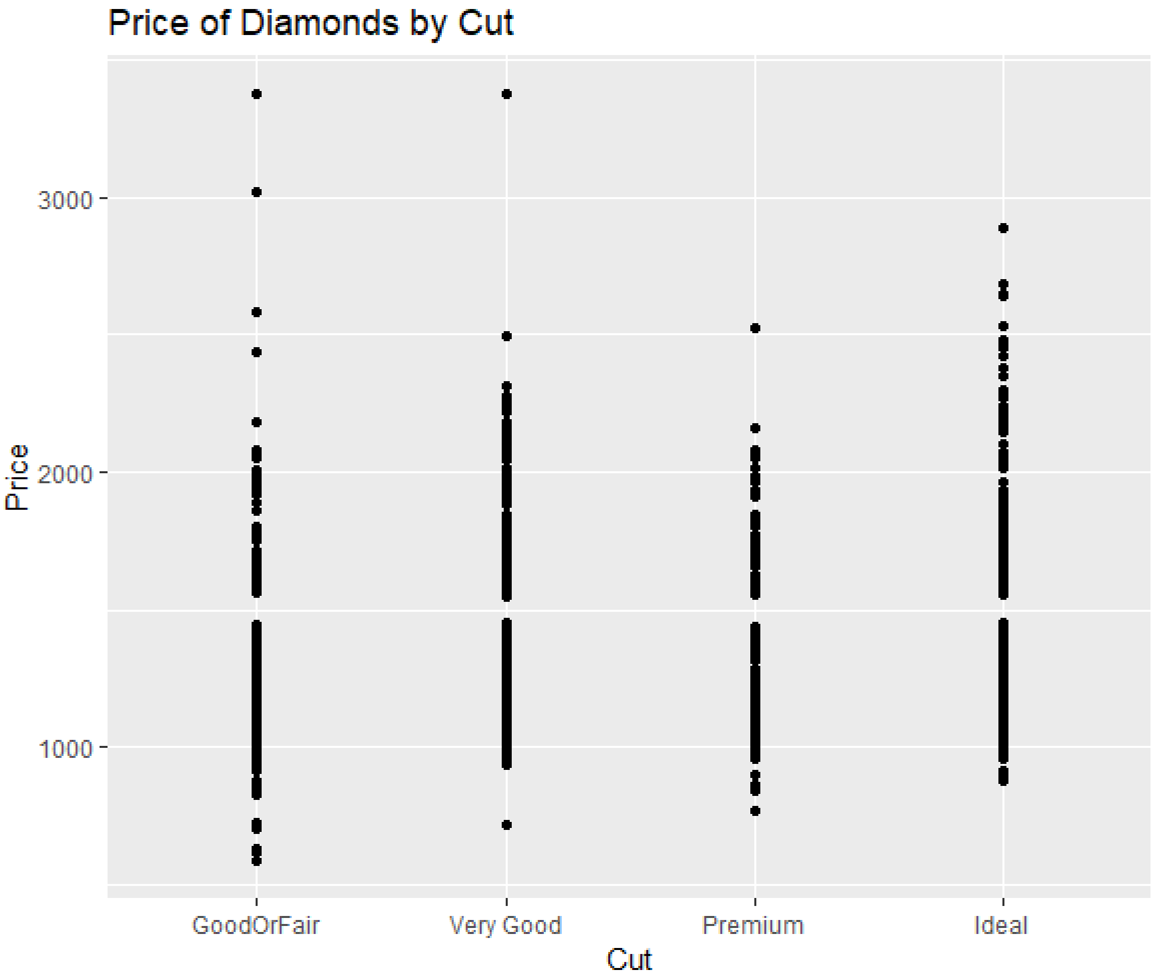
Output: cut n 1 GoodOrFair 261 2 Very Good 332 3 Premium 277 4 Ideal 388

I know that the variable cut was recoded as expected because the GoodOrFair category has the same number of entries as the Good and Fair categories combined.

4. Create a plot to display the relationship between cut and price. Be sure to include informative axes labels and an informative title.

Code for Plot:

```
ggplot(data=diamonds, mapping=aes(x=cut, y = price)) + geom_point() +
  labs(x="Cut", y="Price", title = "Price of Diamonds by Cut")
```



Output:

5. Calculate the number of observations along with the mean and standard deviation of price for each level of cut.

Code for calculating count, mean, and standard deviation of price by cut:

```
cut <- diamonds$cut
price <- diamonds$price

tapply(price, cut, mean)
tapply(price, cut, sd)
count(diamonds, cut)
```

Results: GoodOrFair Very Good Premium Ideal Mean 1340.644 1488.663 1531.776 1608.668 Std Dev 364.5216 339.3630 304.1443 368.3448 Count 261 332 277 388

6. Based on the plots and summary statistics from the previous exercises , does there appear to be a relationship between the cut and price for diamonds that are 0.5 carats? Briefly explain your reasoning.

Basd on the data, it appears that there is a slight correlation between the cut of the diamond and the price for diamonds that are 0.5 carats. The mean steadily increases as the cut increases in quality, while the standard deviation of each cut is relatively the same. The increase between cut levels is also mostly uniform.

Analysis of Variance

7. When using ANOVA to compare means across groups, we make the following assumptions (note how similar they are to the assumptions for regression):

Normality: The distribution of the response, y, is approximately normal within each category of the predictor, x - in the ith category, the y's follow a N(μ_i,σ^2) distribution.

This is the code I used to plot the distributions of price within each category:

```
ideal_diamonds <- filter(diamonds, cut == "Ideal")
hist(ideal_diamonds$price)

verygood_diamonds <- filter(diamonds, cut == "Very Good")
hist(verygood_diamonds$price)

goodorfair_diamonds <- filter(diamonds, cut == "GoodOrFair")
hist(goodorfair_diamonds$price)

premium_diamonds <- filter(diamonds, cut == "Premium")
hist(premium_diamonds$price)
```

These are the histograms:

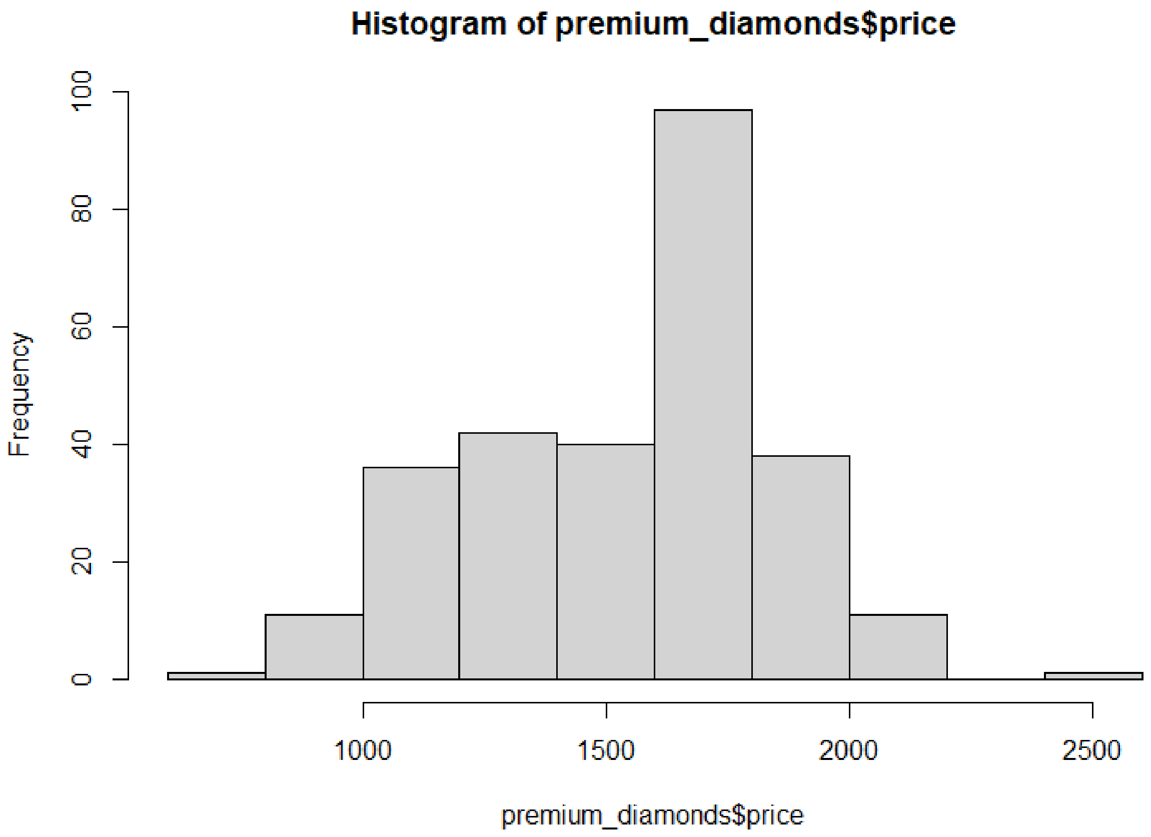


Figure 1: Premium Distribution

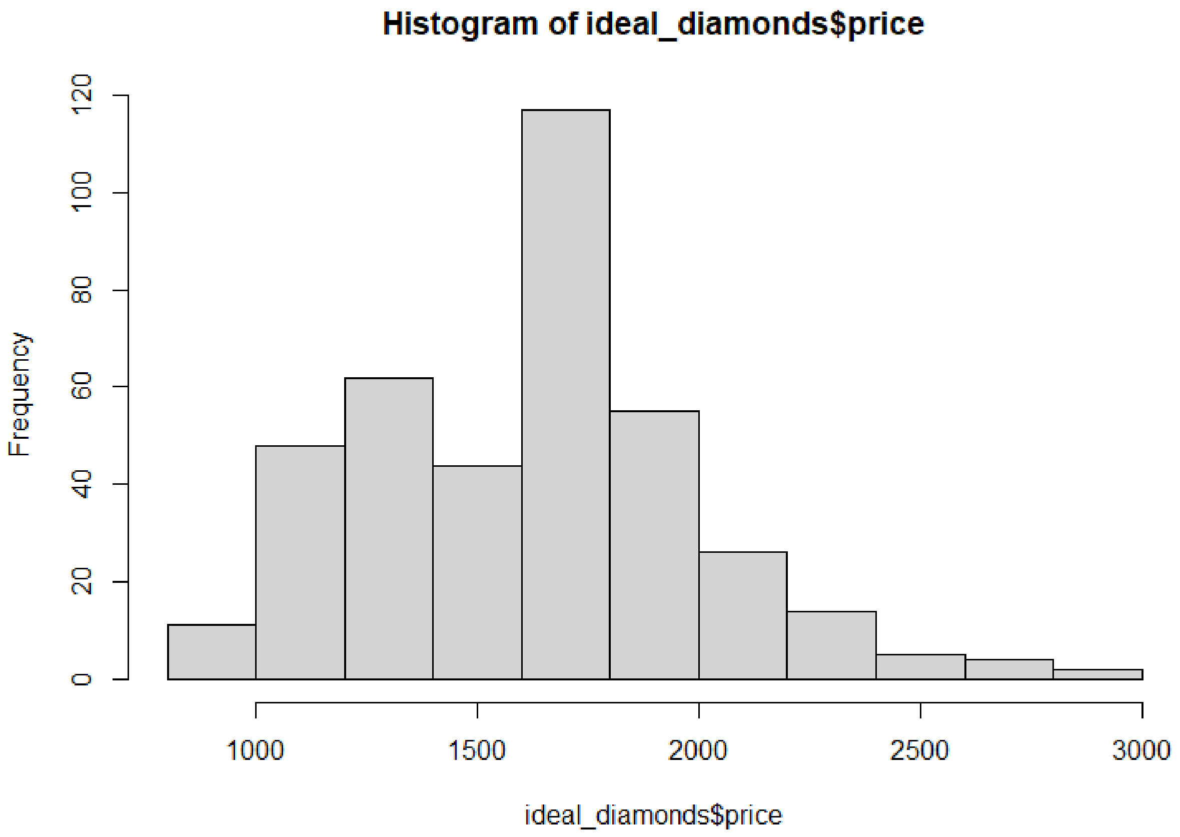
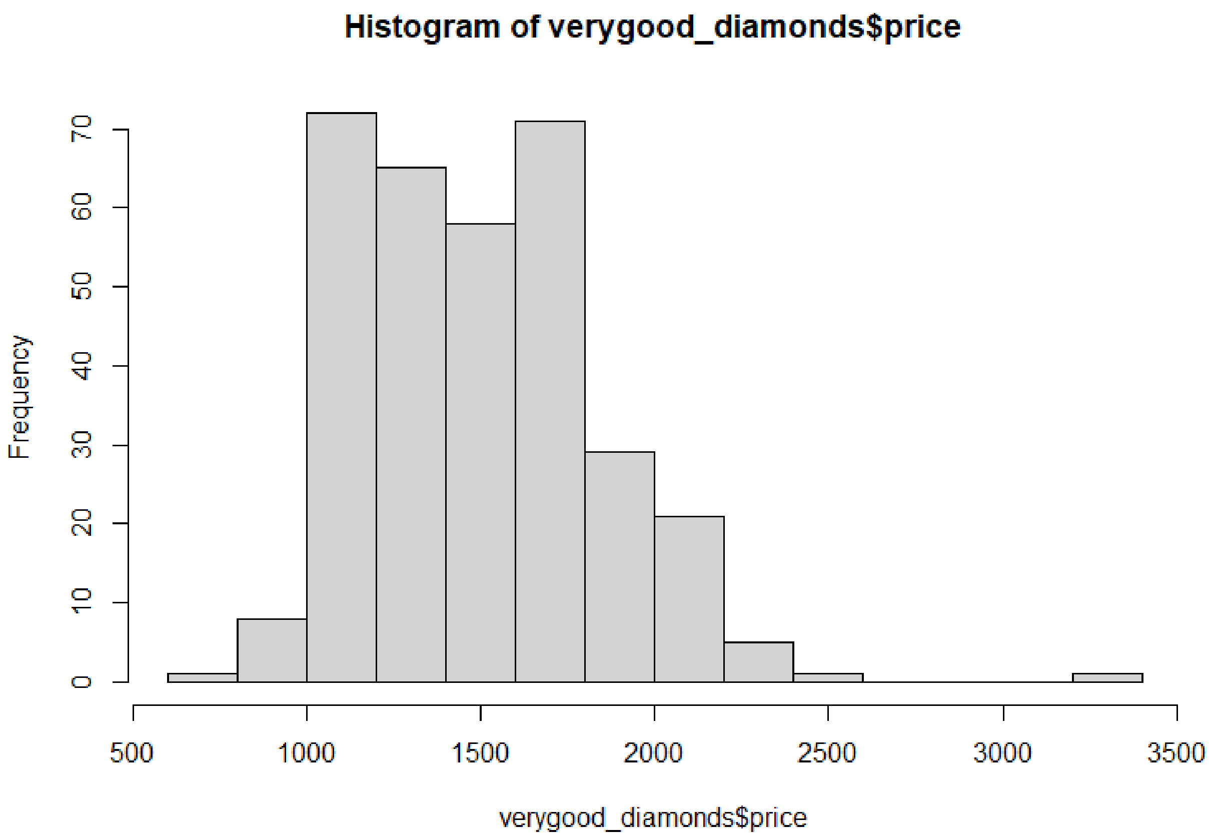
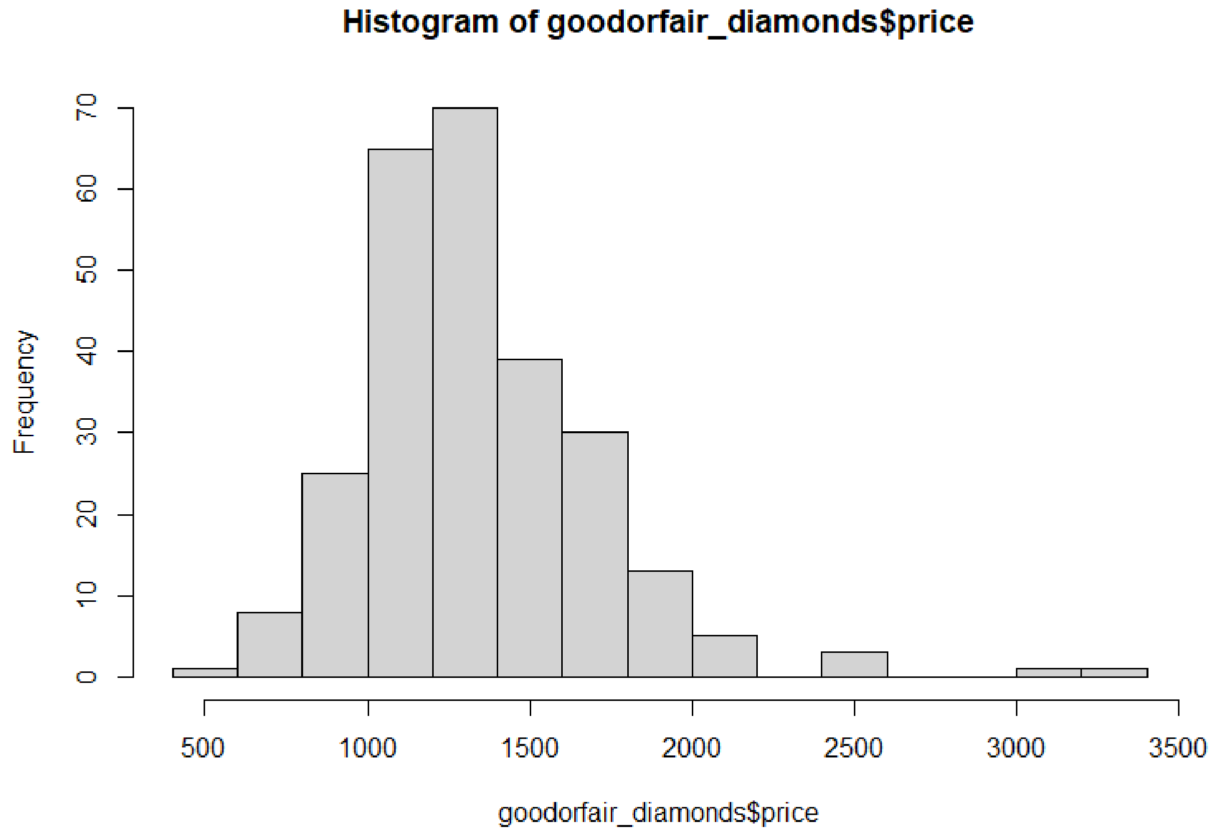


Figure 2: Ideal Distribution


Histogram of verygood_diamonds$price
Frequency
0
10
20
30
40
50
60
70
500
1000
1500
2000
2500
3000
3500
verygood_diamonds$price



The price by each category follows a generally normal distribution with a peak in the middle. So the normality assumption holds.

Independence: All observations are independent from one another, i.e. one observation does not affect another. Constant Variance: The distribution of the response within each category of predictor, x has a common variance, σ^2.

We can assume that the observations are independent since the quality of one diamond or price does not affect the quality or price of another diamond.

Constant Variance: The distribution of the response within each category of predictor, x has a common variance, σ^2.

Code used to find the variance:

```
var(ideal_diamonds$price)
var(verygood_diamonds$price)
var(goodorfair_diamonds$price)
var(premium_diamonds$price)
```

Output: Ideal - 135677.9 Very Good - 115167.2 Good or Fair - 132876 Premium - 92503.75

Given that the variances are all somewhat close together, we can assume constant variance.

Are the assumptions for ANOVA satisfied? Comment on each assumption, including an explanation for your reasoning and any summary statistics and/or plots used to make the conclusion.

Given the evidence above, I believe that the ANOVA assumptions are satisfied.

8. Display the ANOVA table used to examine the relationship between cut and price for diamonds that are 0.5 carats.

```
          Df    Sum Sq Mean Sq F value Pr(>F)
```

cut 3 11507056 3835685 31.92 <2e-16 *** Residuals 1254 150706506 120181

9. Use the ANOVA table from the previous question to calculate the sample variance of price. Show the code / formula used to calculate the sample variance.

The formula for variance is the sum of the square residuals divided by the number of observations, so we can take the Sum Sq column and divide it by n, which is the Mean Sq column. So we know the sample variance of the price is 120,181.

10. What is σ^2, the estimated variance of price within each level of cut.
11. State the null and alternative hypotheses for the test conducted using the ANOVA table in Exercise 8. State the hypotheses using both statistical notation and words in the context of the data.

The null hypothesis is that the the mean price of the diamonds is not affected by the cut of the diamond.

The alternative hypothesis is that the mean price of the diamonds is not affected by the cut of the diamond.

12. What is your conclusion for the test specified in the previous question? State the conclusion in the context of the data.

##Additional Analysis

13. Based on the conclusion of the ANOVA test, conduct further statistical analysis to provide more detail about which level(s) is(are) different and by how much. If further statistical analysis is not required, provide a brief explanation why it isn't based on the conclusion from the ANOVA test.