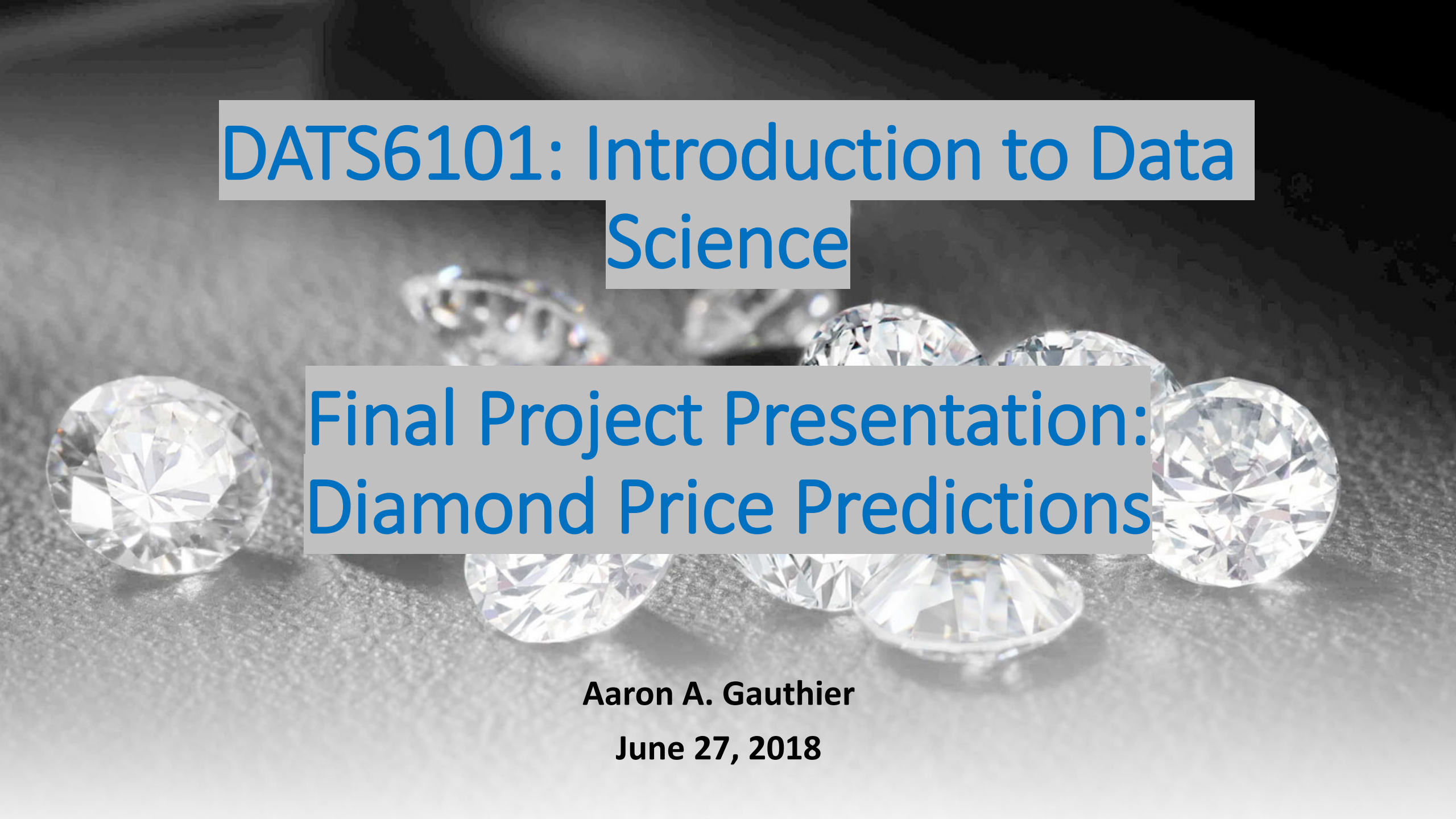


DATS6101: Introduction to Data Science

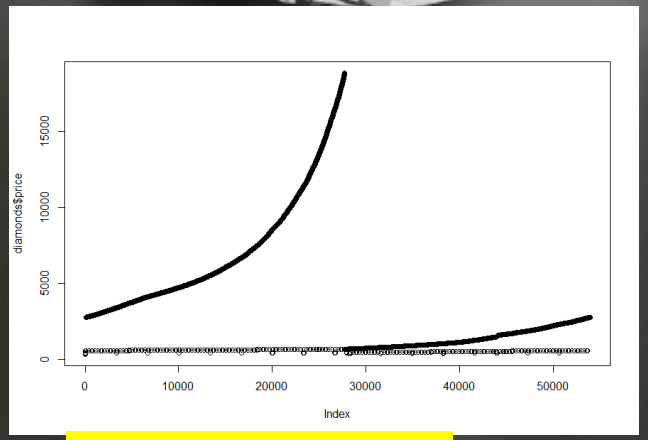
Final Project Presentation: Diamond Price Predictions

Aaron A. Gauthier

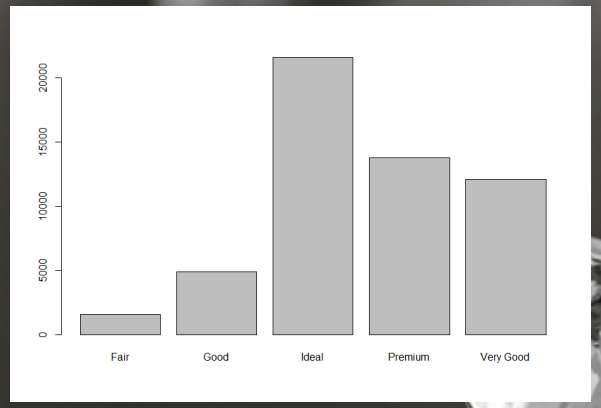
June 27, 2018

Question:

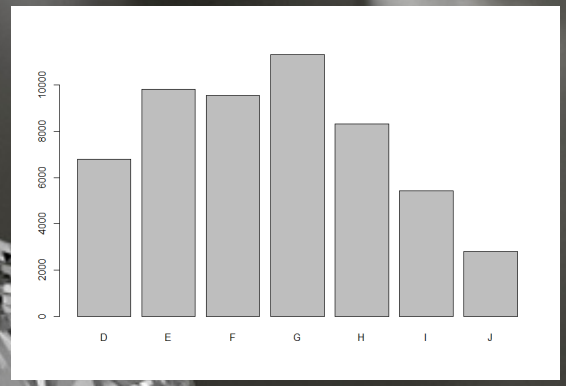
- What is the predicted price of a diamond based on the attributes of cut, color, clarity and carat?
- Data: Looked at how Carat (volume), Color, Clarity, Cut to predict Price
- Data Preparation: Ordinal Factor Variables to Factor Variables
- Exploratory Data Analysis (EDA) / Initial Observations:



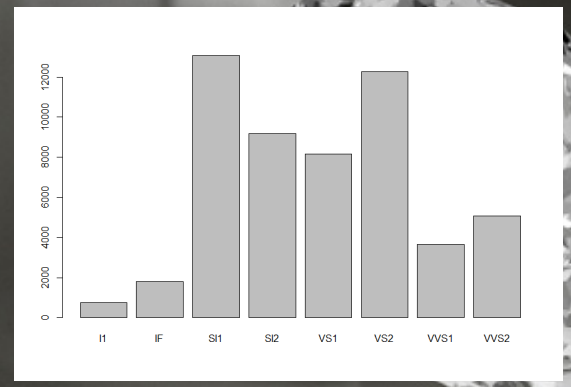
Diamond Price: Looks like an exponential relationship?



Diamond Cut: Very Good, Premium and Ideal – best sellers?

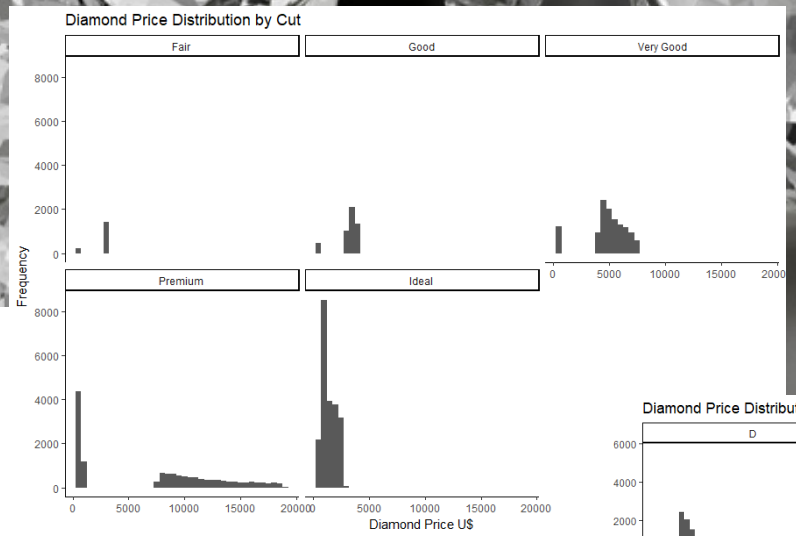


Diamond Color: There's less of The lower quality diamond colors…

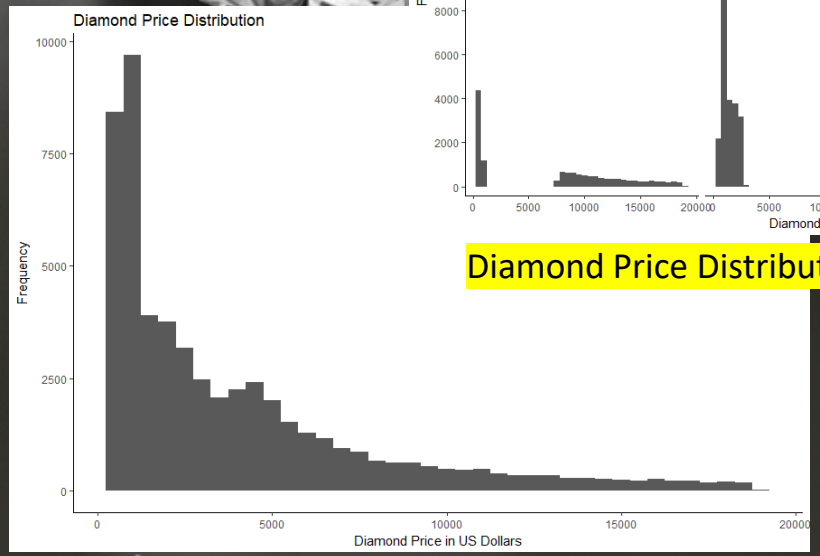


Diamond Clarity: There's more of the lesser quality diamonds…

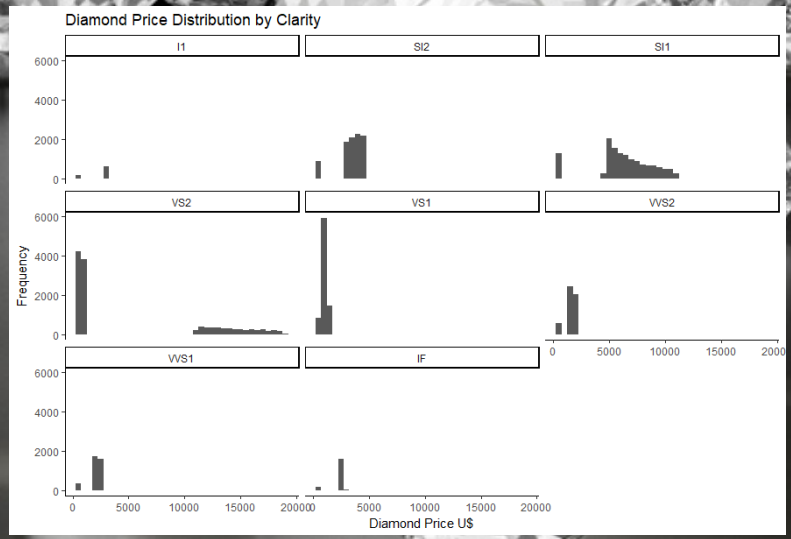
Diamond Data Distribution:



Diamond Price Distribution by Cut



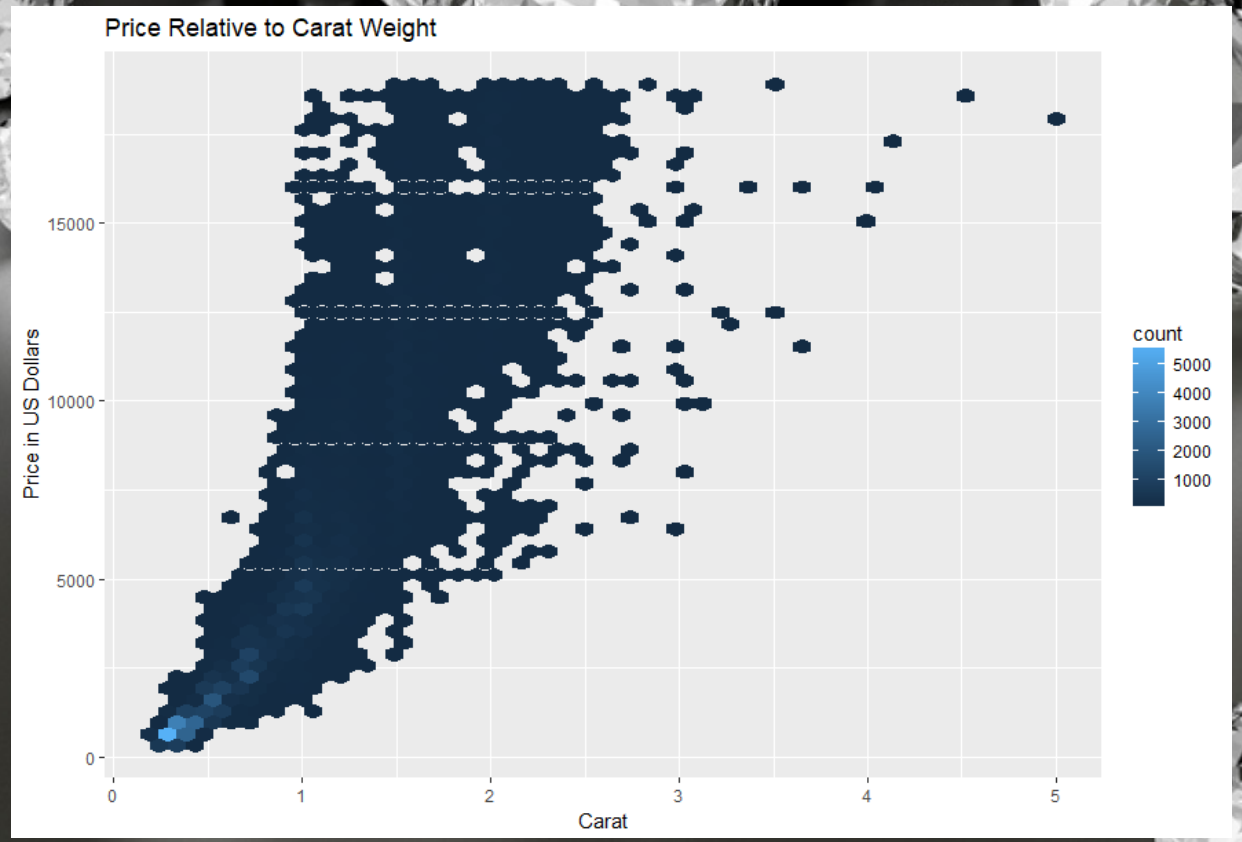
Long Tail Distribution
Mean Price: $3.932.80



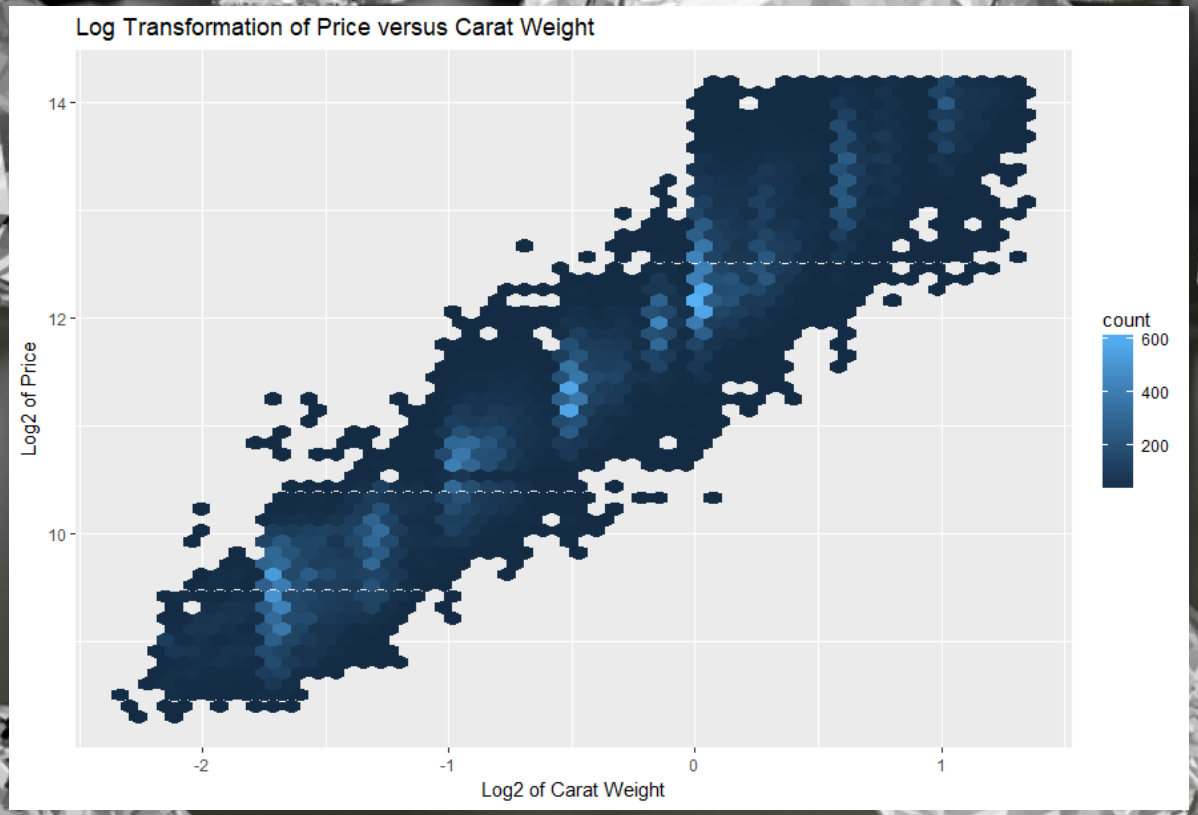
Diamond Price Distribution by Clarity

Diamond Price Distribution by Color

Log Transformation



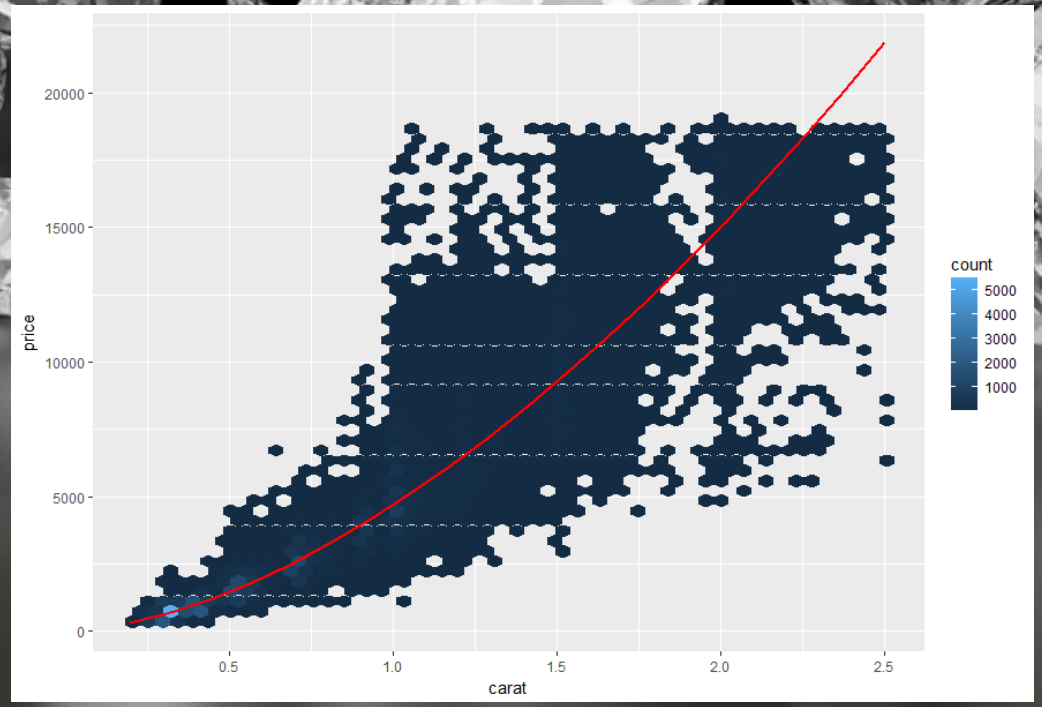
Looks exponential to me…lets take the Log2 to see if we can "flatten it out"…



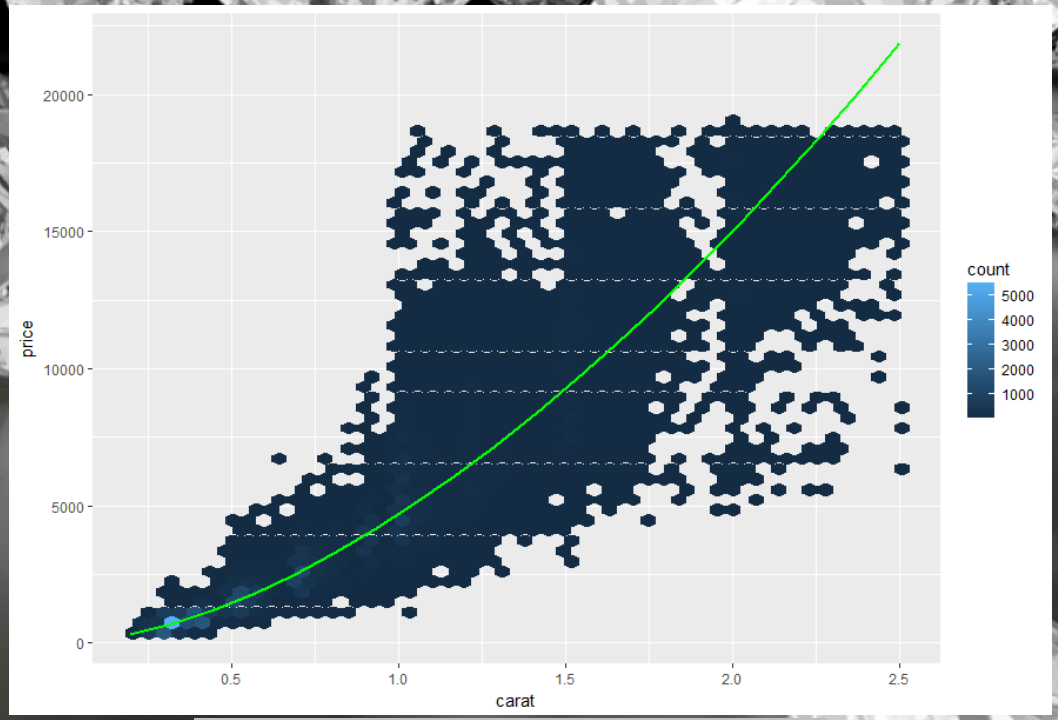
…by taking the Log2 it created a linear relationship between Carat Weight and Price…

There is a strong linear relationship between Price and Carat Weight!

Created A Simple Model…



Price vs. Carat: Linear Model



Price vs. Carat: Generalized Linear Model

```
Call:
lm(formula = lprice ~ lcarat, data = diamonds2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.96407 -0.24549 -0.00844  0.23930  1.93486

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.193863   0.001969  6194.5   <2e-16 ***
lcarat       1.681371   0.001936   868.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3767 on 53812 degrees of freedom
Multiple R-squared:  0.9334,    Adjusted R-squared:  0.9334
F-statistic: 7.542e+05 on 1 and 53812 DF,  p-value: < 2.2e-16
```

Models are exactly the same…

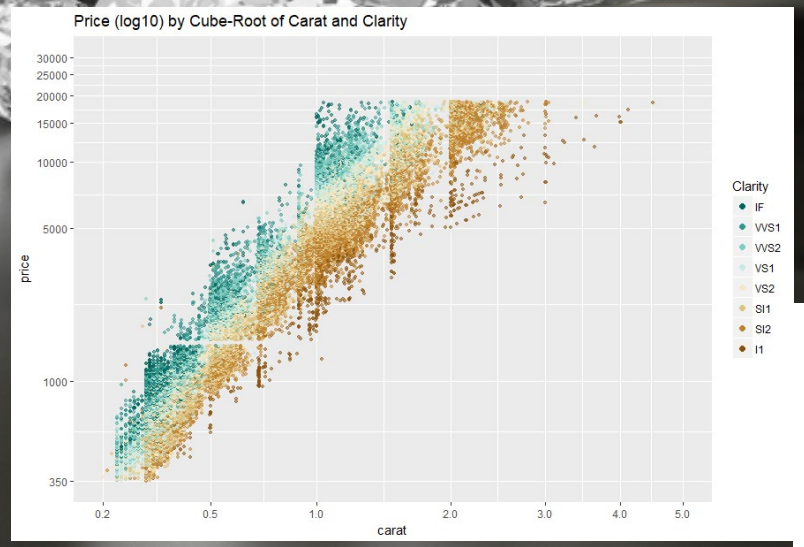
```
Call:
glm(formula = lprice ~ lcarat, data = diamonds2)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.96407  -0.24549  -0.00844   0.23930   1.93486

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.193863   0.001969  6194.5   <2e-16 ***
lcarat       1.681371   0.001936   868.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1419315)

    Null deviance: 114683.4  on 53813  degrees of freedom
Residual deviance:   7637.6  on 53812  degrees of freedom
AIC: 47654
```

Creating A More Complex Model…

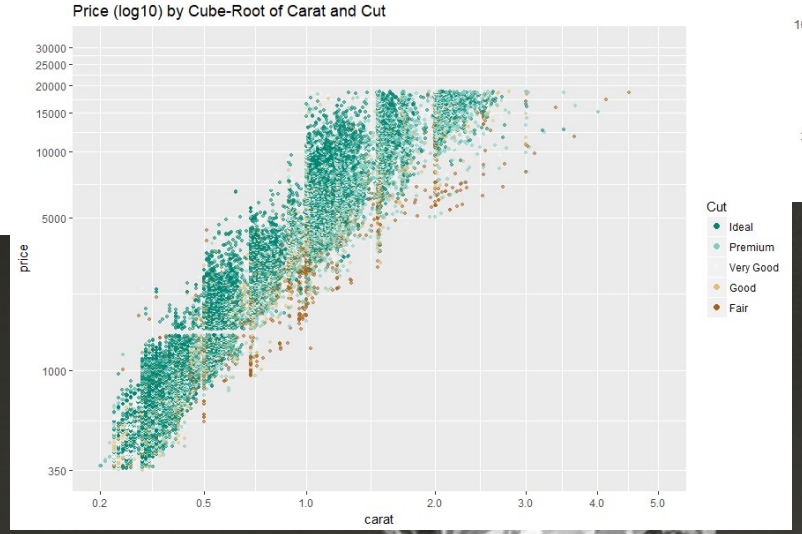
- Remember that volume in 3-D space is cubed…
- Carat Weight = volume
- Backward Selection Confirmed Variable Use
- Notice the Adjusted R^2:
  - Simple Model (Price ~ Carat)- 0.9334
  - Complex Model (Price ~ Carat, Cut, Color, & Clarity- 0.9839
  - ~difference of 0.0505
- **Proves Carat Weight is the most significant factor in the price of a diamond…**

```
Call:
lm(formula = I(log(price)) ~ I(carat^(1/3)) + carat + cut + color +
    clarity, data = diamonds)

Residuals:
     Min       1Q   Median       3Q      Max
-0.81377 -0.08307 -0.00080  0.07976  1.93542

Coefficients:
                Estimate Std. Error  t value Pr(>|t|)
(Intercept)     0.414792   0.009855   42.090  < 2e-16 ***
I(carat^(1/3)) 9.144314   0.016156  565.988  < 2e-16 ***
carat          -1.092551   0.005965 -183.164  < 2e-16 ***
cut.L           0.119825   0.002264   52.926  < 2e-16 ***
cut.Q          -0.031025   0.001992  -15.577  < 2e-16 ***
cut.C           0.013578   0.001730    7.849 4.28e-15 ***
cut^4          -0.001884   0.001385   -1.360   0.1739
color.L        -0.440905   0.001973 -223.494  < 2e-16 ***
color.Q        -0.092790   0.001796  -51.658  < 2e-16 ***
color.C        -0.013299   0.001676   -7.936 2.13e-15 ***
color^4         0.012047   0.001540    7.824 5.20e-15 ***
color^5        -0.003204   0.001454   -2.203   0.0276 *
color^6         0.001330   0.001322    1.006   0.3142
clarity.L       0.907144   0.003438  263.861  < 2e-16 ***
clarity.Q      -0.239602   0.003214  -74.552  < 2e-16 ***
clarity.C       0.130897   0.002749   47.624  < 2e-16 ***
clarity^4      -0.062759   0.002195  -28.593  < 2e-16 ***
clarity^5       0.025752   0.001792   14.371  < 2e-16 ***
clarity^6      -0.002090   0.001561   -1.339   0.1806
clarity^7       0.031982   0.001378   23.213  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1286 on 53920 degrees of freedom
Multiple R-squared:  0.9839,    Adjusted R-squared:  0.9839
F-statistic: 1.738e+05 on 19 and 53920 DF,  p-value: < 2.2e-16
```

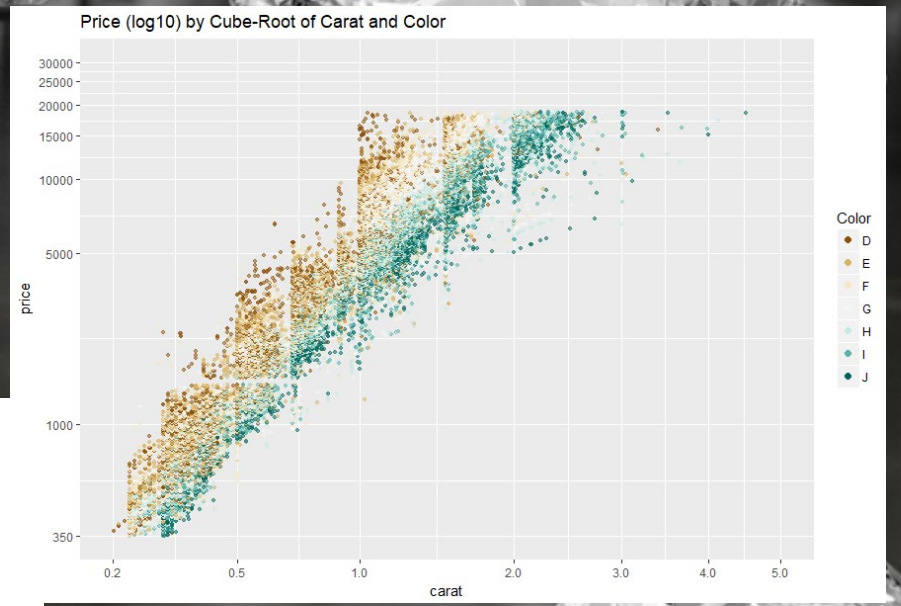
As the quality increases, so does the price…



There are less IF diamonds as the Carat weight gets larger…



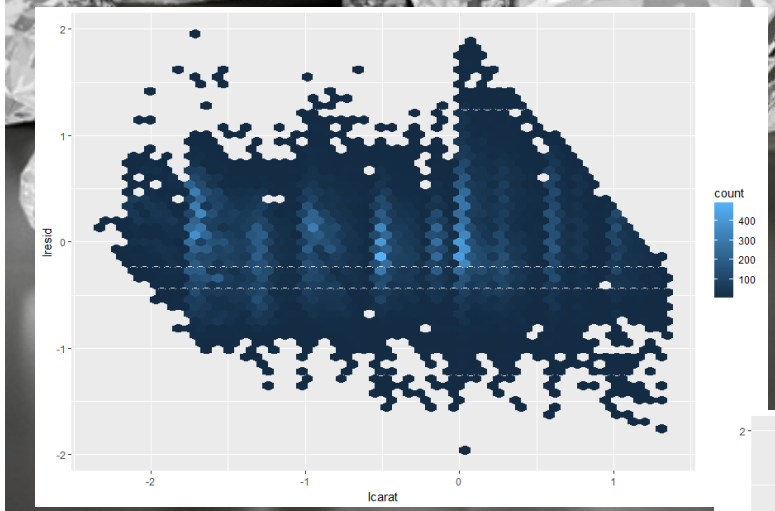
The data suggests that Cut does not influence diamond price…at least not significantly…



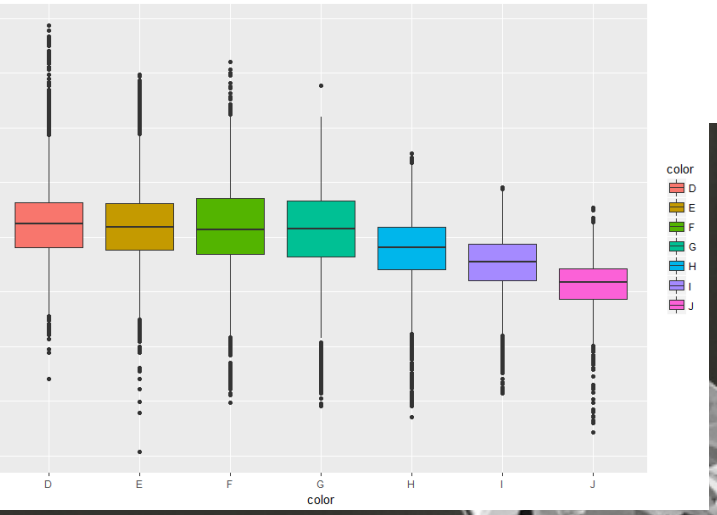
As the quality in Color increases, there are less available = rare

Residual Plots

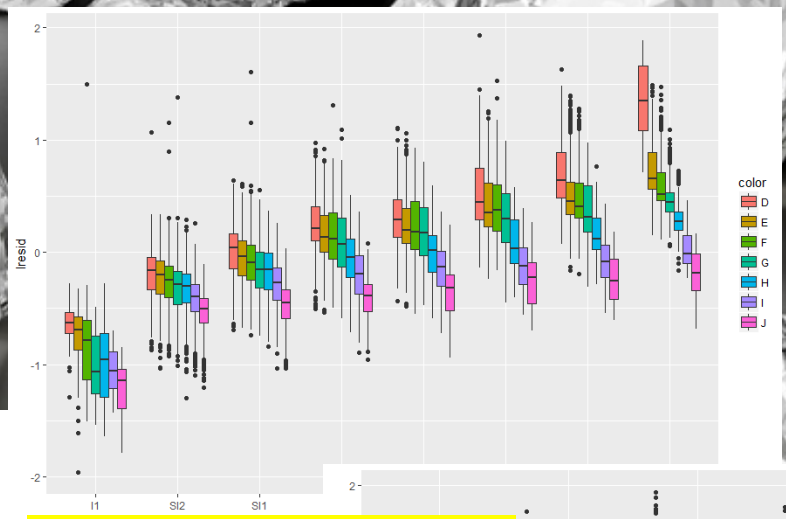
As the quality of the diamond increases, so does it's relative price…



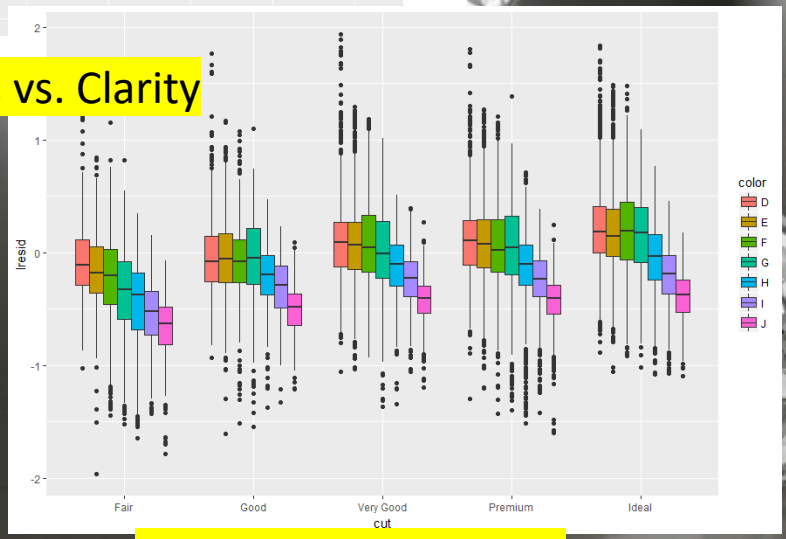
Price Residuals vs. Carat



Price Residuals vs. Color



Price Residuals vs. Clarity



Price Residuals vs. Cut

The stronger the correlation to one, the stronger the relationship between the attribute and price…

Prediction of Two Blue Nile Diamonds…

```
435  #Let's do some predictions with our models to see how we did....
436
437  #Our Example Diamond from BlueNile:
438
439  #1-- Round, Carat: 1.10, Cut: Astor by Blue Nile Ideal, Color: F, Clarity: VVS2, Price: $11,212     Taken from:
     https://www.bluenile.com/build-your-own-ring/diamond-details/LD09800963?refTab=DIAMONDS&track=viewDiamondDetails&action=newTab
440
441  #2-- Round, Carat: 0.90, Cut: Very Good, Color: F, Clarity: VS2, Price: $4,188      Taken from:
     https://www.bluenile.com/build-your-own-ring/diamond-details/LD10490866?refTab=DIAMONDS&track=viewDiamondDetails&action=newTab
442
```

Prediction 1:

```
> #1--Blue Nile Astor Diamond
> BlueNileDiamondLM = data.frame(carat = 1.10, cut = "Ideal",
+                          color = "F", clarity="VVS2")
> # data.Frame creates a data frame, we created a dataframe with one value, BlueNileDiamondLM
> modelEstimate = predict(lm5, newdata = BlueNileDiamondLM,
+                         interval="prediction", level = .95)
> exp(modelEstimate) # this will give us the actual price because our model outputs log 10.
       fit      lwr      upr
1 8635.849 6711.02 11112.75
>
```

The output is based off a 95% chance the diamond will fall within the price range and it did not…possibly attributed to Astor Diamond branding?

Prediction 2:

```
1 8635.849 6711.02 11112.75
> BlueNileDiamondLM1 = data.frame(carat = 0.90, cut = "Very Good",
+                           color = "F", clarity="VS2")
> # data.Frame creates a data frame, we created a dataframe with one value, BlueNileDiamondLM1
> modelEstimate1 = predict(lm5, newdata = BlueNileDiamondLM1,
+                          interval="prediction", level = .95)
> exp(modelEstimate1) # this will give us the actual price because our model outputs log 10.
       fit      lwr      upr
1 4548.009 3534.359 5852.372
>
```

The output is based off a 95% chance the diamond will fall within the price range and it did between the "fit" and "lower" prediction.

Values are within the bounds of the model!

Maybe branding plays a role in the price of a diamond…

Questions?

References:

1. Wickham, Hadley & Grolemund, Garrett. R for Data Science. Sebastopol: O'Reilly Media, Inc., 2017.

2. https://www.youtube.com/watch?v=GgthMorTsn0

3. https://rpubs.com/taylorwhite/diamondPricing

4. https://rpubs.com/ameliji/EDA_lesson3

5. https://rpubs.com/anthonycerna/diamondspredictions

6. www.bluenile.com

7. https://www.Kaggle.com/shivam2503/diamonds

8. https://gohighbrow.com/the-data-analysis-epicycle/

9. https://briatte.github.io/ggcorr/

10. *http://www.sthda.com/english/articles/36-classification-methodsessentials/149-penalized-logistic-regression-essentials-in-r-ridge-lasso-and-elastic-net/*