PERFECT DIAMOND

Capstone Team 4

CADE CULVER, JUSTIN MERRYMAN, KARLA MURPHY, AND PATRICIA BERRY

Executive Summary

Diamonds have been a fascination with throughout the world for centuries. They have always been a coveted possession and are truly a girl's best friend. How do you decide what type of diamond you want and how much should you pay for the one you want?

Our team investigated the diamond industry by looking at the anatomy of diamonds and how those factors affect pricing. With our tools, a user can define what they are looking for in a diamond and run it through the price predictor model to get an idea of what they might expect to pay for the diamond of their dreams.

The issue

When you make the decision to purchase a diamond, many people do not understand what they should be looking for in a diamond. Having to search the internet can be a long and tedious process. Our application makes this search much easier. There is a need to understand what the 4Cs are in diamond selection and how they impact the price you will be paying for the product as well as other features that can contribute to the price. Having this knowledge will help the user in being prepared when shopping for their diamond.

Data

The data used during the project covers the 4Cs – cut, clarity, color, and carat weight as well as length(mm), depth(mm), width (mm) and table size. The data was found on the www.kaggle.com website. The data was exceptionally clean and only needed one column deleted. No null values were found in the dataset.

Visualization Analysis

We developed four dashboards using Tableau to view the data in multiple ways. The end user will be able to look at price and how different features affect the price of the diamond. We looked at the number of times a diamond was purchased at different price points. The greatest sale was at $18,000, most sales were at $2,000 or less. There are several trend charts looking at pricing at different levels of cut, clarity, color, and carat weight.

Machine Learning

To ensure that we chose the right model, several linear regression and decision tree models were explored. The linear regression models provided us with a better model as it had a 91% accuracy and a low R2 score. The decision tree models provided great accuracy and low R2 scores, however, during testing the models overfit the data and would predict values over $14,000 regardless of the diamond carat size.

In conclusion, while the dataset was old, we were able to create a model that gives an end user multiple options in finding that perfect diamond for any occasion. Unfortunately, the data is not current, and it will not provide the most up to date pricing.

Conclusions

Choosing the perfect diamond boils down to choosing the right cut as it is the most important feature. Color, clarity, and carat weight play a large part in the pricing. The larger the weight, the least amount of color and the best clarity will bring the heftiest price. The largest sale in 2017, was $18,000. Most of the diamonds sold in 2017 were under $2,000.

Future Work Recommendations

If we were to continue this project in the future, we would look for an additional data source that would include shape in the dataset. We would also look for a more updated dataset as this was from 2017. Being able to include shape would add a new dimension to the overall tables and give the end user additional choices.

Limitations/Bias

Key limitation to the data was the lack of shape in the data. Diamond shape is an especially important piece of diamond selection, the data did not include this feature. We would have been able to narrow down a user's choice in picking the right diamond for their purposes.

Dataset is a bit dated from 2017.

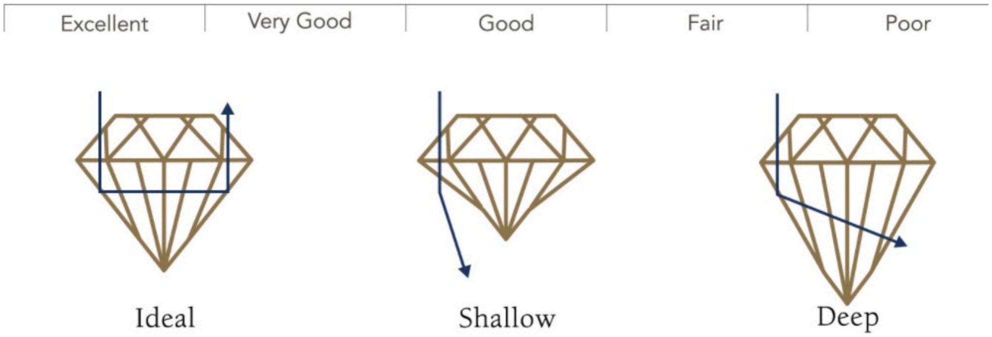
Model selection was limited to the linear regression models as the decision tree models overfit the data and was not as desirable as the regression models.

Table of Contents

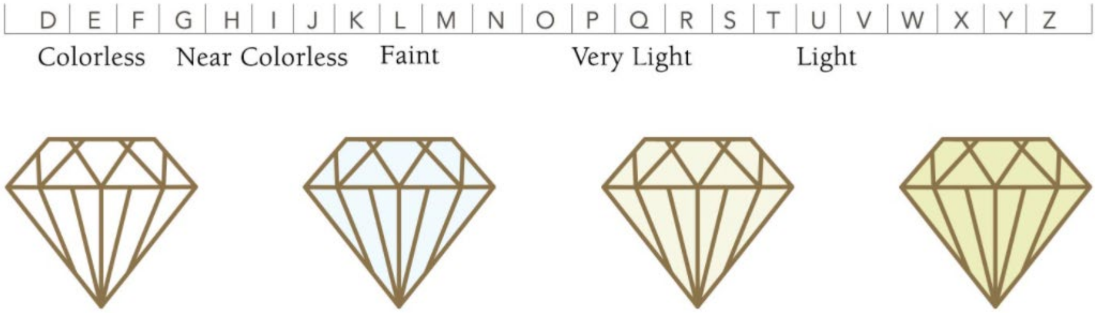
Diamond Introduction

Diamonds have been defined as a girl's best friend, however we wanted to know more about the diamond trade. Our team investigated pricing and the anatomy of a diamond to determine whether we could predict pricing and help the user to buy the best choice for their budget and preference. To choose the perfect diamond, you must follow the 4C's that determine the price of the diamond.

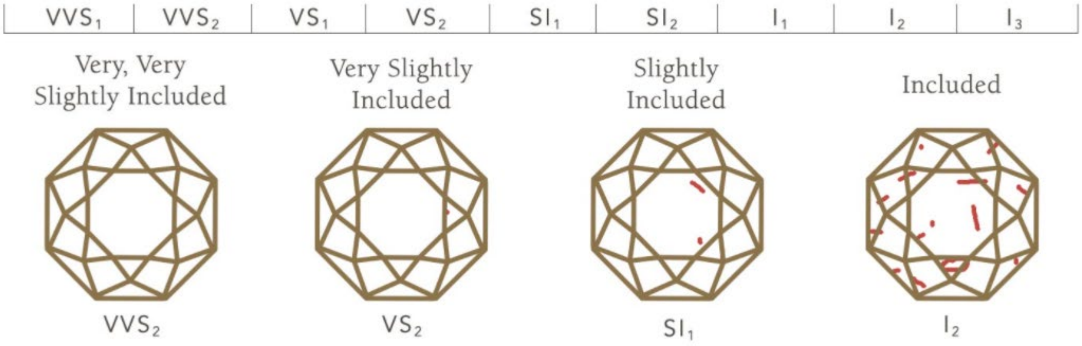
Cut – The most important of the 4Cs is cut. Cut drives how much a diamond sparkles when held up to the light.



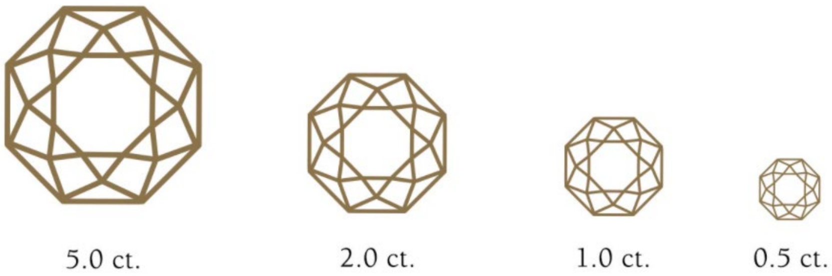
Color – Color is the second most important feature in diamond selection. The less color that a diamond has the higher the grade of the diamond.



Clarity – clarity defines the degree that a diamond is flawed. Clarity grade is based on the size, number, position of, and color of the diamond's flaws.

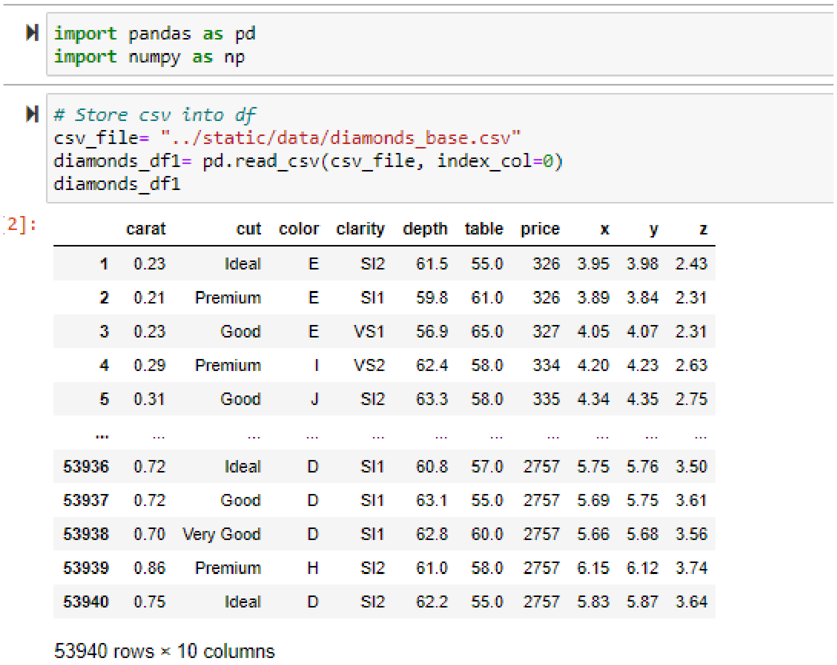


Carat Weight – carat weight refers to a diamond's total weight not to size.



Data Information

The data was selected from Kaggle and was in a .csv format. The data set was medium in size and we were able to use it as is without taking a random sample. The data selected was exceptionally clean and only one column was deleted as it mirrored the index column.



```
import pandas as pd
import numpy as np
```

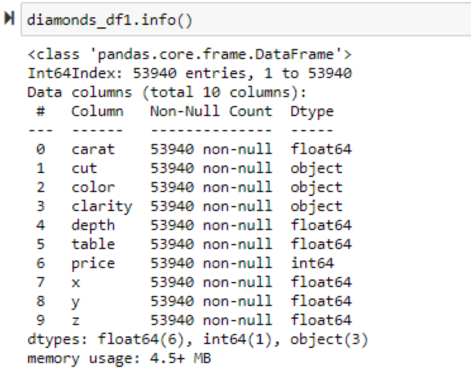
```
# Store csv into df
csv_file= "../static/data/diamonds_base.csv"
diamonds_df1= pd.read_csv(csv_file, index_col=0)
diamonds_df1
```

2]:

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...	...
53936	0.72	Ideal	D	SI1	60.8	57.0	2757	5.75	5.76	3.50
53937	0.72	Good	D	SI1	63.1	55.0	2757	5.69	5.75	3.61
53938	0.70	Very Good	D	SI1	62.8	60.0	2757	5.66	5.68	3.56
53939	0.86	Premium	H	SI2	61.0	58.0	2757	6.15	6.12	3.74
53940	0.75	Ideal	D	SI2	62.2	55.0	2757	5.83	5.87	3.64

53940 rows × 10 columns

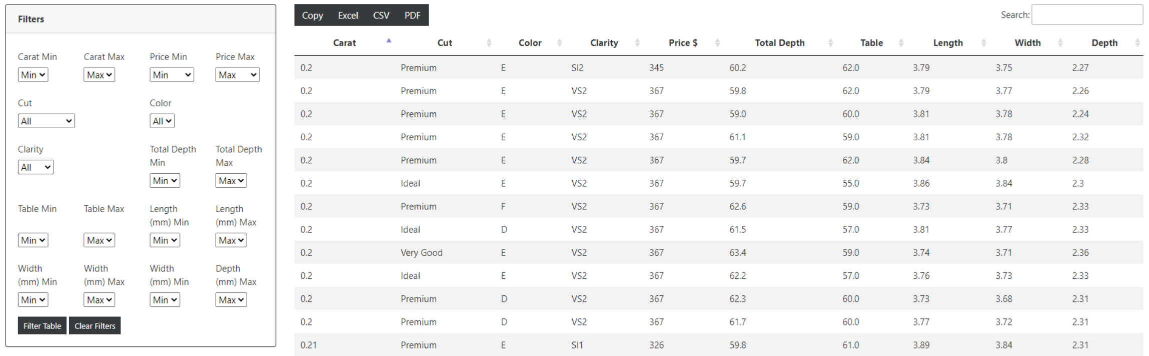
We checked to ensure that the data did not have any null values.



```
diamonds_df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 53940 entries, 1 to 53940
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   carat    53940 non-null  float64
 1   cut      53940 non-null  object 
 2   color    53940 non-null  object 
 3   clarity  53940 non-null  object 
 4   depth    53940 non-null  float64
 5   table    53940 non-null  float64
 6   price    53940 non-null  int64  
 7   x        53940 non-null  float64
 8   y        53940 non-null  float64
 9   z        53940 non-null  float64
dtypes: float64(6), int64(1), object(3)
memory usage: 4.5+ MB
```

Data



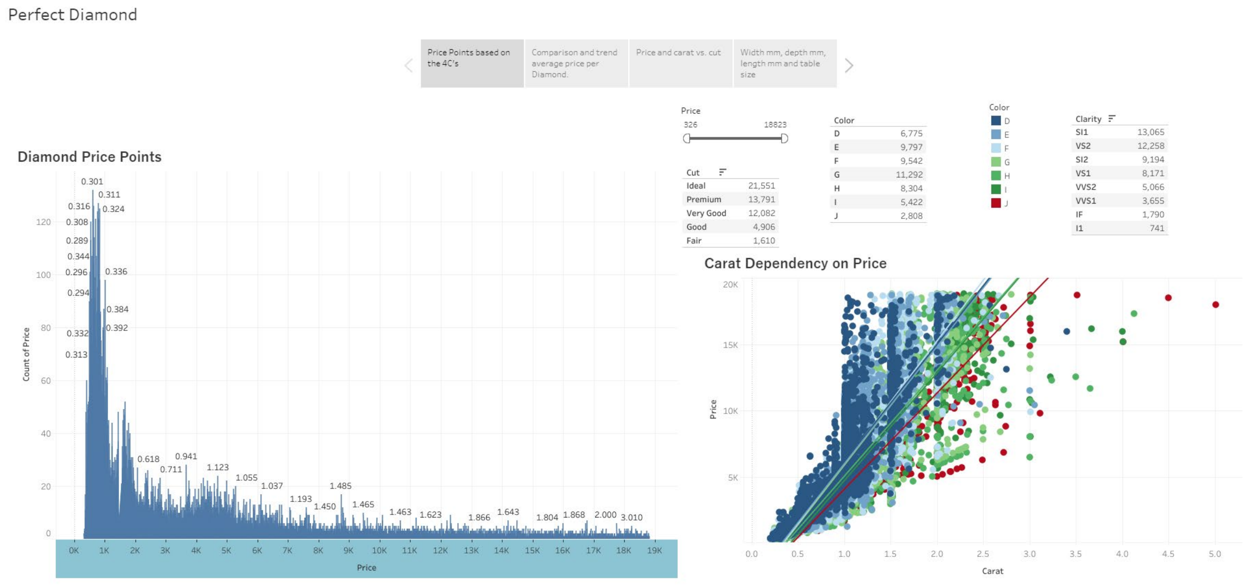
Carat	Cut	Color	Clarity	Price $	Total Depth	Table	Length	Width	Depth
0.2	Premium	E	SI2	345	60.2	62.0	3.79	3.75	2.27
0.2	Premium	E	VS2	367	59.8	62.0	3.79	3.77	2.26
0.2	Premium	E	VS2	367	59.0	60.0	3.81	3.78	2.24
0.2	Premium	E	VS2	367	61.1	59.0	3.81	3.78	2.32
0.2	Premium	E	VS2	367	59.7	62.0	3.84	3.8	2.28
0.2	Ideal	E	VS2	367	59.7	55.0	3.86	3.84	2.3
0.2	Premium	F	VS2	367	62.6	59.0	3.73	3.71	2.33
0.2	Ideal	D	VS2	367	61.5	57.0	3.81	3.77	2.33
0.2	Very Good	E	VS2	367	63.4	59.0	3.74	3.71	2.36
0.2	Ideal	E	VS2	367	62.2	57.0	3.76	3.73	2.33
0.2	Premium	D	VS2	367	62.3	60.0	3.73	3.68	2.31
0.2	Premium	D	VS2	367	61.7	60.0	3.77	3.72	2.31
0.21	Premium	E	SI1	326	59.8	61.0	3.89	3.84	2.31

The Data table in the application gives the end user an opportunity to review all data points in the dataset or they can narrow down to just a few points.

Tableau

Three dashboards were created to visualize the different features of diamonds. We chose the green-blue diverging color scheme for our charts due to the cool appearance and that the color scheme would fit with the chosen bootswatch template we chose for the html.

The first dashboard is broken up into two segments. The chart on the left identifies the number of times a certain price point occurred. The maximum price in this case was around $18 K. Most of the diamonds sold are below $2 K. The second chart highlights the carat dependency on price. Filters at the top allow the user to choose the cut, color, and clarity.

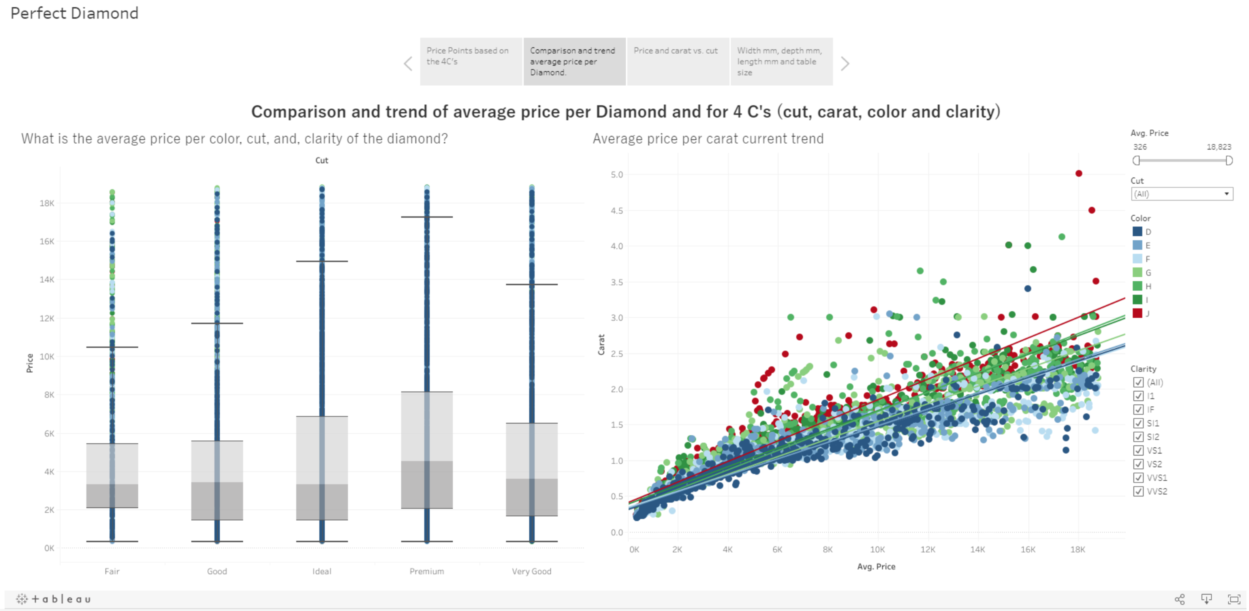


Cut	
Ideal	21,551
Premium	13,791
Very Good	12,082
Good	4,906
Fair	1,610

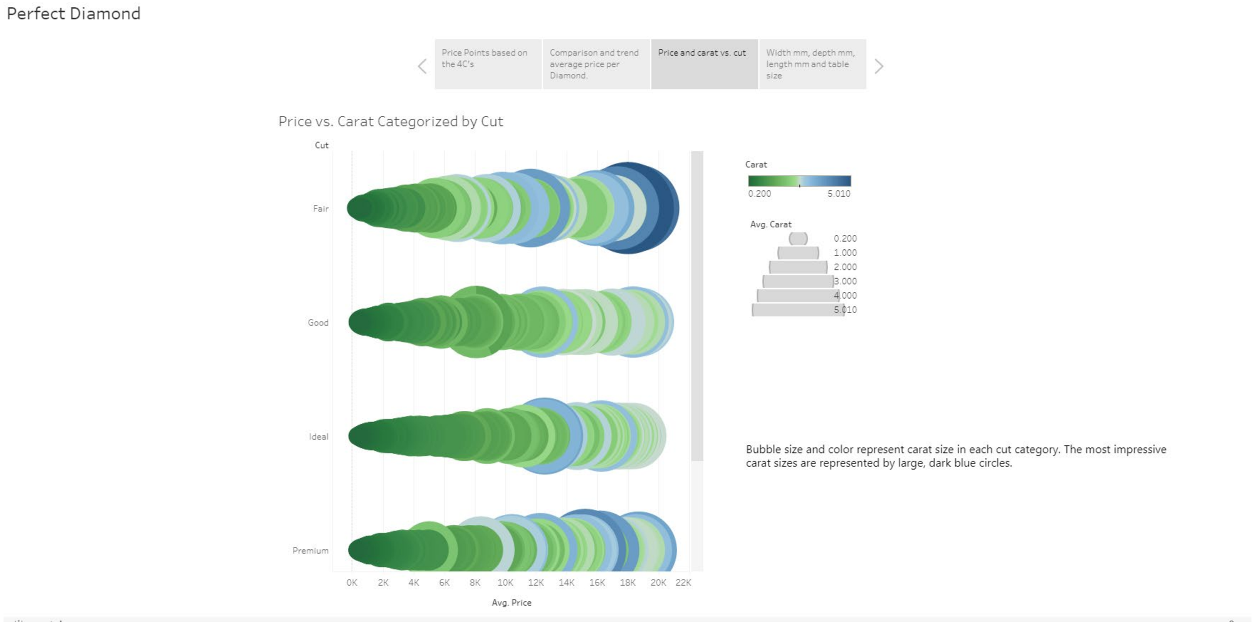
Color	
D	6,775
E	9,797
F	9,542
G	11,292
H	8,304
I	5,422
J	2,808

Clarity	
SI1	13,065
VS2	12,258
SI2	9,194
VS1	8,171
VVS2	5,066
VVS1	3,655
IF	1,790
I1	741

The second dashboard covers the comparison and trend of average price per Diamond and the 4C's (cut, carat, color, and clarity). There are filters to switch between the 4C's to understand the trend between the 4 different features. The first chart looks at how cut affects price. The second chart looks at the average price per carat in a scatter chart and identifies the trends.

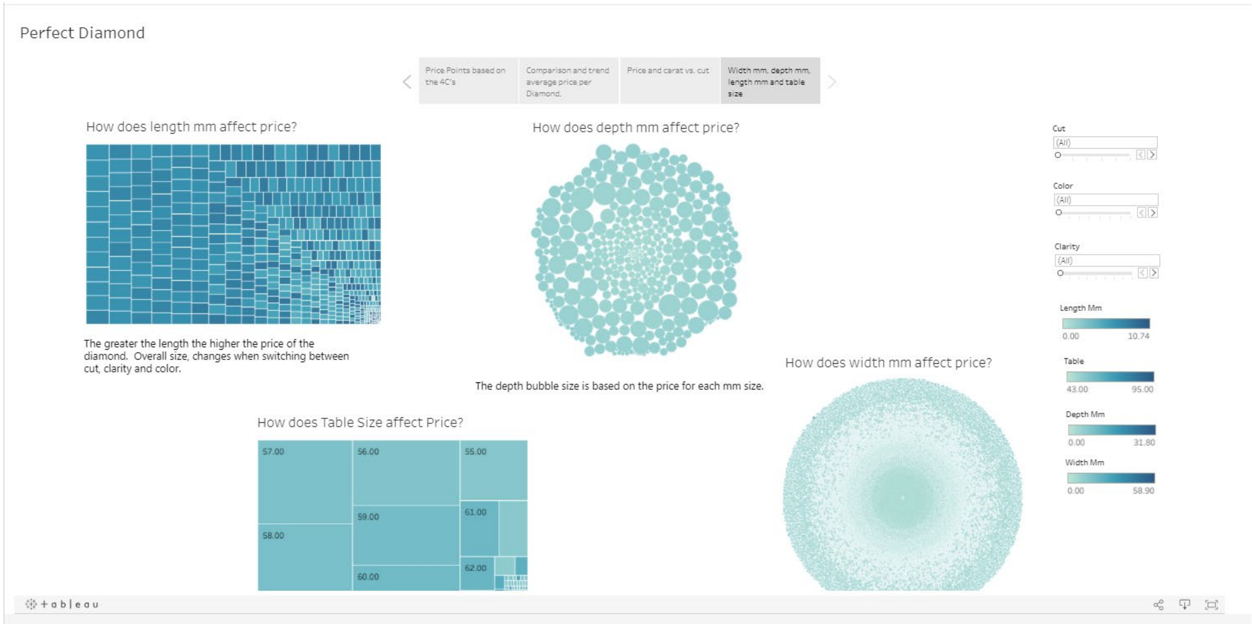

Perfect Diamond
Price Points based on the 4C's
Comparison and trend average price per Diamond.
Price and carat vs. cut
Width mm, depth mm, length mm and table size
Comparison and trend of average price per Diamond and for 4 C's (cut, carat, color and clarity)
What is the average price per color, cut, and, clarity of the diamond?
Cut
Price
0K
2K
4K
6K
8K
10K
12K
14K
16K
18K
Fair
Good
Ideal
Premium
Very Good
Average price per carat current trend
Carat
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
0K
2K
4K
6K
8K
10K
12K
14K
16K
18K
Avg. Price
Avg. Price
326
18,823
Cut
(All)
Color
D
E
F
G
H
I
J
Clarity
☑ (All)
☑ I1
☑ IF
☑ SI1
☑ SI2
☑ VS1
☑ VS2
☑ VVS1
☑ VVS2
+ a b | e a u

The third dashboard is a circle bar chart that highlights how carat weight determines the diamond price.



Bubble size and color represent carat size in each cut category. The most impressive carat sizes are represented by large, dark blue circles.

The final dashboard dives into the remaining data points contained within the data set and how price affects those measures. We look at length mm, width mm, depth mm, carat weight and table size. Filters allow the user to switch between cut, color, and clarity. The length mm chart is a square chart that starts at the largest impact and dives down to the lower point. The table size chart is square chart that starts at the largest impact and dives down to the lowest point. The depth mm chart is a circle scatter chart where the price identifies the size of the bubble. The width mm chart is a circle chart that plots the layers of price driven by the width mm of the diamond.



Perfect Diamond

How does length mm affect price?

The greater the length the higher the price of the diamond. Overall size, changes when switching between cut, clarity and color.

How does depth mm affect price?

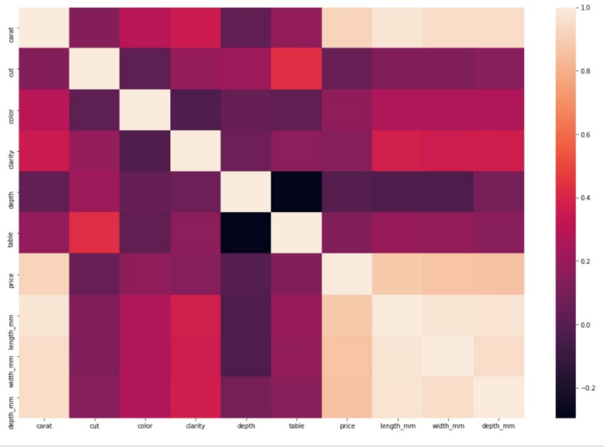
The depth bubble size is based on the price for each mm size.

How does width mm affect price?

How does Table Size affect Price?

Machine Learning

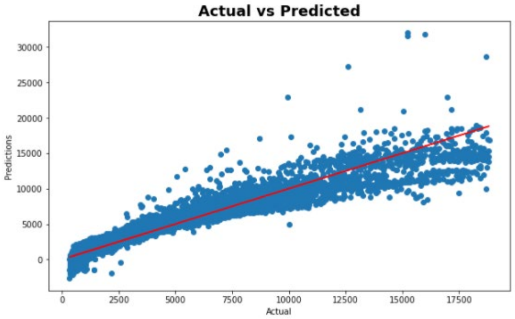
The machine learning module was created to give an end user the opportunity to identify the price of a diamond based on their choices between carat weight, color, clarity and cut. The data was put through several tests to identify the best model to use for the predictive application. The first step was to identify correlations within the data.



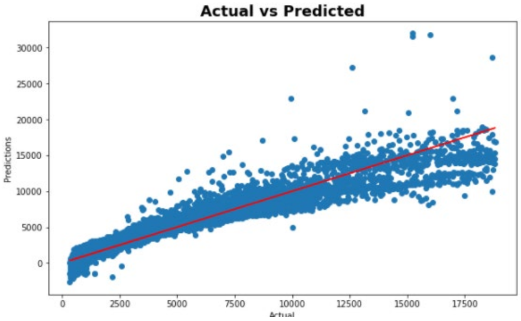
The next step was to look at the different models.

Linear Regression Models

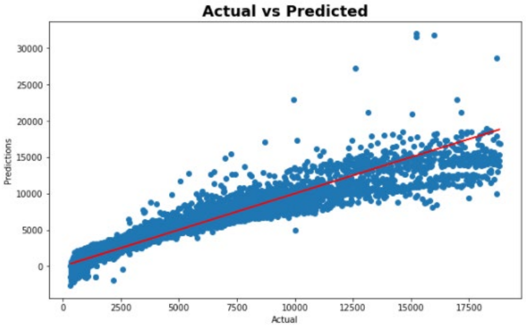
LinearRegression - the in sample R2 score was 0.9092517263784572 and the in sample RMSE was 1201.9520820408259. The out sample R2 score was 0.9076446121106826 and the out sample RMSE was 1211.675332479086.



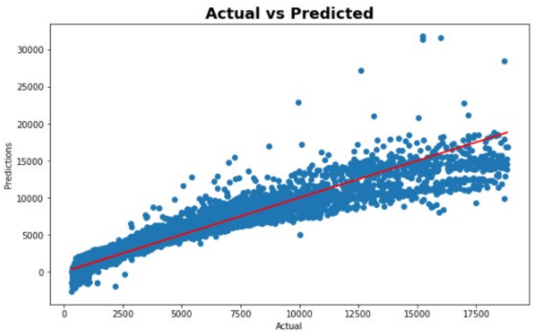
The KNeighborsRegression model - the in sample R2 score was 0.9092517263784572 and the in sample RMSE was 1201.9520820408259. The out sample R2 score was 0.9076446121106826 and the out sample RMSE was 1211.675332479086.



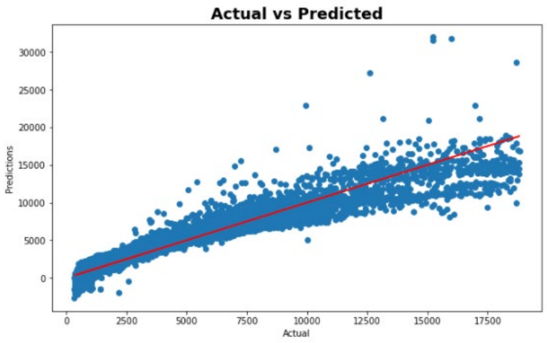
The SVR model - the in sample R2 score was 0.9092517263784572 and the in sample RMSE was 1201.9520820408259. The out sample R2 score was 0.9076446121106826 and the out sample RMSE was 1211.675332479086.



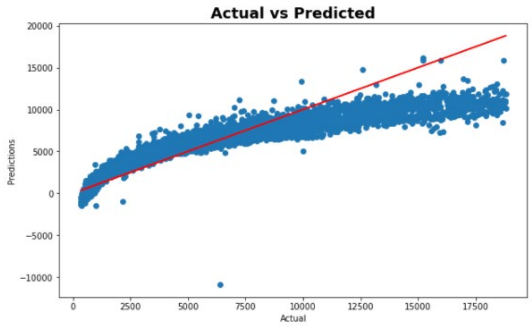
The Lasso model - the in sample R2 score was 0.9092361303192542 and the in sample RMSE was 1202.0553617492924. The out sample R2 score was 0.9076434629479834 and the out sample RMSE was 1211.6828707927464.



The Ridge model - the in sample R2 score was 0.9092513195555165 and the in sample RMSE was 1201.954776203001. The out sample R2 score was 0.9076494175687261 and the out sample RMSE was 1211.6438089759688.

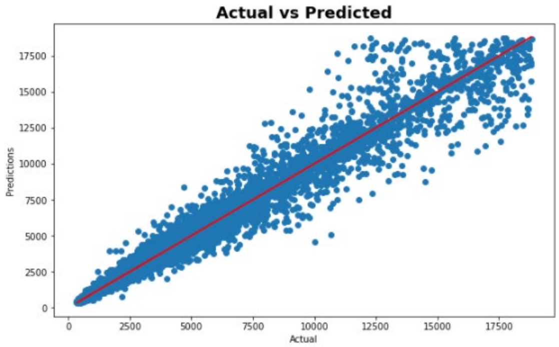


The ElasticNet model - the in sample R2 score was 0.8104628540520513 and the in sample RMSE was 1737.061267132269. The out sample R2 score was 0.8197579525330729 and the out sample RMSE was 1692.7125115007573.

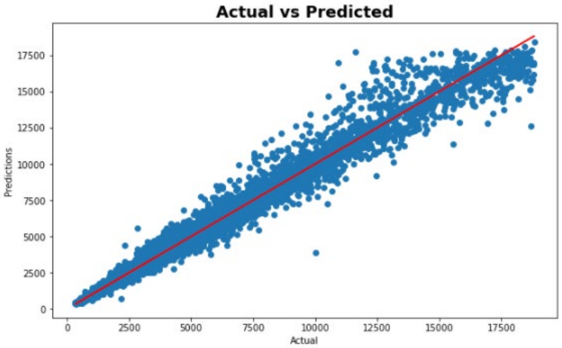


Trees Models

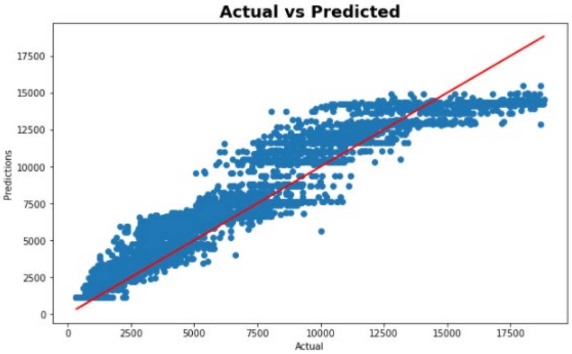
The DecisionTreeRegressor model - the in sample R2 score was 0.9999948355708338 and the in sample RMSE was 9.067323921015339. The out sample R2 score was 0.9666252386506127 and the out sample RMSE was 728.3904324945589.



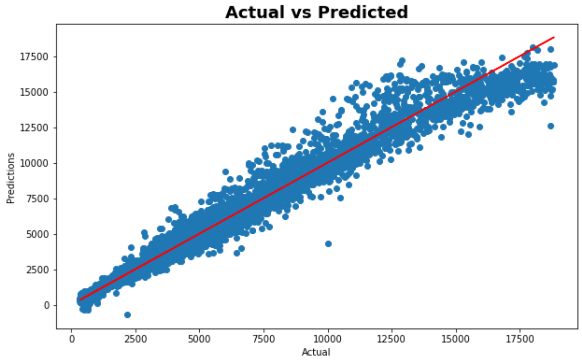
The RandomForestRegressor – the in sample R2 score was 0.997422047746516 and the in sample RMSE was 202.58412171224964. The out sample R2 score was 0.981658235117793 and the out sample RMSE was 539.9791992422503.



The AdaBoostRegressor – the in sample R2 score was 0.9150623620214242 and the in sample RMSE was 1162.8349006414346. The out sample R2 score was 0.9131433607546641 and the out sample RMSE was 1175.0508439806226.

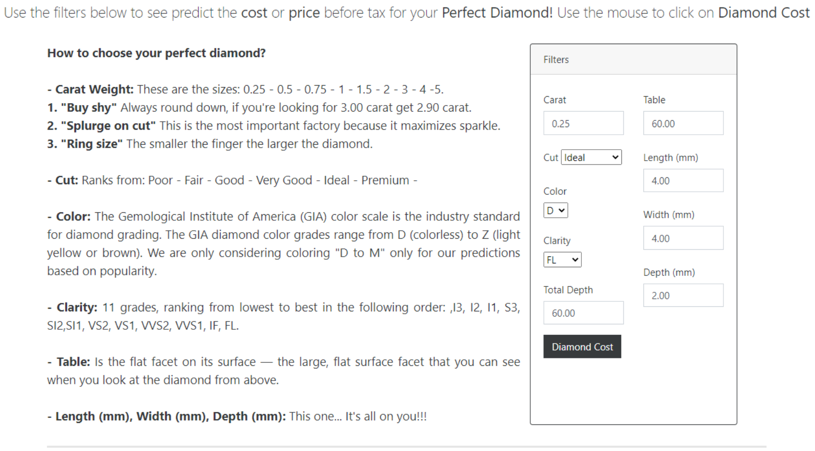


The GradientBoostRegressor – the in sample R2 score was 0.9773703273701563 and the in sample RMSE was 600.2153281248416. The out sample R2 score was 0.9766551379985087 and the out sample RMSE was 609.1872479312117.



Linear Regression model was chosen because it predicted with an accuracy of 91% and has the lowest R2 score of all the linear models. While the Decision Tree models predict a better fit of the data and a lower R2 square, the linear regression was chosen over any of the Decision Trees because all the Decision Tree models overfit the data and would not predict anything over $14,000 regardless of the diamond carat size.

The Price Predictor tool:


Use the filters below to see predict the **cost** or **price** before tax for your **Perfect Diamond!** Use the mouse to click on **Diamond Cost**

How to choose your perfect diamond?

- Carat Weight: These are the sizes: 0.25 - 0.5 - 0.75 - 1 - 1.5 - 2 - 3 - 4 -5.
1. "Buy shy" Always round down, if you're looking for 3.00 carat get 2.90 carat.
2. "Splurge on cut" This is the most important factory because it maximizes sparkle.
3. "Ring size" The smaller the finger the larger the diamond.

- Cut: Ranks from: Poor - Fair - Good - Very Good - Ideal - Premium -

- Color: The Gemological Institute of America (GIA) color scale is the industry standard for diamond grading. The GIA diamond color grades range from D (colorless) to Z (light yellow or brown). We are only considering coloring "D to M" only for our predictions based on popularity.

- Clarity: 11 grades, ranking from lowest to best in the following order: ,I3, I2, I1, S3, SI2,SI1, VS2, VS1, VVS2, VVS1, IF, FL.

- Table: Is the flat facet on its surface — the large, flat surface facet that you can see when you look at the diamond from above.

- Length (mm), Width (mm), Depth (mm): This one... It's all on you!!!

Filters

Carat: 0.25
Table: 60.00
Cut: Ideal
Length (mm): 4.00
Color: D
Width (mm): 4.00
Clarity: FL
Depth (mm): 2.00
Total Depth: 60.00

Conclusions

In conclusion, while the dataset was old, we were able to create a model that gives an end user multiple options in finding that perfect diamond for any occasion. Unfortunately, the data is not current, and it will not provide the most up to date pricing.

Choosing the perfect diamond boils down to choosing the right cut as it is the most important feature. Color, clarity, and carat weight play a large part in the pricing. The larger the weight, the least amount of color and the best clarity will bring the heftiest price. The largest sale in 2017, was $18,000. Most of the diamonds sold in 2017 were under $2,000.

Future Work Recommendations

If we were to continue this project in the future, we would look for an additional data source that would include shape in the dataset. We would also look for a more updated dataset as this was from 2017.

Being able to include shape would add a new dimension to the overall tables and give the end user additional choices.

Limitations/Bias

Key limitation to the data was the lack of shape in the data. Diamond shape is an especially important piece of diamond selection, the data did not include this feature. We would have been able to narrow down a user's choice in picking the right diamond for their purposes.

Dataset is a bit dated from 2017.

Model selection was limited to the linear regression models as the decision tree models overfit the data and was not as desirable as the regression models.

Resources

Diamonds: https://www.kaggle.com/shivam2503/diamonds

Diamonds In-Depth Analysis: https://www.kaggle.com/fuzzywizard/diamonds-in-depth-analysis

Prediction Diamond Prices: https://www.kaggle.com/himahima/prediction-diamond-prices

Regression graph examples: https://datascienceplus.com/wp-content/uploads/2016/02/Screen-Shot-2016-02-15-at-6.39.53-PM.png

Regression graph examples: https://cme195.github.io/assets/lectures/Lecture4_Visualizing_Data_files/figure-revealjs/unnamed-chunk-41-1.png

Tableau example: https://public.tableau.com/profile/zitong.yang#!/vizhome/DiamondAttributes-PriceClarityandColor/Story1

Data cleaning inspiration: https://analyticsindiamag.com/tutorial-get-started-with-exploratory-data-analysis-and-data-preprocessing/

EDA example: https://www.excelr.com/exploratory-data-analysis-in-data-science

Diamond research: https://www.lumeradiamonds.com/diamond-education/index

Diamond research: https://www.bluenile.com/education/diamonds?gclid=da6dfee664eb16716dca9e624e867e33&msclkid=da6dfee664eb16716dca9e624e867e33&click_id=173535549&utm_source=bing&utm_medium=text&utm_campaign=Diamonds+%3A+Education+Prospect_Exact_Desktop&utm_content=Diamonds%3AGrading%3AChartutm_term%3Ddiamond+grading+chart

Diamond research: https://www.ihatestevensinger.com/engagement-rings/101-diamonds-women/?lang=en_US

About Us

Cade Culver graduated from Texas Tech University with a Bachelor of Arts in Economics with a minor in Mathematics. He is currently seeking a position in Data Analysis. Hobbies: In his free time, he enjoys playing guitar and working on his family's ranch.

Justin Merryman graduated from The University of Texas with a Bachelor of Science in Chemistry and an Elements of Computer Programming Certificate. He has worked as Chemist in the oil and gas industry for over 10 years. Currently, he is a deepwater engineer for Schlumberger working with production chemicals. As he has expanded his education, he is using his computer programming and data analysis skills to further solve the chemistry problems he faces daily. Hobbies: He enjoys traveling, golf, fishing, and hunting.

Karla Murphy graduated from Universidad Latina de Costa Rica with a Bachelor of Business Administration in Marketing/Sales Management. She was recently promoted as Sales Excellence Specialist at Emerson Automation Solutions leading cross-functional projects, driving improvement, data analysis and sales projections. Hobbies: Beach, outdoors, read, music (Tropical music!), family and animals.

Patricia Berry graduated from the University of Phoenix-Dallas with a Bachelor of Arts in Business Administration. She has been a Category Manager in the Food Industry for the last 28 years working for a variety of top tiered companies throughout her career. Hobbies: She enjoys spending time traveling to different Disney properties around the world.