Diamond Dataset
Exploratory Analysis

Morgan Hardin, mhardin5@bellarmine.edu
Gavin Trumbull, gtrumbull@bellarmine.edu

I. INTRODUCTION

The largest diamond dataset was 1 file consisting of 26 columns, each with 219,703 entries. There were no null entries in the dataset and there were 8 numerical columns and 18 categorical columns. We focused on the how the cut, color, cut quality, carat weight, length, and table percent impacted the total sales price. We chose this dataset because it had a lot of entries which helped give us a better overview on how all of these variables impact the price. We thought it was very interesting and different than the other datasets we have seen. The dataset focuses on 25 different variables that all impact the price of a diamond. Although we focused on a few, this dataset showed us just how much goes into determining the total sales price of a diamond. The diamond dataset was found on Kaggle and the link can be found here: https://www.kaggle.com/datasets/hrokrin/the-largest-diamond-dataset-currely-on-kaggle.

II. DATA SET DESCRIPTION

The diamond dataset consists of 1 file with 26 columns and 219,703 rows. Within the dataset, there are no null entries in any of the rows or columns. There are 8 numerical columns and 18 categorical columns. For all of the columns, 0% of the data is missing.

'Unnamed: 0' and 'Total Sales Price' are int64 data types that are ratio since there is an absolute zero.
'Cut', 'Lab', 'Eye Clean', 'Culet Condition', 'Fancy Color Dominant Color', 'Fancy Color Secondary Color', 'Fancy Color Overtone', and 'Fancy Color Intensity' which are object data types and nominal data.
'Color', 'Clarity', 'Cut Quality', 'Symmetry', 'Polish', 'Culet', 'Girdle Min', 'Girdle Max', 'Fluor Color', and 'Fluor Intensity' are object data types and ordinal data.
'Carat Weight' is a float64 data type and is interval data because there is no absolute zero.
'Depth Percent', 'Table Percent', 'Meas Length', 'Meas Width', and 'Meas Depth' are float64 data types and ratio data since there is an absolute zero.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
Unnamed: 0	Int64	0%
Cut	Object	0%
Color	Object	0%
Clarity	Object	0%
Carat Weight	Float64	0%
Cut Quality	Object	0%
Lab	Object	0%
Symmetry	Object	0%
Polish	Object	0%
Eye Clean	Object	0%
Culet Size	Object	0%
Culet Condition	Object	0%
Depth Percent	Float64	0%
Table Percent	Float64	0%
Meas Length	Float64	0%
Meas Width	Float64	0%
Meas Depth	Float64	0%
Girdle Min	Object	0%
Girdle Max	Object	0%
Fluor Color	Object	0%

Fluor Intensity	Object	0%
Fancy Color Dominant Color	Object	0%
Fancy Color Secondary Color	Object	0%
Fancy Color Overtone	Object	0%
Fancy Color Intensity	Object	0%
Total Sales Price	Int64	0%

III. Data Set Summary Statistics

The diamond dataset has 219,703 in each column with no null entries. There was a pretty decent distribution across the board for most columns. There were a few categorical columns like 'Cut Quality', 'Symmetry', 'Culet Condition', 'Fancy Color Dominant Color', 'Fancy Color Secondary Color', and 'Fancy Color Overtone' that were skewed a little and had an uneven distribution. We found that there was a strong and positive correlation between 'Meas Length' and 'Meas Width' to 'Carat Weight'. We thought this was interesting since we assumed 'Depth Percent' and 'Table Percent' would have more of a strong positive correlation since it deals with the area of the top of the diamond. We were shocked to find that there was only a weak positive correlation to them and the 'Carat Weight'. Overall, these statistics show that the distributions and the correlations have more of a positive impact on 'Total Sales Price' than we originally thought.

Table 2: Summary Statistics for Diamond Dataset

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
Unnamed: 0	219703	109851.75	63423.26	0	54925.50	109852	164777.5	219703
Carat Weight	219703	0.76	0.85	0.08	0.31	0.5	1	19.35
Depth Percent	219703	61.68	9.92	0	61.2	62.4	63.5	98.7
Table Percent	219703	57.75	9.96	0	57	58	60	94
Meas Length	219703	5.55	1.76	0	4.35	5.06	6.35	93.66
Meas Width	219703	5.14	1.37	0	4.31	4.8	5.7	62.3
Meas Depth	219703	3.29	2.05	0	2.68	3.03	3.63	76.3
Total Sales Price	219703	6.91	2.6	2	9.58	1.97	5.21	1.45

Table 3: Proportions for Diamond Dataset Categorical Variables

Cut

Category	*Frequency*	*Proportion (%)*
Round	158316	72%
Oval	13857	6.3%
Emerald	11091	5.1%
Pear	9860	4.5%
Princess	7050	3.2%
Radiant	5630	2.6%
Heart	4774	2.2%
Cushion Modified	3984	1.8%
Marquise	2916	1.3%
Asscher	1696	0.8%
Cushion	529	0.2%
TOTAL	**219703**	**100%**

Color

Category	*Frequency*	*Proportion (%)*
E	33103	15.1%
F	31566	14.4%
D	30873	14.1%
G	29184	13.3%
H	26073	11.9%

I	22364	10.2%
J	16898	7.6%
K	11750	5.3%
Unknown	9162	4.2%
L	5683	2.5%
M	3047	1.4%
TOTAL	**219703**	**100%**

Clarity

Category	*Frequency*	*Proportion (%)*
SI1	38627	17.6%
VS2	38173	17.4%
VS1	36956	16.8%
SI2	31105	14.2%
VVS2	28985	13.2%
VVS1	27877	12.6%
IF	9974	4.5%
I1	6961	3.2%
I2	944	0.4%
I3	91	0.04%
SI3	10	0.06%
TOTAL	**219703**	**100%**

Cut Quality

Category	*Frequency*	*Proportion (%)*
Excellent	124861	56.8%
Unknown	60607	27.5%
Very Good	34201	15.68%
Good	28	0.017%
Fair	5	0.0025%
Ideal	1	0.0005%
TOTAL	**219703**	**100%**

Lab

Category	*Frequency*	*Proportion (%)*
GIA	200434	91.2%
IGI	15865	7.2%
HRD	3404	1.6%
TOTAL	**219703**	**100%**

Symmetry

Category	*Frequency*	*Proportion (%)*
Excellent	131619	59.9%
Very Good	83143	37.8%
Good	4609	2.19%
Fair	325	0.107%
Poor	7	0.003%
TOTAL	**219703**	**100%**

Polish

Category	*Frequency*	*Proportion (%)*
Excellent	175806	80%
Very Good	42323	19.3%
Good	1565	0.6961%
Fair	7	0.003%
Poor	2	0.0009%
TOTAL	**219703**	**100%**

Eye Clean

Category	*Frequency*	*Proportion (%)*
Unknown	156916	71.4%
Yes	61931	28.2%
Borderline	515	0.28%
E1	300	0.1%
No	41	0.02%
TOTAL	**219703**	**100%**

Culet Size

Category	*Frequency*	*Proportion (%)*
N	131899	60%
Unknown	85740	39%
VS	1345	0.7%
S	476	0.2%
M	163	0.06%
L	58	0.03%
SL	14	0.006%
EL	4	0.002%
VL	4	0.002%
TOTAL	**219703**	**100%**

Culet Condition

Category	*Frequency*	*Proportion (%)*
Unknown	204384	93%
Pointed	15293	6.99%
Chipped	18	0.007%
Abraded	8	0.003%
TOTAL	**219703**	**100%**

Girdle Min

Category	*Frequency*	*Proportion (%)*
Unknown	83432	37.9%
M	74421	33.8%
STK	26335	11.9%
TN	16744	7.6%
TK	10353	4.7%
VTK	4471	2.2%
XTK	1981	0.9%
VTN	1650	0.8%
XTN	292	0.19%
STN	24	0.01%
TOTAL	**219703**	**100%**

Girdle Max

Category	*Frequency*	*Proportion (%)*
Unknown	84295	38.4%
STK	70440	32.1%
TK	25186	11.5%
M	17977	8.2%
VTK	12638	5.8%
XTK	7647	3.5%
TN	1363	0.425%
VTN	111	0.05%
XTN	34	0.02%
STN	12	0.005%
TOTAL	**219703**	**100%**

Fluor Color

Category	*Frequency*	*Proportion (%)*
Unknown	203977	92.8%
Blue	15219	6.9%
Yellow	400	0.25%
Green	55	0.03%
White	42	0.015%
Orange	10	0.005%
TOTAL	**219703**	**100%**

Fluor Intensity

Category	*Frequency*	*Proportion (%)*
None	143491	65.3%
Faint	38302	17.5%
Medium	20705	9.43%
Strong	13243	6%
Very Slight	2729	1.2%
Very Strong	1093	0.5%
Unknown	128	0.065%
Slight	12	0.005%
TOTAL	**219703**	**100%**

Fancy Color Dominant Color

Category	*Frequency*	*Proportion (%)*
Unknown	210539	95.8%
Yellow	6487	2.95%
Pink	1369	0.62%
Brown	531	0.27%
Green	302	0.14%
Orange	271	0.12%
Purple	76	0.034%
Gray	66	0.03%
Blue	38	0.025%
Chameleon	12	0.005%
Black	6	0.003%
Red	4	0.0021%
Other	2	0.0009%
TOTAL	**219703**	**100%**

Fancy Color Secondary Color

Category	*Frequency*	*Proportion (%)*
Unknown	218641	99.5%
Brown	306	0.14%
Yellow	239	0.11%
Orange	155	0.07%
Pink	126	0.06%
Green	105	0.05%
Purple	81	0.04%
Gray	36	0.023%
Blue	11	0.0056%
Violet	2	0.0009%
Red	1	0.0005%
TOTAL	**219703**	**100%**

Fancy Color Overtone

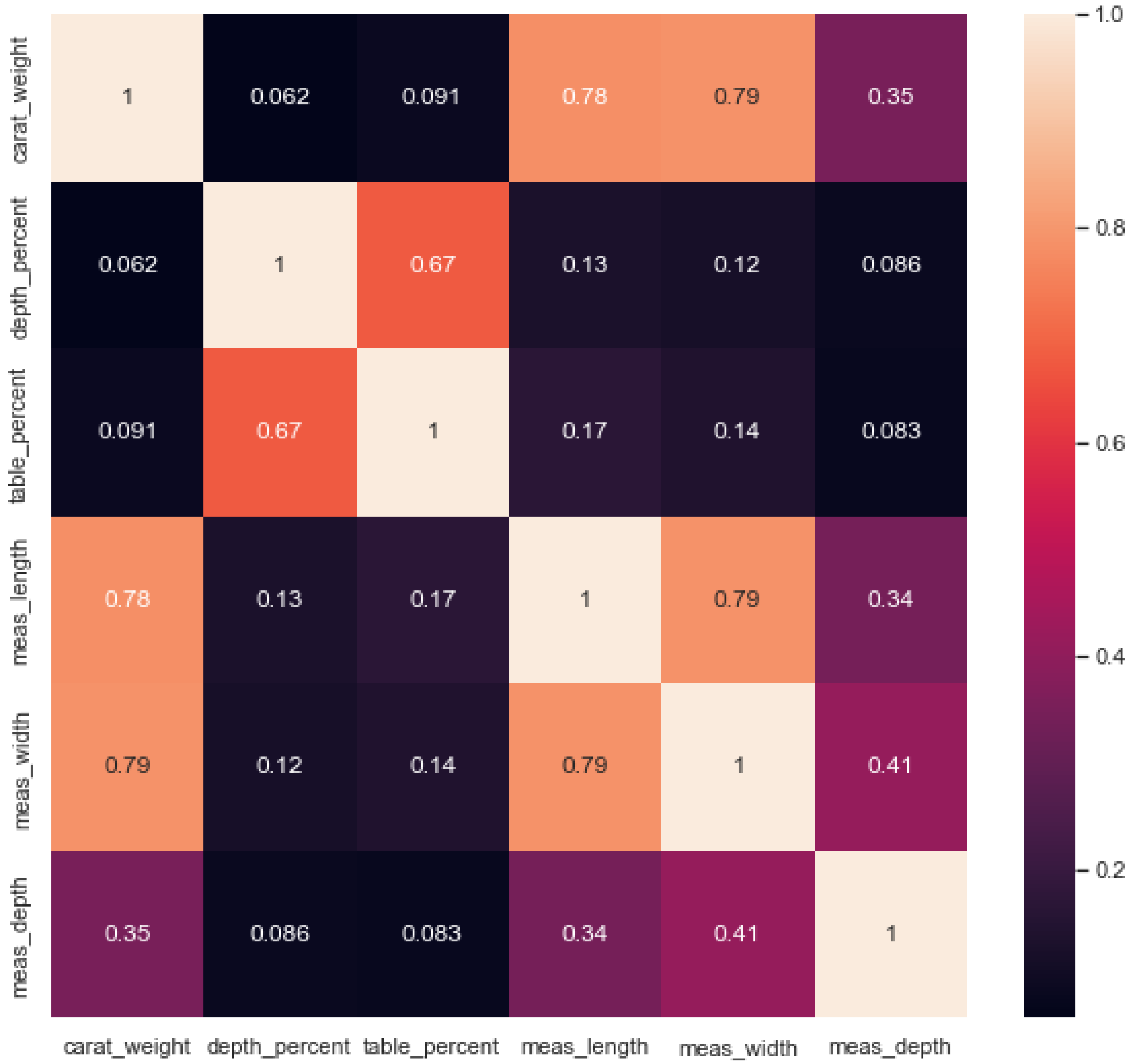
Category	*Frequency*	*Proportion (%)*
Unknown	217665	99.1%
None	1650	0.72%
Brownish	123	0.056%
Yellowish	78	0.034%
Orangey	54	0.025%
Pinkish	51	0.023%
Greenish	47	0.021%
Purplish	34	0.0205%
Grayish	1	0.0005%
TOTAL	**219703**	**100%**

Fancy Color Intensity

Category	*Frequency*	*Proportion (%)*
Unknown	210541	95.8%
Fancy	3447	1.57%
Fancy Intense	1943	0.88%
Fancy Light	1288	0.59%
Fancy Deep	777	0.37%
Fancy Vivid	714	0.34%
Light	318	0.14%
Faint	238	0.11%
Fancy Dark	238	0.11%
Very Light	199	0.09%
TOTAL	**219703**	**100%**

Table 4: Correlation Table/Tables

	Carat Weight	*Depth Percent*	*Table Percent*	*Meas Length*	*Meas Width*	*Meas Depth*
Carat Weight	1.000000	0.061724	0.090697	0.782683	0.788912	0.350719
Depth Percent	0.061724	1.000000	0.673835	0.128791	0.119692	0.086477
Table Percent	0.090697	0.673835	1.000000	0.165742	0.141250	0.082533
Meas Length	0.782683	0.128791	0.165742	1.000000	0.788652	0.342209
Meas Width	0.788912	0.119692	0.141250	0.788652	1.000000	0.412933
Meas Depth	0.350719	0.086477	0.082533	0.342209	0.412933	1.000000



carat_weight
depth_percent
table_percent
meas_length
meas_width
meas_depth
1
0.062
0.091
0.78
0.79
0.35
0.062
1
0.67
0.13
0.12
0.086
0.091
0.67
1
0.17
0.14
0.083
0.78
0.13
0.17
1
0.79
0.34
0.79
0.12
0.14
0.79
1
0.41
0.35
0.086
0.083
0.34
0.41
1
carat_weight
depth_percent
table_percent
meas_length
meas_width
meas_depth
1.0
0.8
0.6
0.4
0.2

IV. DATA SET GRAPHICAL EXPLORATION

A. Distributions

Figure 1: This shows that the cushion cut diamond is the most expensive and round cut is the least expensive.

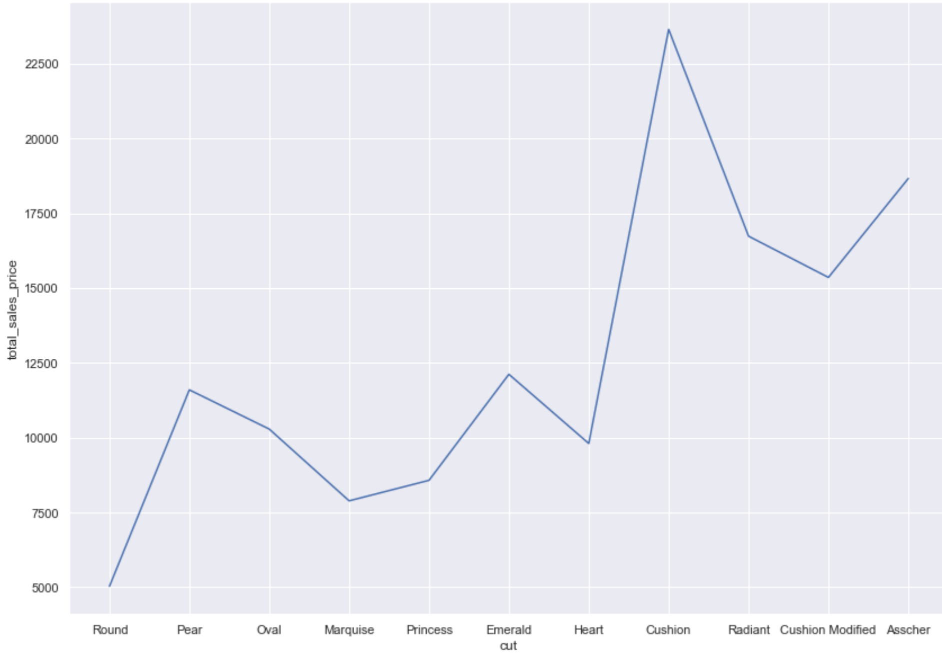


Figure 1: Distribution Comparison of Cut / Total Sales Prices from diamonds dataset (single plot)

Figure 2: This shows that most people purchase a middle of the road 'Good' quality diamond because they are able to get a higher carat weight.

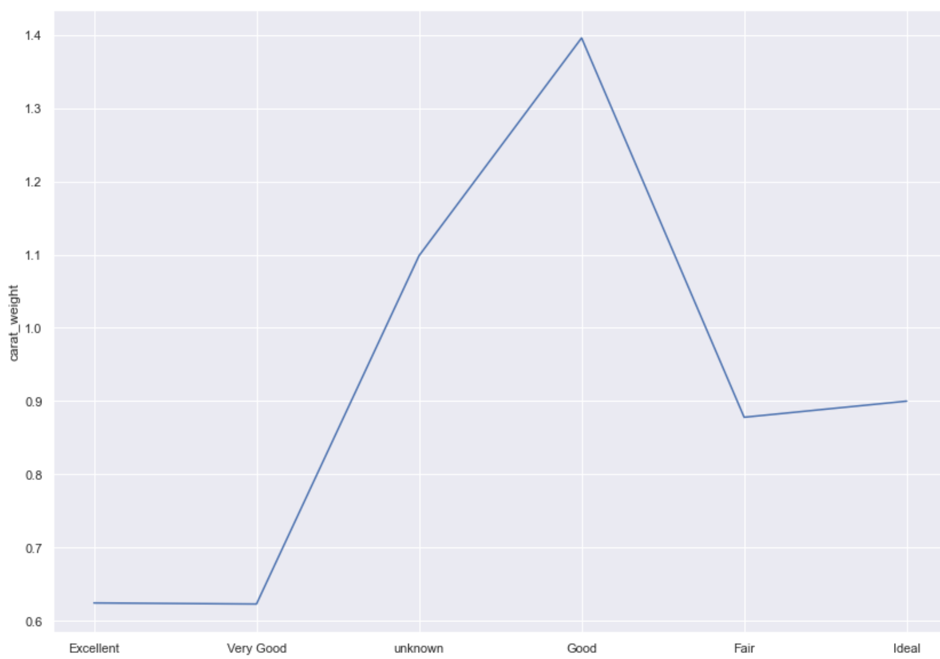


Figure 2: Distribution Comparison of Cut Quality / Carat Weight from diamonds dataset (single plot)

Figure 3: This shows that the HRD certified diamonds are more expensive and make more money than IGI and GIA. This standard is more common than IGI and GIA. GIA, or natural diamonds, make slightly more money and are a little more expensive than IGI, or lab grown diamonds.

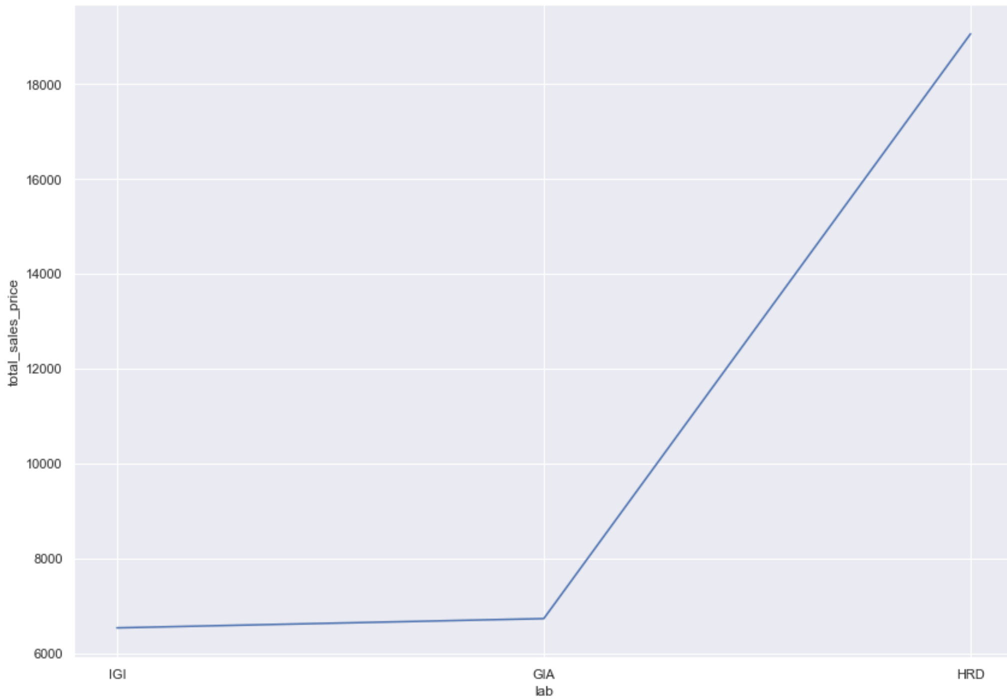


Figure 3: Distribution Comparison of Lab / Total Sales Price from diamonds dataset (single plot)

B. ScatterPlots / Pairwise Plots (continuous variables)

Figure 4: The average table percent of a diamond is between 45 and 80 percent which is average for a diamond.

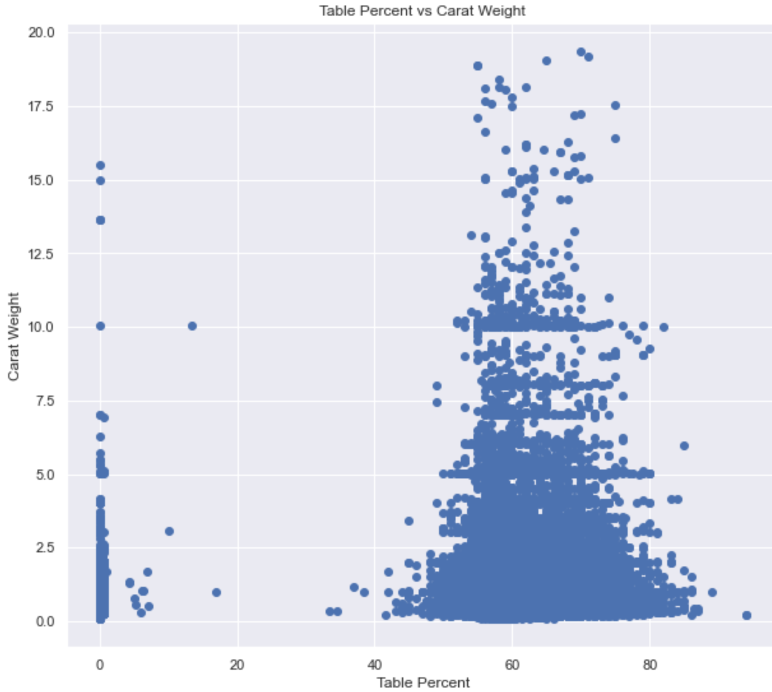


Figure 4: Scatterplot Comparison of Table Percent / Carat Weight from diamonds dataset (single plot)

Figure 5: This shows that the meas length grows exponentially as the carat weight grows bigger, the length also gets bigger. There are a few outliers, but the main cluster shows that these are positively correlated between each other.

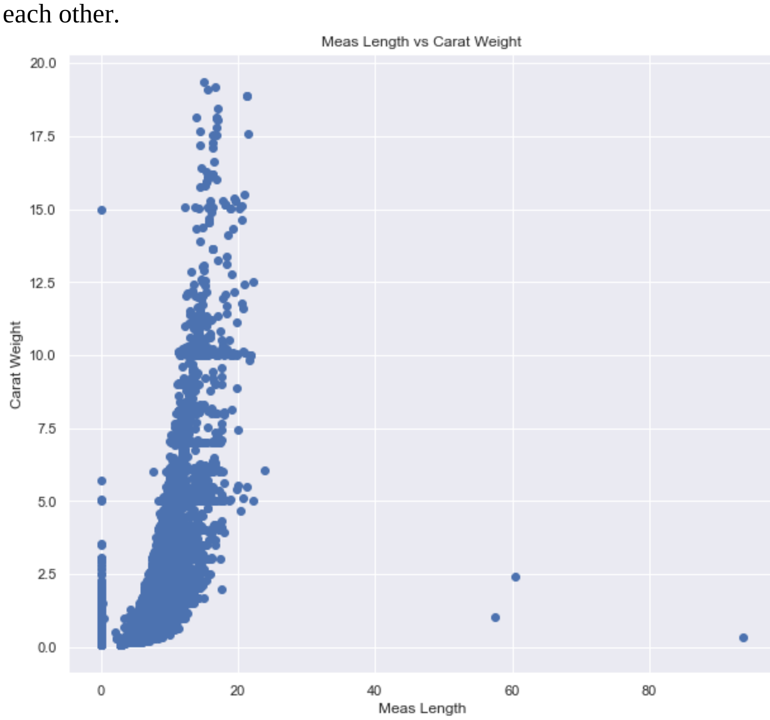


Figure 5: Scatterplot Comparison of Mass Length / Carat Weight from diamonds dataset (single plot)

C. Pie Plot

Figure 6: The pie plot shows that there are more people buying colorless diamonds, G, D, F, E, than those with a little color, like L, M, J, and K.

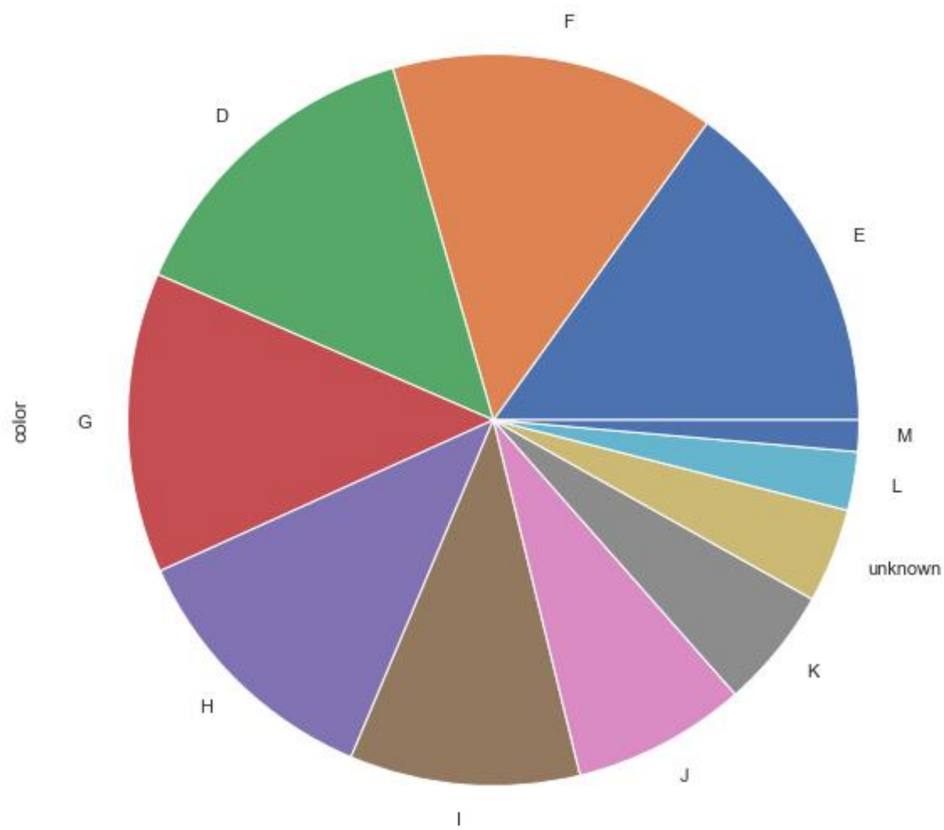


Figure 6: Pie Plot of Color from diamonds dataset (single plot)

D. Bar Charts (categorical variables)

Figure 7: This graph shows that the better cut to a diamond, the more expensive it will be.

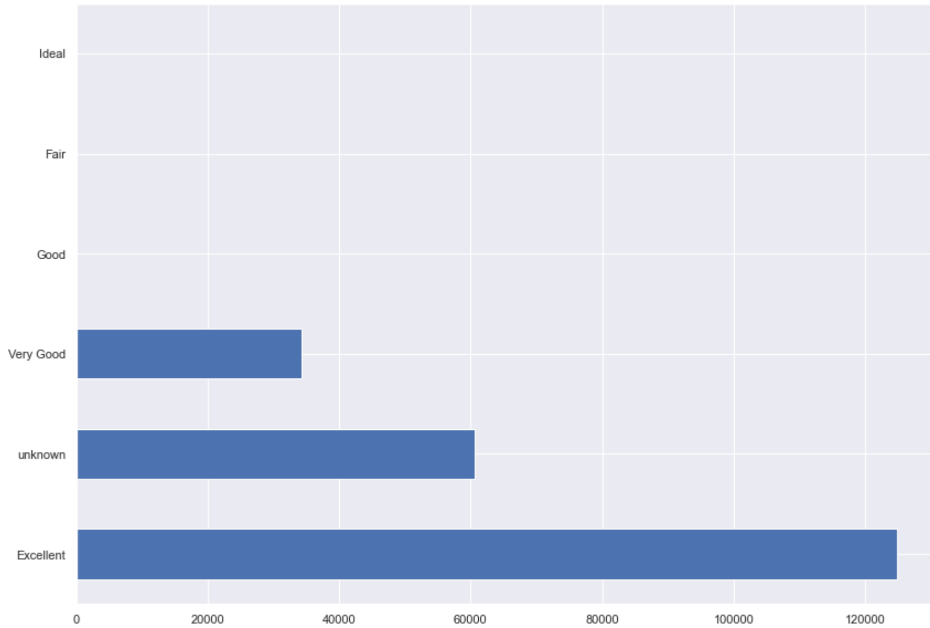


Figure 7: Bar Chart of Cut Quality from diamonds dataset (single plot)

Figure 8: This graph shows that the nicer the color the higher the cost will be for that specific diamond.

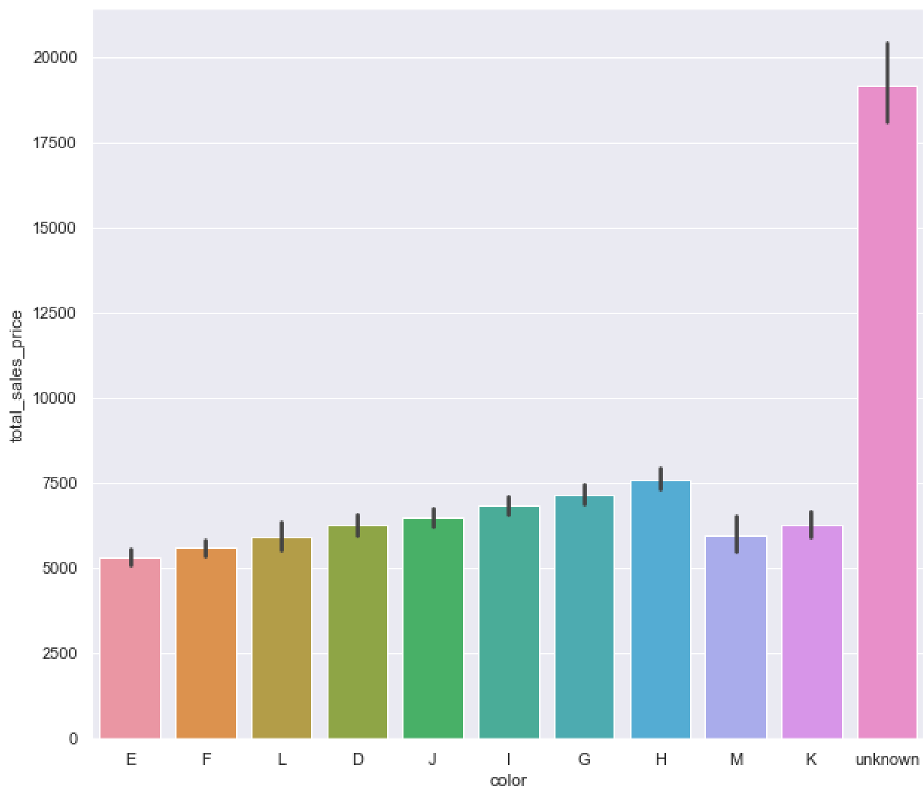


Figure 8: Bar Chart Comparison of Color / Total Sales Price from diamonds dataset (single plot)

E. *Other Plots - don't skimp – there are likely other plots that would be useful that I haven't already specified. Include those in this section.*

Figure 9: This shows that the higher the carat weight the more a diamond will cost.

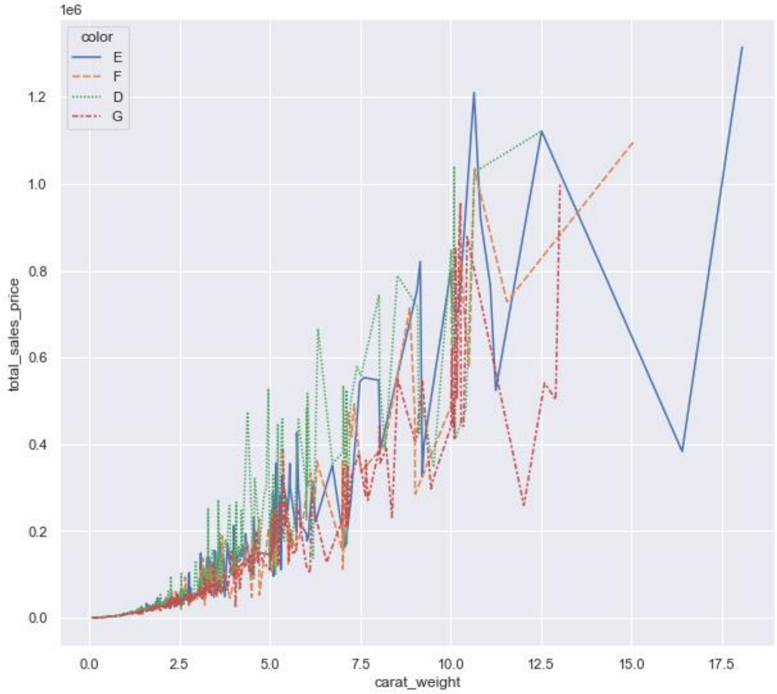


Figure 9: Line Chart Comparison of Carat Weight / Total Sales Price from diamonds dataset (single plot)

Figure 10: This bar plot shows that the diamonds with the highest carat weight are the ones that are in the middle of the color categories where they have a little color but aren't colorless or completely yellow. The best option for the highest carat weight would be colors in the middle like L and M.

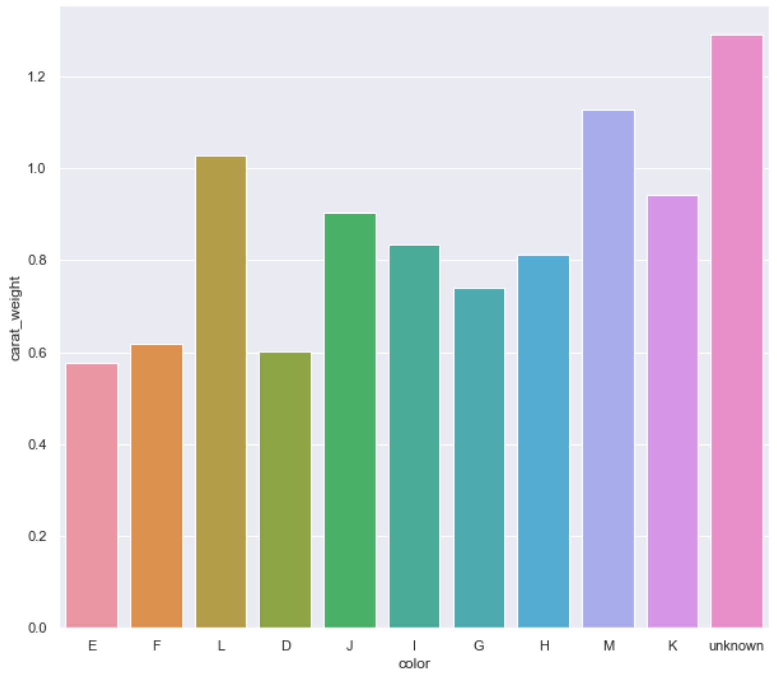


Figure 10: Bar Chart Comparison of Color / Carat Weight from diamonds dataset (single plot)

Figure 11: This goes with the previous graph and shows that a person can buy a middle of the road diamond like L and M for a fraction of the cost with more carat weight. This graph shows that the better color to a diamond the more it will cost, but the previous graph also states that the better the color the less carat weight a person can get. These show that the L and M colors have the best price for the best carat weight.

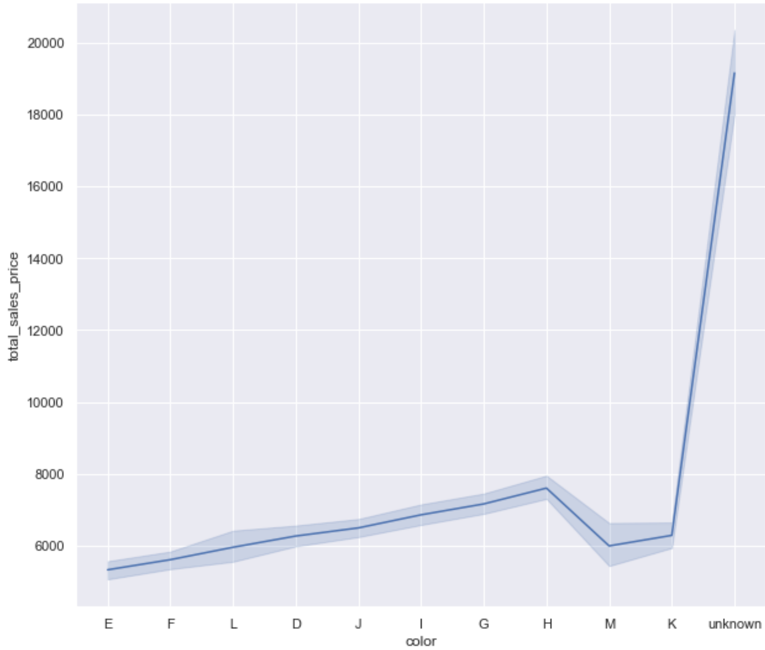


Figure 11: Line Chart Comparison of Color / Total Sales Price from diamonds dataset (single plot)

Figure 12: This shows that a person can buy a middle of the road diamond like L and M for a fraction of the cost with more carat weight and an average cut. This means that the better cut to a diamond, the more it will cost, but the previous graph also states that the better the color the less carat weight a person can get. These show that the L and M colors have the best price for the best carat weight. Along with this, an average cut allows for a person to have a bigger carat weight for a better price.

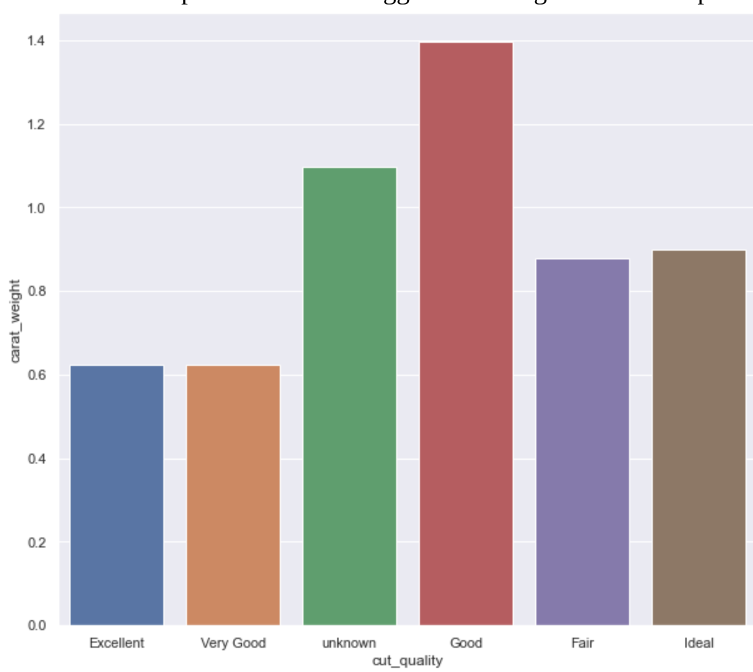


Figure 12: Bar Chart Comparison of Cut Quality / Carat Weight from diamonds dataset (single plot)

V. SUMMARY OF FINDINGS

Our findings show that the better cut quality and color of a diamond along with a high carat weight makes the total sales price higher than a lower cut quality and color of a diamond with a smaller carat weight. From our line distribution graph in Figure 2, we found that an average cut quality allowed someone to get a higher carat weight. The correlation matrix showed that ‘meas_length’ had the strongest positive correlation with carat weight while ‘table_percent’ had an extremely weak positive correlation with carat weight. We thought this was an interesting finding since table percent is the length of the top part of the diamond. We originally thought these two would be heavily and positively correlated, but this was not correct. Figure 3 shows that the average table percent is between 45 and 80 percent for all carat weights, but it was not as consistent as the findings in Figure 4. This figure showed an exponential like scatter plot where when ‘meas_length’ grew larger, the carat weight also grew larger. Figure 7 shows that the better cut quality to a diamond, the more expensive the total sales price will be. Figure 8 shows a bar graph of the color of the diamond compared to the total sales price. We took these two figures and determined that the better cut quality and color of a diamond make the total sales price go up.

These were the basic findings in the diamond dataset, but we wanted to determine which of these variables would be the best option to find the best deal on a diamond and its price. Figure 9 shows that the higher the carat weight, the higher the cost. We already knew this from previous graphs, but wanted to make sure this was clearly shown in our findings. Figure 10 shows that colors L and M allow a person to get the biggest carat weight out of all of the options, excluding the unknown category. Figure 11 shows the total sales price based on the color. In this graph, we were looking for one of the lowest or middle of the road sales price. This figure also solidified our findings since colors L and M both had average sales prices. Finally, Figure 12 indicates that a good cut quality diamond allows for a bigger carat weight. Our findings suggest that a good cut quality diamond that is either L or M, will allow a person to get a higher carat weight for a cheaper price. All of these variables fall right in the middle of the color and quality categories. L and M both have a little color but are still slightly colorless and the good cut quality falls in-between poor and excellent. This indicates that an average color and cut will make for a higher carat weight for a great price.