MCDA 5520
PROJECT REPORT

The Professor Proposes

GROUP 6

MOHD NAWAZ HUSSAIN (A00428036)
VIVEKANAND BOOPATHY (A00425792)
BHAGYA SHREE (A00431152)
RISHAB GUPTA (A00429019)
VINAY GOVINDAN (A00429120)
RAVNEET SINGH OBEROI (A00426623)

Table of Contents

Introduction

This report is about building a model that will help the professor purchase a diamond ring for his girlfriend. He assumed purchasing a diamond ring within $2,000 to $4,000 would be a simple process. However, he is confounded with the various factors involved in the price of a diamond ring. The professor retrieved information about the characteristics of diamonds and what they mean to its value. He had his eye on a diamond ring and wanted to know if it was worth the quoted price. In this report, we will build a statistical model that will help the professor determine if the diamond ring is worth the quoted price.

Characteristics of the Diamond

There are several characteristics that determine the price of the diamond. The most noteworthy of these are the 4 C's:

Characteristic	Scale	Description
Carat	Metric	1 carat = 0.2 grams 2 diamonds of 1 carat have a combined price lesser than that of a single 2 carat diamond
Color	D-F | G-I | J-K | L-N | O-S |T-Z	Colorless | Near colorless | Faint yellow | Very light yellow | Light yellow | Yellow
Cut	Poor | Fair |Good | Very good | Excellent | Ideal	
Clarity	FL | IF | VVS1 | VVS2 | VS1 | VS2 | SI1 | SI2 | SI3 | I1 | I2 | I3	Flawless | Internally Flawless | VV few inclusions at 30x | V few inclusions at 30x | Few inclusions at 30x | Several inclusions at 30x | VV few inclusions at 10x | V few inclusions at 10x | Several inclusions at 10x | V few inclusions visible to naked eye | few inclusions visible to naked eye | several inclusions visible to naked eye

Fig 1. Characteristics of the 4 C's

And some of the other characteristics used to determine the diamond's price are:

Characteristic	Scale	Description
Polish	Poor | Fair |Good | Very good | Excellent | Ideal	
Symmetry	Poor | Fair |Good | Very good | Excellent | Ideal	
Certification	AGS | GIA | EGL | IGI | DOW	GIA and AGS are more respected certifications than EGL, DOW & IGI
Wholesaler	1 | 2 | 3	Determines which wholesaler

Fig 2. Other characteristics of a diamond

Problem Statement

The professor returned from his shopping confused over the various factors used in determining the price of the diamond. The professor's eye caught a diamond ring with the following characteristics:

Price	$3,100
Carat	0.9
Cut	Very Good
Color	J
Clarity	SI2
Polish	Good
Symmetry	Very Good
Certification	GIA

Fig 3. The professor's diamond ring attributes

In order to determine the fair price of the diamond, the professor collected data from three wholesalers online. Based on this data, he needed a way to compile the figures he obtained in a meaningful fashion to value the diamond ring.

Analysis

The price of the diamond can be determined using numerous independent variables. Therefore, we have considered going with a multiple linear regression model as a solution to the professor's dilemma. We have opted to carry individual testing on each independent factor and its effect on the price of the diamond. This will help us in selecting the variables for our multiple linear regression model that will help in determining the price of the diamond.

We carried out Univariate Analysis using frequency distribution for each independent variable on the dataset that were categorical in nature. We wanted to understand the distribution of the attributes in each of these variables and their effect on the dataset as a whole. We only ran it on the categorical data, which excludes price and carat. The frequency distribution for each attribute resulted in a non-uniform pattern.

Scatter Plot of Numerical Variables (Carat vs Price)

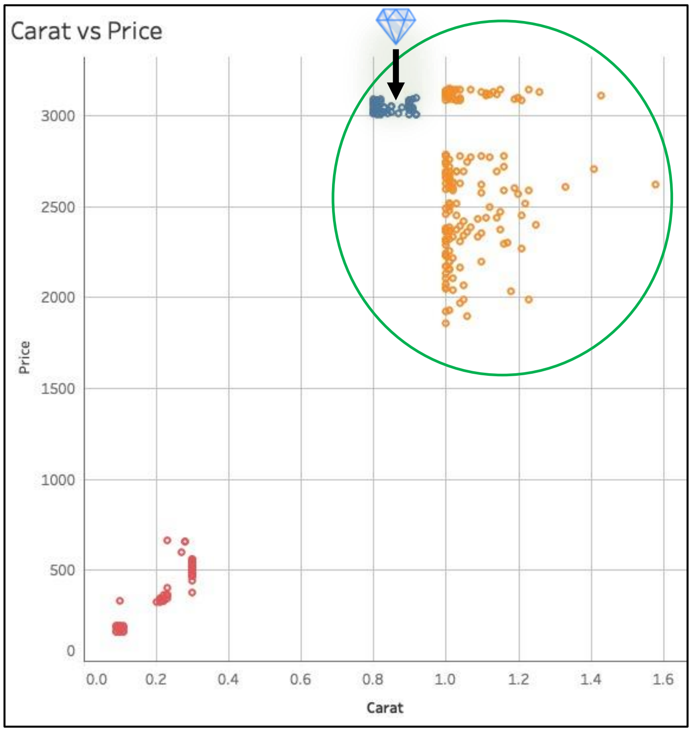


Fig 4. Scatter Plot of all 440 records in Carat vs. Price

After running a scatter plot on Carat vs Price, we see the distribution according to the dataset we have. We realized that the data was spread across three distinct clusters in the graph shown above as first (red), second (blue) and third (orange). The distance across the first cluster versus the second and third was large enough to decide that the first cluster would skew our results. This is because the first cluster will have a significantly different model than the second and third. The diamond the professor is interested in lies in the second cluster.

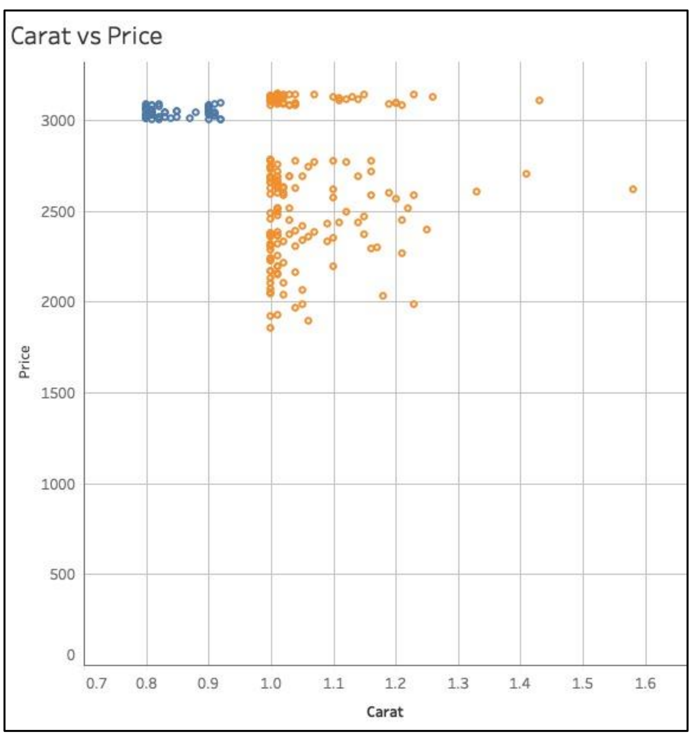


Fig 5. Scatter Plot of 240 records after filtering records in Carat vs. Price

Therefore, for the further analysis, we will only consider the second and third clusters which combined have 240 records in our dataset.

Analysis of the Filtered Data

Carat

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.327[a]	.107	.103	347.868	.107	28.565	1	238	.000

a. Predictors: (Constant), Carat

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	3740.984	185.421		20.176	.000
	Carat	-980.604	183.475	-.327	-5.345	.000

Fig 6. Simple Linear Regression on the Revised Scale of "Carat" vs "Price"

We ran a regression on Carat vs Price to see how price is determined from Carat alone. The model is significant, but the negative coefficient of carat indicates that Carat and Price are inversely proportional. This is the trend we see in the filtered scatter plot of Carat vs Price (see Fig 5). However, when Carat is compared with the other variables in the final model, it gives a high positive coefficient.

Cut

		Cut5			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Fair	56	23.3	23.3	23.3
	Good	34	14.2	14.2	37.5
	Very Good	27	11.3	11.3	48.8
	Excellent	78	32.5	32.5	81.3
	Ideal	45	18.8	18.8	100.0
	Total	240	100.0	100.0	

Fig 7. Frequency distribution of "Cut" after filtering

From the filtered data, we ran the frequency distribution again to have a better understanding regarding the distribution in the second and third cluster (See Fig 5.)

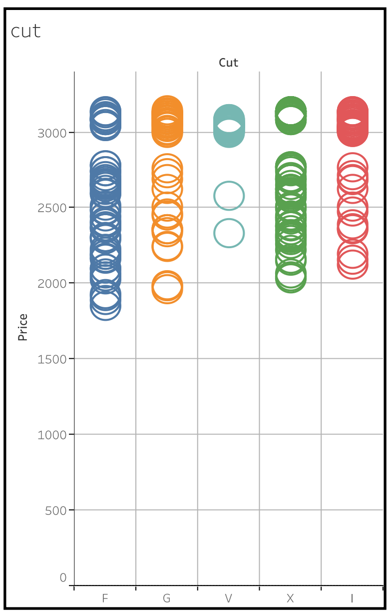


Fig 8. Visual distribution of "Cut" and "Price"

We plotted a scatter plot of "Cut" vs "Price" to visually understand how cut affects the price. We see that from the categories Fair to Ideal, we see a similar range in price (lowest value compared to highest).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.379[a]	.144	.129	342.789	.144	9.881	4	235	.000

a. Predictors: (Constant), Cut5=Ideal, Cut5=Very Good, Cut5=Good, Cut5=Fair

Coefficients[a]

Model		Unstandardized Coefficients: B	Std. Error	Standardized Coefficients: Beta	t	Sig.
1	(Constant)	2739.513	38.813		70.582	.000
	Cut5=Fair	-179.674	60.040	-.207	-2.993	.003
	Cut5=Good	11.252	70.445	.011	.160	.873
	Cut5=Very Good	263.006	76.541	.227	3.436	.001
	Cut5=Ideal	152.021	64.169	.162	2.369	.019

a. Dependent Variable: Price

Fig 9. Simple Linear Regression of "Cut" vs "Price"

We ran a simple Linear Regression on "Cut" vs "Price" to measure the significance and R^2 value of Cut on Price and see the reliability of the model. We see that Cut = "Good" is insignificant in the model.

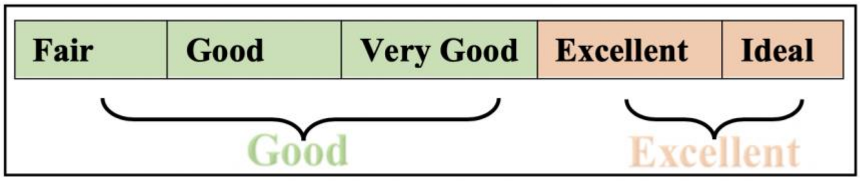


Fig 10. Revised Scale of "Cut"

After multiple attempts trying various combination of clubbing the variables, we concluded this combination will yield us the best results for the final regression model. We ran a frequency distribution for the second time to see if the attributes are equally distributed

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.106[a]	.011	.007	366.083

a. Predictors: (Constant), CutN=Good

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	361558.735	1	361558.735	2.698	.102[b]
	Residual	31896005.1	238	134016.828		
	Total	32257563.9	239			

a. Dependent Variable: Price

b. Predictors: (Constant), CutN=Good

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	2795.130	33.009		84.679	.000
	CutN=Good	-77.651	47.276	-.106	-1.643	.102

a. Dependent Variable: Price

Fig 11. Simple Linear Regression on the Revised Scale of "Cut" vs "Price"

After clubbing the variables together, we ran the regression again to measure the significance and the R^2 values. We see that this model is insignificant yielding a low R^2 value. However, this model gives us a significant and reliable final regression model.

Color

Colour4

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Colourless	65	27.1	27.1	27.1
	Near Colourless	96	40.0	40.0	67.1
	Faint Yellow	67	27.9	27.9	95.0
	Very Light Yellow	12	5.0	5.0	100.0
	Total	240	100.0	100.0	

Fig 12. Frequency distribution of "Color" after filtering

From the filtered data, we ran the frequency distribution again to have a better understanding regarding the distribution in the second and third cluster (See Fig 5.)

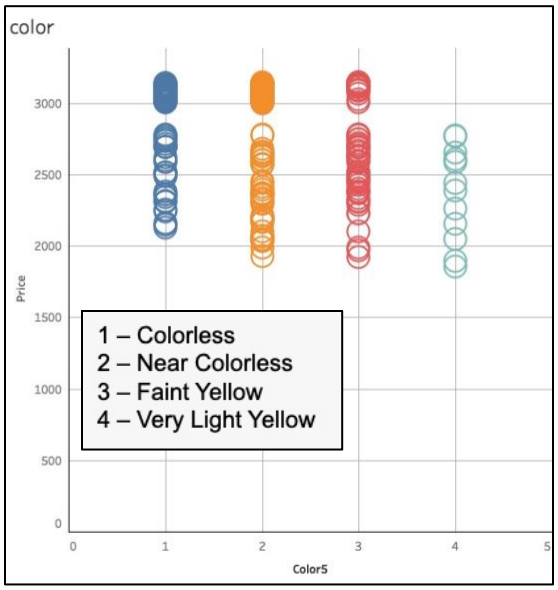


Fig 13. Visual Distribution of "Color" and "Price"

From "Colorless" to "Very Light Yellow", we see the price range decreasing. This follows a natural trend as "Colorless" is more valuable.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.256[a]	.066	.054	357.365	.066	5.528	3	236	.001

a. Predictors: (Constant), Color5=Very Light Yellow, Color5=Colorless, Color5=Faint Yellow

Coefficients[a]

Model		Unstandardized Coefficients: B	Std. Error	Standardized Coefficients: Beta	t	Sig.
1	(Constant)	2782.135	36.473		76.279	.000
	Color5=Colorless	32.480	57.403	.039	.566	.572
	Color5=Faint Yellow	-46.479	56.890	-.057	-.817	.415
	Color5=Very Light Yellow	-413.635	109.420	-.246	-3.780	.000

Fig 14. Simple Linear Regression of "Color" vs "Price"

We ran a simple Linear Regression on "Colour" vs "Price" to measure the significance and R^2 value and see the reliability of the model. We see that Colourless and Faint Yellow are insignificant in the model.

D-F	G-I	J-K	L-N
Colourless	Near colourless	Faint yellow	Very-light yellow

$\underbrace{\text{D-F} \quad \text{G-I}}_{\text{Colourless}} \qquad \underbrace{\text{J-K} \quad \text{L-N}}_{\text{Yellow}}$

After multiple attempts trying various combination of clubbing the variables, we concluded this combination (Fig.15) will yield us the best results for the final regression model. We ran a frequency distribution to see if the attributes are equally distributed.

Fig 15. Revised Scale of "Color"

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.148[a]	.022	.018	364.105

a. Predictors: (Constant), Color5=Yellow

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	705293.813	1	705293.813	5.320	.022[b]
	Residual	31552270.0	238	132572.563		
	Total	32257563.9	239			

a. Dependent Variable: Price

b. Predictors: (Constant), Color5=Yellow

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	2795.248	28.696		97.411	.000		
	Color5=Yellow	-115.362	50.016	-.148	-2.307	.022	1.000	1.000

a. Dependent Variable: Price

Fig 16. Simple Linear Regression on the Revised Scale of "Color" vs "Price"

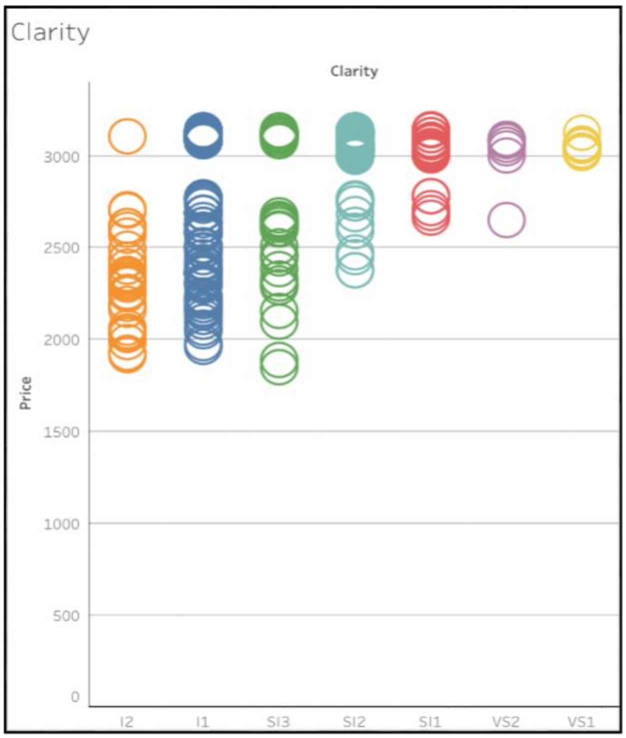
After clubbing the variables together, we ran the regression again to measure the significance and the R^2 values. We see that this model is significant.

Clarity

		Frequency	Percent	Valid Percent	Cumulative Percent
Clarity9					
Valid	Few 30	8	3.3	3.3	3.3
	Several 30	7	2.9	2.9	6.3
	VV few 10	27	11.3	11.3	17.5
	V few 10	65	27.1	27.1	44.6
	Several 10	26	10.8	10.8	55.4
	V few NakedEye	79	32.9	32.9	88.3
	Few NakedEye	28	11.7	11.7	100.0
	Total	240	100.0	100.0	

From the filtered data, we ran the frequency distribution again to have a better understanding regarding the distribution in the second and third cluster (See Fig 5.)

Fig 17. Frequency distribution of "Clarity" after filtering



From "Few inclusions to Naked Eye" to "Flawless", we see the price range increasing. This follows a natural trend as "Flawless" is more valuable.

Fig 18. Visual Distribution of "Clarity" and "Price"

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.635[a]	.403	.388	287.376	.403	26.266	6	233	.000

a. Predictors: (Constant), Clarity9=Few NakedEye, Clarity9=Several 30, Clarity9=Few 30, Clarity9=Several 10, Clarity9=VV few 10, Clarity9=V few 10

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2622.911	32.332		81.124	.000
	Clarity9=Few 30	431.714	106.623	.211	4.049	.000
	Clarity9=Several 30	380.517	113.328	.175	3.358	.001
	Clarity9=VV few 10	376.607	64.063	.325	5.879	.000
	Clarity9=V few 10	368.012	48.124	.446	7.647	.000
	Clarity9=Several 10	-3.527	64.975	-.003	-.054	.957
	Clarity9=Few NakedEye	-280.983	63.205	-.246	-4.446	.000

We ran a simple Linear Regression on "Clarity" vs "Price" to measure the significance and R^2 value and see the reliability of the model. We see that Clarity is an important factor in determining the price of the diamond. However, "Several inclusions at 10x" is highly insignificant.

Fig 19. Simple Linear Regression of "Clarity" vs "Price"

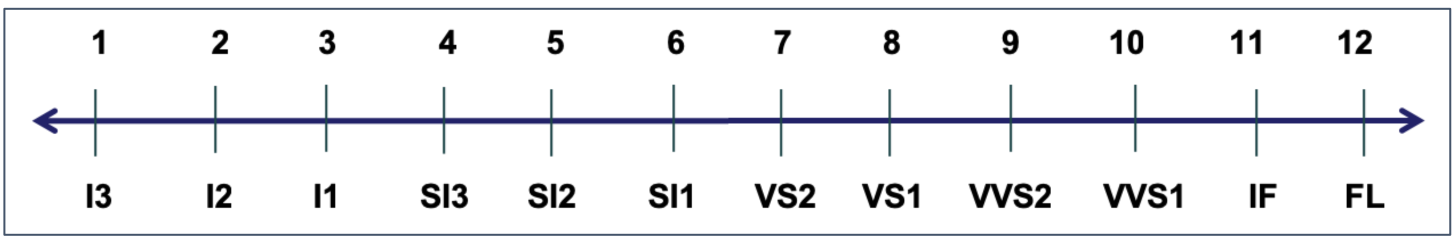


Fig 20. Revised Scale of "Clarity"

After multiple attempts trying various combination of clubbing the variables, we concluded that converting the categorical data of Clarity to metric data by assigning the values from 1 to 10. 1 being "Several Inclusions to the naked eye" and 10 being "Flawless" would give us the best regression result.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.577[a]	.333	.330	300.645	.333	118.881	1	238	.000

a. Predictors: (Constant), Clarity9

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3714.762	89.936		41.305	.000
	Clarity9	-139.864	12.828	-.577	-10.903	.000

After converting the categorical data to metric data, we ran the regression to measure the significance and the R^2 values. We see that this model is significant.

Fig 21. Simple Linear Regression on the Revised Scale of "Clarity" vs "Price"

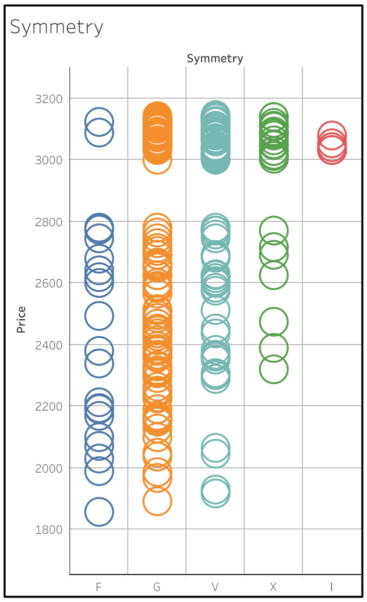
Symmetry

Symmetry5

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Fair	21	8.8	8.8	8.8
	Good	104	43.3	43.3	52.1
	Very Good	84	35.0	35.0	87.1
	Excellent	26	10.8	10.8	97.9
	Ideal	5	2.1	2.1	100.0
	Total	240	100.0	100.0	

From the filtered data, we ran the frequency distribution again to have a better understanding regarding the distribution in the second and third cluster (See Fig 5.)

Fig 22. Frequency Distribution of Symmetry after filtering



From "Fair" to "Ideal", we see the price range increasing. This follows a natural trend as "Ideal" is more valuable.

Fig 23. Visual Distribution of "Symmetry" and "Price"

Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.375[a]	.141	.126	343.393	.141	9.639	4	235	.000

a. Predictors: (Constant), SymmetryNew=Ideal, SymmetryNew=Fair, SymmetryNew=Excellent, SymmetryNew=Good

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	2845.679	37.467		75.951	.000
	SymmetryNew=Fair	-413.393	83.779	-.319	-4.934	.000
	SymmetryNew=Good	-152.602	50.375	-.206	-3.029	.003
	SymmetryNew=Excellent	89.475	77.066	.076	1.161	.247
	SymmetryNew=Ideal	201.721	158.075	.079	1.276	.203

We ran a simple Linear Regression on "Symmetry" vs "Price" to measure the significance and R^2 value and see the reliability of the model. We see that Excellent and Ideal are insignificant in the model.

Fig 24. Simple Linear Regression of "Clarity" vs "Price"

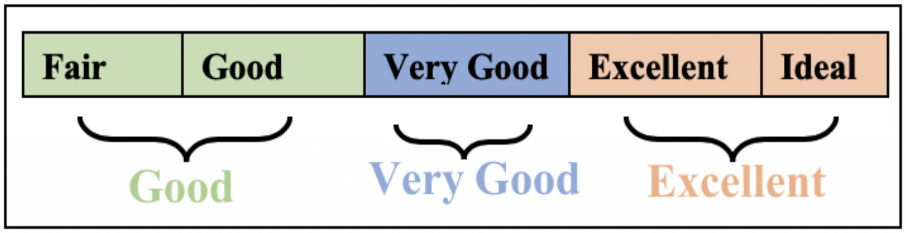


Fig 25. Revised Scale of "Symmetry"

After multiple attempts trying various combination of clubbing the variables, we concluded this combination will yield us the best results for the final regression model. We ran a frequency distribution to see if the attributes are equally distributed.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.320[a]	.102	.095	349.515	.102	13.529	2	237	.000

a. Predictors: (Constant), Symmetry2=Excellent, Symmetry2=Very Good

Coefficients[a]

Model		Unstandardized Coefficients: B	Std. Error	Standardized Coefficients: Beta	t	Sig.
1	(Constant)	2649.264	31.262		84.745	.000
	Symmetry2=Very Good	196.415	49.311	.256	3.983	.000
	Symmetry2=Excellent	303.994	70.128	.278	4.335	.000

Fig 26. Simple Linear Regression on the Revised Scale of "Symmetry" vs "Price"

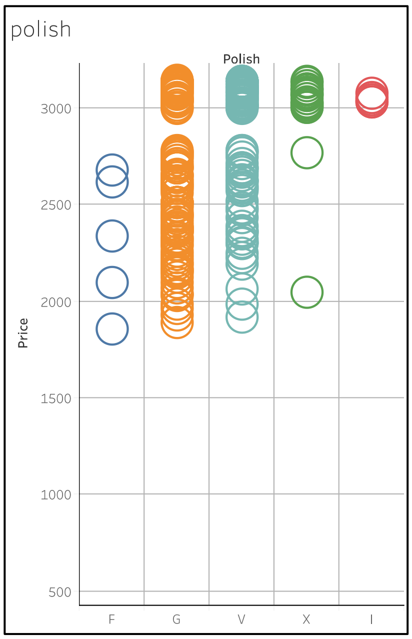
After clubbing the variables together, we ran the regression again to measure the significance and the R^2 values. We see that this model is significant.

Polish

		Polish5			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Fair	5	2.1	2.1	2.1
	Good	112	46.7	46.7	48.8
	Very Good	97	40.4	40.4	89.2
	Excellent	21	8.8	8.8	97.9
	Ideal	5	2.1	2.1	100.0
	Total	240	100.0	100.0	

From the filtered data, we ran the frequency distribution again to have a better understanding regarding the distribution in the second and third cluster (See Fig 5.)

Fig 27. Frequency Distribution of Polish after filtering



From "Fair" to "Ideal", we see the price range increasing. This follows a natural trend as "Ideal" is more valuable.

Fig 28. Visual Distribution of "Polish" and "Price"

Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.386[a]	.149	.134	341.836	.149	10.264	4	235	.000

a. Predictors: (Constant), Polish5=Ideal, Polish5=Fair, Polish5=Excellent, Polish5=Very Good

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	2643.795	32.300		81.850	.000
	Polish5=Fair	-325.195	156.249	-.127	-2.081	.038
	Polish5=Very Good	199.123	47.413	.267	4.200	.000
	Polish5=Excellent	358.491	81.288	.276	4.410	.000
	Polish5=Ideal	403.605	156.249	.157	2.583	.010

Fig 29. Simple Linear Regression of "Polish" vs "Price"

We ran a simple Linear Regression of "Polish" vs "Price" to measure the significance and R^2 value and see the reliability of the model (See Fig. 29). Even though the model is significant at 5%. However, when these variables are used in the final model, they make it insignificant.

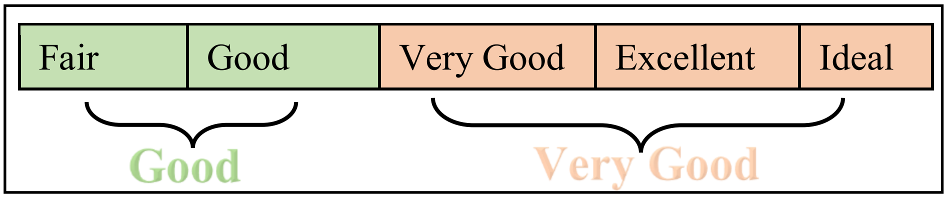


Fig 30. Revised Scale of "Polish"

After multiple attempts trying various combination of clubbing the variables, we concluded this combination (See Fig. 30) will yield us the best results for the final regression model. We ran a frequency distribution to see if the attributes are equally distributed.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.339[a]	.115	.111	346.371	.115	30.874	1	238	.000

a. Predictors: (Constant), Polish3=Very Good

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2629.897	32.022		82.128	.000
	Polish3=Very Good	248.542	44.730	.339	5.556	.000

Fig 31. Simple Linear Regression using revised scale of "Polish" vs "Price"

After clubbing the variables together, we ran the regression again to measure the significance and the R^2 values. We see that this model is significant.

Certification

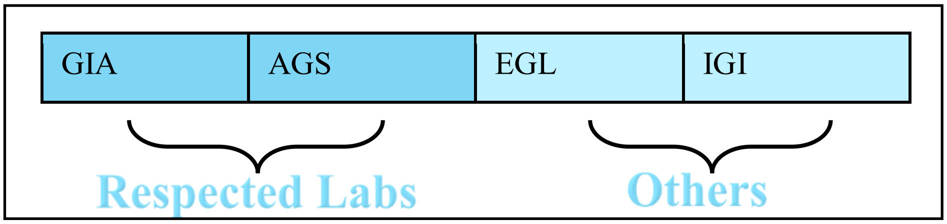


Fig 32. Revised Scale of "Certification"

The requirement in the case clearly stated that the "GIA" and "AGS" are the most respected labs. Some other labs are "EGL", "IGI", etc. Therefore, we clubbed the labs to make them into the following two categories (See Fig 32.).

Certification2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Others	120	50.0	50.0	50.0
	Respected	120	50.0	50.0	100.0
	Total	240	100.0	100.0	

Fig 33. Frequency Distribution of the revised scale of "Certification"

We ran a frequency distribution of the revised scale to gain a better understanding of the distribution

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.231[a]	.054	.050	358.160	.054	13.464	1	238	.000

a. Predictors: (Constant), Certification2=Certified

Coefficients[a]

Model		Unstandardized Coefficients: B	Std. Error	Standardized Coefficients: Beta	t	Sig.
1	(Constant)	2672.442	32.695		81.738	.000
	Certification2=Certified	169.667	46.238	.231	3.669	.000

Fig 34. Simple Linear Regression on the Revised Scale of "Certification" vs "Price"

After clubbing the variables together, we ran the regression again to measure the significance and the R^2 values. We see that this model is significant.

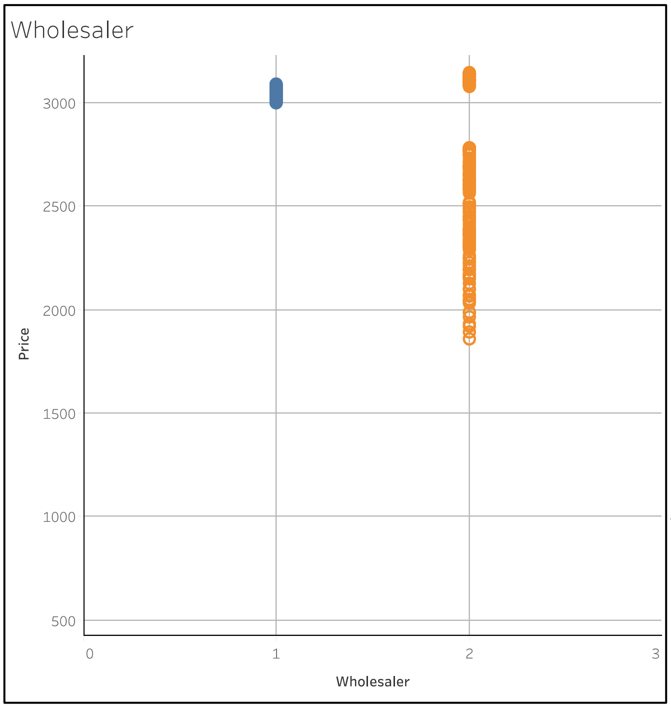
Wholesaler

Wholesaler

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	60	25.0	25.0	25.0
	2	180	75.0	75.0	100.0
	Total	240	100.0	100.0	

Fig 35. Frequency Distribution of Wholesaler after filtering

From the filtered data, we ran the frequency distribution again to have a better understanding regarding the distribution in the second and third cluster (See Fig 12.)



From the scatter plot we see that the wholesaler 1 offers high priced diamonds whereas wholesaler 2 offers diamonds at various price points.

Fig 36. Visual Distribution of "Wholesaler" and "Price"

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.450a	.203	.199	328.724	.203	60.518	1	238	.000

a. Predictors: (Constant), Wholesaler=1.0

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2661.972	24.502		108.645	.000
	Wholesaler=1.0	381.211	49.003	.450	7.779	.000

Fig 37. Simple Linear Regression of "Wholesaler" vs "Price"

We ran a simple Linear Regression of "Wholesaler" vs "Price" to measure the significance and R^2 value and see the reliability of the model. This model is significant at 5%

Initial attempts to the final regression model

Attempt 1

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	79.441	308.314		.258	.797
	Carat	1309.469	241.200	.437	5.429	.000
	ClarityNum	219.439	19.994	.906	10.975	.000
	Wholesaler=2.0	8.975	88.421	.011	.102	.919
	Certification2=Certified	105.881	47.964	.144	2.208	.028
	Polish3=Very Good	112.524	41.147	.153	2.735	.007
	Symmetry2=Good	82.153	67.606	.112	1.215	.226
	Symmetry2=Very Good	93.242	58.762	.121	1.587	.114
	ColorN=Colorless	417.232	56.824	.506	7.343	.000
	ColorN=Near Colorless	245.150	46.468	.328	5.276	.000
	CutN=Excellent	95.234	39.477	.130	2.412	.017

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.726[a]	.527	.507	258.033	.527	25.549	10	229	.000

a. Predictors: (Constant), CutN=Excellent, ColorN=Near Colorless, ClarityNum, Symmetry2=Very Good, Certification2=Certified, Polish3=Very Good, ColorN=Colorless, Carat, Symmetry2=Good, Wholesaler=2.0

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	17010524.3	10	1701052.43	25.549	.000[b]

Fig 38. Attempt #1 of the final regression before revising the scales

From the regression model we see that "Wholesaler" has the lowest coefficient compared to the other variables and also it contributes least towards determining the price and is highly insignificant. Similarly, we see that Symmetry and the constant coefficient are insignificant too. After multiple attempts of running the regressions, we revised our scales to conclude on the aforementioned scales.

Attempt 2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.704[a]	.495	.478	265.471	.495	28.340	8	231	.000

a. Predictors: (Constant), Symmetry2=Good, Color5=Yellow, Certification2=Certified, CutN=Excellent, ClarityNum, Polish3=Very Good, Carat, Symmetry2=Very Good

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	723.794	277.630		2.607	.010
	Color5=Yellow	-288.509	44.535	-.370	-6.478	.000
	Certification2=Certified	86.552	36.767	.118	2.354	.019
	Symmetry2=Very Good	73.557	60.018	.096	1.226	.222
	CutN=Excellent	93.449	38.927	.127	2.401	.017
	Polish3=Very Good	105.030	42.201	.143	2.489	.014
	ClarityNum	200.592	17.360	.828	11.555	.000
	Carat	1090.905	222.762	.364	4.897	.000
	Symmetry2=Good	57.702	68.889	.079	.838	.403

Fig 39. Attempt #2 of the final regression after revising the scales

We removed “Wholesaler” from the model in this attempt. We see that “Symmetry” is insignificant in this model. We tried multiple combinations of each variables to come up with the final regression model

We attempted more combinations to ensure our Final Regression Model has a higher accuracy and significance. In the context of the report, the final regression model is illustrated on the next page.

Final Regression Model

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	880.447	270.458		3.255	.001
	Carat	1104.287	221.210	.369	4.992	.000
	CutN=Excellent	84.404	36.222	.115	2.330	.021
	ClarityNum	197.985	16.942	.817	11.686	.000
	Certification2=Certified	91.559	36.424	.125	2.514	.013
	Color5=Yellow	-285.698	44.295	-.366	-6.450	.000
	Polish3=Good	-99.307	38.278	-.135	-2.594	.010

a. Dependent Variable: Price

Model Summary

Mod(					Change Statistics			df2	Sig. F Change
1	.701[a]	.492	.479	265.207	.492	37.605	6	233	.000

Fig 44. Attempt #2 of the final regression after revising the scales

a. Predictors: (Constant), CutN=Good, Certification2=Certified, ClarityNum, Color5=Yellow, Polish3=Good, Carat

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15869589.4	6	2644931.56	37.605	.000[b]
	Residual	16387974.5	233	70334.654		
	Total	32257563.9	239			

Fig 40. Final regression model

We removed “Symmetry” and “Wholesaler” from the final model as they were insignificant and contributed little comparatively to determine the price of the diamond.

The final model shows all the variables are significant at 5% and we have an R^2 value of 49.2% which tells us that approximately 49% of the diamond prices can be determined from this model.

Regression Equation

$$Y(price) = b_0 + b_1(carat) + b_2(cut) + b_3(color) + b_4(clarity) + b_5(polish) + b_6(symmetry) + b_7(certification)$$

Substituting the values of the final regression model into the regression equation.

VARIABLES	VALUES	VALUE*COEFF	RESULT
Carat	0.9	0.9 * 1104.28	993.6
Cut	Very Good	1 * 84.4	84.4
Color	J	1 * -285.7	-285.7
Clarity	SI2	5 * 197.98	989.9
Polish	Good	1 * -99.3	-99.3
Symmetry	Very Good	1 * 0	0
Certification	GIA	1 * 91.56	91.56
Constant			880.447

Estimated Diamond Price	Cdn$2654.9

Fig 41. Final Regression Model Output

Our regression model predicts that the price of the diamond should be **Cdn$2654.9.**

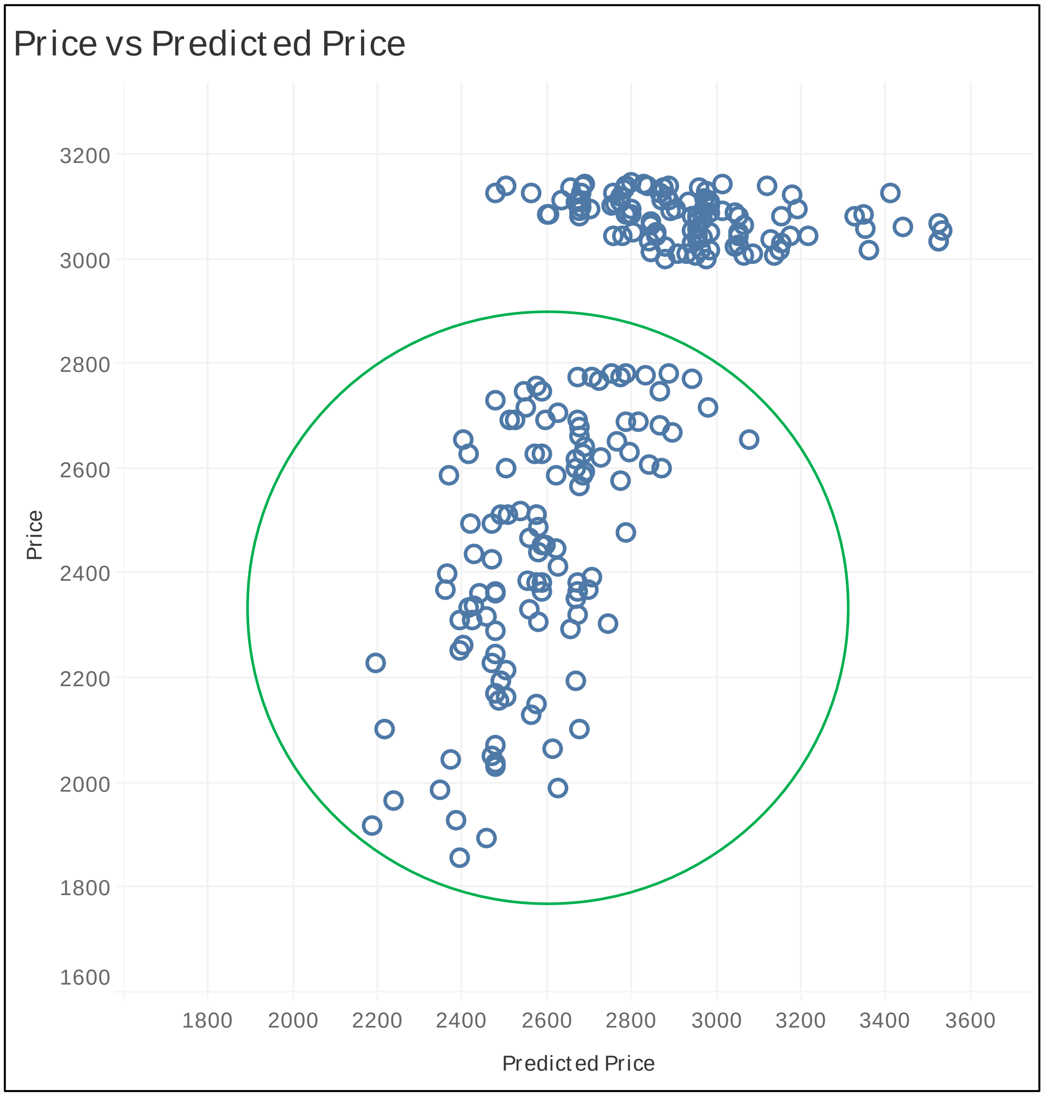
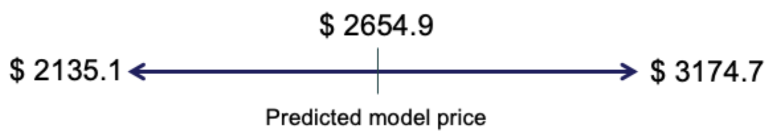


Fig 42. Scatter plot of the Given Price vs the Predicted Price

In the circled region in Fig 42., we see a linearity between the "Predicted Price" and "Price". However, in the region outside the circle, we see a high variance among the values.

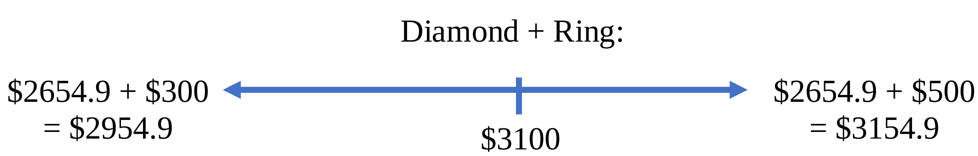
Conclusion

From the Fig.40, the standard error of our Regression Model is 265.2. Therefore, at 95% confidence the price will lie between 2654.9 ± 519.8.



The price that we predicted from the model is only for the diamond and does not take into consideration the *price of the ring*.

Assuming the price of the ring to be between 300$ to 500$ which includes workmanship, guild, finishing, etc. Adding the price of the ring to the price of the diamond gives us:



Since the price quoted to the professor for the diamond ring was $3100, which lies in the range above. Therefore, we suggest the professor to go ahead with the purchase.