Hack.Diversity 2024 Cohort – Tech Dive Project "Diamond.Hackers"

Anmar Abdi, Rhoda Nankabirwa, Amy Boateng, Jose Salerno

Project prompt can be viewed at the bottom of doc

Deliverable 1

Overview

The wholesale diamond dataset spans a period of ten years and serves as a comprehensive repository of diamond sales data for the company. It comprises 407,280 diamonds sold between 2012 and 2023, each represented by 11 attributes: carat, cut, color, depth, table, length (mm), width (mm), height (mm), cost (dollars), clarity, and year of sale.

The goal of the exploratory data analysis is to leverage this dataset to uncover patterns, trends, and correlations that can enhance our understanding of the diamond market. The insights gained from this exploratory analysis will guide our efforts in building predictive models for future diamond prices. By combining numerical and categorical data, we aim to identify factors influencing diamond prices and provide actionable recommendations for strategic decision-making.

Data Quality

Data quality is a critical aspect that underpins the reliability and effectiveness of any analysis or modeling endeavor. In the context of our wholesale diamond dataset, ensuring data quality involved a thorough examination and cleansing process to address potential issues and inaccuracies, which included:

- 2048 blank entries for carat,
- 2037 negative entries for cost/diamond prices,
- 3 zero cost/diamond price,
- Several zero entries for length, width and height of diamonds,
- Column heading nomenclature, among others.
- Finally, after data cleaning, we ended with 403049 entries

Findings

Descriptive statistics

A fundamental aspect of our analysis involves summarizing the distribution of numerical attributes such as carat weight, diamond dimensions (length, width, height), and price. Calculating summary

statistics, including measures such as mean, median, standard deviation, and range, will enable the department to gain insights into the central tendencies and variability of these critical variables. Figure 1 below provides the summary statistics of the cleaned diamond dataset.

	index	carat	depth	table	cost_dollars	length_mm	width_mm	height_mm	year
count	403049.000000	403049.000000	403049.000000	403049.000000	403049.000000	403049.000000	403049.000000	403049.000000	403049.000000
mean	16972.458120	0.797543	61.747950	57.457361	4396.659143	5.730381	5.732401	3.539603	2017.500254
std	9796.286636	0.474654	1.434202	2.240105	4500.786342	1.121121	1.112858	0.709608	3.452029
min	1.000000	0.200000	43.000000	43.000000	304.000000	3.730000	3.680000	1.070000	2012.000000
25%	8491.000000	0.400000	61.000000	56.000000	1053.000000	4.710000	4.720000	2.910000	2015.000000
50%	16974.000000	0.700000	61.800000	57.000000	2676.000000	5.690000	5.710000	3.520000	2018.000000
75%	25454.000000	1.040000	62.500000	59.000000	5983.000000	6.530000	6.530000	4.030000	2021.000000
max	33939.000000	4.130000	79.000000	95.000000	26930.000000	10.140000	10.100000	31.800000	2023.000000

Diamonds description.

The diamonds in the dataset can be classified according to three categorical attributes; cut type, color and clarity.

Cut type

The diamonds were classified into five cut types as indicated in Figure 1 below. Most of the diamonds (40%) fall within the 'Ideal' cut type while 2.9% of the diamonds can be described as fair.

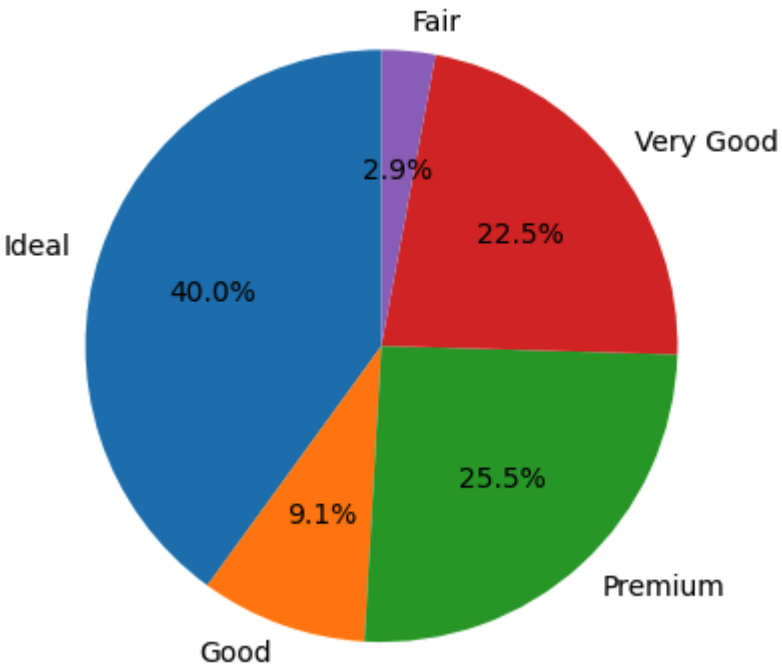


Figure 1: Distribution of diamond according to cut types.

Color

The diamonds were classified into seven colors, see Figure 2 below. Majority (21.1%) of the diamonds belonged to color 'G' while only 5.2% belonged to color 'J'.

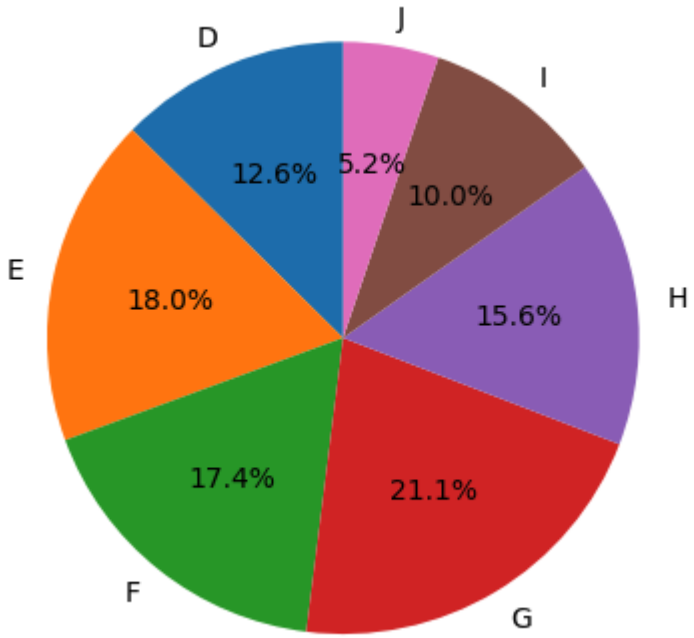


Figure 2: Distribution of diamonds according to color

Clarity

There were eight clarity categories with VS2 being the most common (22.6%) and I1 the least common (1.4%), Figure 3.

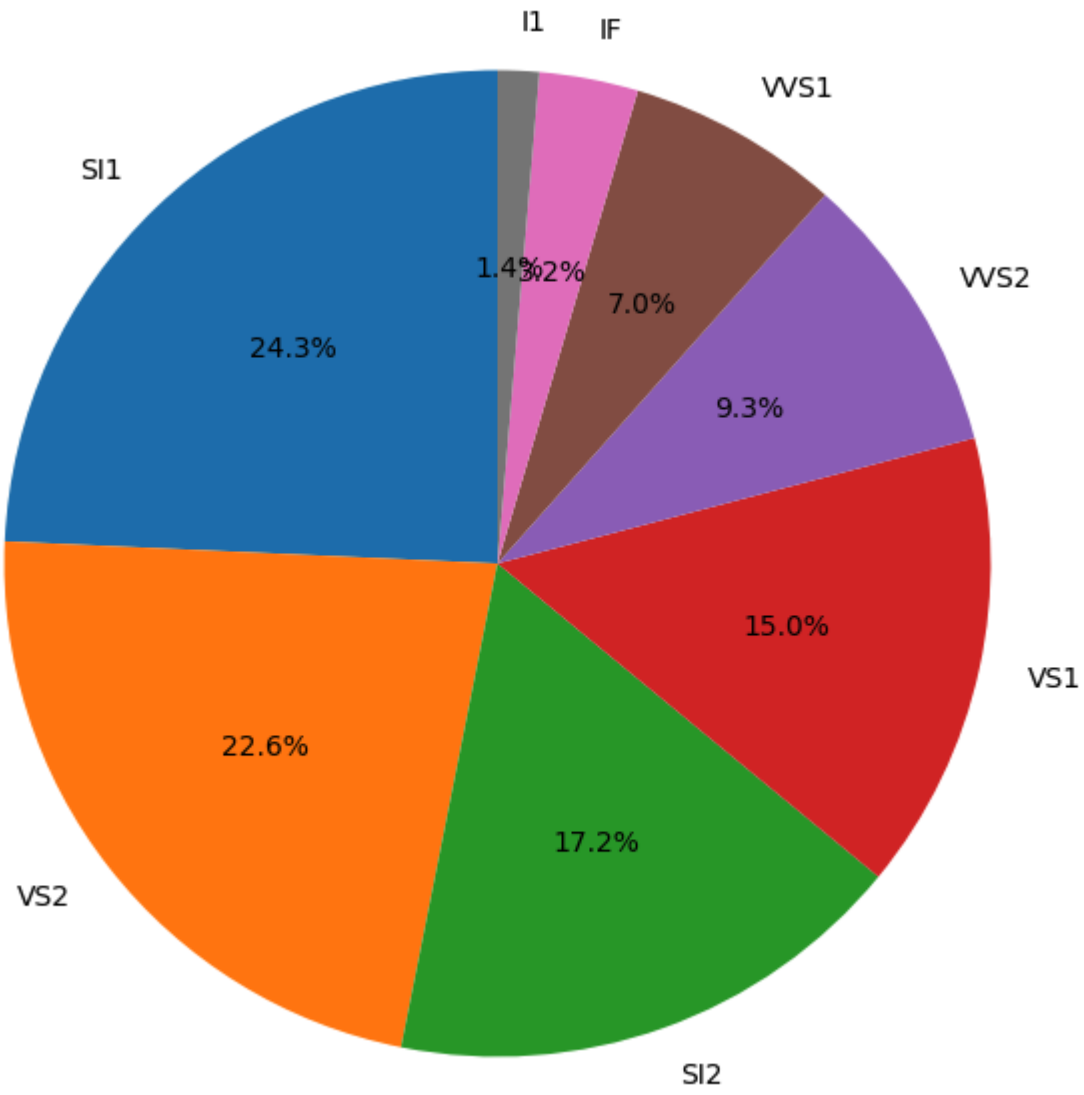


Figure 3: Distribution of diamonds according to clarity

Correlation Analysis

A Pearson correlation matrix was calculated, and the results presented in a heatmap (Figure 4).

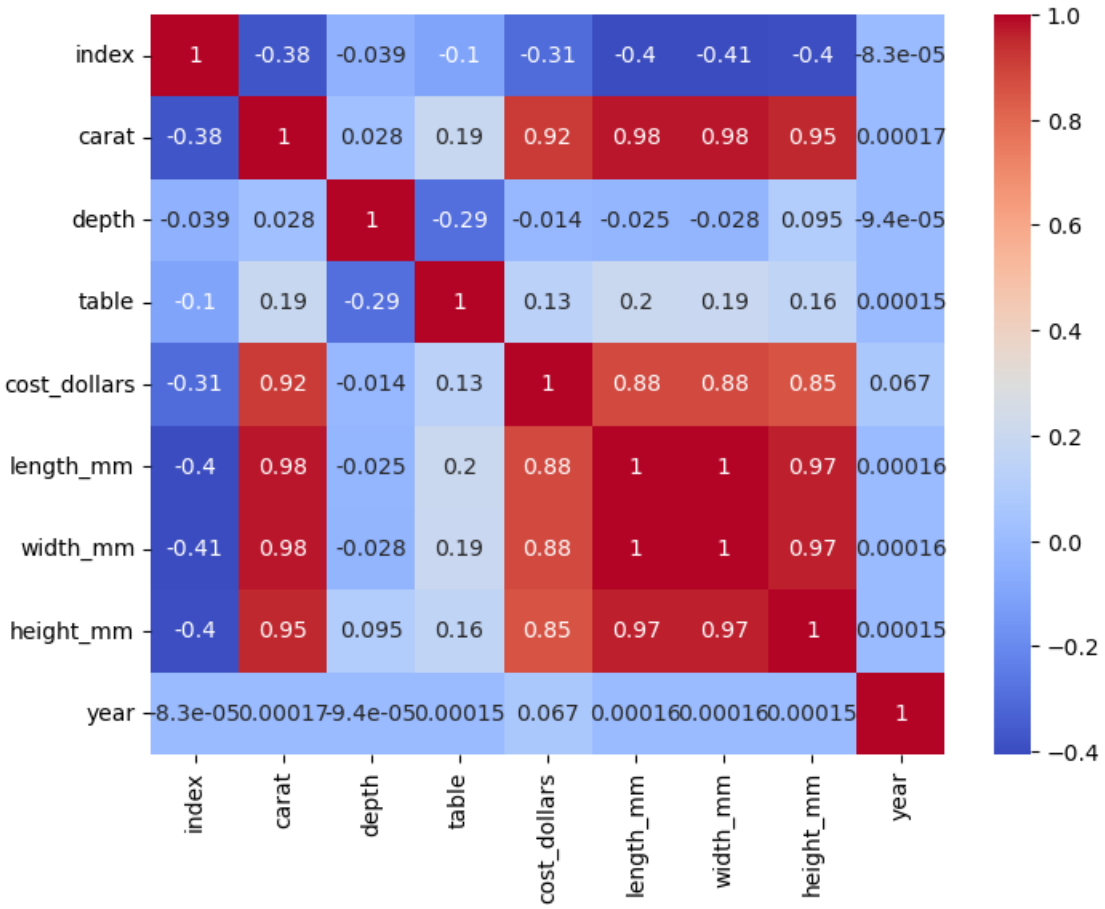


Figure 4: Pearson correlation matrix for numeric variables in the diamond dataset

One notable trend is the positive correlation between carat weight and diamond price “cost(dollars)” indicating that heavier diamonds tend to command higher prices. This observation aligns with the traditional understanding of the diamond market, where larger stones are often associated with increased rarity and value. A strong positive relationship was observed between carat and length, width and height of a diamond. Similarly, there was a strong positive correlation between cost (dollars) and length, width and height of a diamond sold. On the other hand, a negative correlation was observed between the table and depth attributes of a diamond.

Key Trends

Price distribution

Majority of the diamonds sold for less than 5000 dollars. Only a handful sold for above 250,000 dollars, Figure 5, with a price variance of $20,257,077.69 over the ten years. Factors affecting the price of the diamonds included, cut type, color, clarity and year of sale.

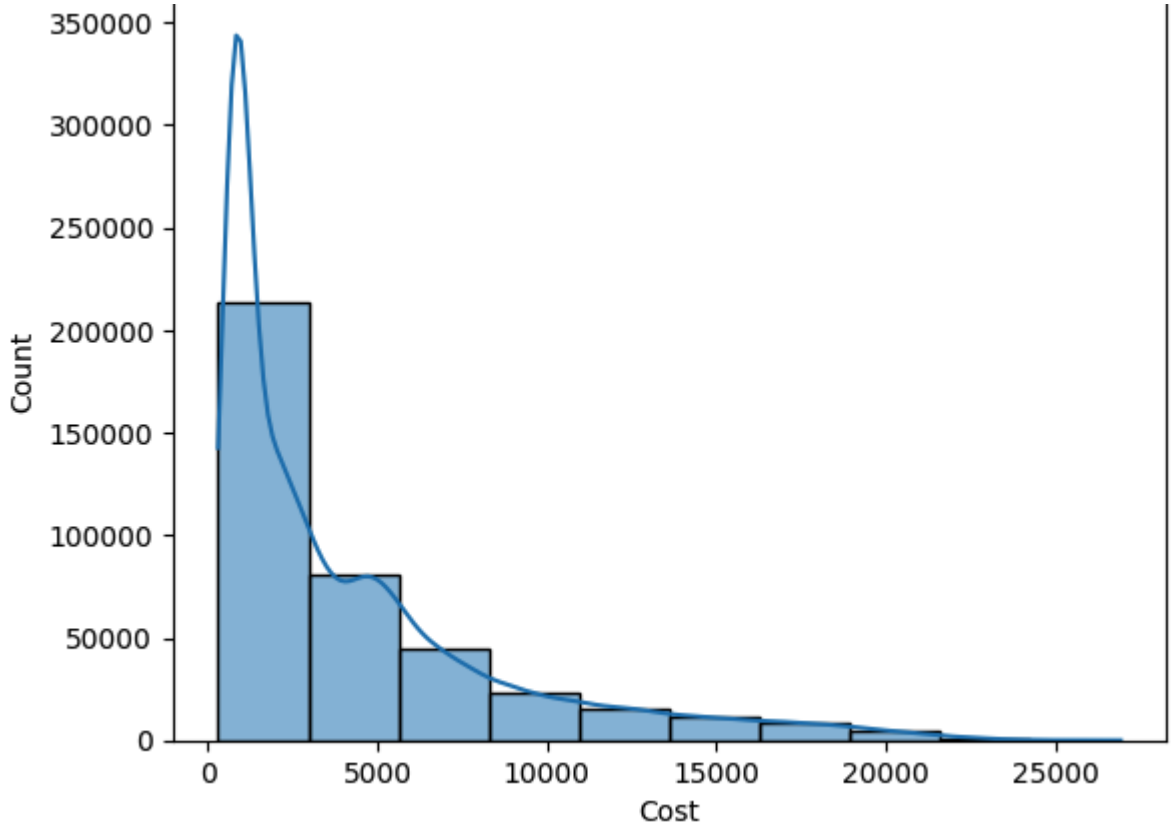


Figure 5: Diamond price distribution over the ten year sale period

The prices varied across cut types, Generally, the 'Premium' cut type cost higher than any other cut type, Figure 6.

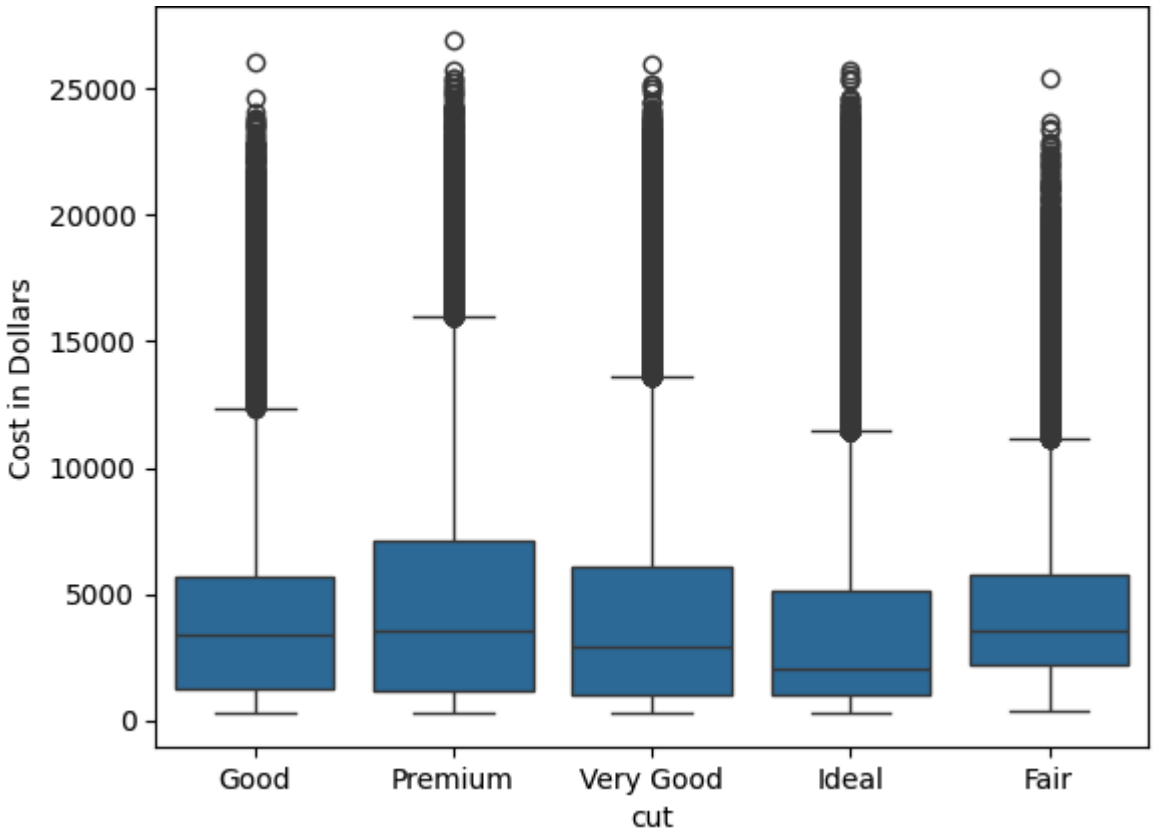


Figure 6: Average price distribution across cut types.

Pricing over the years

Over the years, diamond prices generally increased for each of the cut types, Figure 7.

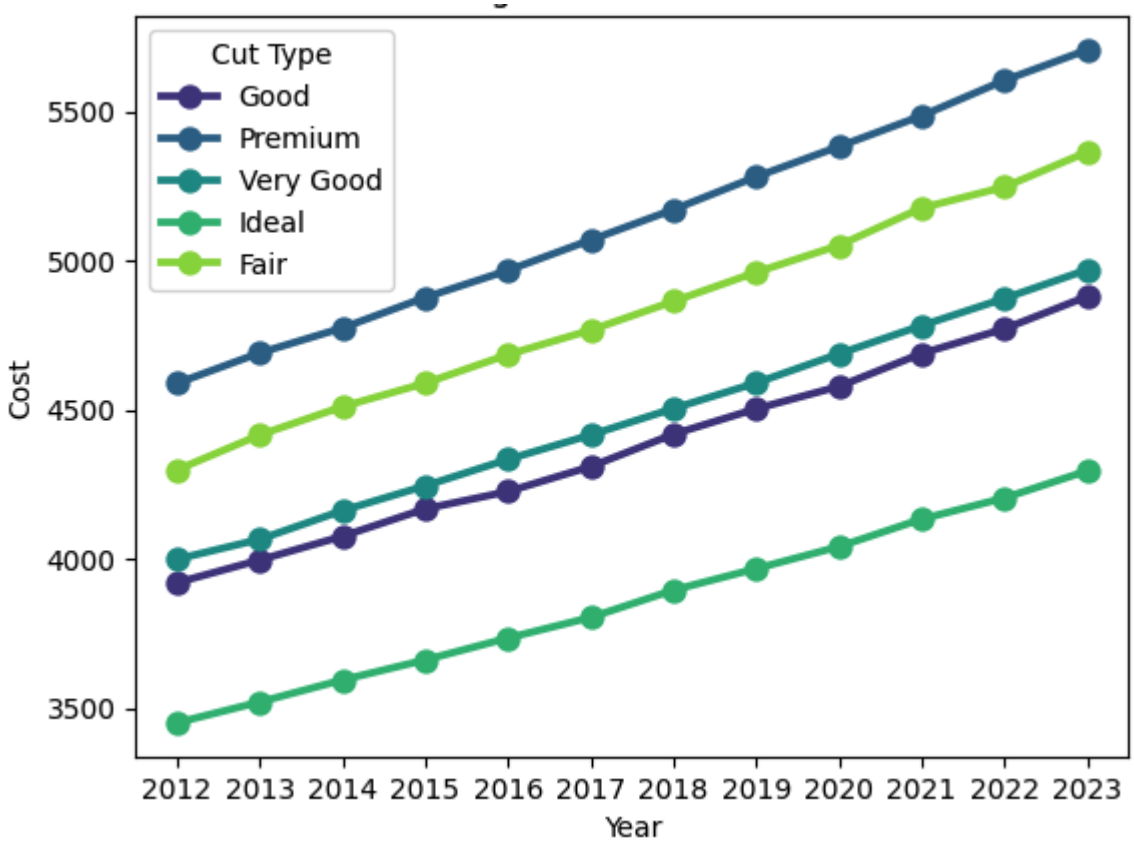


Figure 7: Trend for diamond prices over the ten-year period.

Diamond Clarity

Diamond clarity also affected the price of the diamond with the SI2 category fetching a higher price compared to the rest of the categories, Figure 8.

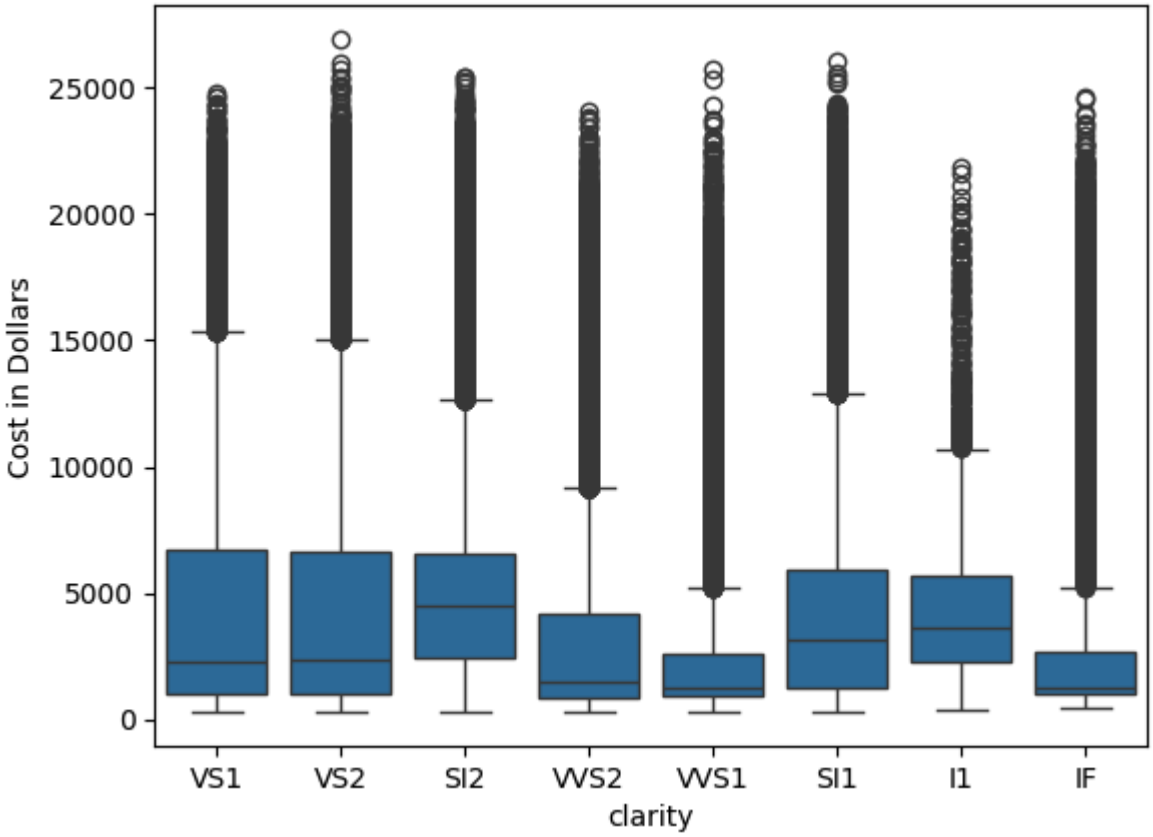


Figure 8: Price variation of diamonds versus clarity types

Carat distribution

Most of the diamonds sold in the ten-year period weighed less than 1 carat, Figure 9.

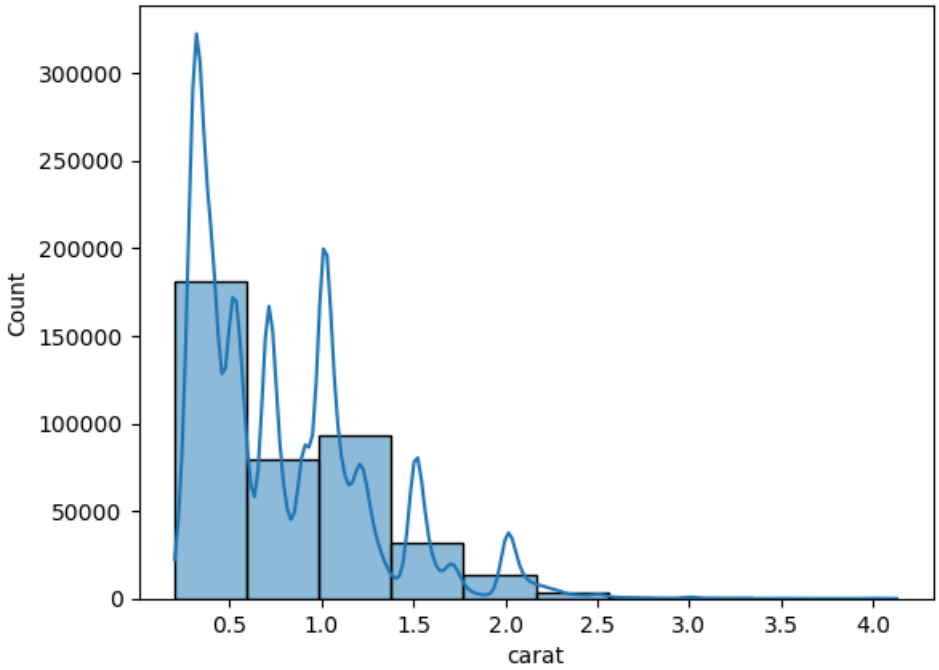


Figure 9: Carat distribution for diamonds sold between 2012 and 2023.

Pairwise Clustering of Diamonds

We conducted pairwise clustering to find natural clusters of diamonds considering attributes perceived to have some type of correlation with price as seen below, Figure 10.

Carat vs Cost: There is a positive correlation between the carat of the diamonds and their cost, confirming the expectation that larger diamonds tend to be more expensive.

Depth vs Cost: This relationship is more complex but there seems to be separation based on depth, but this suggests that this cannot be a strong predictor of cost.

Table vs Cost: This relationship is similar to Depth, indicating it is not a good attribute alone to predict cost.

Cluster 2 seems to have a tighter spread within all 3 attributes compared to cost, which could suggest that this group is more similar in physical characteristics.

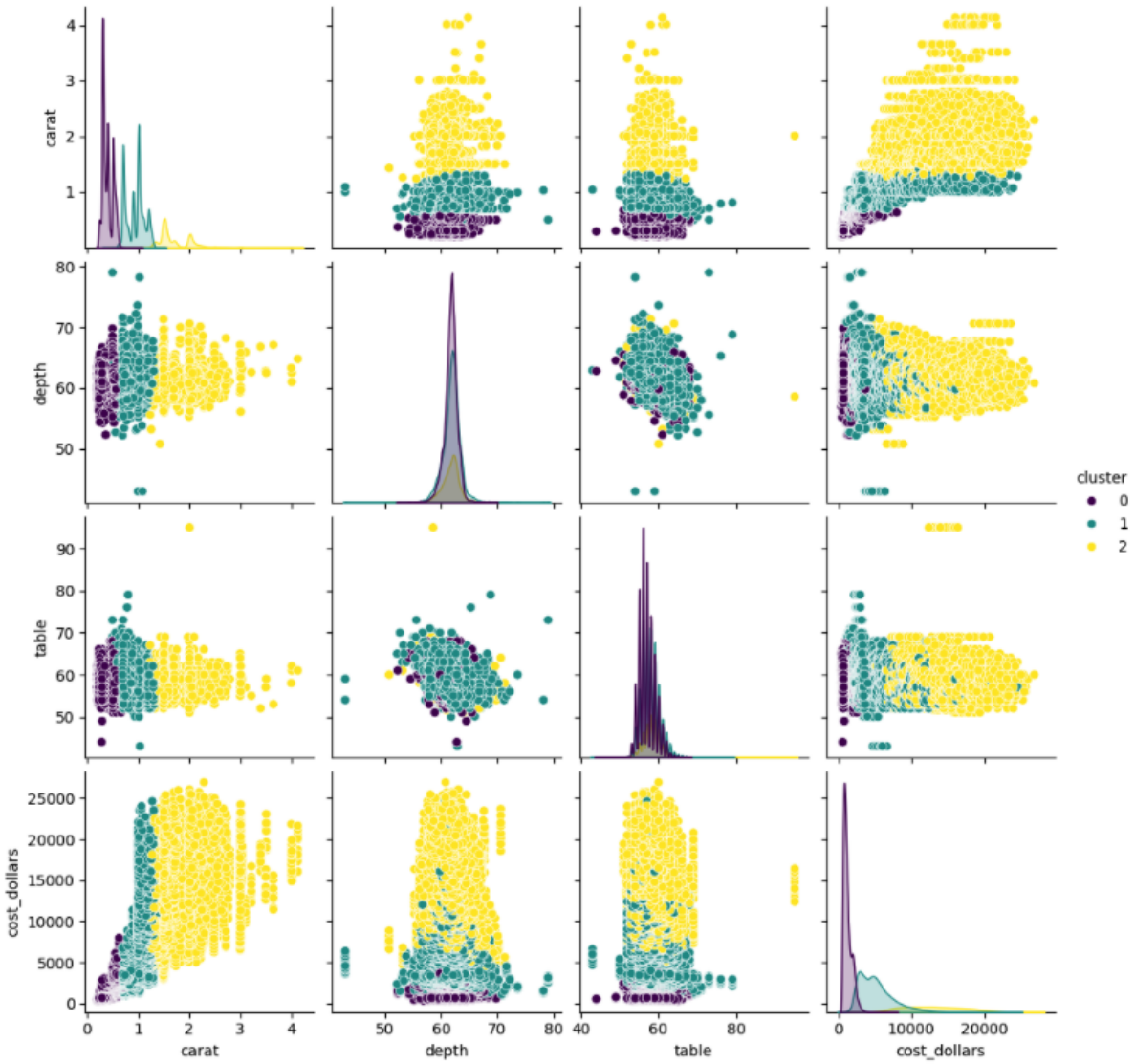


Figure 10: Pairwise plots for Multiple Attributes with natural clusters

Deliverable 2

Regression Model Creation and Execution

This report provides an overview and evaluation of five Regressor models developed for predicting the price of diamonds (target variable) in 2024. The model training and validation process involved several key steps aimed at developing accurate and reliable predictive models.

Data Cleaning

The dataset had been cleaned previously during exploratory data analysis. Consulting with our Alumni lead, Urenna, we realized we missed a few things so we went back and removed the index and year column, and only removed null and 0 columns in the carat column rather than in the dimensions columns.

Encoding categorical variables

Most machine learning algorithms require numerical input data. The categorical variables were encoded to convert them into numerical format suitable for modeling.

Train-test split

The choice of train-validation split depends on several factors, including the size of the dataset, the complexity of the model, and the goal of analysis, for example complex models tend to have a higher risk of overfitting, and in such cases, it's important to set aside a larger portion of the data for testing to ensure that the model generalizes well to new data. Similarly, building a predictive model for deployment in the real world may necessitate simulating real-world performance as closely as possible by allocating a larger portion of the data to the test set.

In practice, common train-test split ratios include 70/30, 80/20, or 90/10, where the first number represents the percentage of data allocated to the training set and the second number represents the percentage allocated to the test set. However, there's no one-size-fits-all answer, and the optimal split ratio may vary depending on the factors mentioned above.

The diamond sales dataset was split into training and validation sets using a 70-30 split ratio. All the models were trained on a dataset containing 10000 diamond samples and ten features, i.e carat, cut, color, depth, table, length, width, height, clarity, year. The training set was used to train the model, the validation set was used to tune hyperparameters and evaluate performance during training, and the test set was used to assess the final model's performance.

Model building

Various machine learning algorithms (Table 1) were employed to train regression models on the training data, including decision tree regression, linear regression, random forest regression, k-nearest neighbors regression, and support vector regression. The respective classification and regression models were built and fitted on the training data by leveraging the respective python packages.

Table 1: Machine learning Algorithms used to train the models

Model	Algorithm used
Decision Tree Regression	Decision Tree
Linear Regression	Ordinary Least Squares
Random Forest Regression	Random Forest
Support Vector Regression	Support Vector Machine
K Nearest Neighbor Regression	kNN
Extreme Gradient Boosting	XGBRegressor

Model Performance, Accuracy, and Limitations

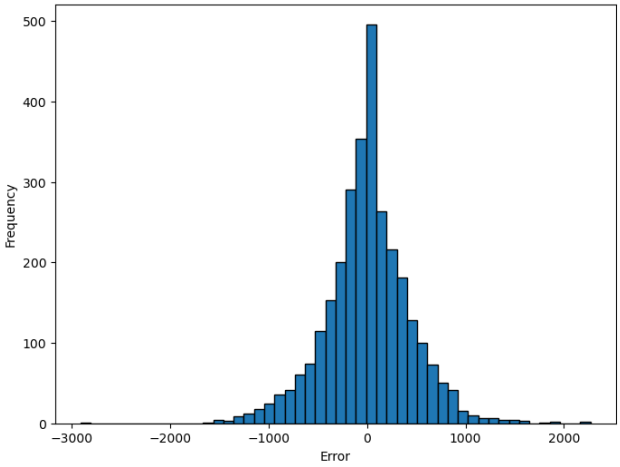
Once the models were trained, they were evaluated on the validation set to assess their performance. Evaluation metrics such as mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R-squared) were used to quantify the models' accuracy and performance. MAE measures the average absolute errors between the actual and predicted values. MSE quantifies the average squared differences between actual and predicted values. Lower MAE and MSE indicate higher accuracy. Additionally, visualization techniques such as histograms of modeling errors were utilized to gain insights into the models' predictive capabilities and identify any potential limitations or areas for improvement.

Table 2: Model Performance, Accuracy, Performance and Limitations

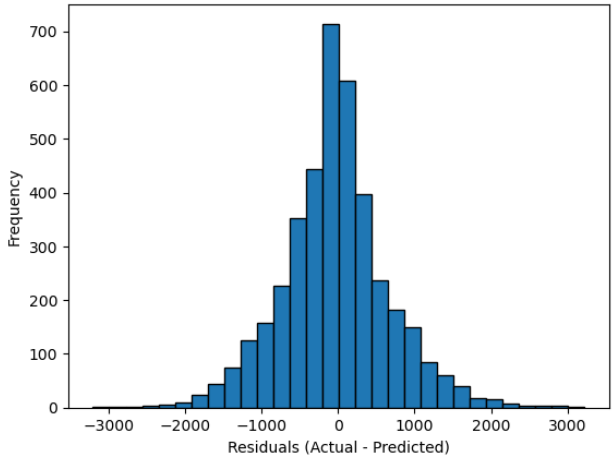
Model	Regression statistics: training			Regression statistics: validation			Model Limitations
	MAE	MSE	R-squared	MAE	MSE	R-squared	
K Nearest Neighbor	305.399	282237.539	0.986	344.639	351275.094	0.983	● Sensitive to outliers in the data ● May not perform well with a large number of predictors ● May not predict well beyond the range of values input in the training data ● Choice of k-value can significantly impact predictions
Random Forest	294.99	258058.667	0.987	336.529	331733.186	0.984	● Requires more memory than other algorithms because it stores multiple trees. This can be a problem if the dataset is large, just like was the case for the diamond wholesale data ● The training time can be longer than other algorithms, especially if the number of trees and the depth of the trees are high ● Random Forest can be less interpretable than a single decision tree because it involves multiple trees. It can be difficult to understand how the algorithm arrived at a particular prediction.

							• Although Random Forest is less prone to overfitting than a single decision tree, it can still overfit the data if the number of trees in the forest is too high or if the trees are too deep
Support Vector	0.0722	0.008	0.98	0.0886	0.0115	0.97	
Linear regression	891.587	1889332	0.90	871.6296	1790658.78	0.91	Model easily affected by multicollinearity Prone to underfitting Sensitive to outliers Non-linearity Constant error variance Simplistic in some cases Linearity constraints
XGBRegressor (Gradient Boosting)	421.89	541929.847	0.973	427.157	556557.6362	0.972	• Very accurate and Efficient with large data sets • Sensitive to outliers • Efficient for diff data types • Effective for complex relationships • Good with overfitting compared to other models especially when dealing with high dimensional data

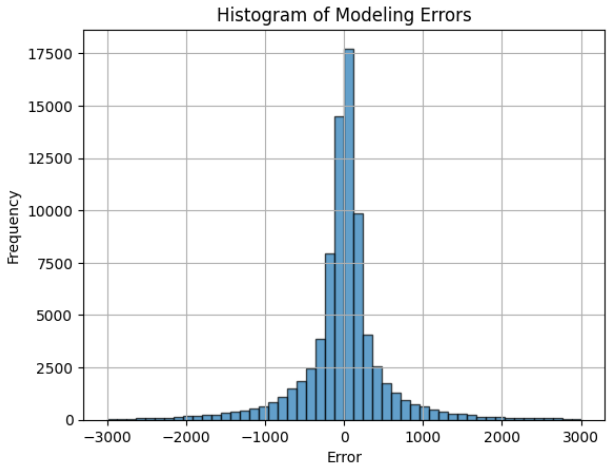
Histogram of modeling errors: Decision Tree Regression



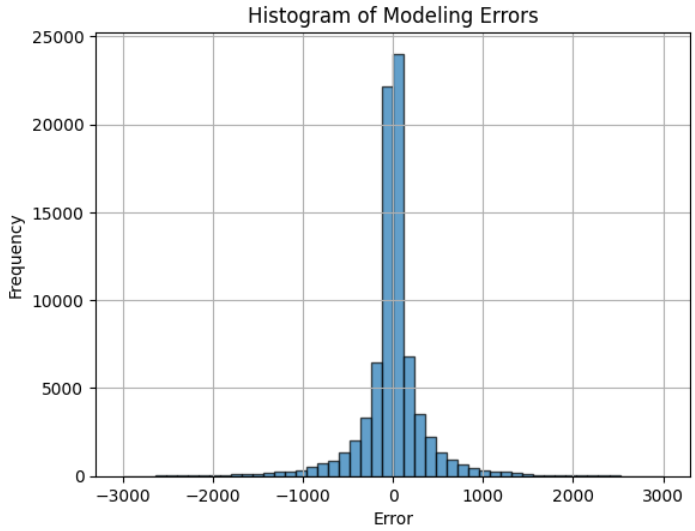
Histogram of modeling errors: k Nearest Neighbor



Histogram of modeling errors: Linear Regression



Histogram for XGB



Histogram for RandomForest

Deliverable 3

Introduction

To predict 2024 diamond prices, we developed several predictive models. The first models were built on data that was trained on 70% of the 10 year diamond sales. Because the dataset contained outliers which could cause errors in our models, a meta-model approach was

taken. The dataset was clustered into 2 sets, Outliers and Normal Data. The models were trained on 70% on outlier data and 70% on normal data respectively and validated using the remaining 30%. If all the company's diamonds are sold in 2024, the total sales will be as indicated in Table 1, based on the selected model. In Table 2, we have mapped out the total sales per cut to get an idea on which cut is the most profitable and should be focused on.

Table 1: Total Predicted Sales and Average Sale Price

Model	**Prediction state**	**Total sales**	**Average diamond price**
Random Forest	Without Hyperparameter tuning	$139,892,669	$4,121
	Based on outlier trained cluster	$390,947,200	$11,518
	Based on cluster without outliers	$139,892,669	$4,121
	Metamodel (combined prediction based on two clusters)	$265,419,934	$7,820
LGBM	Without Hyperparameter tuning	$139,063,307	$4,097
	Based on outlier trained cluster	$388,653,820	$11,451
	Based on cluster without outliers	$139,063,307	$4,097
	Metamodel (combined prediction based on two clusters)	$263,858,563	$7,774
XGB	Without Hyperparameter tuning	$138,996,704	$4,095

	Based on outlier trained cluster	$393,830,048	$11,603
	Based on cluster without outliers	$138,996,704	$4,095
	Metamodel (combined prediction based on two clusters)	$266,413,376	$7,849
KNN	Without Hyperparameter tuning	$137,478,534	$4,050
	Based on outlier trained cluster	$465,417,758	$13,712
	Based on cluster without outliers	$137,478,534	$4,050
	Metamodel (combined prediction based on two clusters)	$301,448,146	$8,881
Decision Tree	Without Hyperparameter tuning	$140,015,783	$4,125
	Based on outlier trained cluster	$389,640,043	$11,480
	Based on cluster without outliers	$140,181,432	$4,130
	Metamodel (combined prediction based on two clusters)	$264,910,738	$7,805
Linear Regression	Without Hyperparameter tuning	$143,733,580	$4,234

	Based on outlier trained cluster	$487,811,444	$14,372
	Based on cluster without outliers	$143,815,339	$4,237
	Metamodel (combined prediction based on two clusters)	$315,813,392	$9,305

Table 2: Predicted Sales for Each Cut Type

Model	**Prediction state**	**Sales for each Cut Type**				
		Fair	**Good**	**Ideal**	**Premium**	**Very good**
Random Forest	Without Hyperparameter tuning	$4,683,530	$12,907,915	$49,281,704	$40,969,363	$32,050,155
	Based on outlier trained cluster	$12,402,131	$36,537,731	$148,141,666	$104,262,180	$89,603,489
	Based on cluster without outliers	$4,683,530	$12,907,915	$49,281,704	$40,969,363	$32,050,155
	Metamodel (combined prediction based on two clusters)	$8,542,830	$24,722,823	$98,711,685	$72,615,772	$60,826,822
LGBM	Without Hyperparameter tuning	$4,599,103	$12,862,761	$49,028,546	$40,645,248	$31,927,647
	Based on outlier trained cluster	$12,346,799	$36,852,956	$146,316,522	$104,012,108	$89,125,433
	Based on cluster without outliers	$4,599,103	$12,862,761	$49,028,546	$40,645,248	$31,927,647

	Metamodel (combined prediction based on two clusters)	$8,472,951	$24,857,858	$97,672,534	$72,328,678	$60,526,540
XGB	Without Hyperparameter tuning	$4,580,829	$12,857,750	$48,973,380	$40,662,768	$31,921,984
	Based on outlier trained cluster	$12,412,139	$37,472,608	$148,127,024	$105,277,800	$90,540,488
	Based on cluster without outliers	$4,580,829	$12,857,750	$48,973,380	$40,662,768	$31,921,984
	Metamodel (combined prediction based on two clusters)	$8,496,484	$25,165,178	$98,550,200	$72,970,280	$61,231,236
KNN	Without Hyperparameter tuning	$4,478,312	$12,409,551	$48,707,819	$40,319,165	$31,563,684
	Based on outlier trained cluster	$9,948,613	$34,863,842	$2,068,077,836	$112,493,276	$101,304,190
	Based on cluster without outliers	$4,478,312	$12,409,551	$48,707,819	$40,319,165	$31,563,684
	Metamodel (combined prediction based on two clusters)	$7,213,462	$23,636,697	$127,757,828	$76,406,220	$66,433,937
Decision Tree	Without Hyperparameter tuning	$4,654,058	$12,919,815	$49,341,901	$49,341,901	$32,104,470
	Based on outlier trained cluster	$12,505,231	$36,878,553	$147,191,947	$104,267,080	$88,797,230
	Based on cluster without outliers	$4,700,033	$12,919,008	$49,398,465	$41,032,096	$32,131,828

	Metamodel (combined prediction based on two clusters)	$8,602,632	$24,898,780	$98,295,206	$72,649,588	$60,464,529
Linear Regression	Without Hyperparameter tuning	$5,137,585	$13,331,399	$49,878,808	$42,602,105	$32,783,681
	Based on outlier trained cluster	$14,749,815	$41,087,048	$191,284,551	$130,419,037	$110,270,992
	Based on cluster without outliers	$51,370	$13,341,313	$49,960,036	$42,637,279	$32,739,687
	Metamodel (combined prediction based on two clusters)	$9,943,418	$27,214,180	$120,622,294	$86,528,158	$71,505,340

Conclusion

Our predictive analysis for 2024 diamond sales highlights significant insights:
The metamodel approach projects an optimal blend of accuracy and insight, with Random Forest and XGBoost models indicating a notable increase in average prices and total sales. The analysis suggests a strong market preference for Ideal and Premium cuts, pointing towards higher value transactions. We suggest the company should focus on these cuts in order to capitalize on potential sales growth in 2024.

In the future, we would like to look at price variation based on year to get an idea on if 2024 would be better performing compared to previous years. This could give the company insight on if there would need to be some type of measures implemented to boost sales and stay on track.

Deliverable 4

Experience Summaries

Anmar - This Tech.Dive project was a solid chance for me to immerse myself in data role. I enjoyed teaming up to tackle our challenges and learned a lot from the experience and my

teammates. I had my first introduction to data visualization platforms and feel like I have a grasp of Tableau now. I'm excited to apply these skills at a company and make a meaningful impact with my work.

Jose - I enjoyed the chance to learn about the unique skills and abilities of all the players in our project. Their diverse range of knowledge and expertise came in handy when completing the project. It was evident that each individual brought a unique and valuable perspective to the table.

Amy- I had a great learning experience from my group and I really want to emphasize the learning portion of it. Being that my expertise is more focused in PowerBI, this project taught me to lean on my group and plenty of clarifying questions from them. With all that being said, I had a great time learning more about python and its many uses whilst being able to encourage my group to keep going when things were looking rough.

Rhoda
I enjoyed getting feedback from the Alumni lead, the Tech.Dive leaders and my mentor regarding the various tasks for the assignment.

Hack.Diversity Data Science
The Diamond Challenge: 2024 Project Prompt

Your new job

Congratulations! You've been hired as a data scientist at a diamond wholesaler. You will be part of the newly established data department. The retailer has collected data on diamond sales over the past 10 years. Your company would like to better understand the diamond market, identify trends, and build a model that can predict diamond prices for next year. You and your team will put your data analysis skills to the test by uncovering insights about the diamond market.

Resources

- Project prompt- This project prompt describes the project and key deliverables for each part of the project.
- Github Page - A github page has been created to host the data sets.
- Jupyter Notebook - introduces the dataset in Python. You can view the notebook on Github, or download and run it for yourself.
- Tech.Dive Leaders - Scott Field & Olga Torres

Milestones

There are Parts 1-4 below, summarized here with due dates for convenience:

1. Provide a cleaned dataset. Document problematic data entries. Make exploratory analysis & summary statistics of the cleaned data. (**Due: 2/2/2024**)
2. Build and assess your price prediction models. (**Due: 2/9/2024**)
3. Use your price prediction models to predict 2024 sales. No model is perfect, and one should expect that the predictions take on a range of plausible values. So include this in your presentation of 2024 sales forecasting, along with possible weaknesses in your model's predictions. (**Due: 2/16/2024**)
4. Build and deploy an executive dashboard to help upper management of your company use and understand your models. (**Due: 3/1/2024**)

The dataset

You will use the dataset wholesale_diamonds.csv, which the company has compiled over the past 10 years. The wholesale diamond dataset includes numerical data (e.g. price, carats, etc) and categorical data (e.g. clarity, quality of cut, etc.). An accompanying Jupyter notebook provides additional information. There are 407280 total diamonds in the dataset covering sales from 2012 to 2023. For each diamond sale we are given 11 attributes:

1. **carat**: the diamonds weight. 1 carat = 200 mg
2. **cut**: rating system from 1 to 5 (poor to ideal)
3. **color**: standardized color code. Each diamond has a color
4. **clarity**: standardized table. The measure of any defects that can impact visual appearance.
5. **depth**: percentage (0 to 100) relating the diamond's depth (top to bottom) with its width
6. **table**: percentage (0 to 100) relating the diamond's overall width to the width of the top part
7. **price**: what the diamond sold for
8. **x**: length in millimeters
9. **y**: width in millimeters
10. **z**: height in millimeters
11. **year**: year of the sale

There are many good internet resources describing these attributes in detail with visuals. You are strongly encouraged to read up on what each of these attributes means.

Your tasks as the new data science department

Before doing any work, please read the entire Project Prompt to get a holistic view of the project. For teamwork, it is suggested that every member of the team be involved with Part 1. While all team members should have some involvement with parts 2 & 3, some members could put more focus on building the price prediction model while others could focus on the initial build of the dashboard (part 4). All members should contribute to the final executive dashboard, final project summary, and presentation of the dashboard at the Project Showcase.

Part 1. Data cleaning, Exploratory analysis & summary statistics

Check the data for correctness. Remove any data entries that appear problematic. Summarize the data that was corrupted and the problems encountered. An example of a problematic data

entry would be a negative sale price, for example. These kinds of data quality problems are commonly found in real-world datasets.

Prepare a summary report for your boss. Include plots, graphs, and other visuals to summarize interesting aspects of the dataset. This will give you and your company insights into the dataset.

Part 1 Deliverables - Summary Report

- Summary of problems with the original dataset
- Summary of attributes cut, color, price, etc.
- Summary correlations (e.g. a correlation matrix) in your dataset
- Summary of any trends discovered in the initial analysis
- Clustering results - use an off-the-shelf clustering algorithm to see if the diamonds in your dataset can be naturally grouped into clusters.
- Save a .csv of your clean dataset

Tip: Start off by thinking about the kinds of figures you want to make. Google around for "exploratory analysis" -- there are many excellent blog posts about the approaches you can take for general datasets. Don't get carried away with making tons of figures. At most 10 well-chosen visuals should be enough to give some insights into your dataset.

Example Questions to Help Guide Your Summary

- What data was corrupted?
- How many corrupted entries?
- How many diamonds of each type of cut are there?
- How much do diamonds cost on average? How does the price change with year, cut, size, etc?
- What's the variance and distribution of prices?
- Generate summary statistics for the attributes.

Part 2. Build Price Prediction Model

The main goal of your work is to build a model to predict the price of diamond sales in 2024. That is, your model should take as input the attributes (carat, cut, color, clarity, depth, table, x, y, z, and year) and predict the price (either as a single number or a probabilistic range of values).

Train a machine-learning algorithm to estimate the price of diamonds based on these attributes. There are many off-the-shelf machine learning algorithms for this task. You should try a few and report on their success.

Some points to keep in mind:

- What are the most important features for predicting the price? Some attributes, such as carat size, should be strongly correlated with price. Other attributes may only be weakly correlated, if at all. Others could be redundant. Because machine learning algorithms work best with a good "feature set", you should report what features you've tried and what works best (and possibly why!).
- Consider building different models for different clusters/kinds of diamonds. Use the cluster identified in Part 1 of the project. A pricing model for large and small carat diamonds could work better than one model for the entire data set, for example.

Tips: The price prediction problem you are solving is known as regression. Some popular machine learning algorithms for regression include

- Decision Tree Regression
- Linear regression models
- Random forest
- Support Vector Regression
- Nearest Neighbors Regression

... and more. Please keep in mind that your dataset contains a mixture of numerical (carats, price, width, etc) and categorical (cut, color, clarity, etc) data types, and your approach to price prediction should keep this in mind.

Tips: If you are looking for a simple way to get started, pick a year (say 2012) and plot carats vs price. You should notice an obvious trend. You could then build a simple model based only on the relationship between the price and the carats. The model won't be very accurate,

but it should give some flavor for how regression models are built and used. You can also use this simple model as you develop code for using your price prediction model.

Part 2 Deliverables - Build Price Prediction Model

- A model that predicts the price of a diamond based on the input attributes,
- Report on the model's accuracy, performance, and limitations by listing
    - Mean Absolute Error
    - Mean Square Error
    - Histogram of modeling Errors
- Summary of how you trained and validated your model.
    - Report on the different models you tried (either different kinds of models or varying the model's hyperparameters)
    - Compute accuracy metrics on both training and validation datasets
- Are there any known limitations to your model? For example, does it fail to give good predictions for certain kinds of diamonds?

Example Questions to Help Guide Your Summary

- What was your test-train split and why?
- How do you assess the accuracy/success of your models?

Part 3. Using your price prediction models

Congrats on building a sophisticated machine learning model to predict prices! Now it's time to deploy it. Build your model in a Jupyter notebook. Run your model on a cloud-based service such as Google Colaboratory or Project Jupyter's binder (both are free platforms that allow your team to import Jupyter notebooks, and train and evaluate models). These tools will allow you to use your prediction model and create a visualization. This visualization will be used in Part 4.

Now use your diamond pricing software to predict your company's sales in 2024:

- The data file for all diamonds your company plans to sell in 2024 is called diamonds_for_sale_2024.csv. Upload this data to your software. Provide a summary report estimating what your model predicts the total diamond sales will be in 2024.

Part 3 Deliverables - Build Price Prediction Model

- If all of your company's diamonds are sold in 2024...
  - What will the total sales be?
  - What will the sales be for each cut?
- What is the average predicted diamond sale price in 2024?
- Provide any additional analysis and insights for the upcoming 2024 year based on your model.

Part 4. Executive Dashboard

You and your team have learned a lot about this data set, and now your boss is asking you to prepare a dashboard for their boss. Throughout the project, you have been summarizing your observations. Now it is time to think back to the start. Why did they hire your team? What did they want to understand?

> ***Goal: Your company would like to better understand the diamond market, identify trends, and build a model that can predict diamond prices for next year.***

Use LookerStudio to create an executive dashboard consisting of up to 5 charts that highlight key observations for the business. One of the charts should be the predictive model developed in Part 3.

Great executive dashboards are easy for people to look at and immediately process the information. Ensure each chart presented clearly tells you something about the business goals. Review your summaries to determine the charts to display. Aim to highlight insights of trends within the chart.

Tips: Read this article for help on creating Business Intelligence dashboard using Google Colab and Google LookerStudio (DataStudio)

Tips on Crafting Charts for Clarity include:

- **Take away anything that you don't need**. Be brutal, as long as the point is still clear.
- **Remove redundancy**. Example titles should not repeat the axis labels.
- **Limit color and eye travel.** Color is powerful and distracting. Use color to focus the eye. Don't make people's eyes dart back and forth for information.
- **Know how people think (heuristics)**. These are the short cuts our brain takes. Red is generally bad, and green good. Future is after the present.
- **Describe ideas not structure**. Use text, headlines, captions, and other visual markers to highlight ideas or insights, rather than to describe the visualization's architecture.

Part 4 Deliverables - Executive Dashboard & Final Summary

- LookerStudio Executive Dashboard
- Summary of your experience working on the project
- Google Doc with all of the summaries from Parts 1-4
- Present your dashboard (Project Showcase)

---

5. Going further (optional - time permitting)

If there's extra time, consider...

- **Data scraping**. Build a data scraping script to get diamond attributes and prices on the internet. Feed this information into your diamond-pricing software to make a "buy" (price is below your estimate) recommendation.
- **Deep networks**: Use deep learning software like tensorflow or pytorch to build a pricing model. Experiment with the number and depth of deep layers, training epochs, and other settings to get a good model. How does this compare with the model you built in step 3?
- **Cut classification**: Your company suspects their newly hired diamond appraiser is incorrectly assigning values of diamond cut. They would like you to devise an algorithm to predict the cut based on other properties, which can then be compared to the appraiser's assignment.

Data science ethics

We are excited to set up a project that allows you to explore a robust and real data set with lots of opportunities for learning. We also recognize that the diamond industry has a complicated history, including involvement in wars and exploitation practices. If you're interested, we encourage you to learn more about this industry and to consider how the field of data science could be used to support strong ethical practices. For example, is there data you would want to see about diamond sourcing? Is there early data about lab-grown diamonds that you would be interested in exploring compared to the data set you have today? Could an algorithm be built to detect diamonds that have not been ethically sourced? All great data work considers ethics in a variety of ways. As you launch your careers in this field, we hope that you do, too.

Appendix: Recommended Python tools

If you are completing this project with Python, consider the following data science tools:

1. **Anaconda** (https://www.anaconda.com/products/individual): Anaconda will install Python on your laptop and allow you to easily install other Python packages like scikit-learn, pandas, matplotlib, and many other data science tools. There are many excellent online resources for how to setup and use Anaconda.
2. **Jupyter notebooks** (https://jupyter.org/): Jupyter notebooks allow for interactive coding with a web browser. You can write and run python code interactively right in your browser! This is a great way to explore data, build models, and do other kinds of interactive computing tasks. Anaconda should automatically install Jupyter notebooks. There are many excellent online resources for how to setup and use Jupyter.