Cairo University
Faculty of Graduate Studies for Statistical Research
Department of Computer Science

Data Science Assignment:
Exploratory Data Analysis and Feature Engineering of Diamonds Data set.

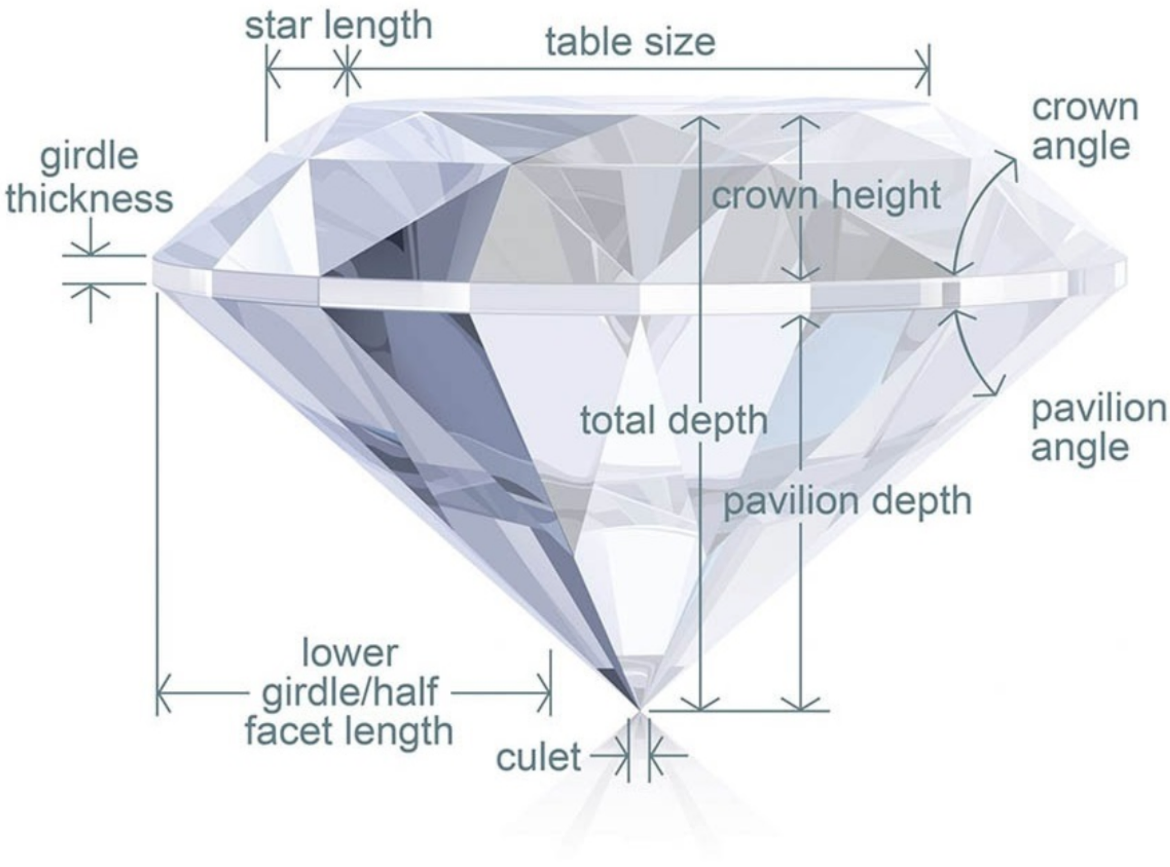
Submitted by

Eslam Fouad Jabr.

Alexandria, Egypt
August, 2023

Introduction

This notebook takes data analysts through the CRISP-DM (Cross-Industry Standard Process for Data Mining) journey by working with the diamonds dataset. Shivam Agrawal's Diamonds dataset is commonly used as a test case for machine learning models and data analysis techniques. It can be used to explore trends in the diamond market, such as the relationship between price and carat weight, or to predict the price of a diamond based on its characteristics.



CRISP-DM Model

The CRISP-DM model will be used to perform the data analytics process start to finish. CRISP-DM is a widely-used data mining process model. It provides a structured approach for identifying the business objectives, understanding the data and its context, preparing the data for modeling, selecting and applying appropriate data mining techniques, evaluating the results, and deploying the results in the business. Note: The data mining process consists of several phases that may be revisited as needed. The dependencies between these phases are important to consider, and it is useful to keep in mind that the data mining process is cyclical. Even after a solution has been implemented, the insights gained through the data mining process can lead to new business questions and inform future data mining efforts.

The CRISP-DM process consists of six main phases:

`Business Understanding`**: In this phase, the business objectives and requirements are defined, and a preliminary plan is developed for achieving them using data mining.**

`Data Understanding`**: In this phase, the data is explored and analyzed to understand its content, quality, and relevance to the business objectives.**

`Data Preparation`**: In this phase, the data is cleaned and transformed to get it ready for modeling.**

`Modeling`**: In this phase, appropriate data mining techniques are selected and applied to the prepared data to create one or more predictive models.**

`Evaluation`**: In this phase, the performance of the predictive models is evaluated and the results are compared to the business objectives.**

`Deployment`**: In this phase, the results of the data mining process are deployed in the business, which may involve creating reports, integrating the results into business processes, or creating a data mining application.**

The CRISP-DM process is designed to be flexible and adaptable, and it can be customized to fit the needs of a particular business or project. It is a widely-used and well-established approach for data mining projects, and it has been adopted by organizations in a variety of industries. Some examples of industries and companies that have adopted the CRISP-DM model include:

- Financial services 🏢: Many financial institutions, such as banks and insurance companies, have used the CRISP-DM model to analyze customer data and identify trends and patterns that can be used to improve products and services.
- Healthcare ⚕: The healthcare industry has used the CRISP-DM model to analyze patient data and identify patterns and trends that can be used to improve patient care and outcomes.
- Retail 🛍: Retail companies have used the CRISP-DM model to analyze customer data and identify trends and patterns that can be used to improve marketing and sales strategies.
- Manufacturing 🏭 : Manufacturing companies have used the CRISP-DM model to analyze production data and identify trends and patterns that can be used to improve efficiency and reduce costs.
- Telecommunications 📡: Telecommunications companies have used the CRISP-DM model to analyze customer data and identify trends and patterns that can be used to improve products and services.

Business Understanding

In our hypothetical business scenario, we are the business analysts of a jewelry store. The CEO is worried about the sales forecast, so he or she asks us to do an analysis of which features are the best predictors of price changes. This will help him or her make a more realistic plan and budget for the coming year.

Define Business Objectives

Here are a few examples of business objectives that a hypothetical diamond retailer might have:

Increase revenue: One objective for a diamond retailer could be to increase revenue by selling more diamonds to consumers. This could be achieved through a variety of strategies, such as expanding the product line, increasing marketing efforts, or improving the customer experience.

Improve profitability: Another objective for a diamond retailer could be to improve profitability by reducing costs or increasing the margin on each sale. This could be achieved through strategies such as negotiating better terms with suppliers, streamlining operations, or implementing more efficient systems and processes.

Expand market share: A diamond retailer might also have an objective to increase its market share by attracting more customers or selling more diamonds to existing customers. This could be achieved through marketing and branding efforts, as well as by offering a wider range of products or more competitive pricing.

Enhance customer satisfaction: Another important objective for a diamond retailer might be to enhance customer satisfaction by providing excellent service and offering high-quality products. This could be achieved through strategies such as training employees to provide excellent customer service, offering after-sales support, and regularly soliciting feedback from customers.

Promote ethical and sustainable practices: In an industry with significant environmental and social impacts, a diamond retailer might also have an objective to promote ethical and sustainable practices throughout the business. This could involve sourcing diamonds from responsible sources, supporting initiatives that minimize environmental impact, and promoting fair labor practices.

Identify Stakeholders

There are many potential stakeholders in the diamonds business, including:

Diamond mining companies: These companies are responsible for extracting diamonds from the earth and are a key stakeholder in the diamonds business.

Diamond wholesalers: These companies purchase diamonds from mining companies and other sources and sell them to retailers.

Diamond retailers: These companies sell diamonds to consumers, either through physical stores or online.

Diamond cutters and polishers: These companies are responsible for cutting and polishing rough diamonds to create finished diamonds.

Consumers: Consumers are the end users of diamonds and are a key stakeholder in the diamonds business.

Governments: Governments may be stakeholders in the diamonds business through the regulation of diamond mining and trade, and through the taxes and other revenues generated by the industry.

Environmental groups: Environmental groups may be stakeholders in the diamonds business due to the environmental impacts of diamond mining, particularly in areas with fragile ecosystems.

Labor unions: Labor unions may be stakeholders in the diamonds business due to the labor practices of diamond mining companies and other industry players.

Investors: Investors may be stakeholders in the diamonds business through their ownership of shares in diamond mining and other industry companies.

Define Data Mining Goals

If you are predicting the price of diamonds, some data mining goals that you might want to have include:

Accurate prediction: A key goal of data mining in this context would be to develop a model that can accurately predict the price of diamonds based on various factors, such as the size, clarity, and color of the diamond.

Understanding the factors that influence price: Another goal might be to understand the factors that influence the price of diamonds, such as the characteristics of the diamond, market conditions, and consumer demand. This could involve analyzing the data to identify patterns and relationships that can help to explain why certain diamonds are more valuable than others.

Identifying trends: Another goal might be to identify trends in the data that can help to anticipate changes in the price of diamonds over time. This could involve analyzing time

series data to identify patterns of demand or other factors that are likely to affect the price of diamonds in the future.

Improving forecasting: Another data mining goal might be to improve the accuracy of forecasting models for the price of diamonds. This could involve testing different models and techniques to see which ones are most effective at predicting the price of diamonds, and using this information to improve the accuracy of future forecasts.

Automating the prediction process: Finally, a data mining goal might be to automate the prediction process as much as possible, so that the model can be used to quickly and accurately predict the price of diamonds without the need for manual intervention. This could involve developing algorithms that can process large amounts of data efficiently, or implementing machine learning techniques to improve the accuracy of the model over time.

Define Success Critera

Success criteria are specific, measurable, attainable, relevant, and time-bound (SMART) goals that are used to evaluate the success of a project or initiative. Defining success criteria is an important step in the project planning process, as it helps to ensure that the project stays focused and aligned with the overall goals of the organization.

SMART Goal: By the end of the project, we will have identified and analyzed at least three key factors that influence prices in diamonds retail using data mining techniques, and we will have achieved an accuracy of at least 95% in predicting price changes based on these factors. We will do this by collecting and cleaning at least 50,000 relevant data points, selecting and applying appropriate data mining algorithms, and verifying the accuracy of the predictions through cross-validation. We will also present the findings and methodology to the CEO of our hypotetical dimanods company.

KPI 1: Percentage of goal completion: Track the number of factors identified and analyzed, as well as the percentage of data points collected and cleaned.

KPI 2: Accuracy of Price Prediction: Calculate the percentage of price predictions that are correct, using cross-validation to verify the accuracy of the model.

KPI 3: Number of Presentations: Track the number of times the individual presents or discusses their findings with colleagues or supervisors.

Communicate the success criteria clearly and effectively to all relevant stakeholders: Make sure that stakeholders understand why the success criteria are important and how they relate to the overall goals of the organization. Share the success criteria with all stakeholders who will be affected by or have a role in achieving the goal. This could include employees, customers, shareholders, and other stakeholders. Make sure that stakeholders fully understand the success criteria and address any questions or concerns they may have. Regularly track progress towards the success criteria and provide feedback to stakeholders as appropriate. This will help ensure that the goal is on track to be achieved and that any necessary adjustments can be made.

Data Quality Check

We will begin data quality checks in the Data Understanding phase of the project. This will include reviewing the data source and evaluating there are no missing values, errors, inconsistencies, or duplicates. Also, necessary data transformations will be performed to prepare the data so that the dataframe is machine readable.

By performing a thorough data quality check, we can ensure that the data used for the project is accurate and relevant, which will help improve the quality of the business understanding.

Data Understanding

Read Data

Pandas is a popular Python library for working with data. It provides functions and methods for reading, manipulating, and analyzing data stored in a variety of formats, such as CSV, Excel, JSON, and SQL databases.

To read data in Pandas, you can use the *read_ functions provided by the pandas module. For example, to read a CSV file, you can use the read_csv() function**

In this notebook, we will be looking at a Diamonds dataset. Kaggle's Diamond dataset is a collection of data on the physical attributes and characteristics of diamonds, including the diamond's price, carat weight, color, clarity, and cut. The dataset includes approximately 54,000 diamonds and is provided in a CSV (comma-separated values) file format.

The columns in the Diamonds dataset include:

price: The price of the diamond in US dollars (326-18,823)

carat: The weight of the diamond, measured in carats (0.2--5.01)

cut: The quality of the diamond cut, with categories including "Ideal", "Premium", "Very Good", "Good", and "Fair".

color: The color of the diamond from J (worst) to D (best)

clarity: The clarity of the diamond (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))

x length in mm (0--10.74)

y width in mm (0--58.9)

z depth in mm (0--31.8)

depth: The depth of the diamond, expressed as a percentage of the total width, z / mean(x, y) = 2 * z / (x + y) (43--79)

table: The width of the diamond's top surface, expressed as a percentage of the total width (43--95)

Data Preparation

Data preparation, also known as data preprocessing, is the process of cleaning, transforming, and organizing data for analysis or modeling. It is a critical step in the data science process, as the quality and structure of the data can significantly impact the accuracy and effectiveness of the analysis or model. Data preparation can be a time-consuming and labor-intensive process, especially if the data is large or complex. However, it is an important step to ensure that the data is of high quality and is suitable for the intended analysis or model.

There are several steps that are typically involved in data preparation including data cleaning, transformation, integration, selection, sampling, and splitting.

Data Cleaning

Data cleaning is the process of identifying and addressing issues with the data that can negatively impact the accuracy and effectiveness of the analysis or model. This can include identifying and correcting errors, filling in missing values, and handling outliers. Some common tasks that are involved in data cleaning include:

Identifying and correcting errors: This can include correcting spelling mistakes, formatting errors, and other types of mistakes in the data.

Filling in missing values: This can involve imputing missing values using statistical methods, such as mean imputation or multiple imputation, or using machine learning algorithms to predict the missing values.

Handling outliers: This can involve identifying and removing extreme values that are outside the range of what is expected in the data, or transforming the data to reduce the impact of the outliers.

Dealing with duplicates: This can involve identifying and removing duplicate records from the data.

Formatting the data: This can involve standardizing the formatting of the data, such as ensuring that all values are in the same format, or converting data between different types, such as converting text to numerical values. Data cleaning can be a time-consuming process, especially if the data is large or complex. However, it is an important step to ensure that the data is of high quality and is suitable for the intended analysis or model.

Data Transformation

Data transformation refers to the process of converting data from one format or structure into another format or structure. Data transformation is often used as a step in data preparation or data wrangling, which involves cleaning, organizing, and preparing data for analysis or machine learning tasks. There are many different types of data transformation, and the specific transformation needed will depend on the characteristics of the data and the requirements of the downstream analysis or modeling tasks. Some common types of data transformation include:

Filtering: Removing unnecessary or irrelevant data from a dataset.

Aggregation: Combining data from multiple sources or grouping data by common characteristics.

Normalization: Scaling data so that it has a common scale or range.

Encoding: Converting categorical data into numerical values.

Imputing missing values: Filling in missing data with a suitable value, such as the mean or median of the data.

Outlier detection and removal: Identifying and removing data points that are significantly different from the rest of the data. Overall, the goal of data transformation is to create a dataset that is more suitable for the intended analysis or modeling tasks, and that accurately represents the underlying data.

Data Integration

Data integration is the process of combining data from multiple sources into a single, coherent view. It involves extracting data from various sources, transforming it into a format that can be integrated, and then loading it into a target system where it can be used for analysis and reporting. Data integration can be used to:

• Combine data from different sources to provide a more complete view of a business or a particular process

• Consolidate data from different systems or departments to provide a single source of truth for decision-making

• Integrate data from external sources, such as suppliers or customers, to gain a better understanding of the business environment

There are several approaches to data integration, including:

Extract, transform, and load (ETL): This involves extracting data from multiple sources, transforming it into a format that can be integrated, and then loading it into a target system.

Real-time integration: This involves integrating data in real-time as it is generated, so that it is always up-to-date.

Batch integration: This involves integrating data on a periodic basis, such as daily or weekly. Data integration can be challenging because it involves dealing with data from multiple sources, which may have different structures, formats, and schemas. It also requires coordinating the movement of data between systems and ensuring that it is consistent and accurate.

Data Selection

Data selection is the process of choosing specific data from a larger dataset for a particular purpose or analysis. It involves identifying the data that is relevant to the task at hand and selecting it for further processing or analysis. There are several reasons why data selection might be necessary:

• The dataset may be too large or complex to analyze in its entirety

• The data may be of poor quality or contain errors that need to be cleaned or corrected

• The data may be in a format that is not compatible with the analysis tools or techniques being used

• The data may be sensitive and need to be protected from unauthorized access or use

There are several methods for data selection, including:

Sampling: This involves selecting a representative sample of the data for analysis, rather than analyzing the entire dataset. Sampling can be useful when the dataset is too large or complex to analyze in its entirety.

Filtering: This involves selecting specific data points or records based on certain criteria, such as a particular date range or a certain value. Filtering can be useful for removing outliers or identifying trends.

Slicing and dicing: This involves selecting specific data points or records based on multiple criteria or dimensions, such as location and product type. Slicing and dicing can be useful for analyzing data from different perspectives or for creating subgroups for analysis. It is important to carefully consider the data selection process, as the data that is chosen will significantly impact the results of the analysis. It is also important to ensure that the data selection process is transparent and well documented, so that others can understand and reproduce the analysis if necessary.

Data Sampling

Data sampling is the process of selecting a subset of data from a larger dataset for the purpose of analysis. It is a common technique used in statistical analysis, as it allows researchers to draw conclusions about a larger population based on a smaller sample of data. There are several types of data sampling techniques, including:

Simple random sampling: This involves selecting a sample of data from the dataset randomly, with each data point having an equal probability of being selected.

Stratified sampling: This involves dividing the dataset into homogeneous subgroups (strata) and then selecting a sample from each stratum. This can be useful when the dataset is not homogeneous and you want to ensure that the sample is representative of the entire population.

Cluster sampling: This involves dividing the dataset into clusters and then selecting a sample of clusters for analysis. This can be useful when it is difficult or expensive to sample the entire population, but the clusters are representative of the population.

There are several factors to consider when selecting a sampling technique, including the size and characteristics of the population, the resources available for sampling, and the purpose of the analysis. It is important to carefully design the sampling process to ensure that the sample is representative of the population and that the conclusions drawn from the sample can be generalized to the larger population.

Data Splitting

Data splitting and data sampling are similar in that they both involve selecting a subset of data from a larger dataset for a particular purpose. However, there are some key differences between the two:

Purpose: Data splitting is primarily used in the development of machine learning models, while data sampling is used for statistical analysis.

Size of the sample: Data splitting involves dividing the dataset into three subsets (training, validation, and test), while data sampling involves selecting a smaller sample from the entire dataset.

Method of selection: Data splitting typically involves randomly dividing the dataset into subsets, while data sampling can use a variety of techniques, such as simple random sampling, stratified sampling, or cluster sampling.

In summary, data splitting is a specific technique used in the development of machine learning models, while data sampling is a more general technique used for statistical analysis. Both techniques involve selecting a subset of data from a larger dataset, but they differ in their purpose and method of selection.

Data Modeling

Model Selection

Build the Model

Evaluate the Model

Fine-tune the Model

Review the Results

Assess the Model Results

Determine the Value of the Results

Review the Project

ata Deployment

Prepare the Results

Communicate the Results

Implement the Results

Monitor the Results

Conclusion