**Data Analysis Visualization CA1**

**Yue Xing Tong**

**Theoretical Framework**

**GitHub : <https://github.com/y-xing01/DAV_CA1>**

**Introduction:**

This document presents the exploratory data analysis (EDA) of a diamond dataset, containing information on almost 54,000 diamonds. The dataset includes various attributes such as price, carat, cut, color, clarity, dimensions, depth, and table. The primary objective of this analysis is to gain insights into the characteristics of the dataset and understand the relationships between different attributes.

1. **Data Selection**

*Data Description:*

* The dataset comprises 10 attributes:
* Price: The price of the diamond (target variable).
* Carat: The weight of the diamond, ranging from 0.2 to 5.01 carats.
* Cut: Quality of the cut, categorized as Fair, Good, Very Good, Premium, and Ideal.
* Color: Color grade of the diamond, ranging from J (worst) to D (best).
* Clarity: Clarity grade of the diamond, ranging from I1 (worst) to IF (best).
* X, Y, Z: Dimensions of the diamond in millimeters.
* Depth: Total depth percentage, calculated as 2 \* z / (x + y), ranging from 43 to 79.
* Table: Width of the top of the diamond relative to the widest point, ranging from 43 to 95.

*Data Preparation:*

* Loaded the dataset into a Pandas DataFrame.
* Removed the index column (Unnamed: 0) as it is redundant.
* Checked the shape and data types of the columns to ensure data integrity.
* Reviewed the descriptive statistics of the dataset to understand the distribution of variables.

*Data Overview:*

* The dataset consists of 53,940 entries and 11 columns.
* There are no missing values in the dataset.
* The price of diamonds ranges from $326 to $18,823.
* The carat weight varies from 0.2 to 5.01 carats.
* The depth percentage ranges from 43 to 79, while the table width ranges from 43 to 95.

1. **Data Cleaning**

*Data Cleaning:*

To ensure the dataset's integrity, I first checked for any missing values using the isnull().sum() function. Fortunately, no missing data points were found. However, I noticed some 0 values in columns 'x', 'y', and 'z', which could indicate inaccuracies, likely representing dimensionless or 2-dimensional diamonds. To address this, I removed the unnecessary first column labeled 'Unnamed: 0'. Then, I filtered out these dimensionless diamonds by dropping rows with 0 values in 'x', 'y', and 'z'. This resulted in a dataset with dimensions of (53,920, 10), having eliminated 20 data points associated with dimensionless diamonds.

*Outlier Detection and Removal:*

Turning to outlier detection, I focused on numerical columns. By computing Z-scores for these columns using zscore, I identified and flagged potential outliers based on a predefined threshold of |Z| > 3. Subsequently, I systematically removed rows containing outliers across any numerical column. This process led to a refined dataset with a revised shape of (51,586, 10), effectively addressing outliers' impact on subsequent analyses.

*Data Label Encoding:*

Transitioning to data label encoding, I started by identifying categorical variables such as 'cut', 'color', and 'clarity'. Utilizing a LabelEncoder object, I encoded these categorical data, mapping original string values to encoded integer values. This facilitated seamless integration of categorical data into analytical processes. Upon completion, I validated the dataset's data types using the info() function, ensuring successful integration of label encoding transformations.

1. **Multivariate Analysis**

*Correlation and Coefficient:*

I started by checking the cleaned and encoded dataset, which contained 53,920 entries and 10 columns. The columns included various features such as carat, cut, color, clarity, depth, table, and dimensions (x, y, z) of diamonds. The data types were appropriately converted, and there were no missing values in the dataset.

Next, I conducted a multivariate analysis to understand the relationships between different variables and their impact on the price of diamonds. Firstly, I calculated the Pearson correlation coefficient between the price (response variable) and other predictor variables such as carat, cut, color, clarity, depth, table, and dimensions (x, y, z) of the diamonds. The correlation coefficients ranged from high positive correlations (e.g., carat, dimensions) to low correlations (e.g., cut, clarity). This indicated that variables like carat and dimensions had a stronger linear relationship with price compared to cut and clarity.

To visualize these relationships, I created scatter plots for each predictor variable against the price. These plots showed the distribution of data points and the fitted regression line indicating the trend between the predictor and response variables. For instance, carat and dimensions (x, y, z) exhibited a clear positive linear relationship with price, while variables like cut and clarity showed less apparent trends.

Furthermore, I generated a correlation matrix and heatmap to provide a comprehensive view of the relationships between all pairs of variables. The heatmap revealed strong correlations between carat and dimensions (x, y, z), while variables like cut, clarity, and depth showed weaker correlations with the price.

Lastly, I investigated collinearity between the price and carat, as carat appeared to have the highest correlation with price. This analysis confirmed that there were various unique values for both price and carat, indicating that they were not identical for all observations. This suggested that while carat had a significant impact on price, it was not the sole determinant, and other factors also influenced diamond pricing.

*Correlation Matrix:*

To begin with, I calculated the correlation matrix using the Pandas DataFrame method corr(). This matrix quantifies the linear relationships between pairs of variables in the dataset. Each cell in the matrix represents the correlation coefficient between two variables, ranging from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

After obtaining the correlation matrix, I visualized it using a heatmap to provide a graphical representation of the correlations. In the heatmap, each cell's color intensity corresponds to the strength of the correlation between the respective pair of variables. Additionally, I included annotations within the heatmap to display the correlation coefficients for better interpretation.

Upon examining the heatmap, several key observations can be made:

High correlation: Variables such as "carat," "x," "y," and "z" exhibit a high correlation with each other. This high correlation is visually represented by the intense coloration in the corresponding cells of the heatmap. This indicates that these variables are closely related to each other, which is expected since they all pertain to the physical dimensions of the diamonds.

Low correlation: On the other hand, variables like "depth," "cut," "color," "clarity," and "table" show relatively low correlations with other variables in the dataset. The heatmap displays lighter colors for these variables, indicating weaker correlations compared to the dimensions-related variables. However, it's important to note that low correlation does not necessarily imply insignificance; these variables may still have important effects on the target variable ("price") individually or in combination with other variables.

Based on the observations, it's suggested to consider the correlations when performing further analysis or modeling. While variables with high correlations may introduce multicollinearity issues in regression models, variables with low correlations may still contribute valuable information to the predictive power of the model. Therefore, careful consideration is needed before deciding whether to drop any variables from the dataset.

*Collinearity and Cross-Tabulation:*

In my analysis, I explored potential collinearity between the predictor variable "price" and several response variables including "carat," "cut," "color," "clarity," "depth," "table," "x," and "y." To conduct this investigation, I iterated through each predictor variable and compared its unique values with those of each response variable. I looped systematically to examine each pair of predictor and response variables.

Beginning with the comparison between "price" and "carat," it was evident that the unique values were not identical for all observations. This indicates variability between the two variables, suggesting that they are not perfectly collinear. Subsequent comparisons with categorical variables such as "cut," "color," and "clarity" also revealed non-identical unique values, reinforcing the absence of perfect collinearity between "price" and these attributes.

Moreover, comparisons with continuous variables such as "depth," "table," "x," and "y" similarly showed non-identical unique values, indicating diverse distributions and further supporting the conclusion of no perfect collinearity between "price" and these features. These findings are crucial for understanding the relationships between variables in the dataset, particularly in predictive modeling or regression analysis where collinearity can distort results or inflate standard errors.

Overall, the systematic examination of collinearity between "price" and the various predictor and response variables provided valuable insights into the dataset's structure and relationships.

1. **Regression Modeling**

*Performing Linear Regression:*

In conducting regression modeling, I first utilized linear regression to analyze the relationship between the predictor variable "carat" and the response variable "price" in a dataset. The "carat" data was reshaped into a 2D array and used to initialize a Linear Regression model. Upon fitting the model, I calculated the coefficient of determination (R-squared), obtaining a value of approximately 0.849, indicating a strong correlation between carat and price.

Subsequently, I performed ordinary least squares (OLS) regression analysis to investigate the relationships between price and other predictor variables such as "cut," "color," "clarity," "depth," and "table." Each variable was individually analyzed against price using OLS regression, with their respective R-squared values providing insights into the goodness of fit for each model.

For instance, when examining the relationship between "cut" and "price," the R-squared value was approximately 0.438, suggesting a moderate degree of correlation between these variables. Similarly, analyses of "color," "clarity," "depth," and "table" against "price" revealed R-squared values of around 0.428, 0.384, 0.492, and 0.497, respectively. These results indicate varying degrees of correlation between each predictor variable and price, with "depth" and "table" showing relatively stronger associations compared to "color" and "clarity."

Moreover, scatter plots were generated for each regression analysis, visually representing the relationship between the predictor variable and price. The regression lines plotted on these graphs depict the trend of the data and help visualize the strength and direction of the relationship between the variables.

Overall, the regression modeling process involved fitting linear regression models and conducting OLS regression analyses to explore the relationships between different predictor variables and the price of diamonds.

*Principal Component Analysis:*

I selected the numeric features from the dataset to perform Principal Component Analysis (PCA). This involved filtering out non-numeric variables and retaining only those suitable for PCA. The numeric data was then standardized using a StandardScaler to ensure that all features had a mean of 0 and a standard deviation of 1, a prerequisite for PCA.

Next, I initialized the PCA object and fitted it to the scaled data. By doing so, I obtained the principal components along with their corresponding explained variance ratios. The scree plot, generated using the explained variance ratios, provided insights into the amount of variance captured by each principal component.

Upon transforming the scaled data using PCA, I visualized the data in a scatter plot to observe the distribution of samples in the reduced-dimensional space defined by the first two principal components (PC1 and PC2). This plot helped visualize any underlying patterns or clusters present in the data.

Then, I extracted the loading scores for the first principal component (PC1) to understand the contributions of each original feature to PC1. The loading scores quantify the strength and direction of the relationship between the original variables and the principal components. By sorting the loading scores in descending order of magnitude, I identified the top 10 variables contributing most to PC1.

The results revealed that "carat" had the highest loading score (0.446), indicating a strong positive correlation with PC1. This suggests that variations in "carat" contribute significantly to the direction of PC1. Additionally, "price" exhibited a relatively high loading score (0.414), implying a positive correlation with PC1, albeit slightly weaker than "carat". Other variables such as "color" and "table" also showed positive loading scores, albeit smaller in magnitude, indicating some level of correlation with PC1. Conversely, "clarity" had a negative loading score (-0.111), indicating a weak negative correlation with PC1.

In summary, the results of the PCA analysis provided valuable insights into the underlying structure of the dataset, highlighting the most influential variables contributing to the principal components and their respective relationships with the original features.

1. **Normalization and Standardization**

*Standardization (Z-score normalization):*

To conduct hierarchical clustering on a dataset consisting of predictor and response variables, I first defined the predictor and response variable subsets. The predictor variable, which serves as the basis for clustering, was identified as "price", while the response variables included "carat", "cut", "color", "clarity", "depth", and "table".

To ensure that variables with different scales contribute equally to the clustering process, I performed Z-score normalization on the response variables. This involved initializing a StandardScaler object and fitting it to the response variable data. Subsequently, the data was transformed to have a mean of 0 and a standard deviation of 1. The standardized data was then converted back to a DataFrame for further analysis.

After standardization, the standardized data exhibited a mean of approximately 0 and a standard deviation of approximately 1 for each variable, ensuring that all variables were on a comparable scale suitable for clustering.

*Normalization (Minmax Scaling):*

I applied Minmax scaling, a form of normalization, to the response variables. This technique scales the data to a fixed range (typically 0 to 1) by subtracting the minimum value and dividing by the range of each feature. Similar to standardization, this step aimed to ensure that variables with different ranges did not disproportionately influence the clustering process.

After normalization, the data was rescaled to fall within the range of 0 to 1 for each variable, facilitating the comparison and interpretation of variable values across different features.

1. **Cluster Analysis**

*Hierarchical Clustering (Complete Linkage):*

First, I selected relevant columns including "carat", "cut", "color", "clarity", "depth", "table", and "price" for clustering. After dropping any rows with missing values, I visualized the distribution of outliers using boxplots and subsequently applied the interquartile range (IQR) method to remove outliers. Next, I standardized the data and performed hierarchical clustering with complete linkage, visualizing the dendrogram to identify potential clusters based on the Euclidean distance metric.

*Hierarchical Clustering (KMeans):*

In this approach, I scaled the data using MinMax scaling and applied the KMeans algorithm with a predefined number of clusters (4 in this case). After fitting the model, I assigned cluster labels to the data and visualized the clusters along with their centroids, highlighting the distribution of data points across the "carat" and "price" dimensions.

*Hierarchical Clustering (DBSCAN):*

Lastly, I employed the DBSCAN algorithm with specified parameters such as epsilon (eps) and minimum samples to detect clusters based on density. Following clustering, I assigned cluster labels to the data and plotted the clusters, enabling the visualization of data points grouped by their proximity in the feature space.

1. **Indexing**

*Final Composite Index:*

For indexing, I constructed a composite index to evaluate the overall quality of diamonds based on various attributes. Initially, I selected relevant columns such as "carat", "cut", "color", "clarity", "depth", "table", and "price" from the dataset. To ensure data integrity, I removed any rows with missing values and randomly sampled 1000 observations for analysis.

After preprocessing the data, I normalized each attribute to bring them to a common scale between 0 and 1. Next, I assigned weights to each attribute based on their relative importance, with a total weight of 1. Using these weights, I calculated sub-indices for each attribute and aggregated them to derive the composite index for each diamond. Finally, I integrated the composite index back into the dataset and showcased the initial entries to illustrate the calculation process**.**

*Top 10 Diamonds based on Composite Index:*

Utilizing the composite index calculated earlier, I ranked the diamonds in descending order of their composite scores and extracted the top 10 diamonds with the highest composite indices. These diamonds represent the highest-quality specimens in the dataset, combining favorable attributes such as carat, cut, color, clarity, and others. I presented these top-ranking diamonds along with their attributes and composite indices in a tabular format for easy interpretation.

*Visualization:*

To provide insights into the distribution and relationships between the composite index and individual features, I employed various visualization techniques. First, I plotted a histogram to depict the distribution of composite indices across the dataset, allowing for an understanding of the overall quality distribution. Additionally, I created a bar plot showcasing the composite indices of the top 10 diamonds, enabling a clear comparison of their quality rankings. Furthermore, I visualized the scatter plots of individual features against the composite index, along with regression lines to identify any trends or correlations between attributes and the composite index. These visualizations offer a comprehensive overview of the dataset's composition and the factors influencing diamond quality.

1. **Conclusion**

In conclusion, the exploratory data analysis of the diamond dataset yielded several key findings. Firstly, correlations and regression analyses revealed significant relationships between attributes like carat and price, indicating their substantial influence on diamond pricing. Additionally, cluster analysis techniques uncovered distinct patterns within the dataset, allowing for the segmentation of diamonds based on their attributes. Moreover, the creation of a composite index provided a comprehensive measure of diamond quality, enabling the ranking of diamonds according to their overall attributes. Visualizations further enhanced the understanding of the dataset, elucidating attribute distributions, price relationships, and quality rankings. These insights are valuable for stakeholders in the diamond industry, offering actionable information for decision-making and strategic planning.