Final Report for INFSCI 2415

Subject: Diamonds
Student: Yutong Tang (YUT89)
Data From:
https://www.kaggle.com/datasets/joebeachcapital/diamonds
GitHub Source Code:
https://github.com/Yutong89/INFSCI_2415_Final

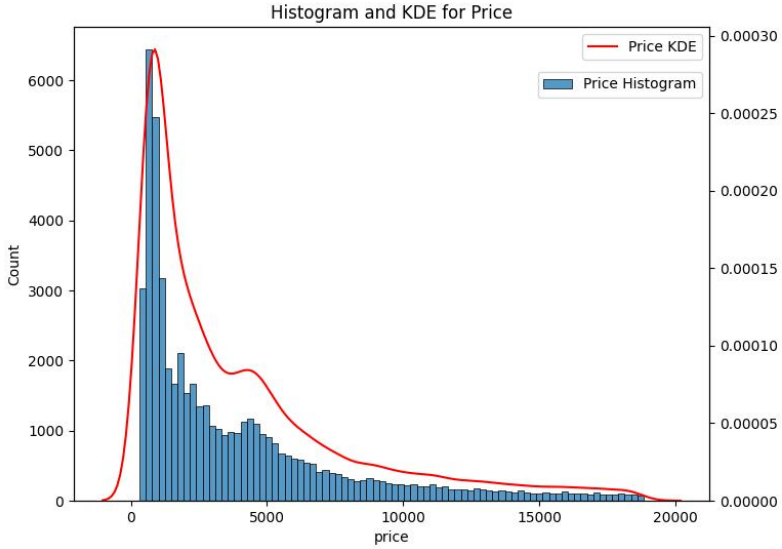
Data and method text describing the data and method used in this process:

- Data derived from Kaggle, related to different dimensions of diamonds. This dataset contains attributes like price, carat, depth, table, cut, color, and clarity of diamonds.
- Method (python package: padas, matplotlib, seaborn, numpy):
  - Histograms and KDE: Frequency distribution of price, carat, depth, table.
  - Heatmap: Average diamond price based on cut and color.
  - Stem Plot: Binned diamond prices against average carat, cut, clarity, and depth.

Significance statement on why the presented figures are important:

- Histograms and KDE: Reveal distribution and smooth representation of attributes.
- Heatmap: Shows how cut and color influence price.
- Stem Plot: Details typical diamond characteristics across price ranges.
- These visuals guide stakeholders in decision-making by highlighting attribute-price relationships.

Figure 1: Histogram and KDE for Diamond Price Analysis (Histogram)



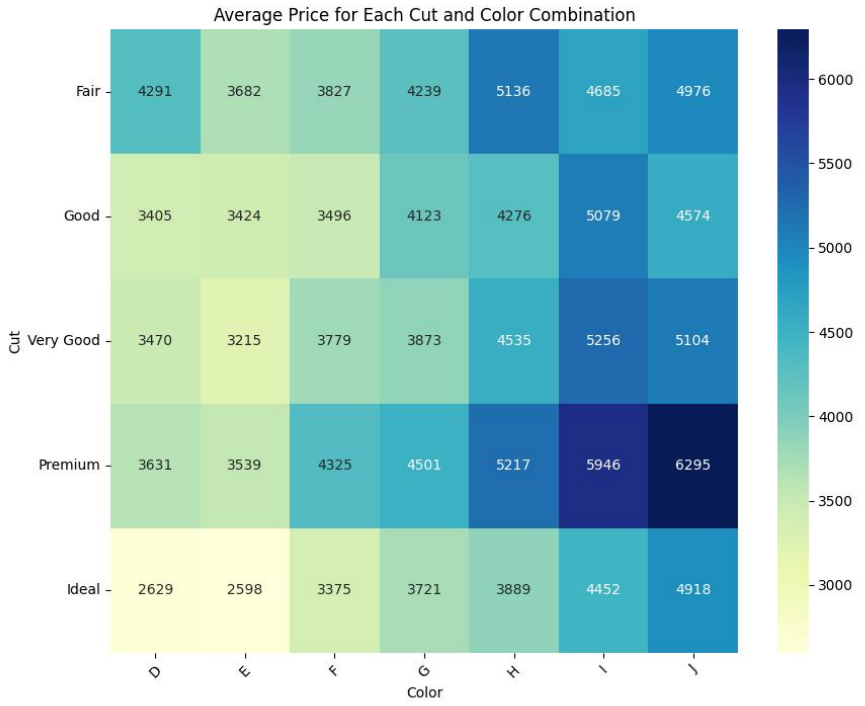
Legend Description:

- Price KDE (red line): Represents the *Kernel Density Estimate*, a smooth curve showing the distribution probability of the diamond prices.
- Price Histogram (blue bars): Shows the actual count of diamonds within various price ranges, represented by bars.

Findings:

- **Most Common Price Range**: The tallest bar indicates the most common price range, which appears to be between $0 to $1000, suggesting that most of the diamonds in this dataset are at the lower end of the price spectrum.
- **High-Price Rarity**: The decreasing height of bars as the price increases indicates that higher-priced diamonds are less common.
- **Smoothness of KDE**: The KDE line provides a smooth approximation of the histogram, helping to visualize the probability density of diamonds across different prices.

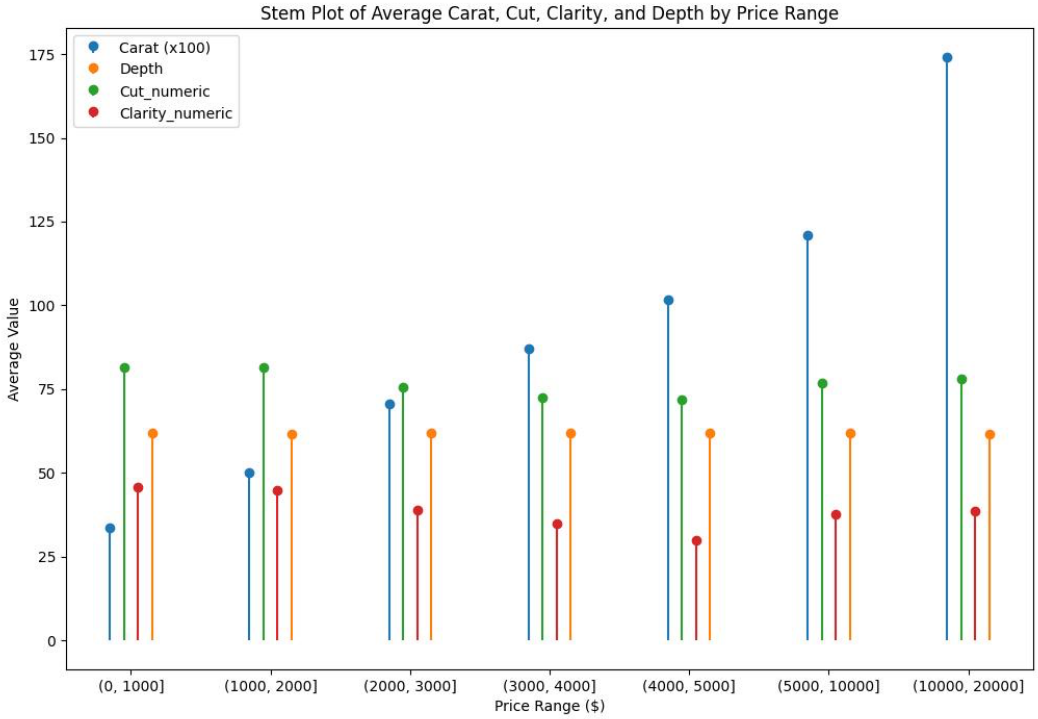
Figure 2: Heatmap of average price of cut and color combination (Heatmap)



Findings:

- **Color Gradient**: The colors range from light green (*lower prices*) to dark blue (*higher prices*), representing the average price for each combination of cut and color.
- **Cut Quality**: The 'Ideal' cut typically has lower average prices across all colors, which might seem counterintuitive as 'Ideal' cut is generally considered to be of higher quality. This could be due to other factors such as carat weight, clarity, or market availability affecting the average price.
- **Price Range**: The premium cut diamonds, particularly in the higher color grades (G through I), show the highest average prices on this heatmap.
- **Color Impact**: The variation in color from D to J shows a fluctuating pattern of prices with certain color grades like H and I consistently commanding higher average prices across various cuts.
- **Inconsistencies**: While one might expect a gradual change in prices from fair to ideal cuts, the heatmap shows inconsistencies. For example, 'Very Good' cuts have a higher average price than 'Premium' cuts in the D color category, indicating that factors other than cut and color are influencing the price.

Figure 3: Stem Plot of Average Carat/Cut/Clarity/Depth by Price Range (Stem Plot)



- **Legend description:**
  - **Blue dots (lines):** Average carat size (multiplied by 100 for scaling purposes).
  - **Orange dots (lines):** Average depth percentage.
  - **Green dots (lines):** Average cut quality, numerically encoded.
  - **Red dots (lines):** Average clarity, numerically encoded.

- **Findings**:
  - **Carat Size:** Carat size increases with price. The largest average carat size is in the highest price range.
  - **Depth:** Depth varies less across price ranges and remains relatively stable.
  - **Cut Quality:** Cut quality, on average, does not show a clear trend with price range, suggesting that cut may be less of a price determinant within this dataset.
  - **Clarity:** Clarity shows a slight increase with price but is not as pronounced as carat size.