Jefferson Veliz

Professor Jairam

Data Mining for Business Analytics

June 28th, 2023

**Lab #1 Report**

**PART I**

Visualizations play a crucial role in data analysis, providing a powerful means to uncover patterns and insights within complex datasets. This report aims to explain how different visualizations assist in revealing patterns and why each visualization is useful in the process.

I chose a FIFA 23 dataset that contains 18,539 rows and 89 columns. To begin my exploratory analysis, I cleaned the data by removing duplicates and identifying missing values. I focused on specific columns such as Full name, age, nationality, overall, potential, club name, value (in Euro), wage (in Euro), and best position, I proceeded with creating visualizations to gain insights from the data.

The first visualization I employed was a heatmap to find missing data. By applying a heatmap specifically to identify missing values, I could quickly locate areas within the dataset that contain NA values. The color scheme used in the heatmap helped distinguish the missing values from other data points, allowing for easy identification. In my dataset, there were no missing values shown in the heatmap, indicating a complete dataset.

Continuing my analysis, I decided to focus on specific nationalities that had over 300 players. This subset allowed me to explore patterns within a specific group. To visualize the relationship between age and potential, I utilized a Boxplot. This visualization displays key statistical measures such as the median, quartiles, and potential outliers. By plotting the age of the players against their potential, I could observe the potential values for each age group and identify any potential outliers. The Boxplot provided a clear summary of the distribution of potential values within each age group, enabling insights into the age-potential relationship. To gain a better understanding of the distribution of player ages, I created a histogram. This visualization allowed me to visualize the frequency distribution of age values. From the histogram, I could observe that most of the players were clustered around the age range of 21 to 23, with a gradual decline in the number of players as age increased. This information provided valuable insights into the age distribution of the players in the dataset. To analyze the relationships between different variables, I generated a correlation matrix. The correlation matrix is a useful visualization that helps in understanding the strength and direction of the associations between variables. By examining the correlation matrix, I could detect positive and negative correlations between variables. For example, there was a negative correlation between age and potential, suggesting that as players grow older, their potential tends to decrease. On the other hand, there was a positive correlation between the overall rating and potential, indicating that players with a higher overall rating tend to have higher potential.

In addition to exploring correlations, I used a count plot visualization to analyze the frequency of players in each position category. Count plots provide a clear and concise representation of the frequency within each category. In my dataset, I observed the number of players in each position, and it was surprising to find a significant number of players to be "CB" (center-back). This information provided valuable insights into the distribution of players across different positions.

To further analyze player positions, I created a new column categorizing players' best positions into "defender," "midfield," and "forward." This categorization allowed for a more meaningful analysis of position-related insights. Using a violin plot, I visualized the distribution of player ages within each position category. Violin plots are significant because they allow us to observe the shape, spread, and skewness of the data in a group while also showing quartiles and the median. The violin plot revealed that the median age for defenders, midfielders, and forwards was around 25 years old, providing insights into the age distribution among different positions.

Furthermore, I created another boxplot to examine the overall rating of players across the three position categories. It was interesting to observe that the median rating for defenders, midfielders, and forwards was around 65, suggesting similar overall performance levels across these positions. However, there were outliers in each category, indicating exceptional performances.

Returning to the previously generated correlation matrix, I focused on the correlation between overall rating and age. This correlation plot revealed a slight positive correlation, suggesting that as players become older, their overall rating tends to increase. This information provided insights into the relationship between age and overall rating and highlighted the potential impact of player experience on their performance.

Finally, to understand the distribution of each position based on players' potential, I created a kdeplot visualization. In the kde plot, the y-axis represents the estimated density of the data and the x-axis represents the estimated potential. The higher the density at a specific y axixs, the more likely it is to find data points within that range. We can clearly see that we are more likely to find players with a potential of 70 in each position category.

In conclusion, the various visualizations employed in this analysis of the FIFA 23 dataset helped uncover patterns and gain valuable insights. Each visualization served a specific purpose, such as identifying missing values, examining distributions, exploring correlations, analyzing frequencies, and understanding the relationship between age and performance.

**PART 2**

My interest is Data science. The big Data revolution as mentioned in the articles has significant implications for the field of data science. There are concepts discussed in the articles that resonate with my future field of work. The use of Big Data and artificial intelligence methods such as machine learning and predictive analytics aligns closely with what a data scientist do daily. The big data revolution brings a couple of positive impacts to the field of data science. For example, Ai can help data scientist with specific code to do a task and it can help data scientist to finish their work faster and obtain precise results, which this can improve business strategies, product development and customer experiences. I also believe that with the help of AI it can increase productivity and cost savings. While the Big Data Revolution brings good opportunities, it also presents challenges and potential negative consequences for data science. One such challenge is ensuring data privacy and security. With the collection and analysis of massive amounts of personal and sensitive information, there is an increased risk of data breaches and privacy violations. Ethical considerations surrounding data usage and potential biases in AI algorithms also need to be addressed to prevent discriminatory practices as learned in ethics in Business Analytics. Additionally, the reliance on big data and AI may lead to overreliance on automated decision-making, potentially reducing the need for human involvement and expertise in certain areas. In conclusion, The Big Data Revolution has profound parallels with the field of data science. It empowers data scientists with the tools and techniques to extract valuable insights and drive innovations. However, it is essential to navigate the potential negative consequences associated with data privacy, biases, and overreliance on automation. By addressing these challenges, data sciencist can harness the power of big data and AI.