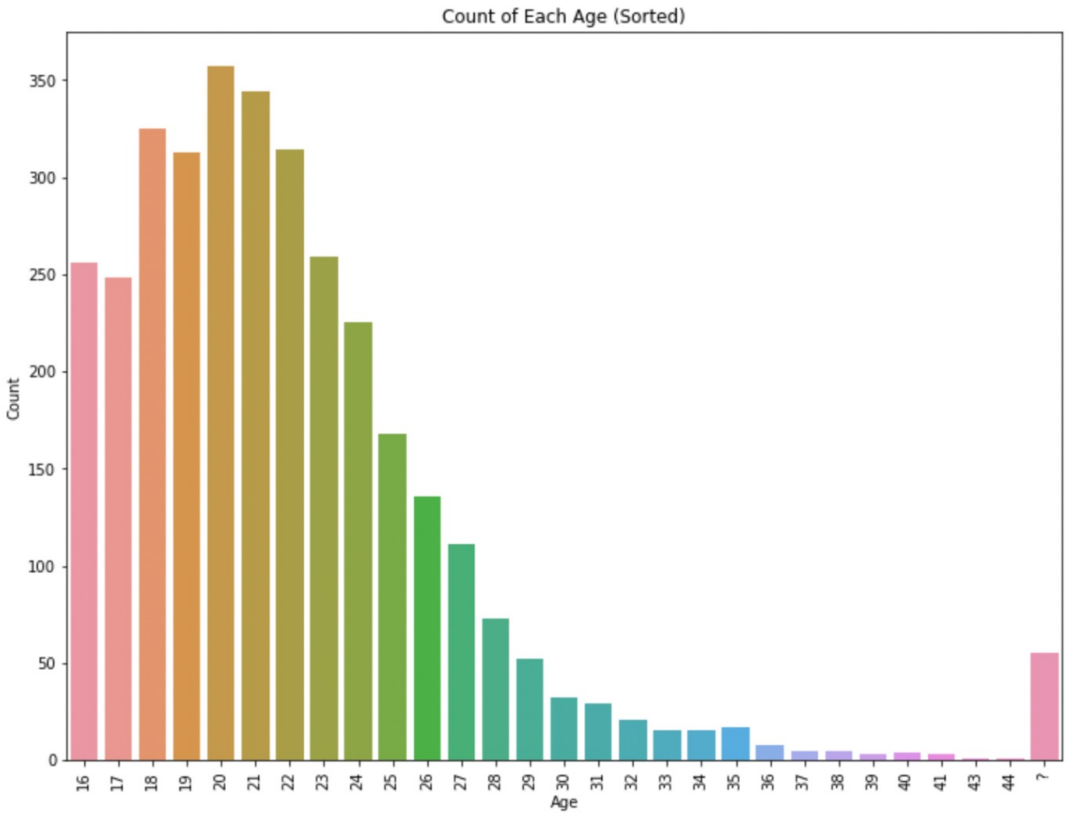
Krishna Barfiwala

Starcraft dataset- Data analytics assessment

About the data:

Data about the starcraft is given with their powers, league rank and a players rank is to be predicted based on it.

Age distribution:

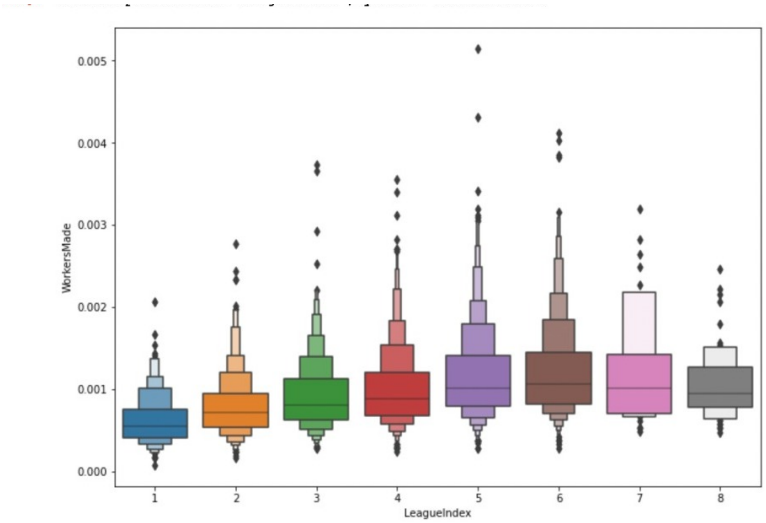


The histogram plot reveals that the age group of 18 to 21 has the highest frequency of players, indicating that players within this age range are the most actively engaged in the StarCraft game. However, when examining the distribution of age across different League Indexes, it is evident that there is a similar distribution pattern among the leagues.

The Master and GrandMaster leagues exhibit lower age distributions, with maximum age observed around 20 and 24 years, respectively. This suggests that players in these higher leagues tend to cluster around these specific age ranges. However, it is important to note that the average age of players in each league is relatively close, ranging from 20.7 to 22.7 years.

Therefore, if you fall within the age range of 18 to 25, it becomes challenging to determine your league placement based solely on age. Other factors, such as skill level, game knowledge, and experience, likely play a more significant role in determining which league a player is placed in.

League Index level and Workers Made by Players:



Labeling categorical data as numerical for visualization is important as it allows for compatibility with data visualization tools and algorithms that often require numerical inputs. This conversion enables leveraging the wide range of chart types, statistical calculations, and data analysis methods available for numerical data.

In the specific case of analyzing the relationship between League Index and Workers Made in StarCraft, labeling the League Index from Bronze to Professional as numerical values (1 to 8) facilitates visualizing and understanding the correlation between a player's skill level and their worker management. By converting the categorical League Index into numerical labels, we can effectively compare and analyze the relationship between these variables.

The relationship between League Index and Workers Made in StarCraft tends to show a positive correlation. As players progress to higher leagues, they typically demonstrate improved game mechanics and a better understanding of gameplay strategies, including the importance of worker production, expansion, and resource management.

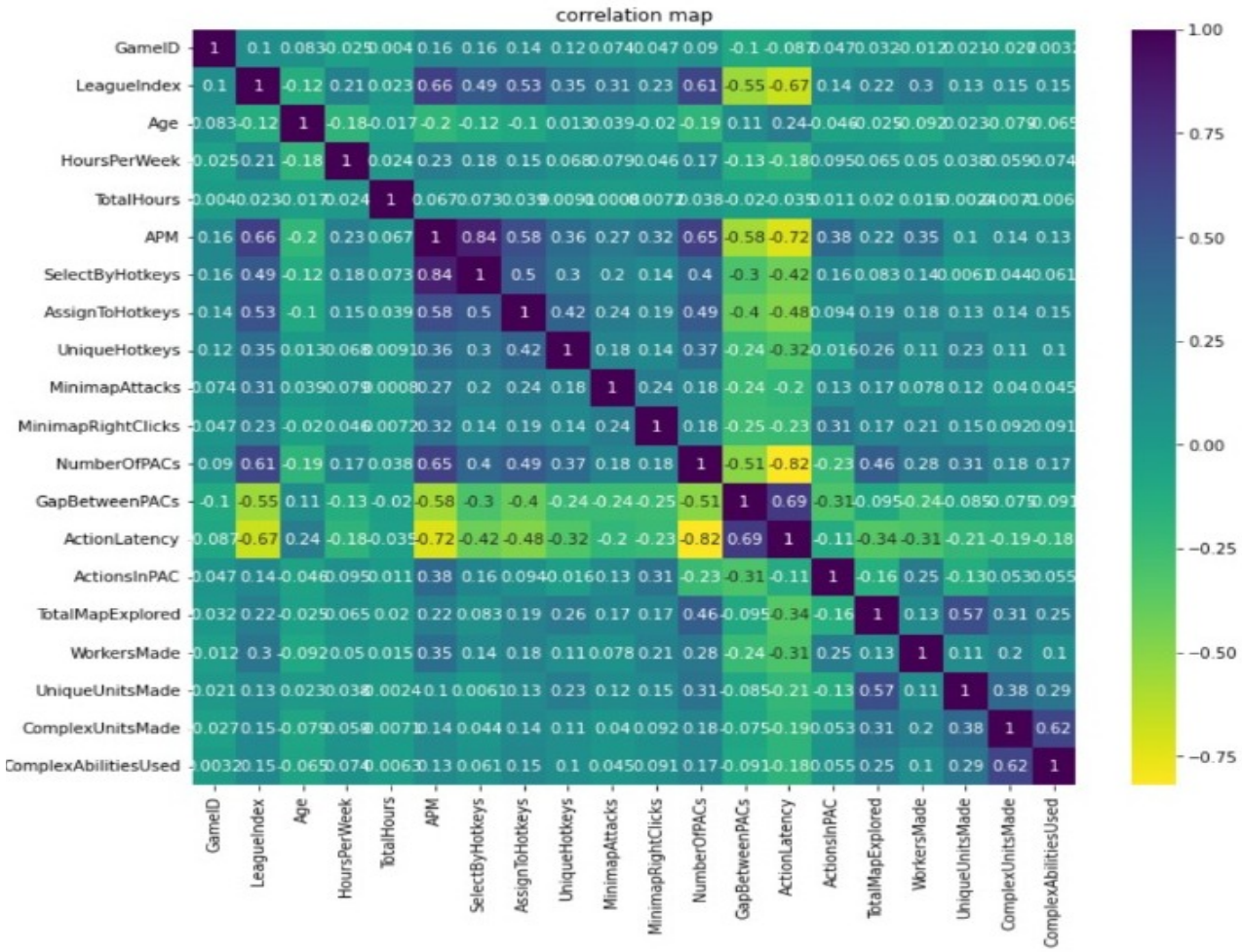
Consequently, higher-ranked players tend to have a higher average number of workers made compared to lower-ranked players. This correlation can be attributed to the fact that efficient worker management and economy optimization are critical aspects of successful gameplay in StarCraft.

Skilled players recognize the significance of maintaining a strong economy, which involves consistently producing and managing a higher number of workers. As a result,

higher-ranked players often exhibit superior worker management skills, leading to a greater number of workers made on average during their matches.

In conclusion, converting categorical data to numerical form for visualization allows for better analysis and understanding of the relationship between variables such as League Index and Workers Made in StarCraft. The positive correlation observed between these variables suggests that higher-ranked players tend to exhibit stronger worker management skills and produce a greater number of workers on average.

Check the correlation of each variables:

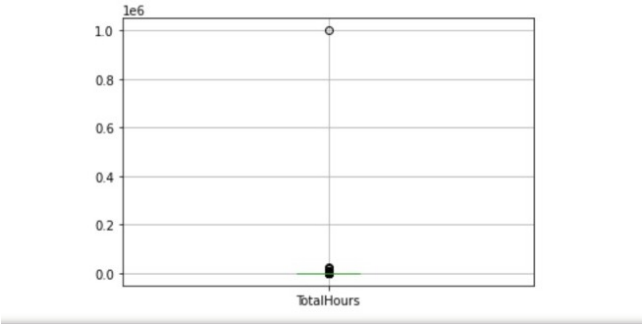


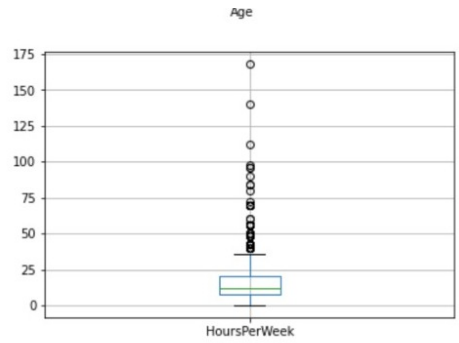
There is a high correlation between the LeagueIndex, our target variable, and APM (Action per minute). More clicks does indicate there are high chances of a player being in the higher league level.

The APM (Action per minute) is also having a high correlation with Number of PACs. More the Actions, the player will have more PACs. The Action latency is having negative correlation with the APM, as the player will be more active and there will be less delay in the game. Similarly with the Gaps Between PACs.

As the total maps explored increases, the Units made also increases with a correlation of 0.57. A player with complex abilities makes more complex units.

Checking the distribution or spread of the Total Hours and Hours per week:





Removing outliers improves data quality by eliminating erroneous or unusual values. It ensures statistical assumptions are met, leading to more reliable analysis and results. It also enables clearer data visualization by reducing distortion and emphasizing the main patterns and trends. The total hours played by a player in the graph is visible to be around 10000. It is practically impossible or impractical to have such a high number of hours by any player. Hence we remove that data point.

While the majority of the values in the dataset are concentrated between 0 and 100, there is a noticeable number of values that fall outside the upper whisker, indicating potential outliers.
Total hours per week is 168 and in the graph it is visible there are few data points close to 140 and 170, it is again impractical assuming it is not a good thing, we remove the data point where hours per week is greater than 140. Eliminating these outliers ensures that the analysis focuses on the central tendencies and reduces the influence of extreme values on statistical measures and visualizations.

Checking Normality of the data:

```
In [53]: #Normality check
         from scipy.stats import shapiro
         for column in data.columns:
             stat, p_value = shapiro((data[column]))
             print(f"Column: {column}, Shapiro-Wilk test statistic: {stat:.4f}, p-value: {p_value:.4f}")
```

```
Column: GameID, Shapiro-Wilk test statistic: 0.9612, p-value: 0.0000
Column: LeagueIndex, Shapiro-Wilk test statistic: 0.9469, p-value: 0.0000
Column: Age, Shapiro-Wilk test statistic: 0.9218, p-value: 0.0000
Column: HoursPerWeek, Shapiro-Wilk test statistic: 0.8410, p-value: 0.0000
Column: TotalHours, Shapiro-Wilk test statistic: 0.5024, p-value: 0.0000
Column: APM, Shapiro-Wilk test statistic: 0.9292, p-value: 0.0000
Column: SelectByHotkeys, Shapiro-Wilk test statistic: 0.6772, p-value: 0.0000
Column: AssignToHotkeys, Shapiro-Wilk test statistic: 0.9378, p-value: 0.0000
Column: UniqueHotkeys, Shapiro-Wilk test statistic: 0.9711, p-value: 0.0000
Column: MinimapAttacks, Shapiro-Wilk test statistic: 0.5856, p-value: 0.0000
Column: MinimapRightClicks, Shapiro-Wilk test statistic: 0.7848, p-value: 0.0000
Column: NumberOfPACs, Shapiro-Wilk test statistic: 0.9835, p-value: 0.0000
Column: GapBetweenPACs, Shapiro-Wilk test statistic: 0.8870, p-value: 0.0000
Column: ActionLatency, Shapiro-Wilk test statistic: 0.9405, p-value: 0.0000
Column: ActionsInPAC, Shapiro-Wilk test statistic: 0.9166, p-value: 0.0000
Column: TotalMapExplored, Shapiro-Wilk test statistic: 0.9763, p-value: 0.0000
Column: WorkersMade, Shapiro-Wilk test statistic: 0.8826, p-value: 0.0000
Column: UniqueUnitsMade, Shapiro-Wilk test statistic: 0.9716, p-value: 0.0000
Column: ComplexUnitsMade, Shapiro-Wilk test statistic: 0.6127, p-value: 0.0000
Column: ComplexAbilitiesUsed, Shapiro-Wilk test statistic: 0.5806, p-value: 0.0000
```

Using Statistical hypothesis testing,
H0: The data follows a normal distribution,
H1: The data does not follow a normal distribution.
Our goal is to assess whether there is sufficient evidence to reject the null hypothesis and accept the alternative hypothesis. If p-value < 0.05 we reject H0.

In this case, we reject the H0, suggesting that the data does not follow a normal distribution. Conversely, if the p-value is greater than the threshold, we fail to reject the null hypothesis, implying that there is not enough evidence to conclude that the data significantly deviates from a normal distribution.

By comparing the p-value to the chosen threshold, we can make an informed decision regarding the normality of the data and whether it is appropriate to proceed with statistical analyses or models that assume a normal distribution.

By trying to transform the data that is log transformation, It turns out many variables follows normal distribution.

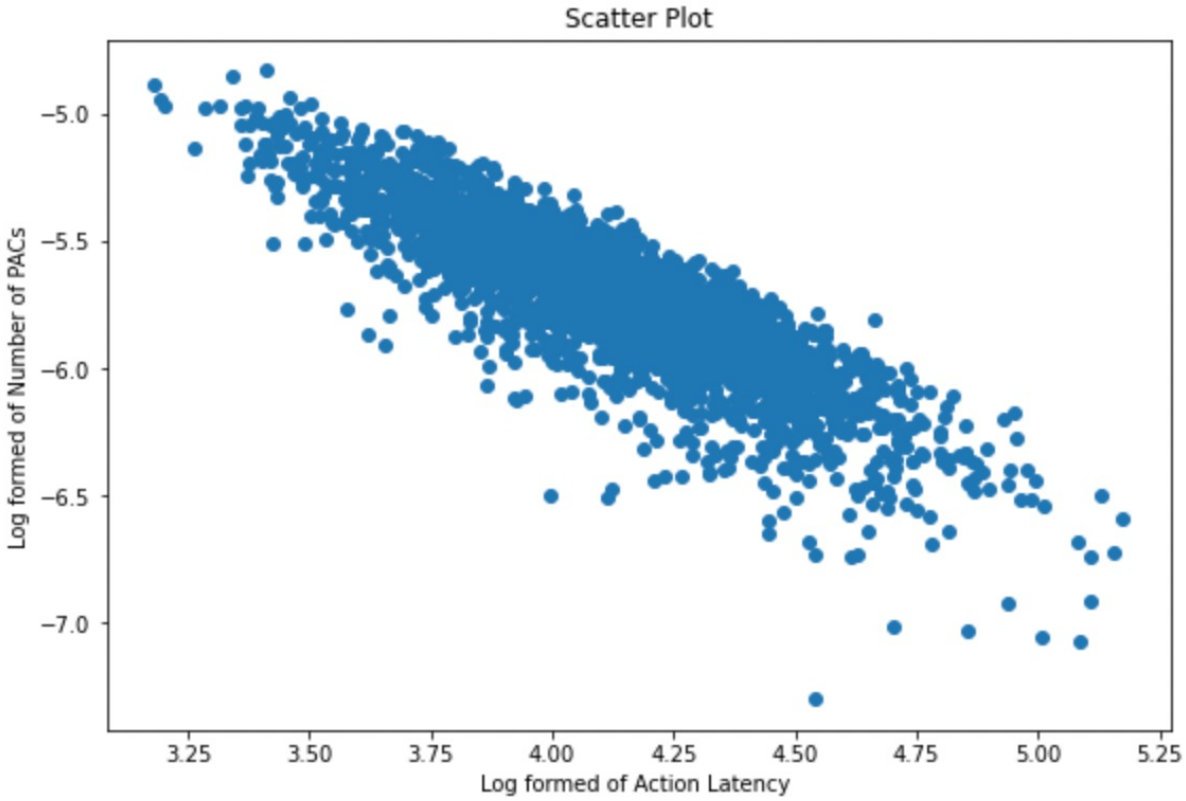
In [54]:
```
#Normality check for log transformed data
from scipy.stats import shapiro
for column in data.columns:
    stat, p_value = shapiro(np.log(data[column]))
    print(f"Column: {column}, Shapiro-Wilk test statistic: {stat:.4f}, p-value: {p_value:.4f}")
```

```
Column: GameID, Shapiro-Wilk test statistic: 0.8492, p-value: 0.0000
Column: LeagueIndex, Shapiro-Wilk test statistic: 0.8605, p-value: 0.0000
Column: Age, Shapiro-Wilk test statistic: 0.9665, p-value: 0.0000
Column: HoursPerWeek, Shapiro-Wilk test statistic: nan, p-value: 1.0000
Column: TotalHours, Shapiro-Wilk test statistic: 0.9500, p-value: 0.0000
Column: APM, Shapiro-Wilk test statistic: 0.9982, p-value: 0.0008
Column: SelectByHotkeys, Shapiro-Wilk test statistic: nan, p-value: 1.0000
Column: AssignToHotkeys, Shapiro-Wilk test statistic: nan, p-value: 1.0000
Column: UniqueHotkeys, Shapiro-Wilk test statistic: nan, p-value: 1.0000
Column: MinimapAttacks, Shapiro-Wilk test statistic: nan, p-value: 1.0000
Column: MinimapRightClicks, Shapiro-Wilk test statistic: nan, p-value: 1.0000
Column: NumberOfPACs, Shapiro-Wilk test statistic: 0.9876, p-value: 0.0000
Column: GapBetweenPACs, Shapiro-Wilk test statistic: 0.9979, p-value: 0.0002
Column: ActionLatency, Shapiro-Wilk test statistic: 0.9986, p-value: 0.0045
Column: ActionsInPAC, Shapiro-Wilk test statistic: 0.9948, p-value: 0.0000
Column: TotalMapExplored, Shapiro-Wilk test statistic: 0.9866, p-value: 0.0000
Column: WorkersMade, Shapiro-Wilk test statistic: 0.9977, p-value: 0.0001
Column: UniqueUnitsMade, Shapiro-Wilk test statistic: 0.9451, p-value: 0.0000
Column: ComplexUnitsMade, Shapiro-Wilk test statistic: nan, p-value: 1.0000
Column: ComplexAbilitiesUsed, Shapiro-Wilk test statistic: nan, p-value: 1.0000
```

```
/Users/krishnabarfiwala/opt/anaconda3/lib/python3.9/site-packages/pandas/core/arraylike.py:397: RuntimeWarning: divid
e by zero encountered in log
  result = getattr(ufunc, method)(*inputs, **kwargs)
```

After applying a log transformation to the data, several columns, like HoursPerWeek, SelectByHotkeys, AssignToHotkeys, UniqueHotkeys, MinimapAttacks, MinimapRightClicks, ComplexUnitsMade, and ComplexAbilitiesUsed, exhibit a distribution that closely resembles a normal distribution.

The log transformation helps in reducing skewness and compressing extreme values, thereby making the data more symmetrical and bell-shaped. This finding indicates that these variables may be suitable for statistical analyses or modeling techniques that assume a normal distribution.



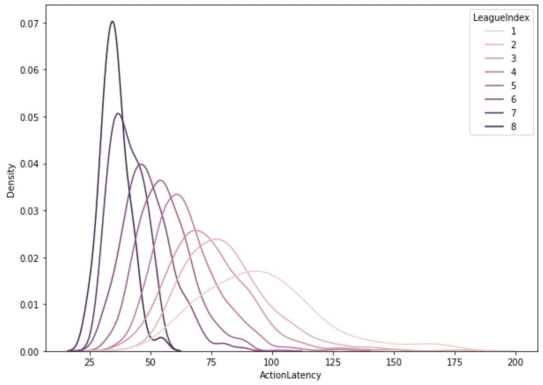
The scatterplot between the logarithm-transformed values of "Action Latency" and "Number of PACs" reveals an inverse relationship. This means that as the logarithm of the action latency increases, the logarithm of the number of PACs decreases, or vice versa. This inverse relationship suggests that there is a trade-off between action latency and the number of PACs. It indicates that players who exhibit lower action latency tend to have a higher number of PACs, reflecting their ability to perform more actions within a given time frame.

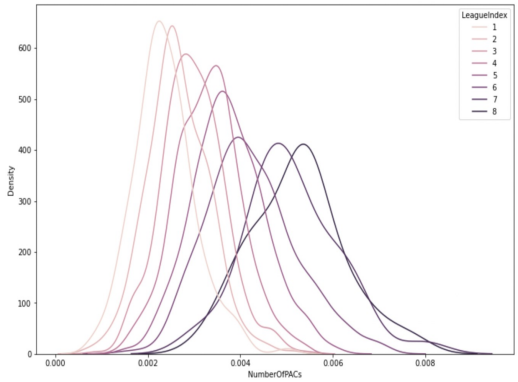
On the other hand, players with higher action latency tend to have a lower number of PACs, indicating a slower response time and potentially limited ability to execute a large number of actions. The observed inverse relationship can have implications for gameplay strategies and performance in StarCraft. Players with faster response times (lower action latency) may have a competitive advantage by being able to execute more actions and make quicker decisions. Conversely, players with higher action latency may need to focus on optimizing their actions and decision-making to compensate for the reduced number of PACs.

- The variables 'LeagueIndex', 'Age', 'TotalHours', 'APM', 'NumberOfPACs', 'GapBetweenPACs', 'ActionLatency', 'ActionsInPAC', 'TotalMapExplored', 'WorkersMade', and 'UniqueUnitsMade' do not exhibit a normal distribution. To address this, a log transformation was applied to these variables, allowing for a more symmetrical and bell-shaped distribution. Log transformation helps reduce skewness and compress extreme values, making the data more amenable to statistical analyses and modeling techniques that assume normality.

EDA:

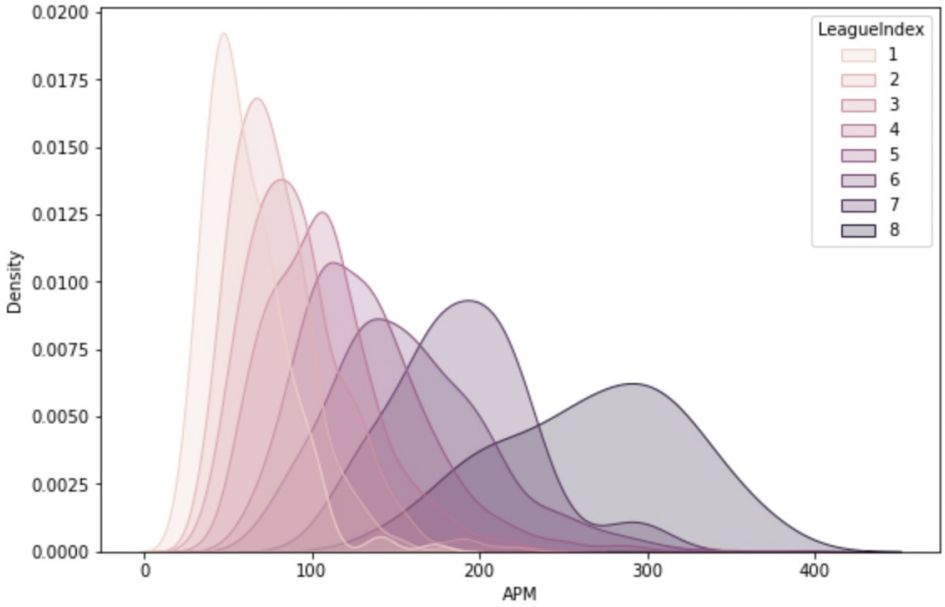
- Action latency and Number of PACs





Players at higher skill levels in StarCraft exhibit faster reaction times and lower action latency compared to players in lower leagues. This could be a result of their experience, training, and familiarity with the game mechanics, allowing them to respond more swiftly and execute actions efficiently. The lower latency observed in higher league players may contribute to their ability to make quick decisions, perform actions promptly, and gain a competitive advantage in gameplay. The number of PACs made is also proportionate to the league index. Higher the league index, higher is the Number of PACs made. The distribution of PACs is spread at a visible level at higher league levels.

- APM (Action Per minute)



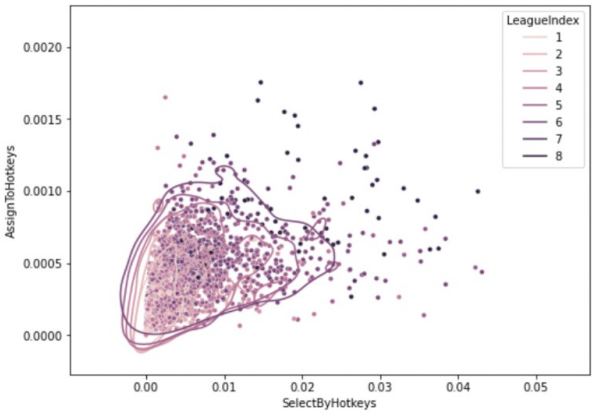
Relationship between density and APM (Actions Per Minute) suggests that as APM increases, the density decreases. This means that players with higher APM tend to have a lower density, indicating a lower concentration of data points in specific regions. This observation can be interpreted in a couple of ways.

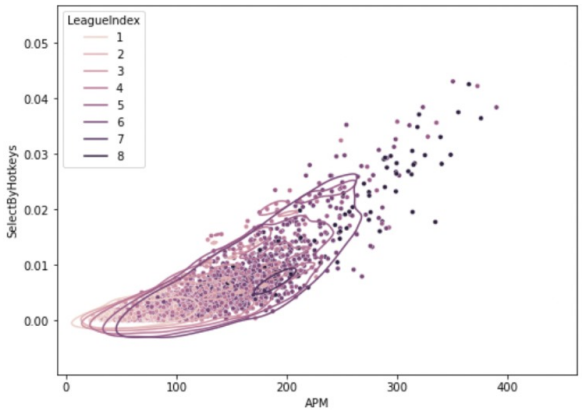
Firstly, higher APM implies faster and more frequent actions by the player. This increased speed and activity may lead to a broader distribution of actions across the game, resulting in a lower density in specific areas.

Secondly, players with higher APM are often more skilled and experienced, capable of executing complex strategies and tactics efficiently. As a result, their actions may be more diverse and spread out, reducing the density within specific gameplay regions.

The decreasing density with higher APM suggests that players with faster and more proficient gameplay tend to explore a wider range of actions and strategies, rather than focusing on specific areas. This observation highlights the dynamic and versatile nature of gameplay at higher APM levels.

- SelectByHotKeys and APM distribution





The probability distributions of "Assign To Hotkeys," "Select By Hotkeys," and APM (Actions Per Minute) indicate interesting correlations. The positive correlation between "Assign To Hotkeys" and "Select By Hotkeys" is intuitive since players who take the effort to assign a hotkey are likely to utilize it frequently during gameplay.

Furthermore, the strong correlation between "Select By Hotkeys" and APM can be attributed to the fact that each selection made using a hotkey contributes to the APM count. Therefore, players with higher APM tend to have a larger number of hotkey selections, reflecting the efficient use of hotkeys in their gameplay.

Probability distributions of these variables exhibit wider ranges in higher leagues compared to the Bronze league. This implies that players in higher leagues have a greater diversity of actions and strategies, allowing for a broader distribution of values. In contrast, the Bronze league exhibits narrower ranges, indicating more restricted gameplay patterns and a reduced variety of actions.

The widening distributions in higher leagues suggest that players at advanced skill levels possess a wider repertoire of gameplay techniques and are more adaptable to various situations. This emphasizes the dynamic and versatile nature of gameplay in top leagues, where players showcase a broader range of actions and strategies compared to their counterparts in the Bronze league.

These observations highlight the progression of gameplay complexity and diversity as players ascend through different leagues. The widening distributions reflect the increased skill, knowledge, and adaptability of players in higher leagues, contributing to a more varied and nuanced gameplay experience.

Understanding these patterns in the probability distributions provides valuable insights into the characteristics and gameplay dynamics across different skill levels, aiding in the assessment and comparison of player performance within the game.

Modeling:

Using Logistic regression, random forest, naive bayes, I tried to predict the accuracy and league index of the players. A better accuracy can be achieved with better and cleaned data.

Conclusion:

After conducting thorough exploratory data analysis (EDA), several significant variables have emerged as influential factors in determining a player's rank. The "LeagueIndex" column stands out as a direct indicator of a player's rank, representing their league index or ranking within the game. This variable serves as a fundamental factor in assessing a player's overall skill level and determining their rank.

Additionally, the "APM (Actions Per Minute)" metric plays a crucial role. Higher APM values generally correlate with faster and more efficient gameplay, showcasing superior multitasking abilities and quicker strategy execution. A high APM positively influences a player's rank by demonstrating their proficiency in gameplay mechanics.

The "TotalHours" column reflects the total number of hours a player has dedicated to the game. This metric serves as an indicator of experience and dedication, as players who invest more time in the game gain a deeper understanding of its mechanics and strategies, leading to an improvement in their rank.

The "UniqueHotkeys" variable is another important consideration. It represents a player's ability to effectively control and manage units and actions using unique hotkeys. A broader range of unique hotkeys demonstrates advanced skill in executing complex strategies, which can positively impact a player's rank.

Efficient worker production, represented by the "WorkersMade" column, is crucial for resource gathering and maintaining a strong in-game economy. Excelling in worker management contributes to achieving a higher rank, as it enhances a player's ability to efficiently allocate resources and maximize productivity.

Furthermore, the "UniqueUnitsMade" column signifies a player's versatility and adaptability in gameplay. Effectively utilizing a variety of units demonstrates skill in countering opponents' strategies, leading to improved performance and a higher rank.

The "TotalMapExplored" column represents the extent to which a player has explored the game map. A higher value indicates that he has explored more areas, allowing him to have better knowledge of the game environment and make informed decisions. This factor can positively influence my rank by demonstrating my map awareness and strategic planning.

While modeling, I have chosen league index as the target variable, as that variable is the closest to predict the rank of any player. Doing basic modeling with Logistic regression, Random forest and naive bayes, I predicted the target variable. If the league index falls in higher categories like Master, GrandMaster or Professional leagues along with parameters mentioned above like high APM, higher SelectByHotkeys, lower ActionLatency, WorkersMade then the player can be classified in higher ranks or the top rankers.

Based on the provided dataset, the proposed approach for determining a player's rank involves predicting their League based on available data and subsequently applying a rule-based engine to differentiate and rank players within each league. This rule-based engine, developed using domain knowledge and key factors identified through EDA, enables accurate assessment and ranking based on individual performance, skill level, and overall gameplay strategies. By considering these significant variables and implementing the rule-based engine, a

comprehensive evaluation of a player's rank can be achieved, providing valuable insights into their performance and positioning within the game.