**Final Project**

Muhammad U. Mirza

College of Professional Studies, Northeastern University Toronto

ALY 6010: Probability Theory and Introductory Statistics

Dr. Steven Habbous

December 17, 2023

I selected the "Football Wages Data" dataset from www.kaggle.com for my final project. This dataset contains the 2022 wage information for almost 3,900 players from six European leagues. The data was collected from the Football Manager 2022 game, which is a popular tool used by football enthusiasts and professionals to simulate football management. My initial analysis involved exploratory data analysis, leading to hypothesis testing to understand the factors influencing player wages. The report focuses on the relationship between a player’s wage, league, position, and club and international appearances frequency.

**Data Inspection and Cleaning**

The dataset comprises 3,907 entries across 8 columns detailing players' wages, age, club, league, nationality, position, appearances, and international caps. In response to feedback from previous milestones, I refined the dataset to enhance its compatibility with R. This involved renaming columns for clarity 'apps' to 'club\_apps' and 'caps' to 'international\_apps' and formatting the 'wage' column cleaned of commas and converted to numeric values; 'league' and 'position' set as categorical factors (see Appendix 1). No missing values were detected. Heeding earlier suggestions and the fact that football wage contracts are short-term and highly dependent on current form, I excluded anomalously high wages, specifically those above $10 million, resulting in a 1.7% data reduction. The final, cleaned dataset contains 3,841 records.

**Exploratory Data Analysis**

The descriptive statistics indicate a median wage of $381,000 against an average of $1,139,612, indicating a skewed distribution. However, since the data appears to be skewed, the outliers identified by the box plot are not considered true outliers. Players' ages cluster around a median of 23, with a range from 18 to 41. The median club appearances stand at 111, with a majority having under 220 appearances. International appearances are less frequent, with a median of 0 and 75% of players having 5 or fewer caps. The Premier League and Primiera Liga are the most represented leagues, with defenders and midfielders being the most common positions (see Appendix 2).

The descriptive statistics reveal the financial complexities of European football, with wages varying widely, as evidenced by the significant skew in their distribution. Median values for age and club appearances hint at a cohort of players not yet at the peak of their careers, possibly reflecting the dynamic nature of the sport where emerging talents rise alongside established professionals. This diversity in career stages mirrors the financial stratification within the industry, where a few high earners may inflate average wage figures. Such disparities underscore the multifaceted economic environment that footballers navigate.

It is important to note that, despite the right skew in wage distribution, the dataset's size justifies parametric testing under the central limit theorem. This theorem indicates that for large samples, typically over 30, the mean will roughly follow a normal distribution, allowing for valid statistical analysis.

**Data Analysis**

The EDA highlighted a predominant representation of players from the Premier League and Primiera Liga, as well as a higher frequency of defenders and midfielders within the dataset. This observation prompted further investigation into the wage dynamics across these subsets. The ensuing research questions aimed to dissect whether these categorical distinctions in the league and playing position bear a statistically significant impact on the average wages of the players in question.

The regression analysis was a natural progression from the EDA, seeking to quantify the influence of age, club appearances, and international caps on players' wages. These variables were highlighted in the EDA as potential indicators of a player's value. By applying regression, the goal is to move beyond descriptive statistics and explore the predictive power of these factors, offering insights into how each one could be leveraged to understand wage variations within the professional football landscape.

***Research Question: At a 0.05 alpha, is there a statistically significant difference in average wages between defenders and midfielders?***

H0: μ\_defenders = μ\_midfielders

H1: μ\_defenders ≠ μ\_midfielders

An independent two-tailed t-test was performed to compare the average wages of defenders and midfielders. This approach was suitable as the wages between these two categories are independent of each other. Preliminary F-test results suggested unequal variances between groups, justifying the use of Welch's t-test, which does not assume equal population variances (see Appendix 3).

In the t-test comparing defenders' and midfielders' wages, the point estimate, reflecting the average wage difference, is $168,142, with midfielders earning more. The t-test yielded a t-value of -2.348 and a p-value of 0.01896, below the alpha level of 0.05, leading to the rejection of the null hypothesis. The 95% confidence interval ranges from approximately -$308,570 to -$27,714, affirming the significance of this wage disparity. This outcome suggests that the observed wage difference between these positions is statistically significant and not a result of random variation.

***Research Question: At a 0.05 alpha, is there a statistically significant difference in average wages between the Premier League and Primiera Liga?***

H0: μ\_premier\_league = μ\_primiera\_liga

H1: μ\_premier\_league ≠ μ\_primiera\_liga

An independent two-tailed t-test was performed to compare the average wages of Premier League and Primiera Liga players. This approach was suitable as the wages between these two categories are independent of each other. Preliminary F-test results suggested unequal variances between groups, justifying the use of Welch's t-test, which does not assume equal population variances (see Appendix 4).

In the Welch two-sample t-test comparing Premier League and Primiera Liga wages, the point estimate for the average wage difference is approximately $1,814,414, with Premier League players earning more. The t-test gives a t-value of 22.483 and a p-value significantly less than 0.05, warranting a rejection of the null hypothesis. The 95% confidence interval, ranging from about $1,656,031 to $1,972,796, supports this significant wage difference. This finding confirms a statistically significant difference in wages between Premier League and Primiera Liga, highlighting the economic contrasts within European football.

***Research Question: Does age, number of club appearances, and number of international appearances significantly predict the wages of football players in top European leagues?***

The scatterplot matrix reveals visual correlations between variables, where age and wages show a positive trend, indicating higher wages with increasing age (see Appendix 5). The correlation matrix quantifies these relationships; for instance, a moderate correlation of 0.45 between club appearances and wages suggests players with more appearances may earn higher wages. Other notable correlations include age with international appearances (0.44) and a very strong correlation between age and club appearances (0.93), indicating a relationship between experience and longevity in a player's career. However, these are purely observational and warrant further investigation to understand causality.

***Model 1: Wage = ß0 + ß1\*age + ß2\*club\_apps + ß3\*international\_apps***

The regression model further explores these relationships by quantifying the impact of each independent variable on wages (see Appendix 5). Each variable's impact on wages is statistically significant, as indicated by p-values far less than the alpha level of 0.05. This confirms that age, club appearances, and international appearances are all predictors of a player's wage within this dataset, with the data providing strong evidence against the null hypothesis that these variables have no effect. The model indicates that each additional year of age is associated with a decrease in wage by approximately $125,673, holding other factors constant. Conversely, both club and international appearances have a positive impact on wages, with each club appearance adding about $8,150 to a player’s wage and each international appearance adding around $32,100, suggesting that more on-field experience correlates with higher wages.

The 95% confidence intervals from the regression model provide ranges for each parameter's estimated value, offering a measure of their precision. For instance, the interval for 'age' suggests that each additional year is linked to a wage decrease between $150,430 and $100,915. Similarly, each extra club appearance could increase wages by $7,175 to $9,125, while each international appearance might add between $29,242 and $34,957 to wages. These intervals help gauge the reliability of the model's estimates, indicating the likely range within which the true values of these coefficients fall.

***Wage = 2789971.8 − 125672.8\*(age) + 8150.1\*(club\_apps) + 32099.6\*(international\_apps)***

The regression model, while quantifying the relationships between age, club appearances, international appearances, and wages, reveals significant unexplained variance. The sizeable residuals suggest that factors outside the model may be influencing wages, indicating the model's limitations. With the observed wide range in residuals (minimum of -7,056,141 to a maximum of 8,566,302) and a weak R-squared value of 0.3151, it's clear that the model does not fully capture wage dynamics. This necessitates a cautious approach in interpreting the model and considering additional variables that may affect a player's wage.

Overall the data analysis suggests a significant influence of playing position and league on wages, with midfielders and Premier League players earning higher wages. The regression model confirmed the positive impact of experience on wages, although it also indicated a potential peak wage age. These insights could inform contract negotiations and career management for players.

**Conclusion**

The analysis began by exploring the influence of age, playing position, league, and club and national appearances on players' wages. Utilizing hypothesis testing, the data revealed distinct financial patterns across positions and leagues. For instance, a two-tailed t-test demonstrated significant wage differences between defenders and midfielders, reflecting positional value disparities in European football. Similarly, a comparison between the Premier League and Primiera Liga wages highlighted the economic divides among leagues.

Regression analysis further quantified the impacts of age and appearance. The model indicated that each additional international appearance correlates with a wage increase, showcasing the premium on international experience. Age, had a subtler effect, suggesting that peak earning potential might not align strictly with age.

These findings narrate the economic story of European football, where market factors such as position, performance, and reputation interplay to determine a player's financial worth. Despite the range in wages and careers, the central limit theorem assured the robustness of this statistical journey, guiding the conclusions drawn from this diverse dataset.

This project has been an illuminating challenge, a true synthesis of the theoretical and practical aspects of probability and statistics. It pushed the boundaries of my understanding, allowing me to delve into the intricacies of data analysis and apply statistical principles to real-world scenarios. The journey from conceptual frameworks to the tangible narratives of European football's financial dynamics has underscored the profound capabilities of data-driven storytelling in sports economics

**References**

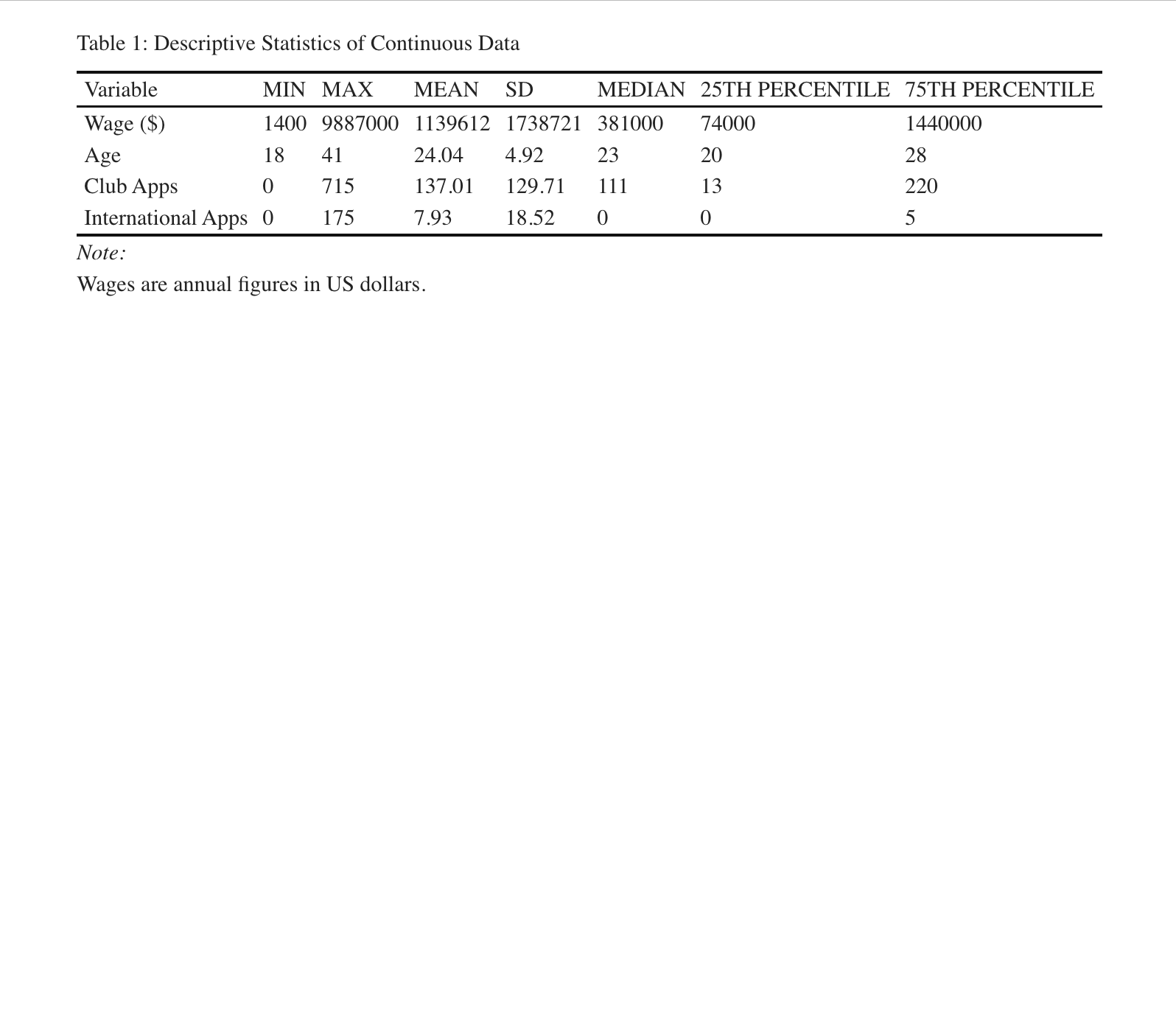
1. Bluman, A. (2018). Elementary statistics: A step by step approach (10th ed.). McGraw Hill.ISBN 13: 978-1-259-755330.
2. Habbous, S. (2023). Module 1-6. Canvas. <https://northeastern.instructure.com/courses/160615/modules>
3. Yash. (2023). Football Wages Dataset. Www.kaggle.com. <https://www.kaggle.com/datasets/ultimus/football-wages-prediction/data>
4. Yaldo, L., & Shamir, L. (2017). Computational Estimation of Football Player Wages. International Journal of Computer Science in Sport, 16(1), 18–38. <https://doi.org/10.1515/ijcss-2017-0002>
5. Paetzold, Ramona. (2016). Re: Low R-squared values in multiple regression analysis?. Retrieved from: <https://www.researchgate.net/post/Low-R-squared-values-in-multiple-regression-analysis/56f40be793553b1f9a11fcb3/citation/download>.
6. Frost, J. (2019, June 13). How to Interpret Regression Models that Have Significant Variables but a Low R-squared - Statistics By Jim. Statistics by Jim. <https://statisticsbyjim.com/regression/low-r-squared-regression/>
7. Valchanov, I. (2018, November 7). Point Estimate and Confidence Interval. 365 Data Science. <https://365datascience.com/tutorials/statistics-tutorials/point-estimates-confidence-intervals/>

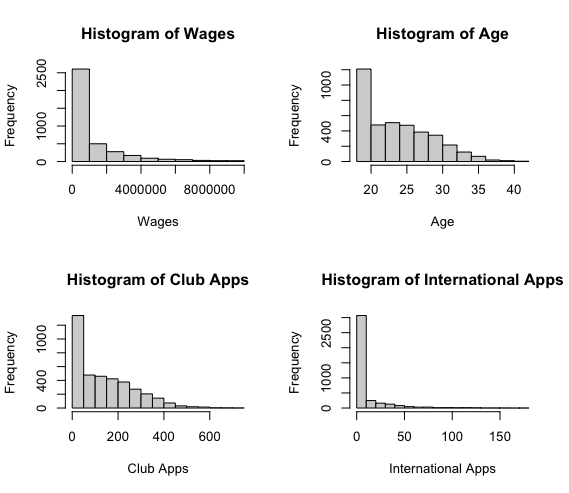
**Appendix**

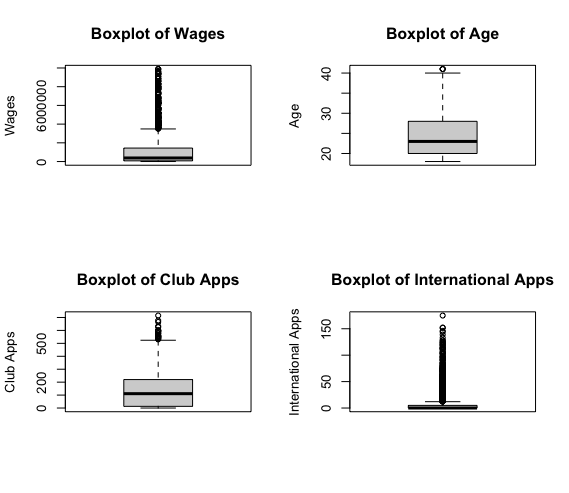
1. Variable description and structure.

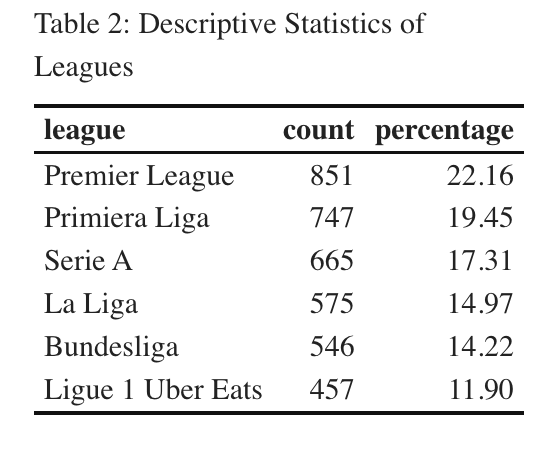
| **Variable** | **Description** | **Data Type** |
| --- | --- | --- |
| wage | The annual salary of each player in US dollars. | Numerical |
| age | Age of each player in years. | Integer |
| club | Name of each club in the top 6 leagues of Europe. | Character |
| league | Name of the top 6 leagues in Europe. | Factor |
| nation | Players' country of origin. | Character |
| position | The four playing positions in football. | Factor |
| club\_apps | Number of club appearances for each player. | Integer |
| international\_caps | Number of international appearances for each player. | Integer |

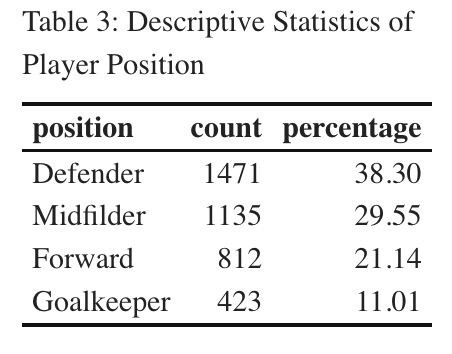
1. Descriptive statistics and distribution of variables



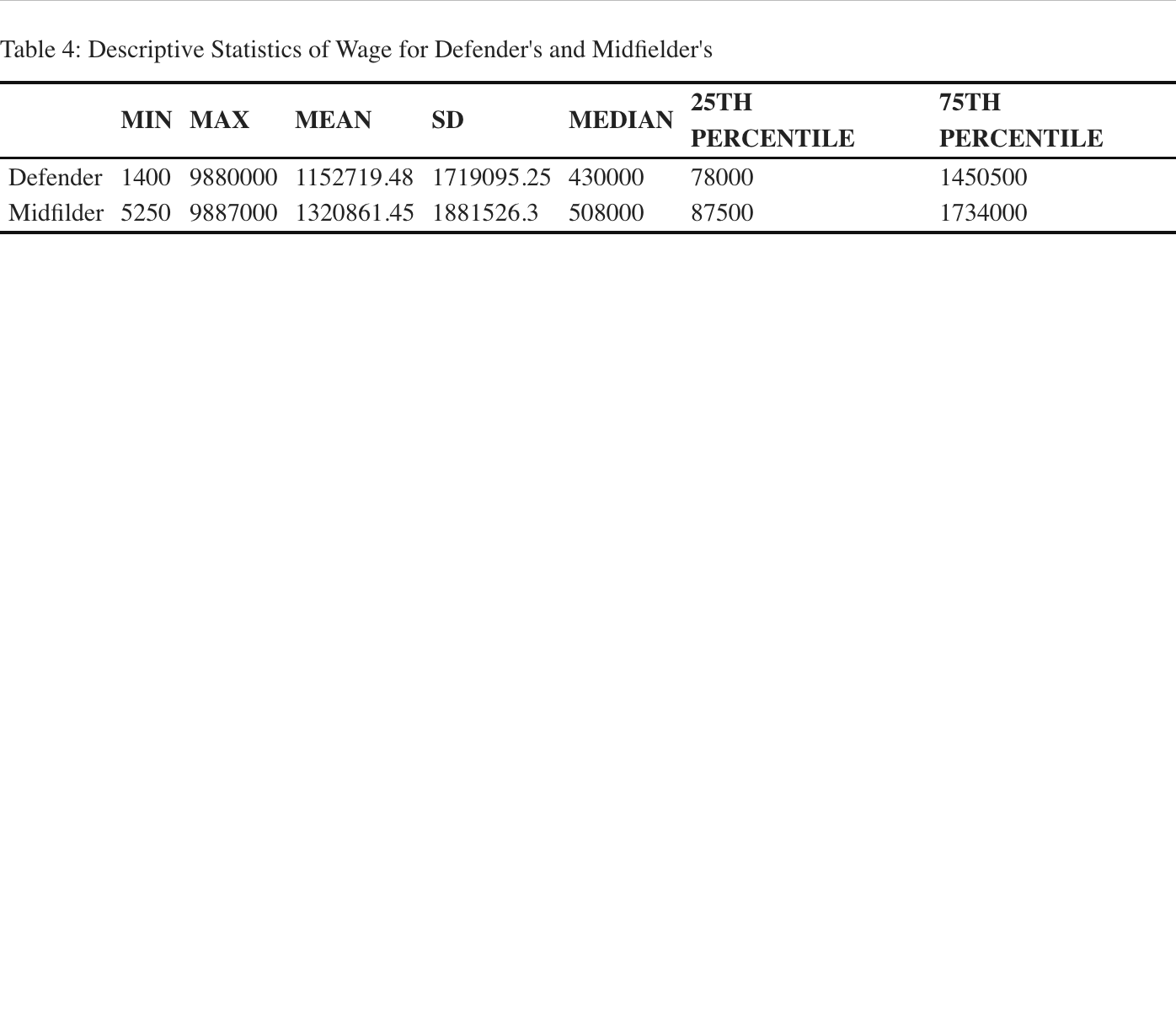


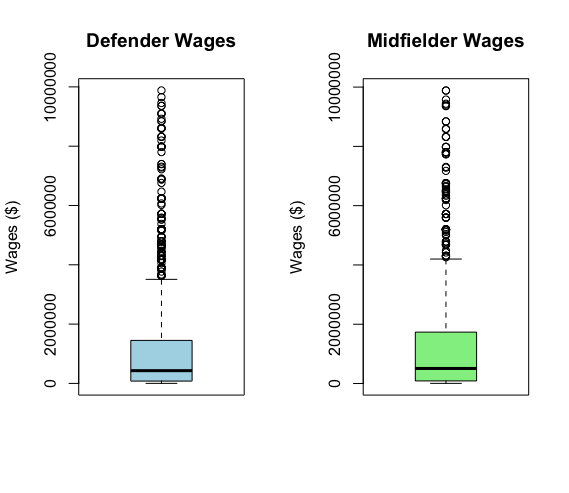


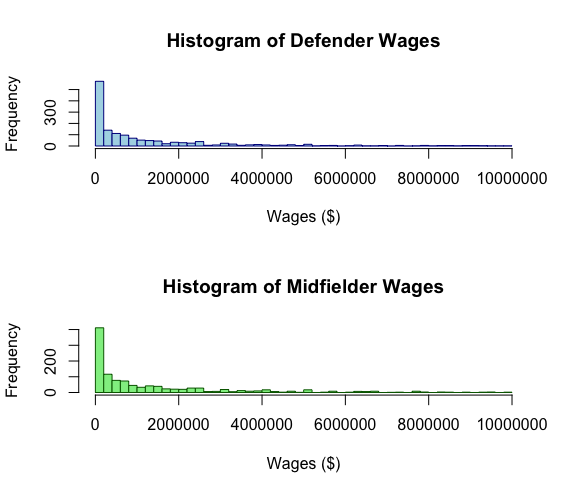


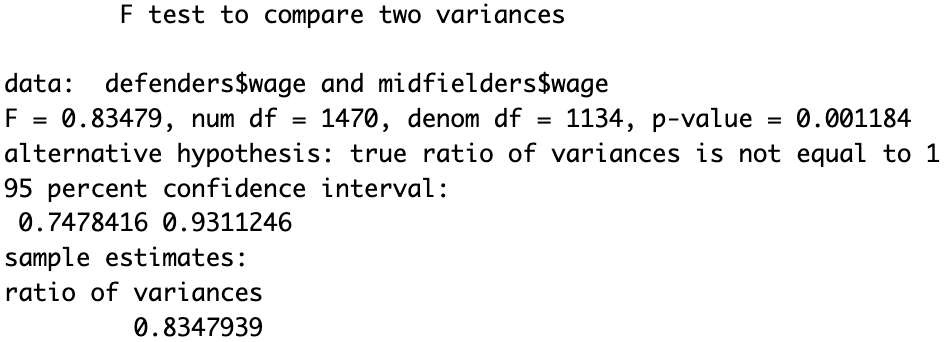


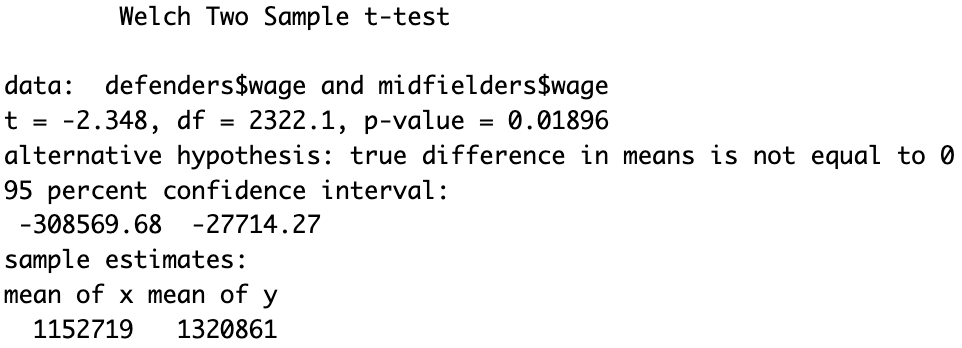
1. 1st research question: descriptive statistics, distribution, f-test, and t-test



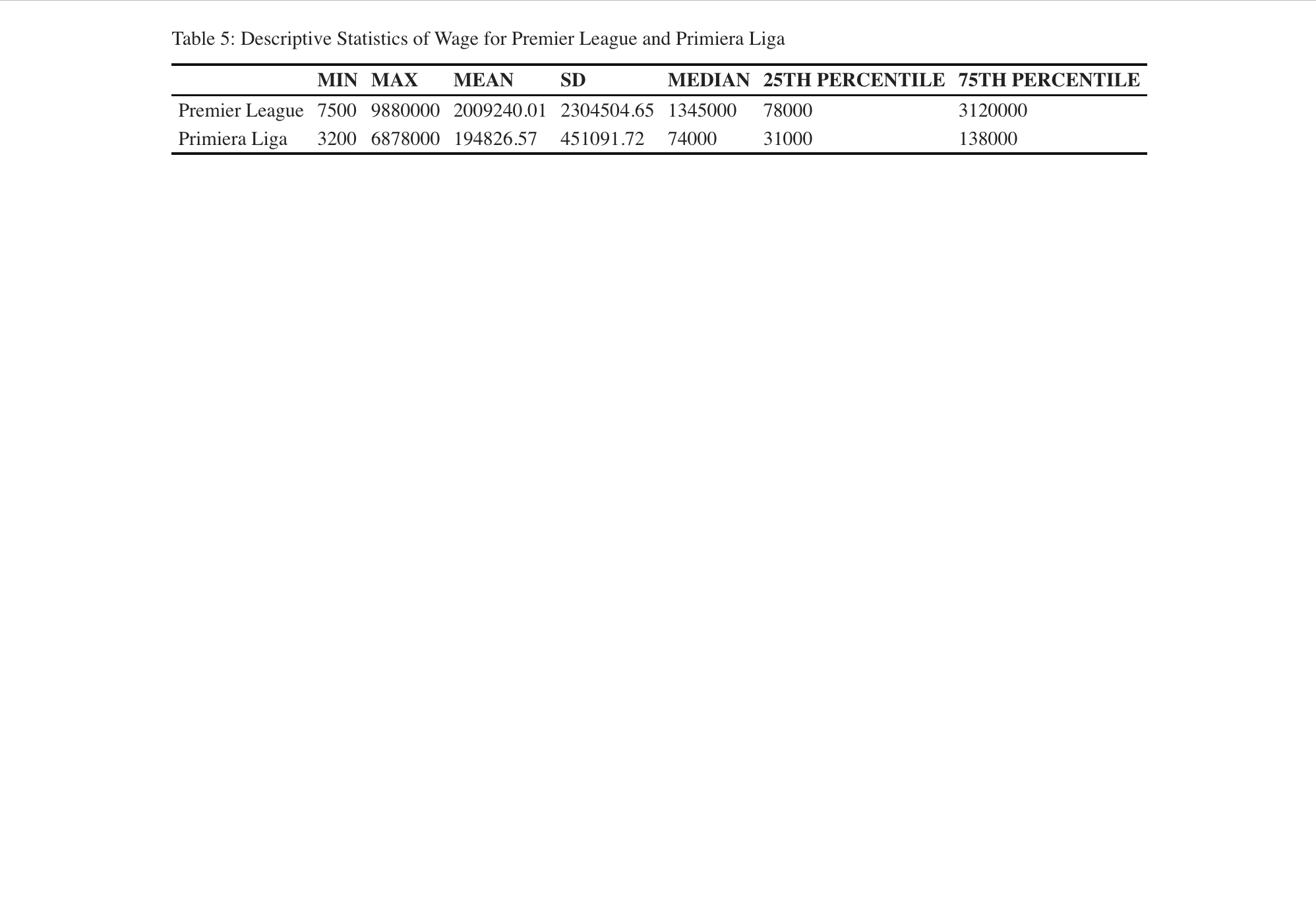


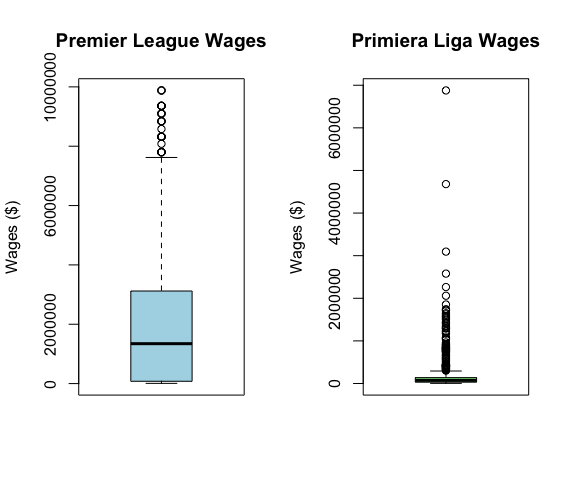


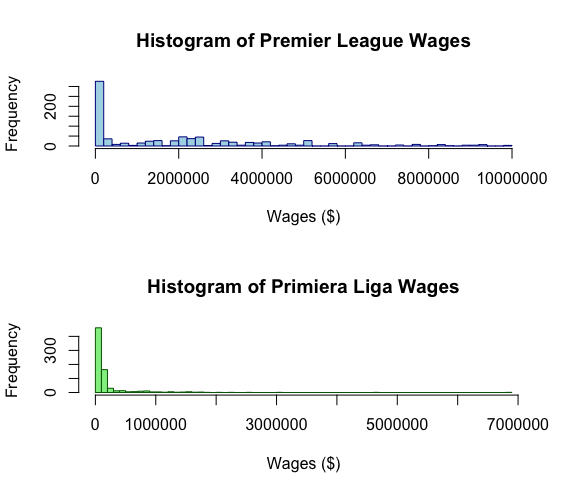


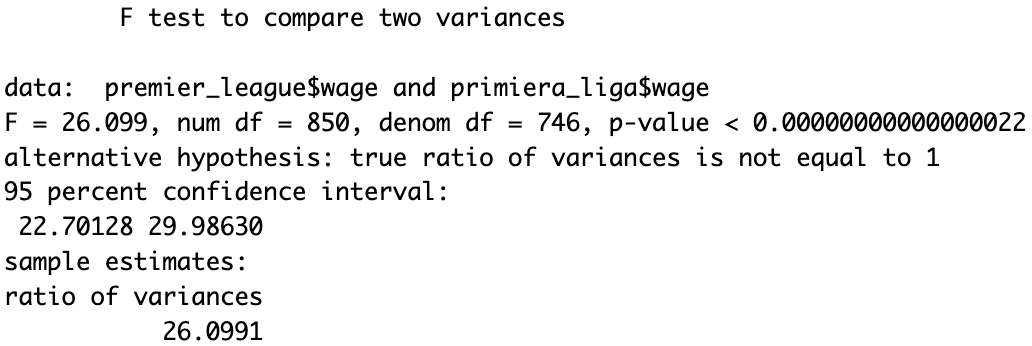


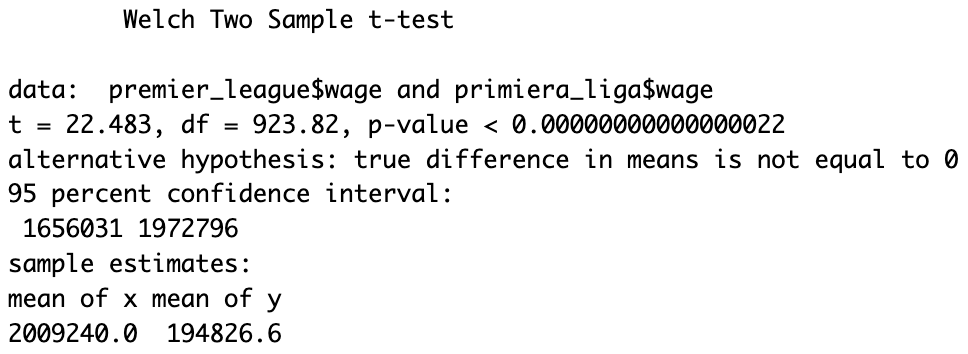
1. 2nd research question: descriptive statistics, distribution, f-test, and t-test











1. 3rd research question: correlation matrix, scatterplot matrix, regression table, and separate scatter plot with gradient for each predictor variable