Project Diary

Data422-Data Wrangling Group Project

Group Members:

- Anirudh Revathi-36899816(**Anirudh**)
- Gus Pawson-63949878(**Gus**)
- Julian Maranan-81456731(**Julian**)
- Pretty Mathew Punnoor -34541658(**Pretty**)
- Uditha Mayadunna-39768450(**Uditha**)

Date	**Estimated Hours**

19/09/2023

- The group was created by the academic staff according to the lab sessions we are participating.

21/09/2023

- During the lab session, all group members had the first meeting. The initial conversation took place, and many topics were introduced as the group project's topic. Finally, all the members agreed to examine two of the world's main football leagues. The LaLiga (Spanish league) and the Premier League (English league) were chosen as two leagues since they are the most popular football leagues in the world. — 2 hours

28/09/2023

- Our second meeting focused on this lab session, and the task was distributed among group members. — 2 hours
    - Scraping Data – **Gus**
    - Wrangling and Plotting Fantasy Data Using R – **Pretty**
    - Wrangling and Plotting Wage Data Using R – **Anirudh**
    - Wrangling and Plotting Wage Data Using R - **Julian**
    - Wrangling and Plotting Data Using Julia – **Uditha**
    - Presentation Slides – **All**
    - Project Report **- All**

02/10/2023

- **Gus:** Began researching potential data sources to collect football data from. Discovered an abundance of websites that contained high quality data and very detailed statistics. *Fbref.com* and *Statbunker.com* were chosen as the best sites. — 1.5 hours

- **Gus**: Researching and planning overall design of the relational database and how to create keys that could connect the several individual tables. 1 hour

03/10/2023

- **Gus:** Started working on scraping data, from *fbref.com,* for the summary table. This required collecting data from various elements found on each page. Found the biggest problem faced when data scraping was the lack of league and year information. Had to spend a lot of time figuring out how to incorporate this into all dataframes as this would be used to create foreign and primary keys for our relational database. 6 hours
- **Gus:** Creating ID columns from year and league information and wrangling some of the data to the correct type. 1 hour
- **Gus:** Began working on collecting end of year tables from *fbref.com* . Was only able to collect data for the Premier League. 3 hours

04/10/2023

- **Gus:** Completed collecting end of year tables for the La Liga championship. This required updating a few previously written functions and creating new ones to deal with slight differences in the web links. 2 hours
- **Gus:** Creating ID columns and additional wrangling steps for end of season tables. 0.5 hour
- **Gus:** Began working on collecting player statistics collected from *statbunker.com.* Dealing with a new website created new difficulties. Collecting year and league information was much more difficult. Datasets were also much larger, so, scraping took a lot longer. Due to small errors in my code this had to attempted more than once. 7 hours
- **Gus:** Creating ID columns and additional wrangling steps for fantasy data. 0.5 hour
- **Pretty:** Began researching on the website and tried to understand the columns and details present in it. 2 hours
- **Pretty**: Tried to have a clear idea about the functions that can be made from the csv files-focused on Premier League and LaLiga Fantasy points table

05/10/2023

- **Gus:** After the lab we had decided on collecting wage information. This was available on *fbref.com* so could utilise many of my previous functions. 2 hours
- **Gus:** Wage information was extremely messy and regex expressions were needed to collect only the relevant data needed for our analyses. This took a long time to get right and two different expressions were needed for each league we wanted information on. 1.5 hours
- **Gus:** Creating ID columns and additional wrangling steps for wage data. 0.5 hour

- **Anirudh:** Started studying the received CSV file on Jupyter Notebook. Researching on how wages influence players in LaLiga and Premier League on the internet. 2 hours

- **Julian:** It was decided to scrape, wrangle, and analyse football stats for the top two football leagues La Liga and Premier league. I started to do some research on football to get some domain knowledge, as I am not very familiar with the important stats in this game.
- **Julian**: I studied the data scraped by Gus, which was published as a csv file. There are two datasets: La Liga stats and Premier league stats. Some functions used to explore the datasets were glimpse, head, tail. The glimpse function showed that n/a values were present in the columns that I am not interested in. 2 hours
- **Julian**: At this point, I am interested in looking at stats related to the win-loss-draw record of each team. I looked at teams which has the highest winning percentage for each season. I wrangled the data to filter the wins and create a new factor called games played which is just the sum of total wins, loses, and draws.
- **Julian**: The group decided to include wages. This was again scraped by Gus. Once this was done, I studied the data. The key stats we are interested in are wins, weekly wages, and annual wages. I wrangled the data to keep the relevant columns only. 0.5 hour
- **Pretty**: Started looking the lab notebooks for getting knowledge about data processing 1.5 hours

08/10/2023

- **Gus:** Had been informed earlier in the week that the code used to scrape the Fantasy data was not executing. This was due *statbunker.com* having crashed. Had to recollect this data once the website was up and convert to csv once collected. 0.5 hour
- **Anirudh:** During the group meeting, I undertook the responsibility of plotting graphs related to wages and wins. Hence did some basic wrangling and plotted some graphs to get a better understanding. 2 hours
- **Pretty**: Started with data processing of both Premier League and Fantasy tables. 2 hours
- **Pretty**: Checked the other tables for the possibility of combining to get a better plot

09/10/2023

- **Uditha:** After obtaining the scrape data, I began working on wrangling with Julia. Altogether there were seven csv files and initially started working on two of them to identify the top seven teams from both leagues (Premier League and LaLiga). 3 hours

10/10/2023

- Had a meeting at the university to discuss more about our project objectives. 2 hours

- **Uditha:** Started making graphs out of the wrangled data regarding the top seven teams in both leagues. 3 hours
- **Julian**: I wanted to find out if there is a correlation between the annual wages and wins. Using the relational data frames we have, I combined the wages and stats data, starting with the two La Liga data sets. I did some wrangling to keep only the relevant columns. 1 hour
- **Julian**: Using this, I created a scatterplot to observe the correlation between annual wages and wins. The scatterplot showed a positive trend, indicating a positive correlation between these two factors. The correlation factor indicates this as well. I did the same process for Premier League stats and found that there is also a positive correlation between annual wages and wins.
- **Anirudh:** Created stacked bar graphs and bar graphs for all 7 seasons for both the leagues. 3.5 hours
- **Pretty**: Continued with data processing and worked on aggregating the data to get the desired data frame 1.5hours

11/10/2023

- **Uditha:** Started processing data to determine how much money each football team has spent to achieve a single league and fantasy point. To get the desired result, all the relational data frames were consolidated into a single data frame and mutated columns were added. 6 hours

12/10/2023

- Attended the semester's last lab session and had a conversation with another group. 2 hours
- **Uditha:** The outcome of the previous data frame was visualised using Julia. 4 hours
- **Anirudh:** Graphs produced previously could not be used for the presentation as there were too many of them. Had to come up with other ideas and try other graphs which could show 7 seasons of data in one plot. Hence researched more about the types of graphs that could be used to present this data. 1 hour
- **Anirudh:** Tried making circular stacked bar plots, and clustered bar graphs. did not see much success in representing the data. Worked online graphs which showed a much better representation of the over 7 years. 4 hours
- **Julian**: I wanted to see if there is a correlation between annual wage and attendance in the games. Using the same process, I did for analysing wages and wins, I found that there is a positive correlation between the two. This is true for both leagues. While this is true, I am not sure if it is very insightful. This would have been better if I scraped data to get the home and away stats for each team. However, I did not have enough time to do this at this point of the semester. 1 hour

- **Pretty**: Had lab discussion to get clear idea about the wrangling each one of us are doing. Some more ideas came and started working on it. — 2 hours

14/10/2023

- **Gus:** As I had completed the data scraping component the majority of the report was written by me due to my understanding of the code and knowing the issues I had. I completed my side of the report this day. — 6 hours
- **Anirudh:** Worked more on adjusting the graph's visual appeal. — 1 hour
- **Julian**: I wanted to see which league allowed more goals. This is for the purpose of seeing if there is a more offensive-minded or defensive-minded league. I wrangled the stats data frame for both leagues and found some summary statistics for both leagues. I found that more goals are scored in the Premier League (mean and median goals) from 2009 – 2022. However, the difference is not significant. I also wanted to confirm the correlation between the goal difference and wins. As expected, there is positive correlation between the two. — 1.5 hours

15/10/2023

- **All**: Catch up meeting on Zoom. Gus walked the team through the report template and parts he already wrote. There are two issues were discussed at this stage: there are too many plots, and they are not cohesive. — 0.5 hour
- **Pretty**: Worked on data wrangling part and plotted bar graphs for the wrangled data that I have made. — 3 hours
- **Pretty**: Plotted graphs for both leagues to get an idea of top scoring teams and plotted in terms of position too

16/10/2023

- **Uditha:** Slides were created to show the results of Julia data wrangling. — 2 hours
- **Uditha:** Added descriptions to the Julia notebook and commented on the work done in Julia. Added details to the Julia notebook and commented on the codes which were written in Julia. — 3 hours
- **Uditha:** Work on the group report began, and Julia wrangling was included. — 3 hours
- **Julian**: Catch up meeting with available team members at Jack Erskine with the goal to improve the cohesiveness of our graphs and report. The goal for next meeting is to have our slides ready for completion. — 0.5 hour

17/10/2023

- **Gus:** Created the data scraping slides to be used in the presentation. — 1.5 hours
- **Gus:** Tidied up my data collection notebook, adding more information about the tables and writing more detailed comments. — 1 hour
- **Uditha:** Continued to work on the group report. — 3 hours

18/10/2023

- **Gus:** Added in another member's work to the report. Adjusted some of the formatting to fit the rest of the report. — 1 hour
- **Anirudh:** Worked on my slides. — 1 hour
- **Julian**: The group met to finalize the slides. The group discussed which plots to include and which ones to leave out. It is tough to choose which plots to include as there is only limited time in the presentation. I have three slides in total. I decide to include 4 plots in two slides, which are the correlation plots for annual wages and wins for both leagues, and the correlation plots for goal difference and wins. Also, I made one slide for the conclusions part which summarizes the key findings of our analysis. After discussing with the team, we finalized the objectives of our project. I wrote this section of the report. — 3 hours
- **Pretty**: Started working on presentation slides and merged all the presentation slides, made it look more presentable. — 3 hours
- **Pretty**: Prepared for the presentation and made changes in the slides and tried to create some more useful graphs.

19/10/2023

- **Julian**: Before the presentation, the group met up to finalize the slides. We had 25 slides in total, so we had to take out some slides to not go over time in the presentation. Now I have two slides left, which is the annual wages and wins correlation and the summary. The team did a practice run to see if we will not go over time. — 2 hours
- **Pretty**: Done with the presentation and discussed how to work on report and journal — 1 hour

21/10/2023

- **Gus:** Created a relational database diagram and wrote in more detail what our final relational database looked like. — 1.5 hours

22/10/2023

- **Uditha:** Continued to work on the group report. — 4 hours
- **Anirudh:** Cleaned up my Jupyter Notebook to make it more readable. — 1.5 hours
- **Pretty**: Worked on making the report of the wrangling part that I have done Added the necessary details — 2 hours

23/10/2023

- **Uditha:** Continued to work on the group report. — 4 hours

24/10/2023

- **Anirudh:** Made my part of the report for the final addition. — 2.5 hours
- **Julian**: Cleaned up jupyter notebook for looking at correlation of certain factors such as wage, wins, attendance, and goal difference. Added some comments supplementing the code. Got rid of the part where I looked at the highest winning percentage. Finalised my part of the report and group diary for compiling. Did a quick review of the report. — 2.5 hours

25/10/2023

- **Uditha:** Continued to work on the group report. — 4 hours
- **Pretty**: Completed the report and cleaned the jupyter notebook to make it presentable, added comments as well — 2 hours

26/10/2023

- **Uditha:** Continued to work on the group report. — 4 hours