**Github URL:**

**Abstract:**

In the following project, I aim to explore and garner insights from a dataset with over 42, 000 points of data regarding the top 250 football transfers from the year 2000 to 2009. I hope to find out some of the attributes that make a player valuable, which leagues have the most buying power, and identify some trends that have developed over time. To do this I’m going to use numerous Python packages and functions, notably Pandas, numpy and matplotlib, to upload and manipulate the data, and produce visualisations from which I’ll draw some insights.

**Intro:**

I’m going to be using python and its associated packages to explore, manipulate and visualise a dataset I obtained from Kaggle, which consists of data relating to the top 250 football transfers from the season 2000/2001 to 2018/2019. I aim to find out some of the factors that determine a footballers transfer fee, establish which leagues and clubs have the most financial power in the transfer market, and identify the long term trends of footballer’s transfer fees: have they inflated over time? I got the dataset from Kaggle in the form of a csv excel file, and I’ll use the Pandas package in Python to upload and manipulate the data, with the help of the numpy package. I’ll create various subsets to isolate potential factors that affect the transfers in the dataset, and use packages matplotlib and seaborn to clearly visualise my dataframes and tables, and help me garner greater insight into the data.

**Dataset:**

The dataset I chose to analyse is a record of the top 250 football transfers completed each season from the year 2000 to 2018. The dataset provides a good variety of variables and values for each transfer, including their age, position, clubs, transfer fee, season and more. One of the things that drew me to this data set is the wide range of values, across wide range of years. 42, 300 points of data across 19 years gives me the opportunity to visualise and track trends across time. It is noted in the description of the data set, that the data for the final season (2018 – 2019) is incomplete, as that transfer window was still ongoing as the data set was being compiled, so this should be kept in mind. I decided to use Kaggle because it’s a respected site that has a huge range of datasets on a wide range of topics.

**Implementation Process:**

Once I had identified the dataset I wanted to analyse, I imported to Pycharm using Pandas. I immediately cleaned it up by replacing any NaN values with 0, and then I quickly explored it to get a better sense of its contents.

The first thing I wanted to analyse about my data, was the effect age has on a player’s value. I sorted the dataset by age and transfer fee, and then created a subset called player\_age\_value, a data frame with just the player’s age name and fee. When I printed out the head, I immediately noted an error in the dataset. One of the players had been attributed an age of 0. I decided to drop this player from the data. I used matplotlib to create a histogram, so I could clearly show the entire distribution of values within the age variable. The histogram was extremely useful in visualising how a players transfer value rises and falls over time.

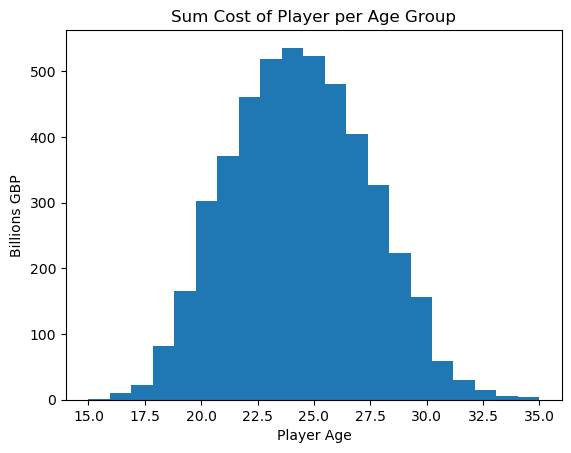
I then turned my focus to how a players position might effect his value. I once again sorted the original data (File), this time Position and Transfer Fee. I wanted to isolate the sum transfer fee per position, so I created a pivot table with position as the index, and applied the numpy sum function to the transfer fee column, and stored this as a new subset. I created a bar chart with this info using matplotlib, and it clearly showed that the “Forwards” had accumulated the largest amount of transfer fee’s in the dataset, but I found that the chart offered little other insights to me. I thought it would be better to group all the positions into three groups representing broader attributes of their position. So, in order to create three new subsets, called Attack, Midfield and Defense, I set the column “Position” as my index, and used the loc function to subset the different positions. I decided it would be interesting to see not just which position was most valuable overall, but whether that value had changed over time. I used matplotlib to plot the Sum of transfer fee’s for each grouped position over each of the 19 season documented in the data set. I repeated this, but using the mean function instead of sum, in order to see how the *average* price had changed over time too.

Next, I wanted to know more about the leagues that are buying these players. Which are the most powerful leagues in the football world? Is there a large disparity in purchasing power? I again used a pivot table to create a data frame, this time displaying the sum of transfer fees per league. I examined the top 15 highest spending leagues, and displayed the info on a bar chart. The disparity between the top 5 leagues and the rest seemed pretty big, so I sliced the top 5 leagues into its own data frame, top5, and stored the rest separately in data frame, world. I manually created a new dataframe with both top5 and world, and their respective sum of transfer fees, by creating a dictionary with the required values, and converting it to a pandas dataframe. I then displayed the result on a bar chart, which perfectly visualised the gulf in spending between the top 5 football leagues, and the rest of the world.

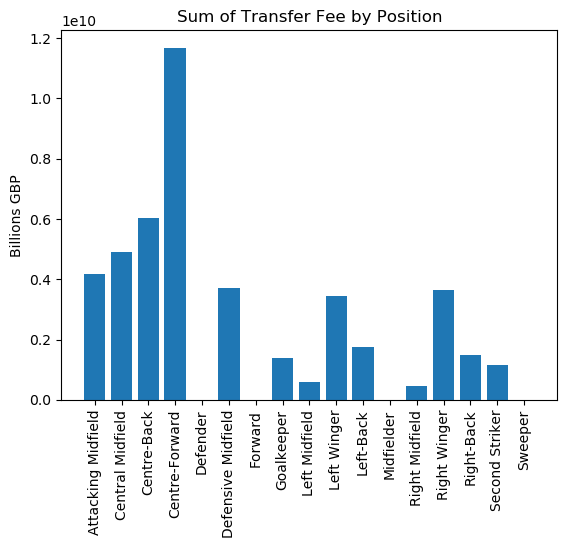
I once again wanted to track performance over time, to reiterate the rate of inflation the football transfer market has experiencing. I used pivot tables to create the subset epl\_avg\_fee, and then used matplot lib to plot it.

The following sections will detail the results and insights I gathered from all of this.

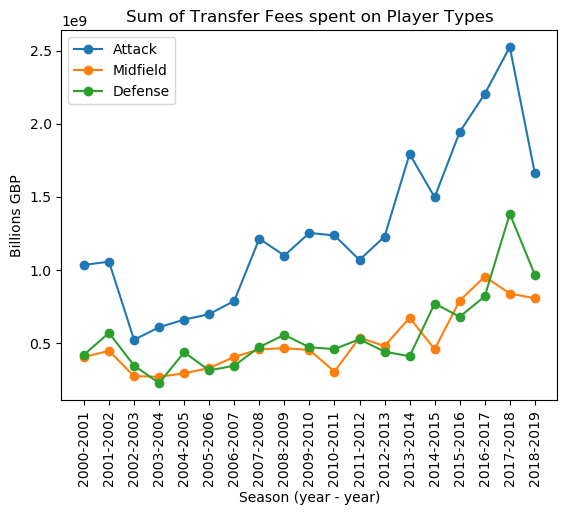
**Results:**



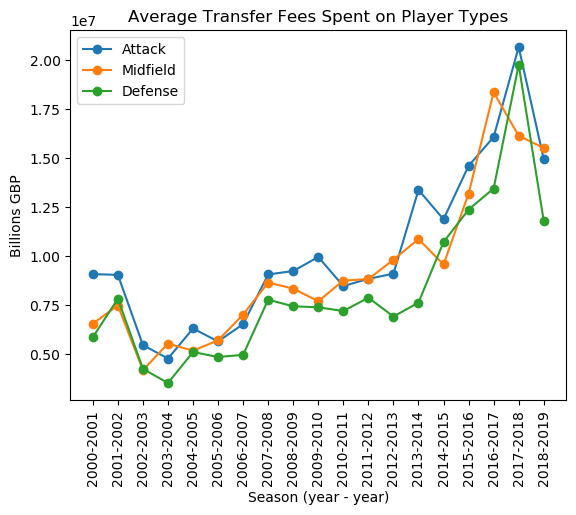
The first visualisation is a histogram that displays the total transfer fees paid for a player in each age group.



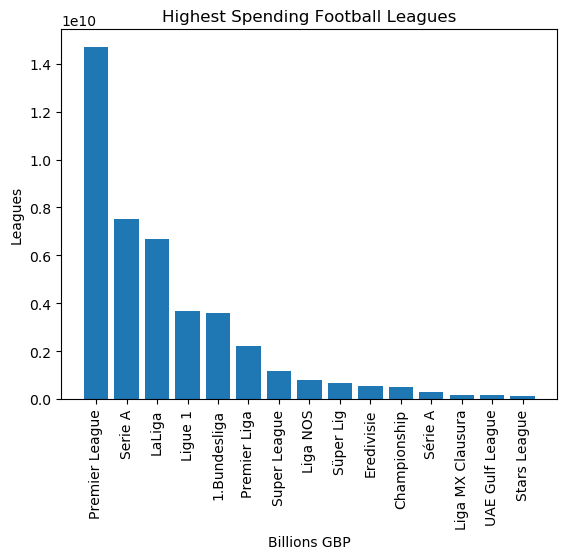
My second visualisation is a simple bar chart, with all the player positions featured in the dataset along the x axis, and the total amount of money paid for players in that position, in billions of Great British Pounds, along the y axis.



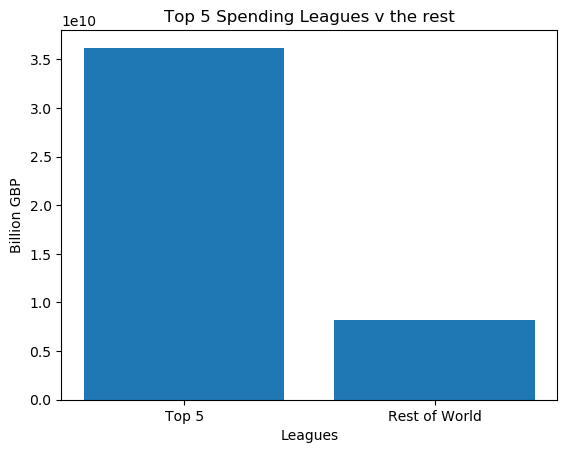
The third visualisation is a plot graph, that shows the sum of all the money paid for players of the positional attribute group defined in the legend. This graph was much more informative than the previous bar chart, and also charts the rising cost of transfer fees over time



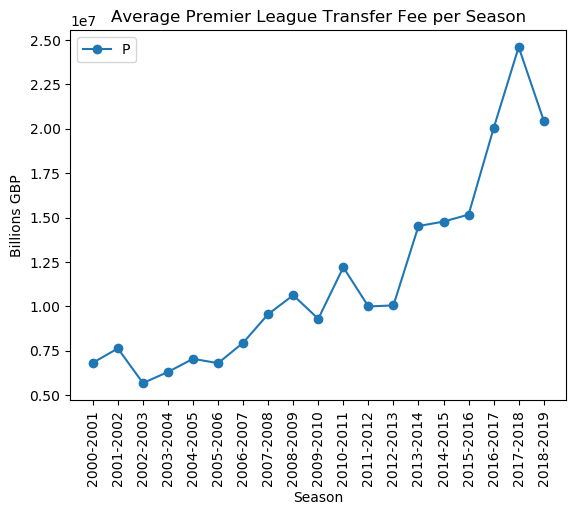
Visualisation 4 is similar to the previous graph, but this charts the mean cost of transfers each season over the course of the 19 seasons.



This bar chart was meant to simply display the spending power of the top 15 leagues in the dataset.



This bar chart shows the sum transfer fees over the full 19 season from the top 5 league, vs the rest of the world.



The above plot show the average total transfer spend by all premier league clubs, per season. It clearly displays the rapid inflation of the market.

**Insights:**

Graph 1:

* From the first graph, we can clearly see the relationship a players age has with their transfer fee. The most valuable players are in their mid 20’s, and there is a sharp fall off in value as the player gets older.

Graphs 2, 3 and 4: Position

* Graphs 2 and 3 show clearly that over the 19 season’s featured in the dataset, more money has been spent on players with attacking attributes. Forward positions attract the largest fees
* The attacking positions have attracted the biggest sums of money over all the seasons detailed in the dataset.
* While attacking positions attract the largest sums of money, on average over the course of the 19 seasons, there is a much smaller difference between the three attributes, though on average attack is still the most valuable, and defense the least valuable. As detailed in Graph 4.
* Graphs 3 and 4 also display the consistent and rapid inflation of transfer fees over the 19 seasons. The fall in 2018-2019 is due the dataset being incomplete for that season, as detailed in previously in my report.

Graphs 5 and 6:

* Graph 5 illustrates the absolute financial dominance of the premier league, spending almost double that of its nearest rival, Serie A.
* Europes top 5 leagues have always had the most money, and this is clear in graph 6. The Total spend of the top 5 leagues completely dwarfs that of every other league featured in the dataset.

Using the analysis and visualisations throughout the project, we could adequately identify age and position as two key factors that drive a player’s transfer fee. Europes top 5 leagues have financial superiority over the rest of the world, with the premier league being particularly dominant over the time frame presented in the dataset. I have also shown that that the top transfer fees have been inflating a massice rate since 2000 and show no signs of slowing down.

**References:**