Ashutosh Singh
M13433470

IS 6030: Data Management

Homework 5 Report

Dataset:

Since 'Football' (or 'Soccer' as they say in US) is one of my favorite sports, I decided to dive into one of such datasets exploring which are the best players in the world, what are their ratings, which countries are most prominently represented and various other interesting questions. I am using the 'Fifa 19 Complete Player Dataset' from Kaggle.

The link for the dataset is here: https://www.kaggle.com/karangadiya/fifa19

The dataset was somewhat messy, some of the shortcomings which I found in data was:

- The presence of some redundant columns.
- The players wages, value and release clause were stored in £ currency and also in millions.
- Some European player names had special characters in their names.

The dataset seemed normalized. No column had more than one value and there were no duplicates resulting in data redundancy.

Before importing the data to SQL Server, I used R for Data Cleaning. I was able to solve the second point and I used *write.csv* to save in a new file. The NAs which are introduced by R were replaced with empty commas so that SQL server would treat them as missing values.

The final dataset which I imported in the SQL Server contained some of the following Attributes:

ID	*Overall (Overall rating)*	*Release Clause*
Name	*Club*	*Joined*
Age	*Value*	*ContractValidUntil*
Nationality	*Position*	*Ratings in each football skill*

In total there are **1872017 observations and 88 columns**.

Exploring the Data in SQL Server:

After importing the data in SQL server, some sql queries were used to do basic data exploration. However, before writing any query to get descriptive statistic I verified whether ID column was unique. After executing the following query: '`SELECT COUNT(DISTINCT ID) FROM dbo.fifaData`', the number of records come out to be same as total. So, it was verified that ID was a unique identifier for each player.

Some of the key questions I tried to answer were:

- Total number of players present in the database.
- Top 10 countries represented in the world football.
- Top 10 players in overall rating.
- Top 10 Forwards and Top 10 Defenders.
- The clubs which pay their players the most.

The following are the snips of my results:

NumOfRecords
18207

	Nationality	NumberOfPlayers
1	England	1662
2	Germany	1198
3	Spain	1072
4	Argentina	937
5	France	914
6	Brazil	827
7	Italy	702
8	Colombia	618
9	Japan	478
10	Netherla...	453

	Name	Overall
1	L. Messi	94
2	Cristiano Ronaldo	94
3	Neymar Jr	92
4	De Gea	91
5	K. De Bruyne	91
6	E. Hazard	91
7	L. SuÃ¡rez	91
8	L. ModriÄ‡	91
9	Sergio Ramos	91
10	J. Oblak	90

	Name
1	L. Messi
2	Cristiano Ronaldo
3	E. Hazard
4	R. Lewandowski
5	P. Dybala
6	H. Kane
7	S. AgÃ¼ero
8	G. Bale
9	M. Icardi
10	R. Lukaku

	Name
1	Sergio Ramos
2	D. GodÃ-n
3	G. Chiellini
4	M. Hummels
5	Marcelo
6	Thiago Silva
7	S. Umtiti
8	K. Koulibaly
9	Jordi Alba
10	J. Vertonghen

	Club
1	Real Madrid
2	FC Barcelona
3	Manchester City
4	Manchester United
5	Juventus
6	Chelsea
7	Liverpool
8	Tottenham Hotspur
9	Arsenal
10	FC Bayern MÃ¼nchen

A last finding which I wanted to make in SQL was to know whether the wages of players depend on the position they play. So, I decided to take Top 10 highest paid players for each position and compare their average wages. Interestingly, Forwards and Midfielders have higher wages as compared to defenders and specially goalkeepers.

AvgWageDefender
232500

AvgWageForward
305000

AvgWageMidfield
311000

AvgWageGK
171500

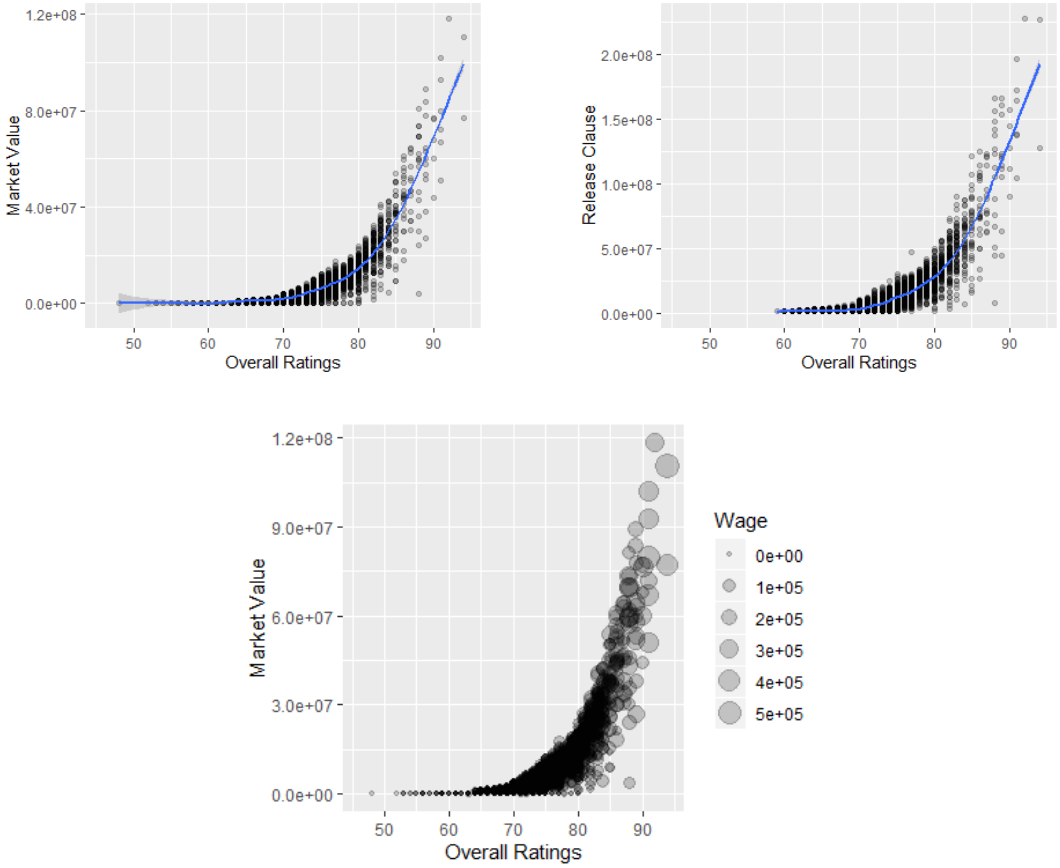
Analyzing the Data further in R:

One interesting observation which came out while working on SQL was that the top 10 players in terms of ratings, wages and market values are quite similar. So, I wanted to figure out whether there is any correlation between market value of a player and their ratings.

After running some correlation commands in R, I get the following as the output:

```
cor(fifaData$Value,fifaData$Overall,use = "complete.obs")
# 0.7811392
cor(fifaData$ReleaseClause,fifaData$Overall,use = "complete.obs")
# 0.7398952
```

From the above results, my perception about the correlation between Rating of a player and his market value and release clause is close to right. The correlation values are quite high which suggest that the players which are highest rated on Fifa tend to have high values and high release clause mentioned in their contracts.



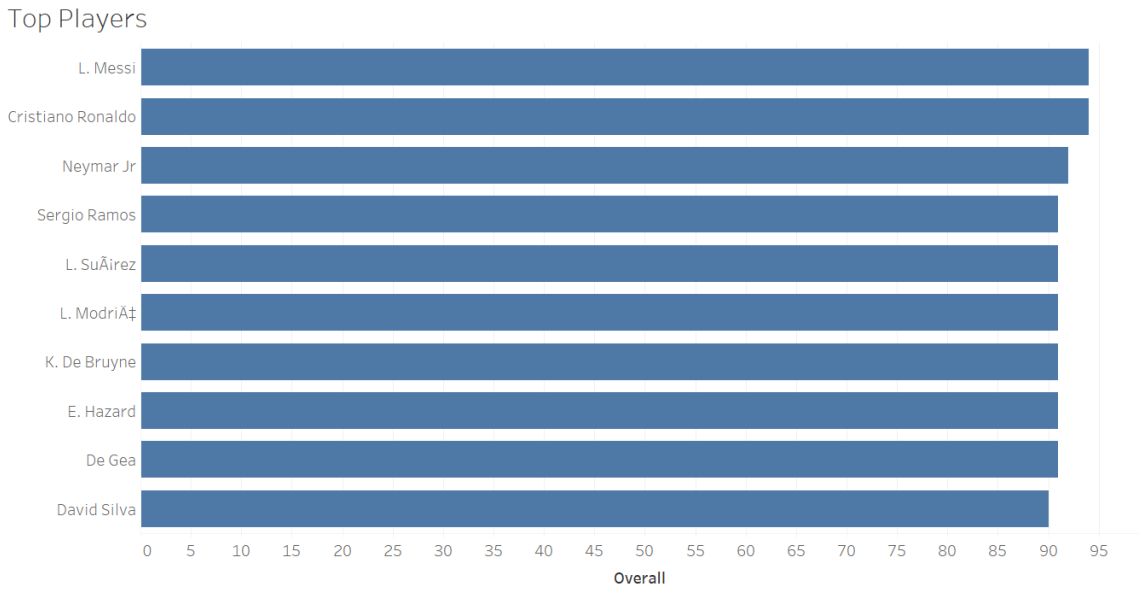
As we can see from the plots above, each of the trends rise exponentially at higher ratings (particularly after 85). In the third figure, I have introduced a third variable Wage in order to see how these three correlate with each other. Here we can see the size of dots is quite large at higher ratings which tells that person with higher ratings like Lionel Messi, Cristiano Ronaldo, Neymar etc. have high wages and predictably high market values.

Visualizing the Data in Tableau:

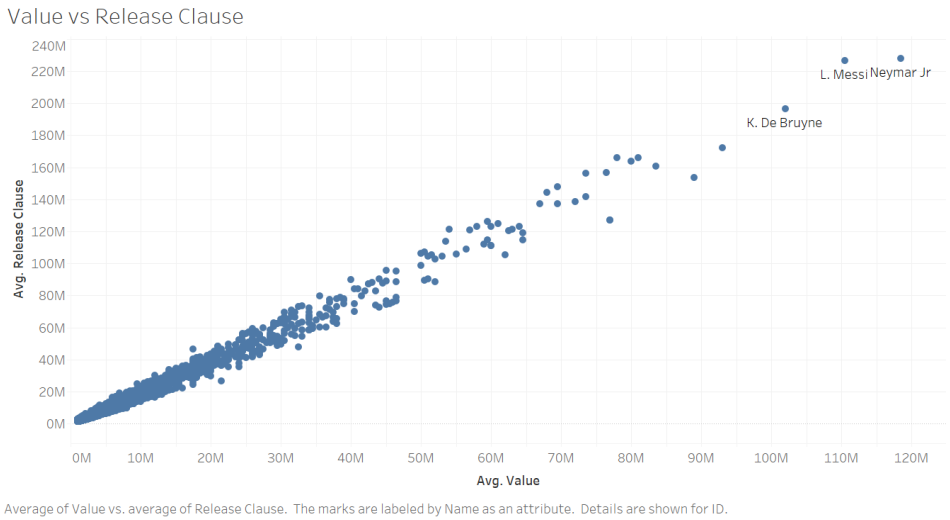
After finding some meaningful inferences, I used tableau to create some insightful visualizations about our dataset.

Ashutosh Singh
M13433470

The first figure shows the same visualization as it was shown earlier by executing Sql query and the bars are ordered in decreasing ratings.



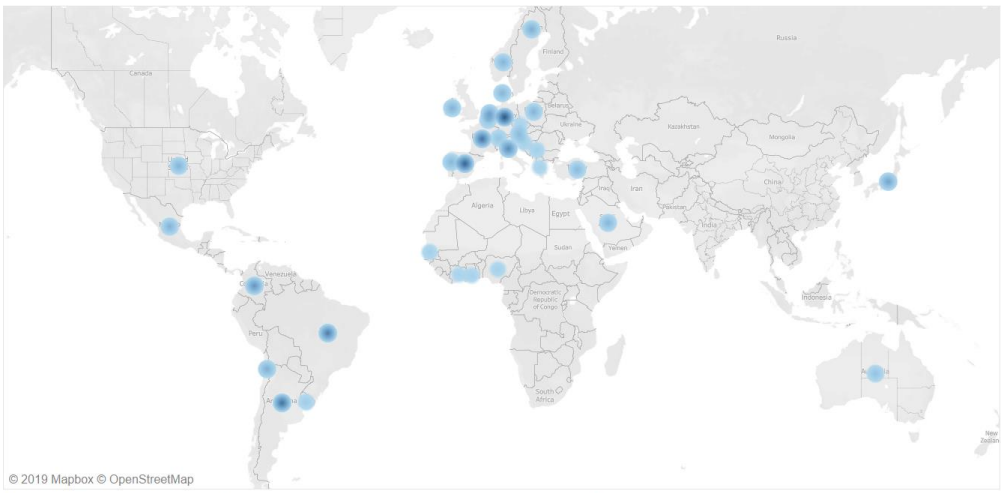
The next plot displays almost linear relationship between average value of a player and the release clause mentioned in his contract. As expected, the players with high market value tend to have higher release clauses in their contract as the clubs don't want to lose them on cheap.



The last figure shows the popularity of football across the globe. The circles with high density denote high number of players representing that country as compared to less dense ones. European and South

American Nations have the most number of players and the sport is quite popular there as compared to Asia and North America.

Popularity Accross the Globe



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Number of Records. Details are shown for Nationality. The view is filtered on sum of Number of Records, which ranges from 99 to 1,662.

Key Findings:

- The players with high ratings tend to have high market values and high wages since these are one of the best players in the sport.
- The wages and the market value of a player depends on the position they play. Forwards and Midfielders perform comparatively better in this aspect.
- The game of football is more popular in European and South American countries.

Challenges Faced:

- The dataset had some inconsistent values like certain numeric fields stored data as string. I used R to replace the currency symbol in these fields and multiply with the corresponding factor (like 10^6 if it's in Million) to convert it into numerical values.
- Similarly, for Tableau I had to convert the data types of certain fields *like Value, Release Clause, Overall etc.* and convert them to Measures in order to create the necessary plots.