**Practical data Science:**

**Assignment 1:**

**Student id: S3814520:**

**Introduction:**

The main aim of the data set is to examine a NBA Player dataset and carry the first and initial step of data science process, including the cleaning and exploring of the dataset given to us.

**Dataset*:***

The dataset given to is of various characteristics of a NBA player such as points, three-point percentages, etc. In total we can see there are 29 attributes and 512 observations but according to the given instruction there are observations of 492 players as some players have played from more than one teams in a season.

The attributes present in the data set are mostly numerical, but we can see some few categorical data type such as position and teams.

**Task 1:**

**Data Preparation:**

**Importing Dataset:**

First, we import the raw dataset given to our Jupyter Notebook workspace. We the read the given CSV file and then we name it, in my case I have named it df which stands for data frame. We then display a random sample to make sure that we have our correct data.

Also, we make sure we make of copy of the data frame in my case I have named it df_copy and work on that and leave the original data frame untouched.

**Cleaning categorical data**:

First, I started cleaning the data from categorical data types. I have simply gone through the categorical data and cross checked with the information given to us in the specifications (the correct attributes). We here also check for typos such as upper case and lower case such as ‘SF’ and ‘sF’ although this both are same python treats them differently.

Another common error we find while cleaning is the redundant whitespaces, although this problem does not seem to big but apparently these errors cause quiet a problem for our data set. Here a simple string function is used to remove all the whitespaces so as in we do not have any unnecessary inconvenience in the future use.

**Cleaning Numerical data:**

First, we check if there are any negative values, luckily, we do not find any negative values in our data set. Next, we check for outliers in our data set that is impossible values such as points which cannot be greater than 2000. I also found some negative values where I made them absolute values. For age I kept the threshold to 120 and then replaced the greater ones with the median of the age.

Most importantly we check for null (Nan) values in the data set its one common and important error we need to correct in a data set. There are many ways to deal with NAN values some case we drop it accordingly sometimes we replace it with the mean or the median.

But in our case, we see the Nan values are because we have divided 0 by another 0 mathematically its undefined and infinite but when we see nan value, we see in in 3P%,2P% and FT% then the point scored and point attempts are divided so we can see when both values are 0 we get Nan so we just replace the Nan here with 0 logically.

After all this fair amount of cleaning we save out dataset in cleaned dataset file.csv.

**TASK 2:**

**Data Exploration:**

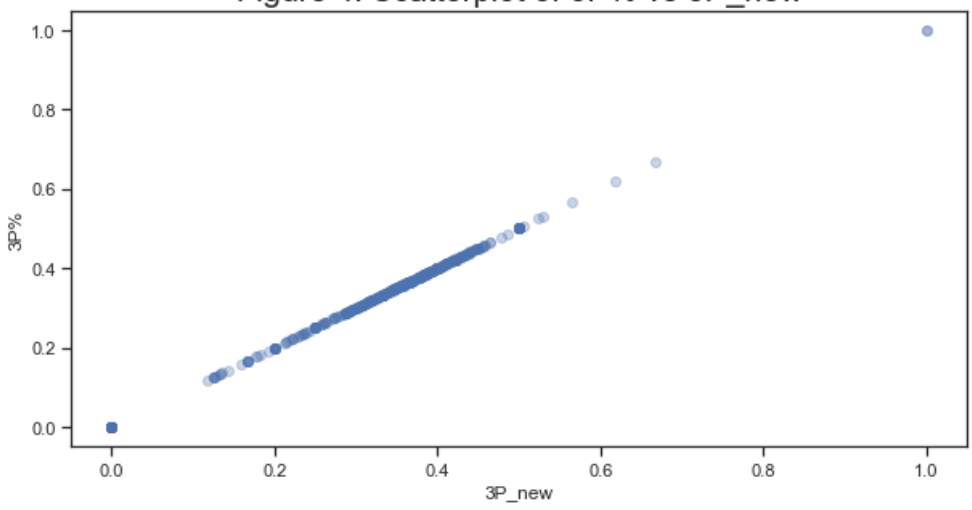
**Task2.1:**

Here in this task, I have basically sorted the entire data in descending order and then just taken the top five players points.

Then when we analyse the composition of the points. We see that the composition of 2P and FT were comparatively almost same and more than that of 3P. Also, the top five players points were very close to one another.

**Task2.2:**

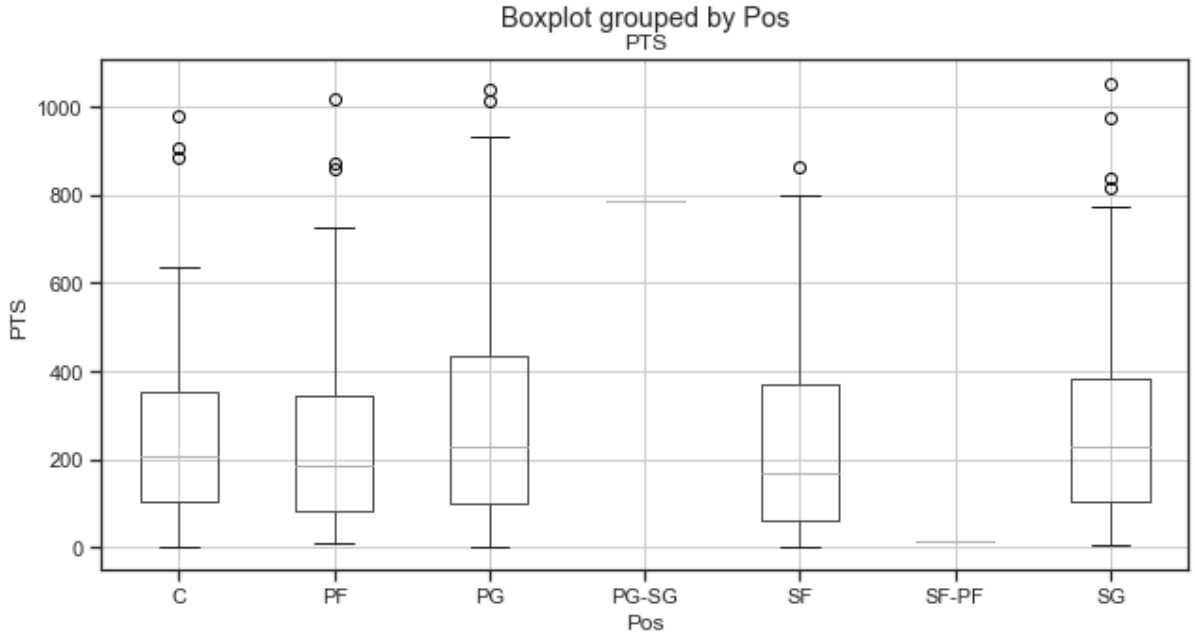
Figure 4: Scatterplot of 3P% vs 3P_new



Here in Task 2.2 I have simply compared the 3P% with the 3P% we have calculated by dividing 3P/3PA and then we have plotted it in the above given scatter plot and we can see that we have got a straight line as both of them are same so we see that while we did the cleaning all the 3P% were cleaned properly so no need to correct as per our visualisation observation.

**Task2.3:**

We have been asked to compare the points with at least 3 other features.




1. I have compared the points with the position (Pos) of the players using a box plot. Here we can observe that the player with position PG i.e., point guards have comparatively higher median points above 200 points hence higher points scored than rest of the position.
   While the SF-PF position have the lowest median points. Also the PG box seems to be the biggest which means has a lot of variance.

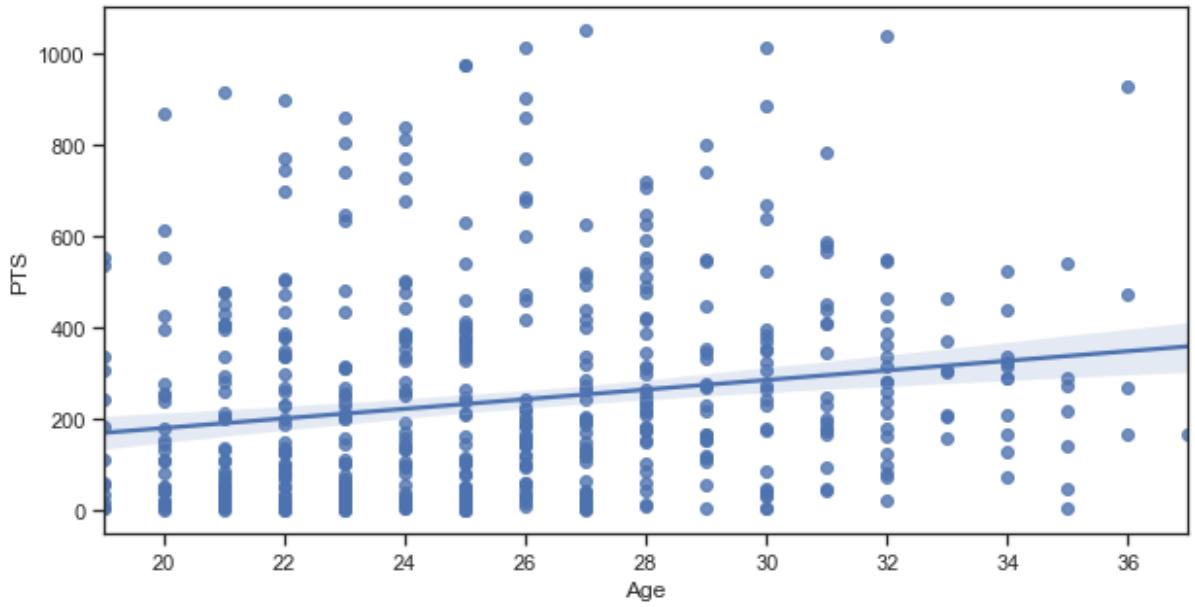
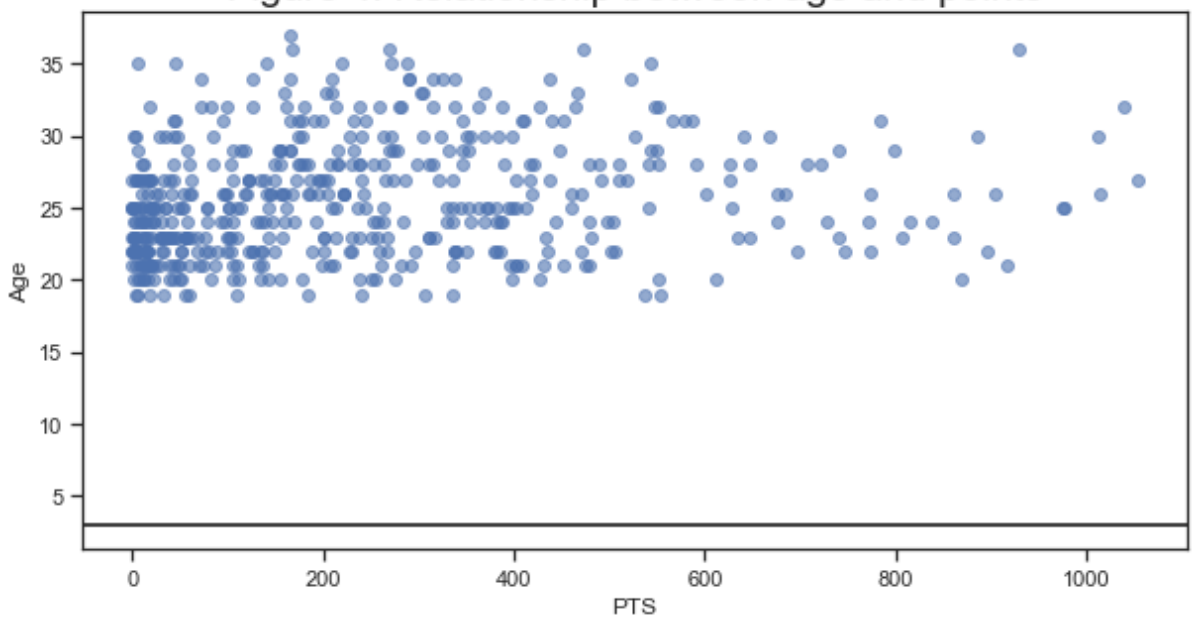
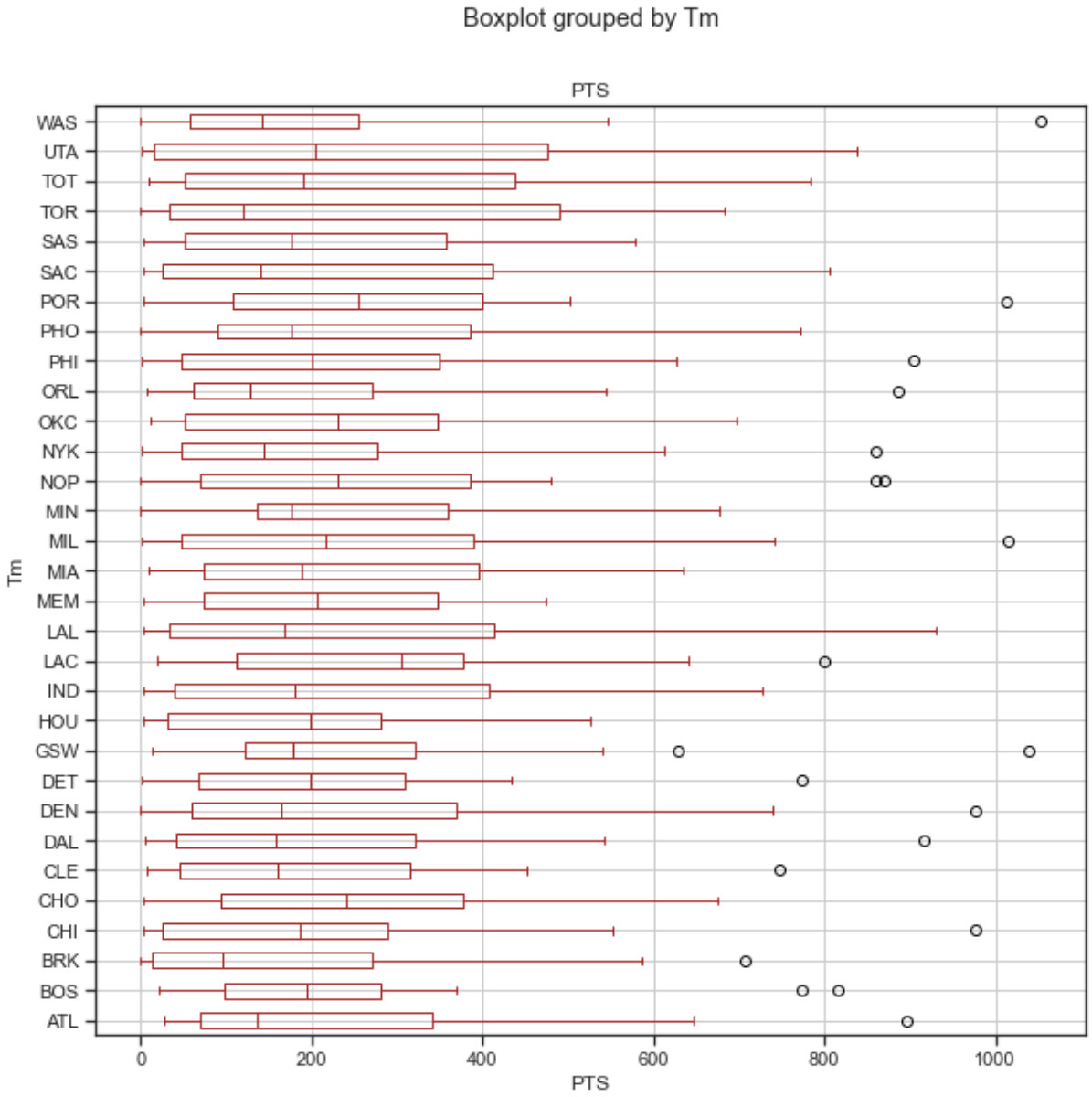



Figure 4: Relationship between age and points




**2.** In here I have plotted points with age in a regression graph and a scatter plot. As per my observation I see that player with the age of 26 have comparatively good points as we can see a very small gap, we can also consider age 22,23,24 and 28 as well as they have some handsome number of points. We can see that players whose age group are between 22 to 30 have managed to score good points, where after they cross the age of 30 the points scored are quite inconsistent.




3. Here we can observe the team LAC which is Los Angeles Clippers has the highest median points amongst all the teams present. Also, most of the median points are close to 200 points.