**Data Science Workshop - NBA Free Throws Prediction**

**Dataset description**

In January 2015 a data-set of 600K NBA free-throws was upload to Kaggle by Sebastian-Mantey. The data was scraped from the website ESPN.com which belongs to an entertainment and sports programming network.

Dataset contains 11 variables:

* end\_result: host total score - guest total score
* game: host team vs guest team
* game\_id: id of specific game
* period: which quarter
* play: who make free throw, make or miss free throw
* player: player name
* playoffs: whether a playoff game or regular game
* score: host team score - guest team score at that time
* season: NBA season
* shot\_made: whether player got the free throw
* time: time left in that quarter

Number of free-throws: 618,019 in which 575893 regular games and 42126 playoffs. Number of unique games: 12,874.

**collecting more data**

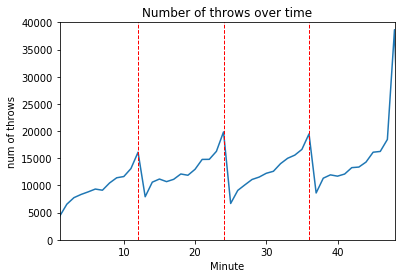
In order to expand our dataset, we decided to use an open source python library PandasBasketball, and use a webscrapper in order to get more players stats from [https://www.basketball-reference.com](https://www.basketball-reference.com/) website.

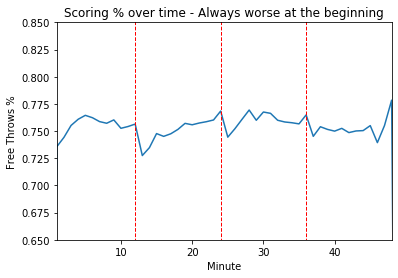
for each player we added the following columns:

* Position: The most common position for the player over his seasons.
* FG%
* 3P%
* FT%
* Height
* Weight
* ShootingHand
* draftRank

We merged both datasets according our collected data from the internet. Some players stats had been inserted manually because some bugs found on PandasBasketball library. We found that we have only two columns with missing data: First - the draftRank column. These values are missing because the players performed the free throw didn't have a draft rank and not because we couldn't collect the data. Second - 3P% has missing values since there are players that have never thrown a 3-pointer. We then changed the relevant categorical variable (position) to numeric and changed the binary and numeric variables to be in the appropriate type.

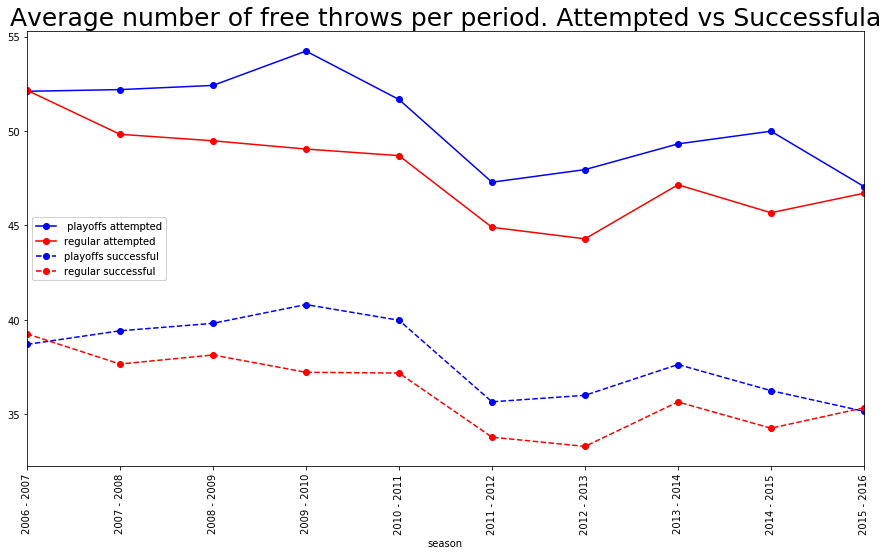
**Data analysis**

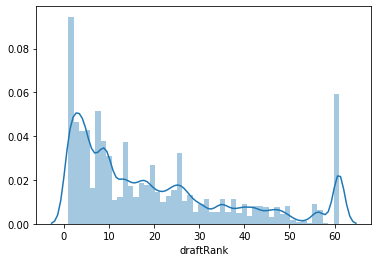
We would like to show the free throws distribution throughout the game time, in our current dataset, the time column represents the time left in that quarter and the period column represents the quarter of the game, so we'll add a new column to our data which calculate the absolute time in the game that the throw was made.

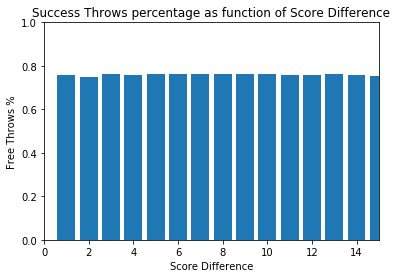
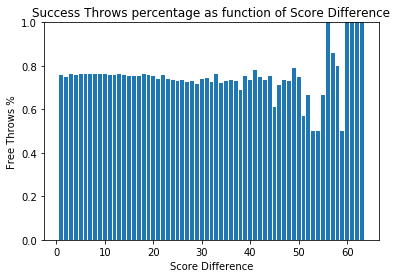


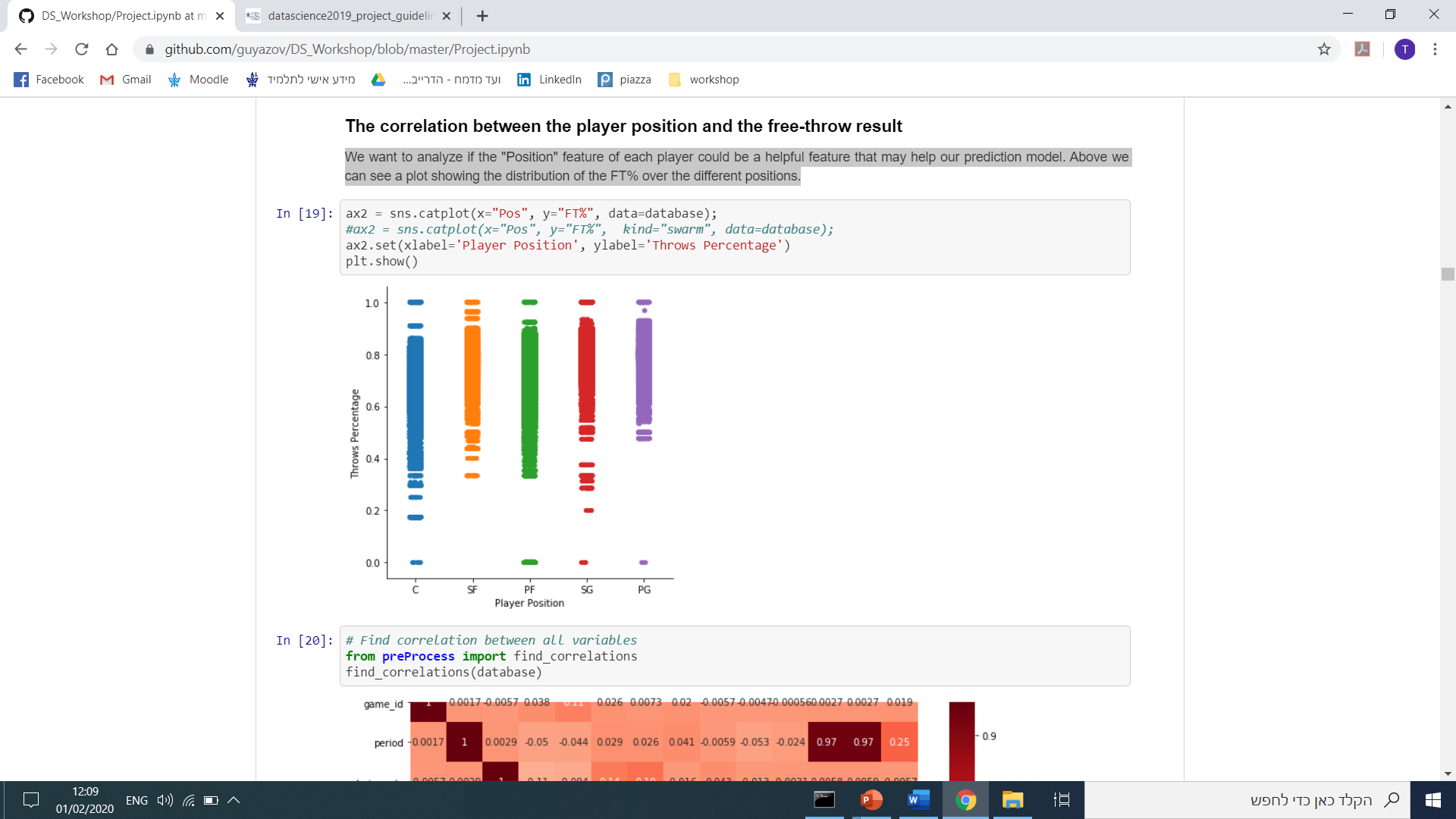
From the plots we can observe that at the begining of every querter, both the number of free throws and the success precentage drops. Moreover, at the end of a quarter, and especially at the end of the game, both plots increase. We can explain this behaivor, as basketball rules when a team made more than 5 fouls, every another foul made in that quarter will be penalty with a free throw.

We want to see how the average of free throws attempted and succeed in games over our differents season in our dataset are distributed:

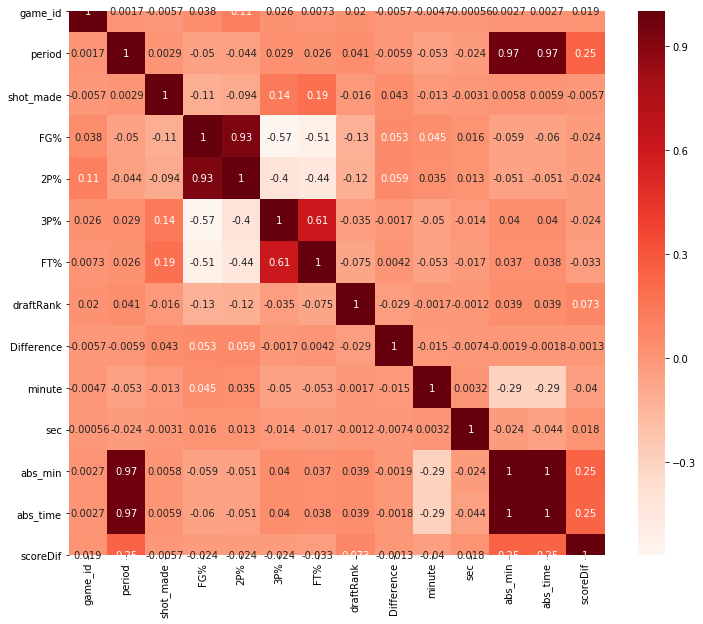
We can see that are not very big differences over the seasons at both attempted and successful shots. But there is a clear difference between the amount of made shots and and successful shots at playoffs and regular season.

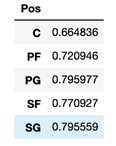
The correlation between the player's draft rank and the free-throw result (after replacing the field "undrafted" with a 0 and filling the NAN cells with 61 and normalization):

The correlation between the score difference and the free-throw result: We will add a column calculating the difference. We assume that if the difference is small in the time of the shot, the player would be more stressed and his shooting percentage would drop:

It seems that our hypothesis that as long as the score difference gets bigger, so as the free-throw success percentage is partly true. We can see a trend in the graph but it is not continuous.

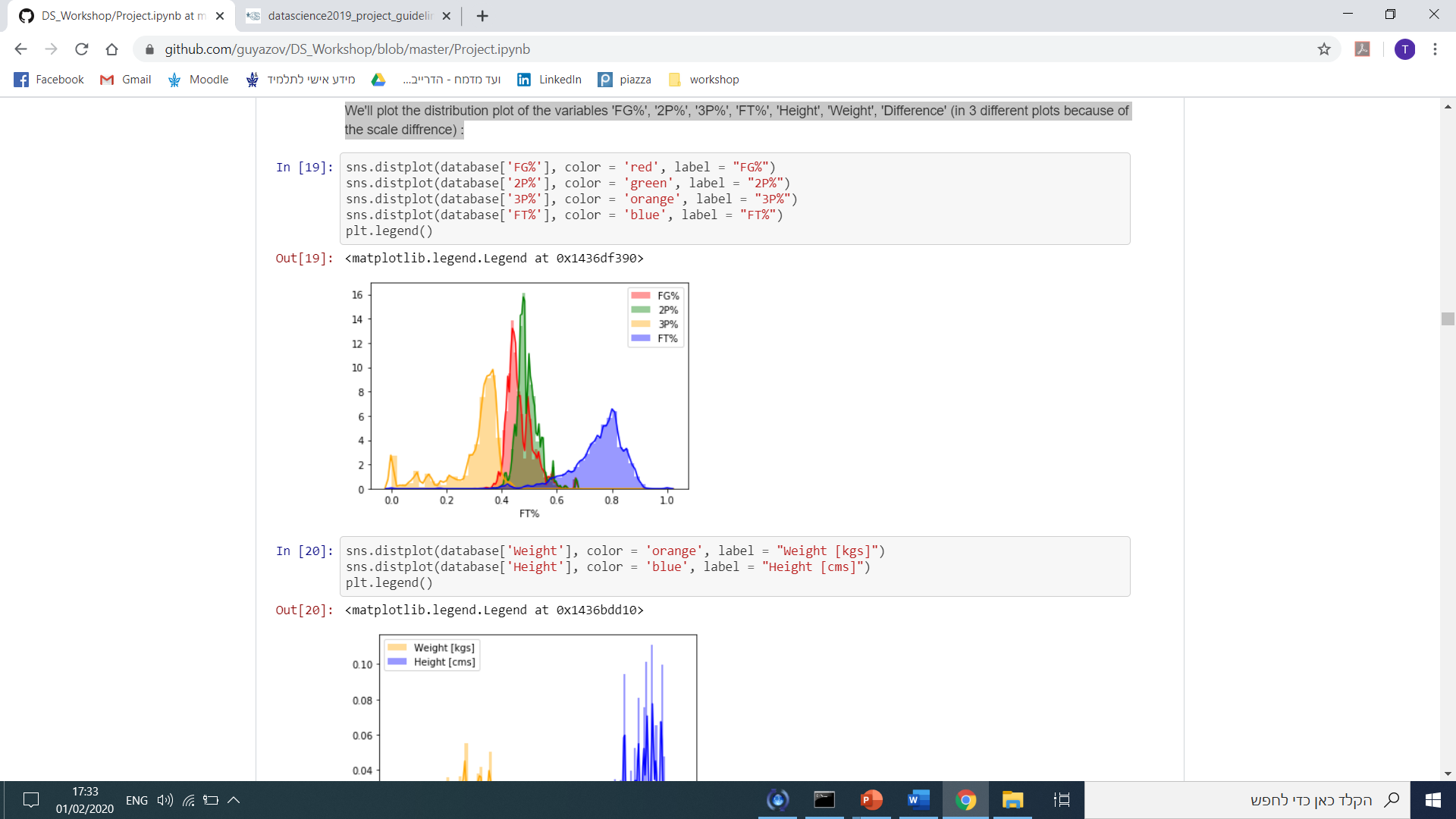
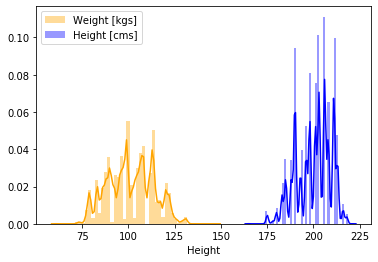
The correlation between the player position and the free-throw result: We want to analyze if the "Position" feature of each player could be a helpful feature that may help our prediction model. Here is a plot showing the distribution of the FT% over the different positions.

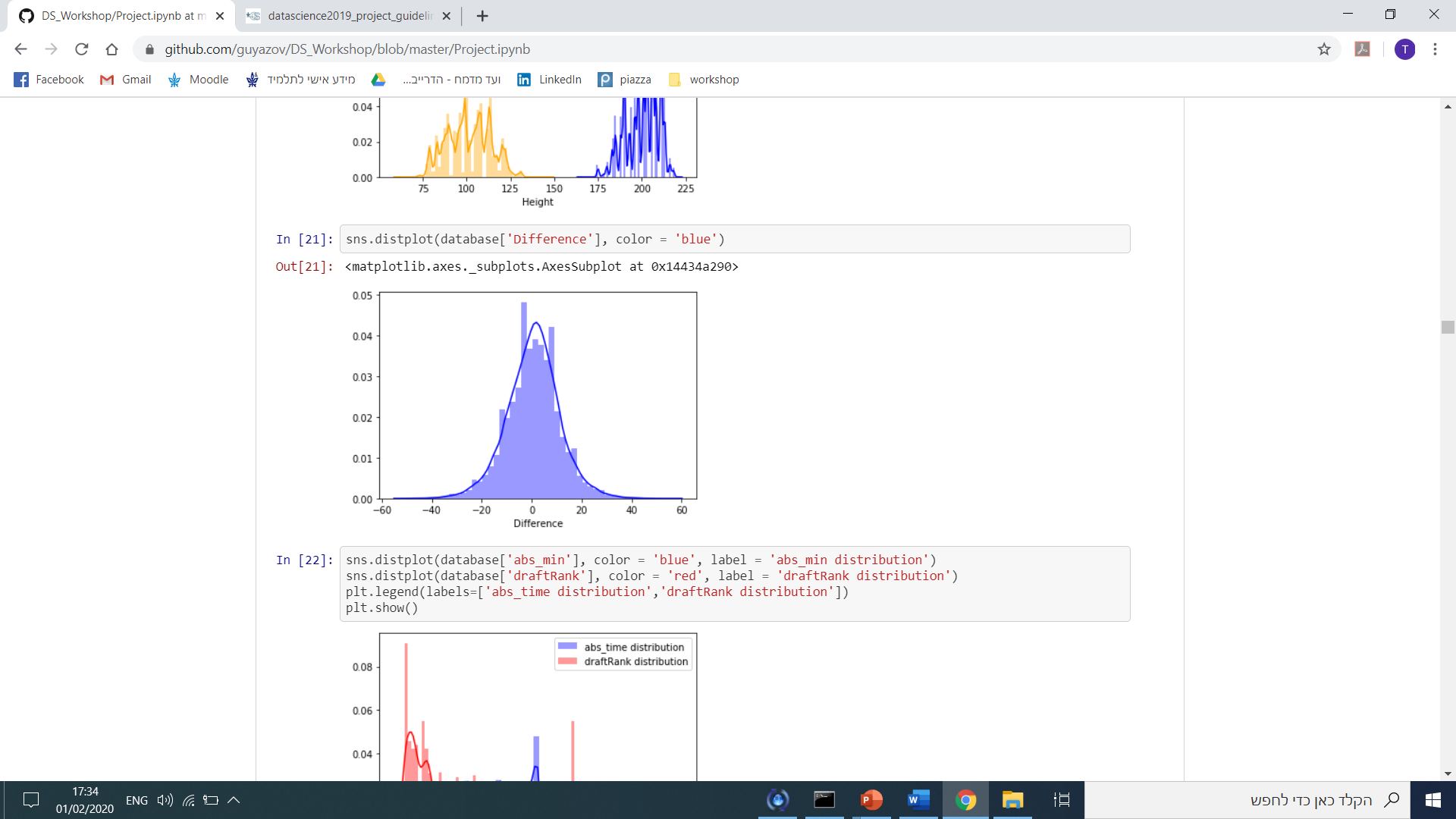
Correlation between all variables: 

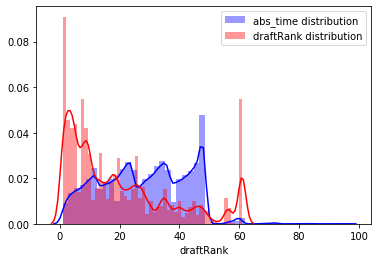
Average FT% at each position: We can see that Centers players owns the worst percentage. Point Guards and Shouting Guards players holds the best percentage over all the positions.

**Problem formulation & the point of focus**

* Main Target: Try to predict whether a player will score a **certain free throw** (online prediction).
* Secondary Target: Predict the players success percentage for the **current** **season** based on his past performance and give the coach an insight about the player (offline prediction).

Our data is **unbalanced**: 76% of the throws with shot made and 24% with shot missed. We split our data to train and test randomly and made sure we keep this ratio in the train and test groups. We chose the relevant variables for our model: playoffs, Pos, ShootingHand, FG%, 2P%, 3P%, FT%, Height, Weight, draftRank, abs\_min, Difference, shot\_made. we assume that our variables are normally distributed. We will examine this hypothesis and normalize the variables. We'll plot the distribution plot of the variables 'FG%', '2P%', '3P%', 'FT%', 'Height', 'Weight', 'Difference' (in different plots because of the scale difference):





**Baseline Model**