COSC2670 Practical Data Science with Python
Assignment 1 – Data Cleaning and Summarising

Joyal Joy Madeckal - s3860476

Data Preparation

The data for the assignment is loaded from the provided CSV file **NBA_players_stats.csv** using **pandas**. Once the data is loaded, head of the data (First 5 observations) was printed out and compared with CSV file to ensure data import is proper. **Number of observations** is also **validated** against the **CSV data**. The name of the players is containing some special characters. But didn't consider performing any actions on names as we are exploring the total points of the players and player name has no role and won't cause any issue for our analysis. Identified the character columns as 'Player', 'Pos' and 'Tm' using **dtypes ()**.

Following sanity checks have been carried out on different columns of the dataset:

Columns	Sanity Checks
Rank	As there are 512 observations, sanity check has been performed to ensure that rank stays in the range 1 to 512
Player, Position, Team	Ensured there are no missing values in these columns using **count ().** Ensured the values inside the columns are not some dummy values using **value_counts ()**
All numeric columns	Plotted a **scatter plot** for each of the column against 'Rank' to find potential deviations. Also, checked the **maximum** and **minimum** values of each of the columns using **describe ()** to find if impossible values are existing.
Games and Minutes Played	As one game is only 48 minutes, total minutes played should not be greater than Games * 48
Field Goal and Field Goal Attempts, 3-Point and 3-Point Attempts, 2-Point and 2-Point Attempts, Free Throw and Free Throw Attempts	Ensured attempts are always greater than or equal to the goals scored
Offensive, Defensive and Total Rebounds	Ensured total rebounds is the sum of defensive and offensive rebounds
Personal Fouls	In a game player can make 6 fouls. Checked Personal Fouls is lesser than Games * 6
Total Points and Field Goals	Checked Total Points is the sum of 2-Point, 3-Point and Free Throw goals. Ensured Field Goals is the sum if 2-Point and 3-Point Field Goals.

Details of how the potential errors have been found out and treated are stated below:

Redundant White Space

Character columns are identified as ‘Player’, ‘Pos’ and ‘Tm’. Possibility of redundant white space cannot be ruled out here. Hence, applied the **strip ()** method on every value on each of these columns using **apply ()** to get rid of the redundant white spaces. This helps to eliminate trailing and leading white spaces, if exists.

Data Entry Errors

From the specifications, we know the possible values for **Position** and **Team** columns. Using **nunique ()** and **value_counts ()** on both Position and Team columns found out that there are data entry errors for the columns. For Positions, instead of **7** distinct values, there were **14** distinct values in the dataset and by observing we could find out the erroneous data. The **erroneous data** is **eliminated** by **direct assignment** of the correct values. For Team, instead of **31** distinct values, there were **33** distinct values and here also by **direct assignment**, the wrong values are replaced with the correct values.

Missing Values

Identified there are missing values present as **FG% - 3, 3P% - 33, 2P% - 7 and FT% - 32** using **count ()**. Found out whenever the FGA/3PA/2PA/FTA are ‘0’ we are having no values on the corresponding columns as **division by zero** leads to error. For the analysis we are going to perform here, it’s better if we consider the missing values as ‘0’ since it won’t be an outlier as ‘0’ already exists and won’t cause issue for the data analysis going to be performed. Using **fillna ()** filled all the missing values as ‘0’ so that that observation need not be dropped from the analysis.

Impossible Values

By common knowledge, we were able to find out the errors with Age column. Two observations were showing ages of players as **-19** and **280**. These values of age are impossible for the players. There is high probability that age **-19 can be 19** considering the facts **negative age is impossible** and **minimum age** of an **NBA player is 19** (Wikipedia, 2021). Based on the statistics of the players having age around 28, concluded that age of the player can be probably 28 instead of removing the observation. So, by **direct assignment the impossible ages were changed**.

Outliers

For finding the outliers best option is **visualisation** and **max** and **min** values of each column. So, plotted the **scatter plots** of all the numeric columns against Rank column. Found outliers for **Personal Fouls, 3P%, FG%, 2P%** and **Total Points** columns. Analysed the case by going deep into the observation to find the reason for the outliers. For Personal Fouls, **one** player’s value was showing as **228** which is the maximum that can be attained by the player. The player has played **38 games** and **228 personal fouls** is achievable only if the player has committed **6 fouls in all the games**. This is kind of an **improbable** scenario. Hence, this observation is removed using **drop ()**. There are 2 outliers for Total Points – **20000** and **28800**. These values are also not correct as its stated in the specification that a player’s Total Points should be lesser than 2000. Total Points of the player is the sum of 2-Point, 3-Point and Free Throw goals. Summing up the values it’s observed that **2** and **288** were the respective values for the outliers. By **direct assignment** these values were changed accordingly. **8 outliers** present from **3P%, FG% and 2P% were removed** since almost all the **numeric column values** were **between 0 to 5 except Age**. But there was only a **difference of 0.5 in mean age of players after dropping the observations** which also showed that these observations don’t contribute much for data exploration.

Data Exploration

Task 2.1

Following bar chart indicates the composition of the Total Points of the top 5 players.

2P_Points – Points obtained by the player by scoring 2-Point Field Goals (2P * 2)
3P_Points – Points obtained by the player by scoring 3-Point Field Goals (3P * 3)
FT – Total Free Throw points from Free Throw goals

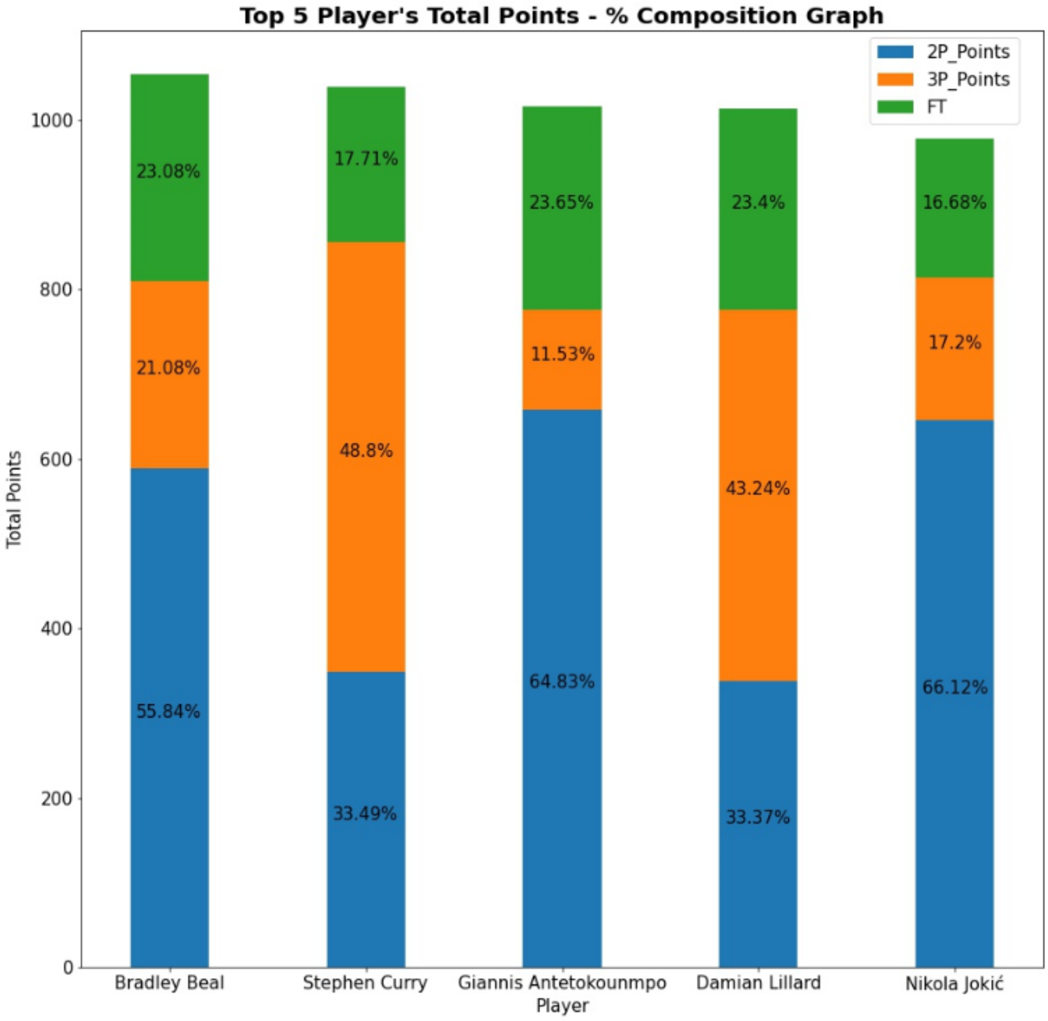


Figure 1

From Figure 1, we can clearly see that majority of the points scored by the players are in 2-Point category. The minimum percentage of 2-Point Field goal is 33.37% for the player Damian Lillard.

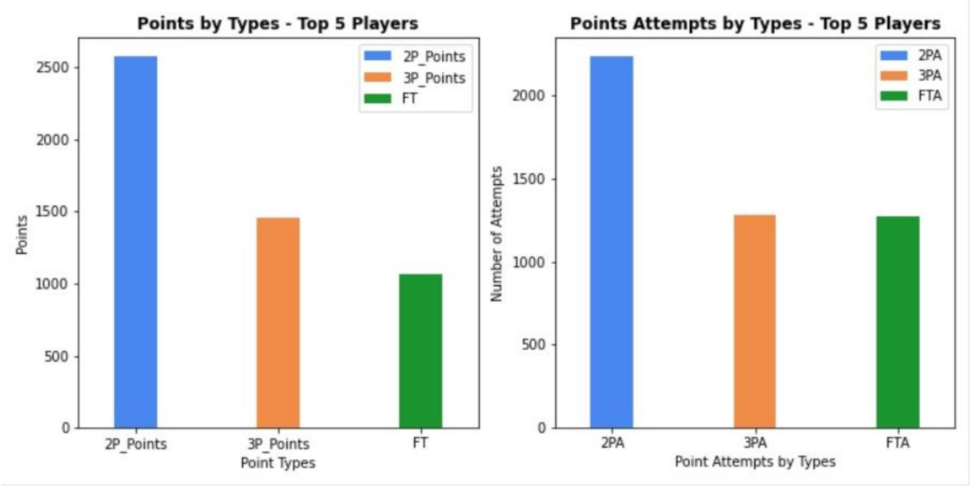
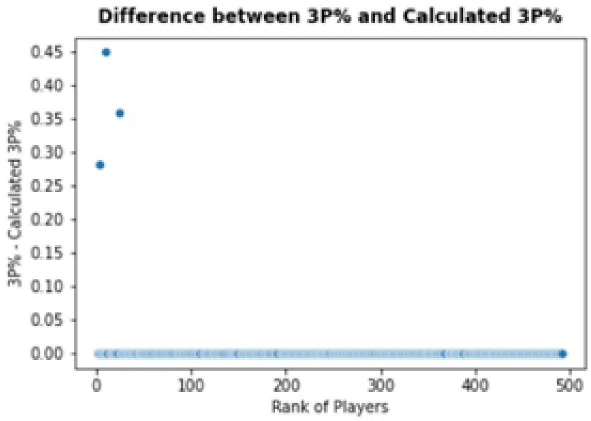


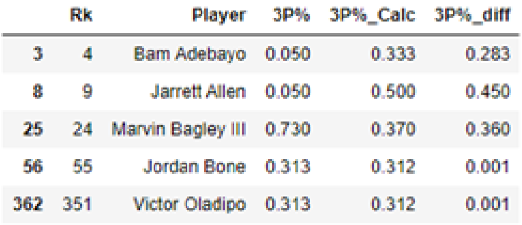
Figure 2

After analysing Figure 2 we can see that major contribution for the Total Points is coming from 2-Point Field goals followed by 3-Point Field goals and Free Throws. And, from the attempts we can see chances for scoring 2-Point goals is more for a player and 3-Point and Free Throw chances follows thereafter. It's not possible to deduce that getting 2-Point Field goals are easy to score. But, for a player the chances for scoring 2-Point Field goal is considerably more when compared with the other two categories.

Task 2.2

The task asks us to explore errors lying in 3P, 3PA and 3P% columns. From the specification provided **3P% = 3P / 3PA**. For finding out the error, I have added a **new column 3P%_Calc** and did the **calculation 3P / 3PA**. Added another column **3P%_diff = abs (3P% - 3P%_Calc)**, where I calculated the difference between the given value in the dataset and value calculated. Plotted a scatter plot for 3P%_diff against Rank. **3P%_diff value should be '0' for every player – ideal case**.





	Rk	Player	3P%	3P%_Calc	3P%_diff
3	4	Bam Adebayo	0.050	0.333	0.283
8	9	Jarrett Allen	0.050	0.500	0.450
25	24	Marvin Bagley III	0.730	0.370	0.360
56	55	Jordan Bone	0.313	0.312	0.001
362	351	Victor Oladipo	0.313	0.312	0.001

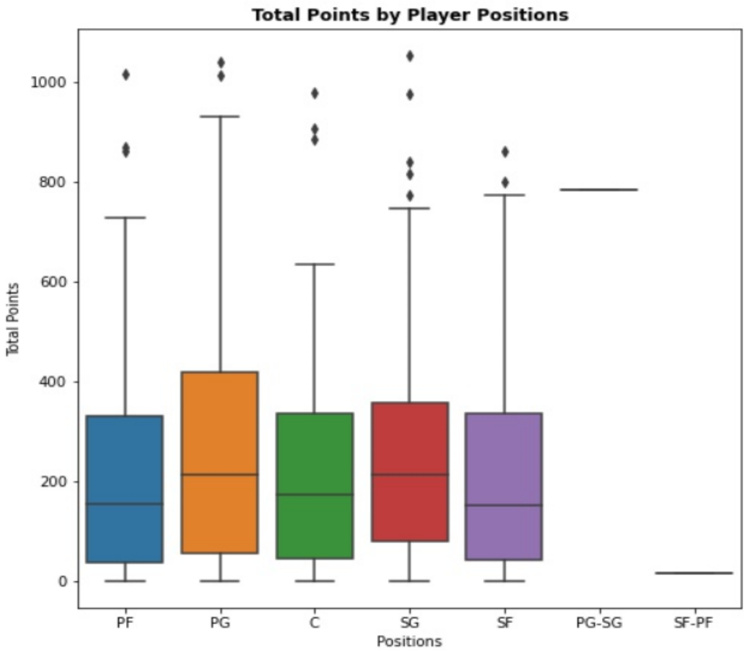
Figure 3

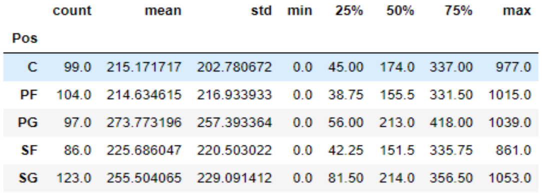
The scatter plot (Figure 3) indicates that there are **three** values greater than '0'. It's clear that these are calculation mistakes and I have **resolved** the issue by **direct assignment of 3P%_Calc values to 3P%**. For the last two observations (with value 0.001, From Figure 3), we can consider it as **rounding off error**. The actual 3P% value for these observations is 0.3125 and in the given dataset it shown as 0.313. **Properly rounding off we can consider the value as 0.312** (Using **round ()** in python.)

Task 2.3

This task asks us to analyse the relationship of different features with Total Points of players. I have considered the following features for this task:

1. Relationship between **Total Points** and **Positions** of players.





	count	mean	std	min	25%	50%	75%	max
Pos								
C	99.0	215.171717	202.780672	0.0	45.00	174.0	337.00	977.0
PF	104.0	214.634615	216.933933	0.0	38.75	155.5	331.50	1015.0
PG	97.0	273.773196	257.393364	0.0	56.00	213.0	418.00	1039.0
SF	86.0	225.686047	220.503022	0.0	42.25	151.5	335.75	861.0
SG	123.0	255.504065	229.091412	0.0	81.50	214.0	356.50	1053.0

Figure 4

Hypotheses: PF position scores most followed by Guards and Centre. (Quora, 2020)

From the boxplot and statistics table shown (Figure 4), we can infer that a player in position PG has chance to get higher points. The top whisker and central 50% values for PG is more spread out than other positions which shows there is more variability in the player's point for position PG. SG position is having a slightly higher value for median (From table, Figure 4) compared with PG but, the variability is less since the box and whisker is more condensed. As per the boxplot we can rank players based on, probability that a chosen player has higher point as **PG > SG > PF = SF > C** slightly different from the hypotheses. Positions PF and SF has almost the same distribution characteristics. Since, position C has a smaller top 25% value whisker spread, it can be considered at the end of the chain. I didn't consider the positions PG-SG and SF-PF since only one observation is present.

2. Relationship between **Games**, **Minutes played** and **Total Points**.

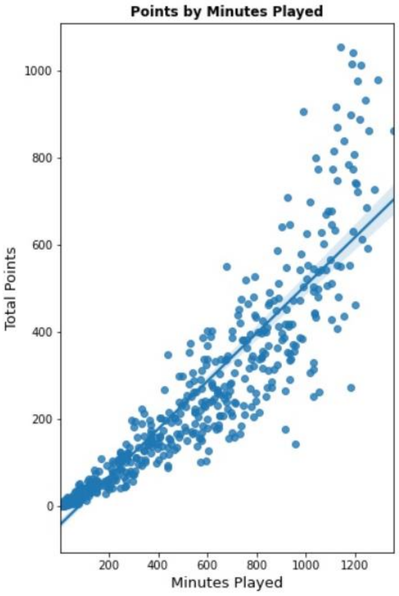


Figure 5.1

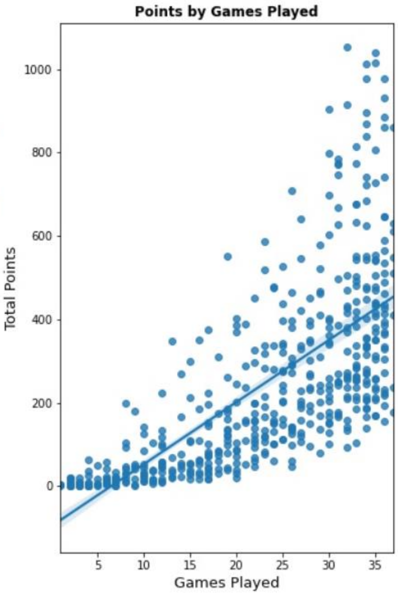


Figure 5.2

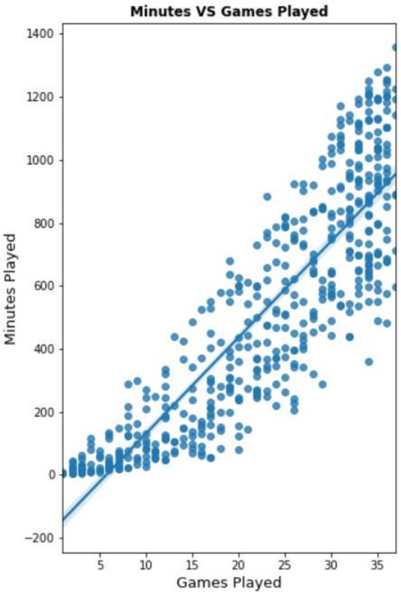
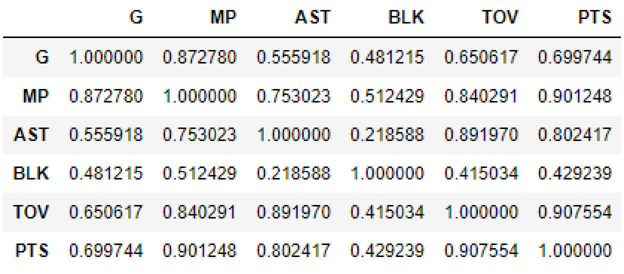


Figure 5.3

__Hypotheses:__ Games and Minutes played should follow a linear relationship with Total Points. The regression plots (Figure 5) show us that there are linear relationships existing between the variables Games, Minutes played and Total Points. Here, we can **check the linearity of these relationships**. **Linearity between Total Points and Games (Figure 5.2) is lesser than the linearity between Total Points and Minutes played (Figure 5.1)**. Practically, if we think there can be cases where a player can play several games but, the time he was on court is less. If a player plays for 10 mins in a game and 30 mins in another game, total Games = 2 whereas Minutes played = 40 (96 mins is possible). As the involvement in the game is less chances for scoring gets reduced. This explains why the scatter is more in 2nd graph compared with 1st. And from the table below we see the correlation between Minutes and Points is **0.9** and between Games and Points is **0.7** indicating the same. The 3rd graph (Figure 5.3) shows us that Games and Minutes played also follows a healthy linearity of **0.87** (Table 1). Analysis goes with the hypotheses.

3. Relationship between **Assists**, **Blocks** and **Turnovers** with **Total Points**

__Assists VS Points__: **Hypotheses -** Assists and Points will form a linear relationship.
Assists and Points forms a linear relationship with a **linearity** of **0.8** (Table 1). This can also be practically explained. A player who is good in assists (Passing ball to teammate leading to a field goal) should also be a good goal scorer. If a player is having good tactics joined with shooting accuracy the player can make more assists and score more. This same trend can be seen in other games like Football as well. As we can infer from the 1st graph (Figure 6.1) the top players are positioned at the top end of the graph and there is clustering at the start which goes with the general notion that number of good players will be less. Analysis goes with the hypotheses.



	G	MP	AST	BLK	TOV	PTS
G	1.000000	0.872780	0.555918	0.481215	0.650617	0.699744
MP	0.872780	1.000000	0.753023	0.512429	0.840291	0.901248
AST	0.555918	0.753023	1.000000	0.218588	0.891970	0.802417
BLK	0.481215	0.512429	0.218588	1.000000	0.415034	0.429239
TOV	0.650617	0.840291	0.891970	0.415034	1.000000	0.907554
PTS	0.699744	0.901248	0.802417	0.429239	0.907554	1.000000

Table 1: Correlation Table

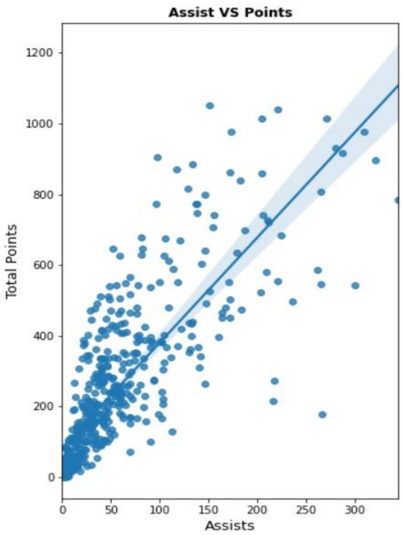


Figure 6.1

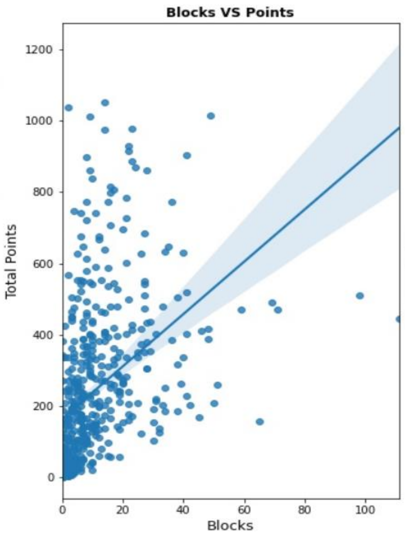


Figure 6.2

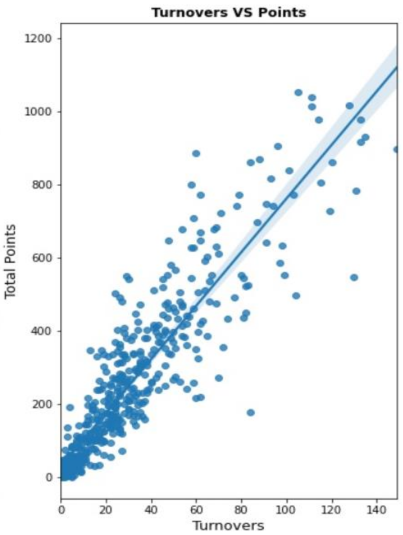


Figure 6.3

Blocks VS Points: **Hypotheses -** Blocks and Points will have no relationship.
Here, we see the **linearity** is **0.43** (Table 1) which suggests there is not much linear relationship between the two variables. Along with this, the **huge confidence interval** on the graph (Figure 6.2) also suggests the same. We can interpret the idea in this manner: **A player who is good at blocks means the player is skilful in defending his own basket rather than attacking the other basket**. So, this player probably might be passing the ball to the forwards for scoring. And **blocks are not contributing anything to the total points**. These ideas explain the graph and goes with the hypotheses.

Turnovers VS Points: **Hypotheses -** Turnovers and Points will have negative relationship.
The **linearity** here is **0.91** (Table 1). Turnovers are not good for a player in basketball. When the player loses the ball it's a turnover. The plot is suggesting that a player with more points will be having more turnovers. This seem to be little confusing because a player who scores more means he should be skilful, good shooter etc. Let us try to think in a different direction. When a player is proceeding for a basket, the defence of the other team will try to intercept and as the game will be very quick - in and around the basket, there are chances that the player might go out of bounds losing the ball to the other team. This leads us to see that the probability for turnovers to happen is more when a player is proceeding to score. So, the player with higher points might be involved in more turnovers. Hypotheses was completely wrong here as the data exploration suggests it forms a linear relationship.

Reference List

- Portilla, J 2021, Python for Data Science and Machine Learning Bootcamp, streaming video, Udemy for Business, viewed 22nd March to 10th April 2021, < https://hexagoncci.udemy.com/course/python-for-data-science-and-machine-learning-bootcamp/learn/lecture/5440650#overview>
- Cielen, D, Meysman, A & Ali, M 2020, *Introducing Data Science,* Dreamtech Press, Delhi.
- Hunter, J, Dale, D, Firing, E, & Droettboom, M 2021, *Matplotlib: Visualization with Python*, viewed from 22nd March to 10th April 2021, < https://matplotlib.org/stable/index.html#matplotlib-visualization-with-python>
- *2020-21 NBA Player Stats: Advanced,* 2021, <https://www.basketball-reference.com/leagues/NBA_2021_advanced.html >
- Yongli, R 2021, 'COSC2670 -> Modules', PowerPoint slides & Pre-recorded lectures, COSC2670, RMIT University, viewed from 22nd March to 10th April 2021, < https://rmit.instructure.com/courses/79792/modules>
- Wikipedia 2021, *Eligibility for the NBA draft,* viewed on 10th April 2021, <https://en.wikipedia.org/wiki/Eligibility_for_the_NBA_draft>
- Quora 2020, *Which position (guard, forward, centre) of an NBA player could have a higher chance of scoring each of these: points, rebounds, assists?*, viewed on 13th April 2021, < https://www.quora.com/Which-position-guard-forward-center-of-an-NBA-player-could-have-a-higher-chance-of-scoring-each-of-these-points-rebounds-assists>