Project Report

Liam Glennie

December 2, 2021

Contents

1 Introduction

Since Tiger Woods became a professional back in 1996 the way how golf is played has been changed forever. Over the past 25 years, how far a player hits the ball off the tee has become increasingly important. It seems that longer-hitters have been gaining more and more success while shorter-hitters have become a dying breed. I believe part of this is because the greater a player's average distance off the tee is, the closer their shot into the green will be. And the closer the ball is to the hole on the green, the easier it is to hole the next shot.

Therefore, the main objective of this study is to investigate whether there is a positive correlation between a player's average driving distance and their strokes gained approach-the-green. That is, does a longer-hitter tend to have more strokes gained approach-the-green than a shorter-hitter.
Additionally, we will also study the relationship between how far a player drives the ball and how many fairways they find. In other words, does the distance a player hits the ball off the tee on average effect the dispersion of their tee shots.

2 Data set

The data set [1] selected for this study contains 2313 instances of PGA Tour players' stats from 2010 until 2018. There are 12 attributes:

- *name*: A player's full name. (string)
- *rounds*: Number of rounds of golf a player has played in a season. (numerical)
- *fPerc*: Percentage of fairways found on par-4s and par-5s. (numerical)
- *year*: Year in which the rest of the stats have been collected from. (numerical)
- *avgDist*: A player's average distance off the tee on par-4s or par-5s. (numerical)
- *gir*: Percentage of greens in regulation found. (numerical)
- *avgPutt*: Average putts per round. (numerical)
- *avgScram*: Average scrambling percentage (Percentage of times a player can get PAR when they miss a green in regulation). (numerical)
- *avgSco*: A player's average score. (numerical)
- *points*: Number of FedEx Cup points scored. (numerical)
- *wins*: Number of wins. (numerical)
- *top10*: Number of times a player has finished tied 10th or better. (numerical)
- *avgSGPutts*: Measures how many strokes a player gains (or loses) on the greens compared to the PGA Tour average. (numerical)

- *avgSGTot*: Strokes gained: off-the-tee + strokes gained: approach-the-green + strokes gained: around-the-green + strokes gained: putting. (numerical)

- *SG:OTT*: Strokes gained off-the-tee measures player performance off the tee on all par-4s and par-5s. This statistic looks at how much better or worse a player's drive is than the average PGA Tour player. (numerical)

- *SG:APR*: Strokes gained approach-the-green measures player performance on approach shots and other shots that are NOT included in strokes gained: around-the-green and strokes gained: putting. It does include tee shots on par-3s. (numerical)

- *SG:ARG*: Strokes gained around-the-green: Measures player performance on any shot within 30 yards of the edge of the green without measuring putting. (numerical)

- *money*: Amount of money won in dollars. (numerical)

For this study, there are 4 interesting attributes to investigate: *avgDist*, *SG:APR*, *fPerc* and *rounds* to only take into account those instances that have played at least 15 rounds.

3 Exploratory analysis

In this study we will perform exploratory analysis on a subset of the attributes of the full dataset. The attributes in the subset were chosen based on their possible relation to *avgDist* and *SG:APR*. The subset is formed by: *fPerc*, *avgDist*, *gir*, *avgSco* and *SG:APR*. It is worth mentioning that there were 634 instances that had null values for these attributes. They were cleaned and removed for the rest of the study. See annex: 6.1 6.2 6.3

3.1 Data description

In the following table we can find some of the main statistics for each of the attributes selected:

	fPerc	avgDist	gir	avgSco	SG:APR
count	1678	1678	1678	1678	1678
mean	61.44	290.81	65.66	70.92	0.07
std	5.06	8.92	2.75	0.70	0.38
skew	-0.01	0.25	-0.32	0.17	-0.27
kurt	-0.11	0.03	0.42	1.08	0.84
min	43.02	266.40	53.54	68.70	-1.68
25%	57.94	284.90	63.83	70.49	-0.18
50%	61.43	290.55	65.79	70.90	0.08
75%	64.91	296.40	67.58	71.34	0.31
max	76.88	319.70	73.52	74.40	1.53

Table 1: Table of statistics. See annex: 6.4

There are some interesting statistics in this table. For example, we can see that the data for the attributes *fPerc* & *avgSco* will follow a normal distribution due to them having a low

skewness value. However, the rest of the attributes have moderately skewed data as their skewness value is greater or equal to 0.25. In terms of the spread of these distribution, by analysing both the standard deviation and the Kurtosis values, we can see that *avgSco* and *SG:APR* will have very narrow distributions when we visualise them later on in the report.

However, it's worth pointing out the difference the shortest-hitter and the longest on the PGA Tour. So, between their average driving distance there is a total of 53.3 yards. With that difference in driving distance, we can see that the longer-hitter, Rory McIlroy, has a higher percentage of *gir*, slightly more *SG:APR* and, most importantly, a lower *avgSco* meaning that McIlroy averages 1.27 shots less than Gay per round. This is a total of 5.08 shots per tournament. Nonetheless, it's not all better for the longer-hitter because, as we can see in Table 2, the shorter-hitter has a much higher *fPerc* finding an average of 13.11 more fairways per tournament than Rory McIlroy.

	fPerc	**avgDist**	**gir**	**avgSco**	**SG:APR**
Rory McIlroy	55.79	319.70	66.30	69.30	0.269
Brian Gay	74.00	266.40	63.44	70.57	0.159

Table 2: Comparing longest and shortest-hitter. See annex: 6.5

3.2 Density Plots

Given the density plots in Figure 1, we can confirm that the data of the attributes *fPerc* & *avgSco* follow a normal distribution with little to no skew. On the other hand, we can see that *gir* & *SG:APR* have a moderate negative or left skew and *avgDist* has a moderate positive or right skew. Additionally, we can observe that *avgSco* & *SG:APR* have narrow distributions meaning that there is very little spread in the data for these 2 attributes.

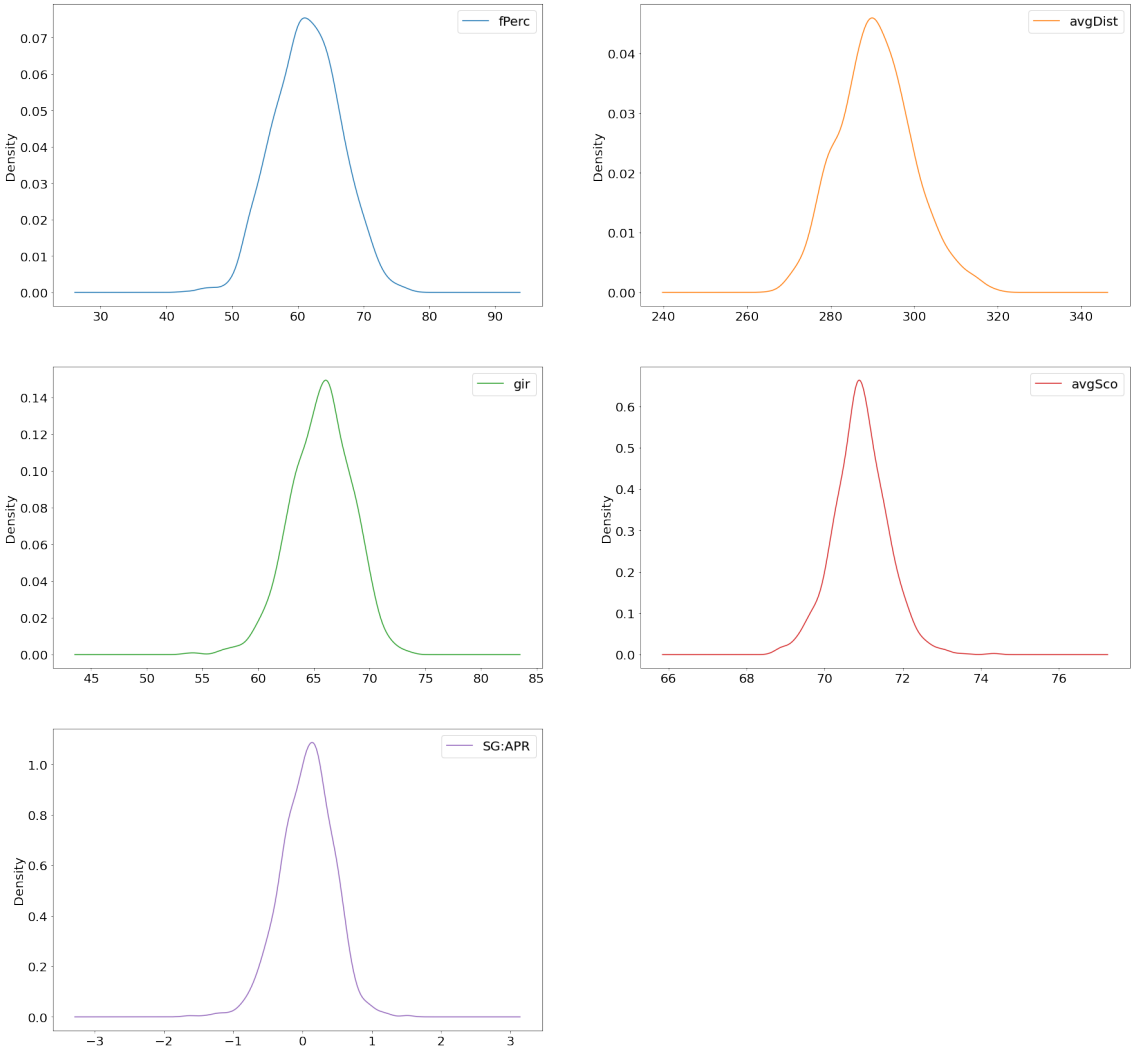

Figure 1: Density plots. See annex: 6.7

3.3 Histograms

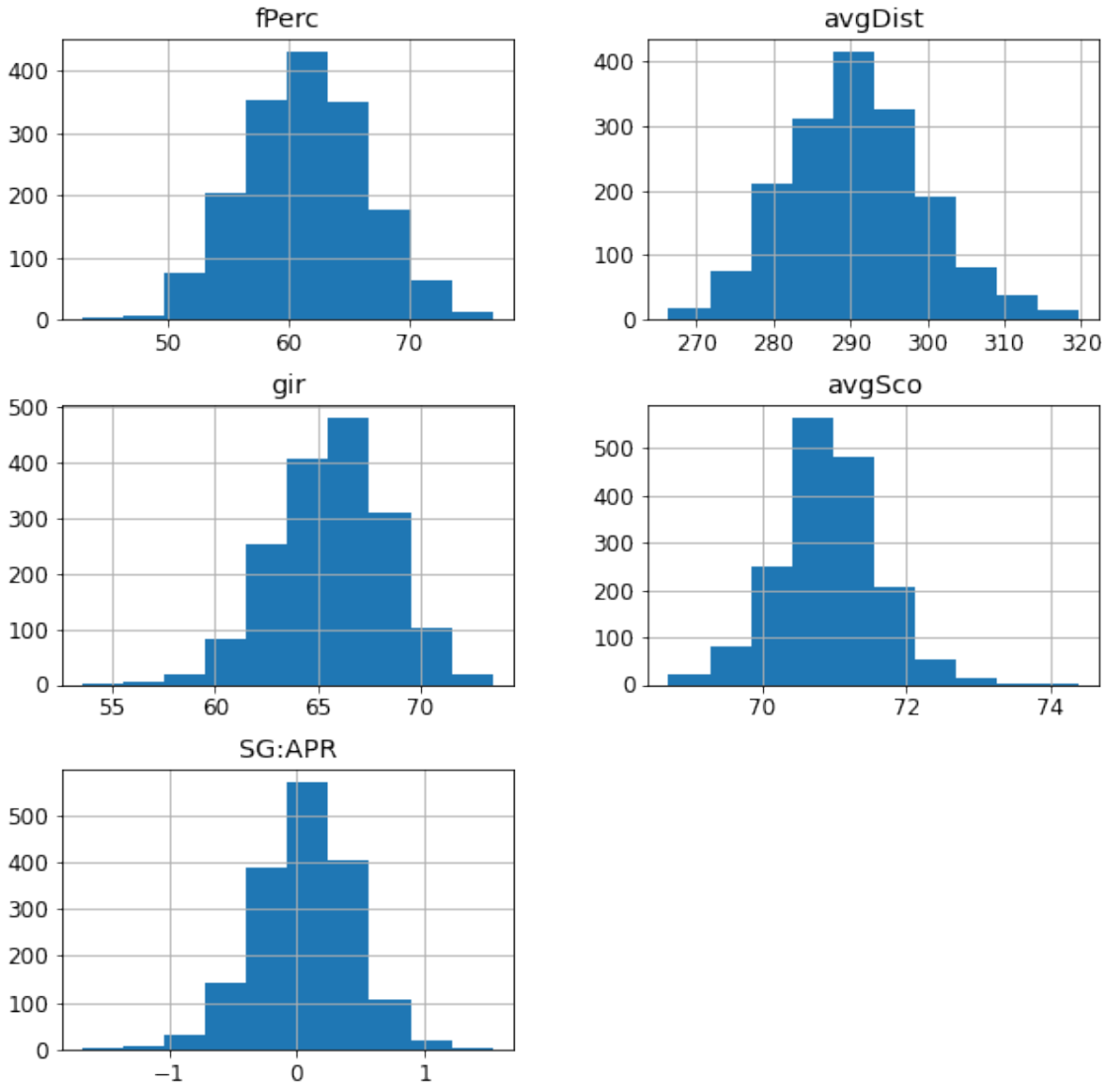

Figure 2: Histogram plots. See annex: 6.6

The reason we have chosen to visualise the attributes using both density plots and histograms is because we believe that density plots represent the spread of the data better than histograms. On the other hand, histograms represent skewness better than density plots. As we can see in Figure 2, the skewness of *gir* is much more appreciable than in Figure 1.

3.4 Comparing average distance over the years

There's one last plot that contains at least one of the main attributes of this study. We believed it would be interesting to analyse the evolution of the average driving distance over the years 2010-2018.

There does seem to be an upward trend in distance off the tee. Now, this could be because of the evolution of technology in the world of golf. Golf clubs are now faster and longer than they were before. And technology has had a big impact on how some player's train, helping

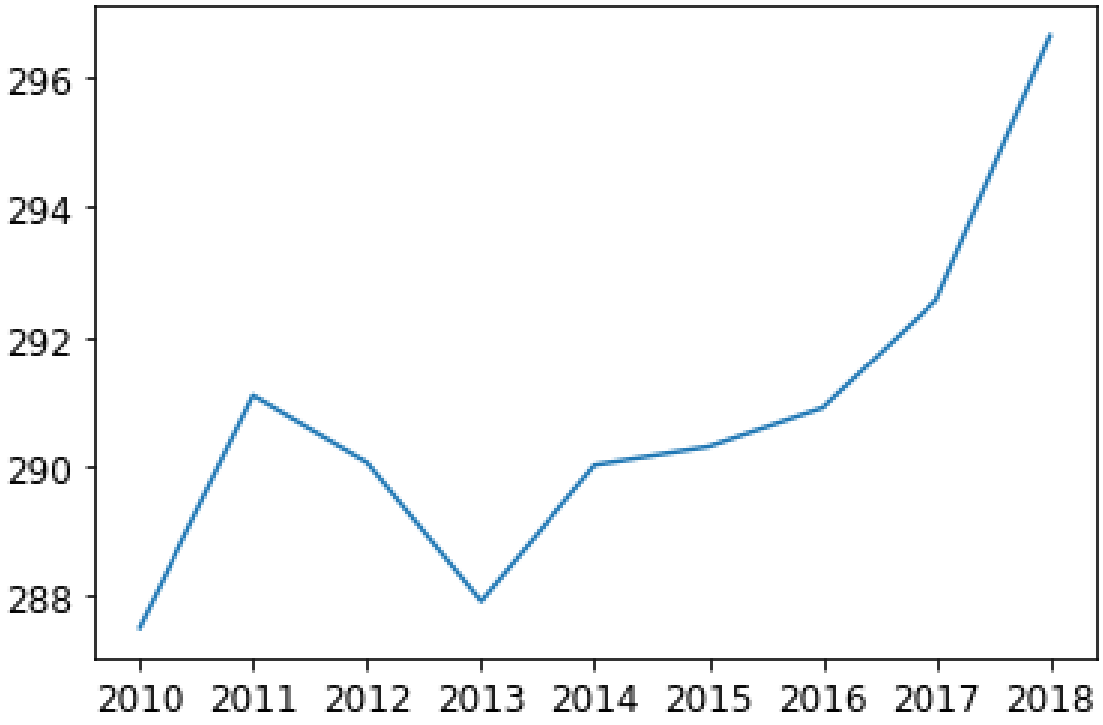

Figure 3: Average distance over the years. See annex: 6.9

players to get stronger and to get the most distance out of their swing. I would be interesting to investigate the cause of this upwards trend in distance in a future project.

4 Relationship Analysis

In this section, we will be performing correlation and hypothesis tests on the attributes *avgDist*, *SG:APR* & *fPerc* to see if there exists a correlation between them.

4.1 Scatter Plots

Let's focus on the scatter plot between the two main attributes of this study: *avgDist* and *SG:APR*. Despite our initial hypothesis, it doesn't seem like there's much of a relationship between these two attributes. Nonetheless, we can still find some interesting plots in Figure 5.

For example, there seems to be a negative correlation between *fPerc* and *avgDist*. Therefore, it is likely that the higher a player's average distance off the tee is, the lower their percentage of fairways found will be. And this could be because with more power, comes less control. So, the further a player can hit the ball, the harder it is for them to keep the ball dispersion down. Another scatter plot worth noting is the one between *avgSco* and *gir*. It seems that there is a slight negative correlation between these 2 attributes, meaning that it is possible that a player with a high percentage of greens in regulation, could have a lower average score than a player with a lower percentage of greens in regulation.

We should mention that some attributes do have a relatively strong correlation but that is because one of the attributes may be used to calculate the other. This is the case, for example, with *gir* and *SG:APR*.

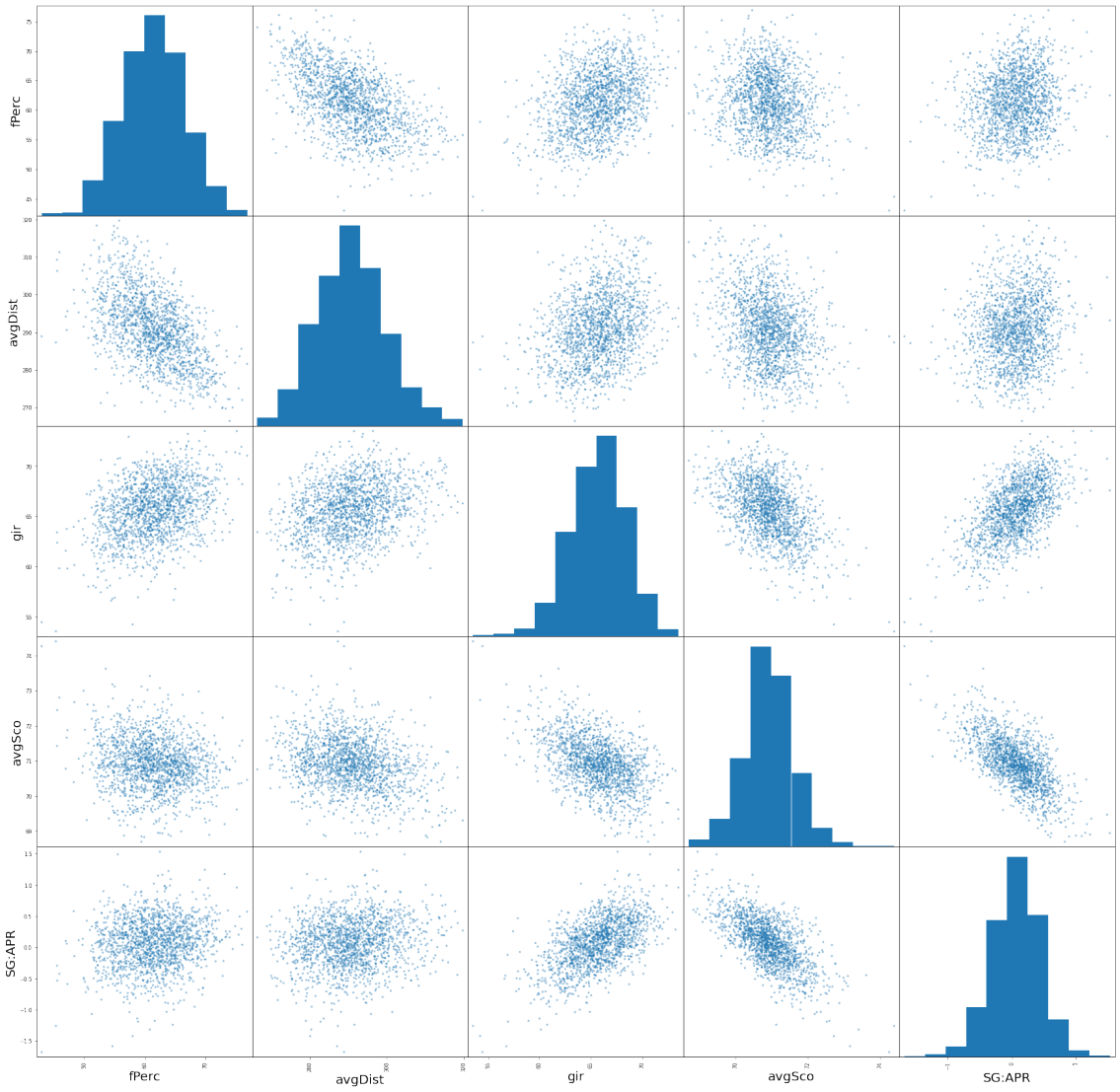

Figure 4: Scatter plots. See annex: 6.8

4.2 Hypothesis Tests

To test our hypothesis we have opted for the Student's T-test as the three attributes are independent and, identically and normally distributed.

4.2.1 *avgDist* & *SG:APR* Student's T-test

For this test, our hypothesis are the following:

- H0: The means of the distributions of *avgDist* & *SG:APR* are equal.
- H1: The means of the distributions of *avgDist* & *SG:APR* are NOT equal.

Upon conducting the hypothesis test for the mean, we get a p-value < 0.000. Therefore, since the p-value is below the significance level of 0.05, we have sufficient evidence to reject the null hypothesis and say that it is statistically probable that the means of *avgDist* & *SG:APR* are different. See annex: 6.10

4.2.2 *avgDist* & *fPerc* Student's T-test

For this test, our hypothesis are the following:

- H0: The means of the distributions of *avgDist* & *fPerc* are equal.
- H1: The means of the distributions of *avgDist* & *fPerc* are NOT equal.

Upon conducting the hypothesis test for the mean, we get a p-value < 0.000. Therefore, since the p-value is below the significance level of 0.05, we have sufficient evidence to reject the null hypothesis and say that it is statistically probable that the means of *avgDist* & *fPerc* are different. See annex: 6.11

4.3 Correlation Tests

In order to see if there exist a correlation between our variables we have opted for the Spearman's Rank Correlation. The reason why we chose this test instead of Pearson's Correlation Coefficient is because the distribution's variance of the attributes are not equal. On the other hand, the distributions of the attributes do meet the requirements of the Spearman's Rank Correlation.

4.4 *avgDist* & *SG:APR* Spearman's Rank Correlation Test

For this test, our hypothesis are the following:

- H0: The attributes *avgDist* & *SG:APR* are independent.
- H1: There exists a dependency between the attributes *avgDist* & *SG:APR*.

Upon conducting the correlation test between the attributes *avgDist* & *SG:APR*, we get a p-value < 0.000. Therefore, since the p-value is below the significance level of 0.05, we have sufficient evidence to reject the null hypothesis and say that it is statistically probable that there exists a dependency between *avgDist* & *SG:APR*. See annex: 6.10

4.4.1 *avgDist* & *fPerc* Spearman's Rank Correlation Test

For this test, our hypothesis are the following:

- H0: The attributes *avgDist* & *fPerc* are independent.
- H1: There exists a dependency between the attributes *avgDist* & *fperc*.

Upon conducting the correlation test between the attributes *avgDist* & *fPerc*, we get a p-value < 0.000. Therefore, since the p-value is below the significance level of 0.05, we have sufficient evidence to reject the null hypothesis and say that it is statistically probable that there exists a dependency between *avgDist* & *fPerc*. See annex: 6.11

4.5 Analysing Correlation Test Results

After visualising the scatter plots on these attributes it seemed unlikely that there would be correlation between them, particularly in the case of *avgDist* & *SG:APR*. And as you can see in Figure 5 & Table 3, there seems to be little to no correlation between these 2 attributes. On the other hand, there is a moderate negative correlation between the attributes *avgDist* & *fPerc*.

	avgDist	**SG:APR**	**fPerc**
avgDist	1.00	0.14	-0.53
SG:APR	0.14	1.00	0.16
fPerc	-0.53	0.16	1.00

Table 3: Correlation Table. See annex: 6.12

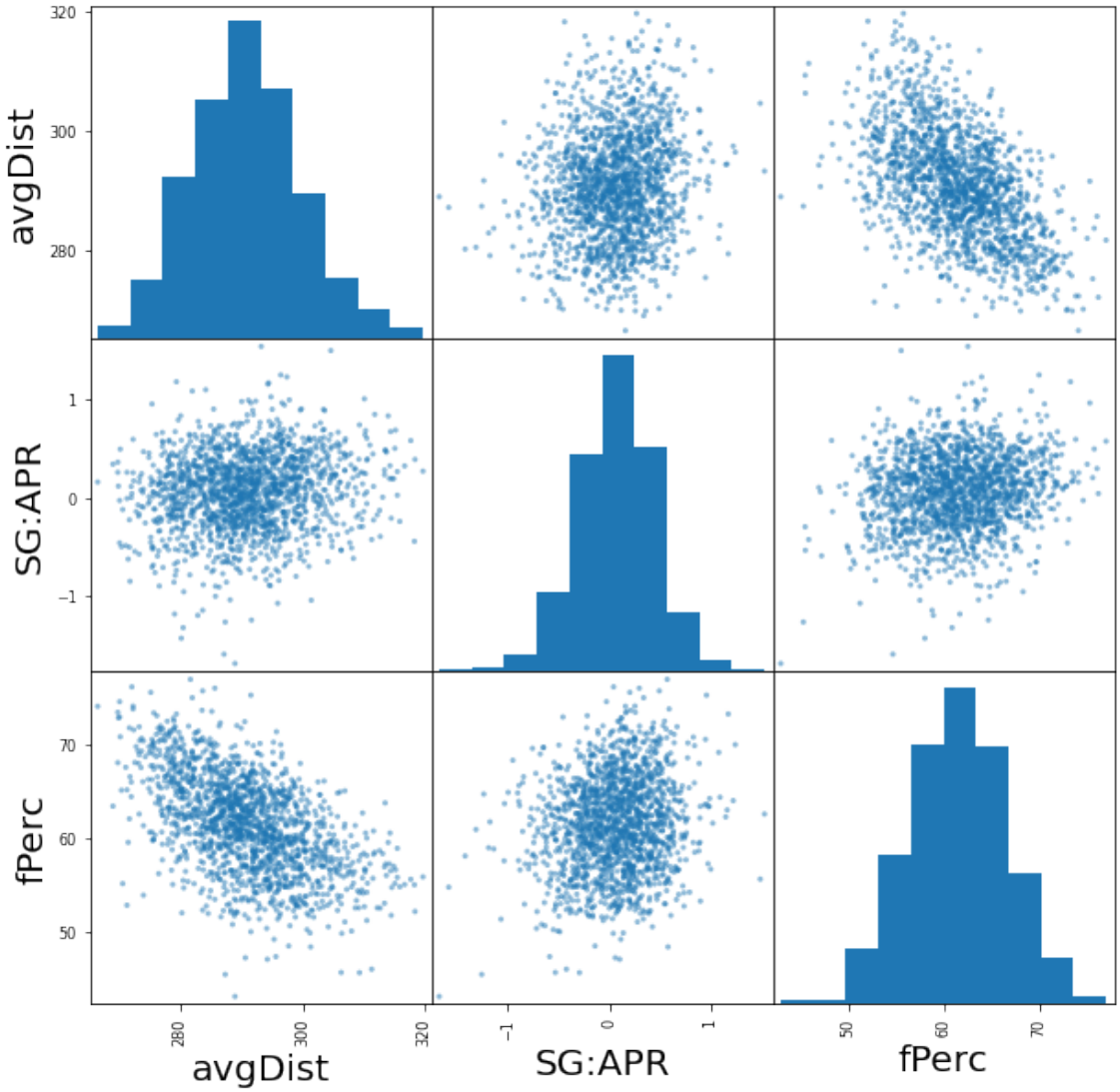

Figure 5: Scatter plots. See annex: 6.13

Having seen the correlations and the scatter plots, it is likely that the test we chose to

test correlation coefficient wasn't the most adequate. Or perhaps, the Spearman's Correlation Coefficient test is problematic when it comes to distributions of a different scale.

5 Conclusion

Throughout this study, we have performed a number of statistical tests and visualisations on the data set [1] and we have obtained some interesting and helpful insights and some unhelpful ones too.

Let's start with our initial hypothesis. This study aimed to discover if longer-hitters on the PGA Tour had a higher number of strokes gained approach-the-green. In other words, if players that hit the ball further off the tee, hit their next shot closer to the hole than players that don't hit the ball as far.
After performing hypothesis tests on the attributes *avgDist*, *SG:APR* & *fPerc* and we have concluded that it is statistically probable that the means of *avgDist* & *SG:APR* are different. In terms of the correlation between the attributes, despite the results obtained in Section 3.1, we have established there is little to no relationship between the attributes *avgDist* & *SG:APR*. Therefore, we cannot say that a player's average distance off the tee has an effect on how close their next shot to the hole will be.

We also had a secondary hypothesis that involves the attributes *avgDist fPerc*. We wanted to investigate whether the distance a player hits the ball off the tee has an effect on how many fairways in regulation they find. Initially, we believed that there would be a negative relationship between these two attributes, because one of the main factors of the distance a player can hit the ball is power. And, in general, with more power often comes less control.
After performing a hypothesis test on the attributes *avgDist* & *fPerc*, we have concluded that it is statistically probable that their means are different. In terms of the correlation between the attributes, we have established that it is likely that there is a negative dependency between *avgDist* & *fPerc*. This is the result of performing a Spearman's rank correlation test. But, we can also see the dependency between the attributes in Table 3 as the correlation coefficient of *avgDist* & *fPerc* is -0.53 meaning that there is a slightly negative correlation.
Therefore, we can conclude that it is likely that a longer-hitter on the PGA Tour will find less fairways in regulation than a shorter-hitter. If we refer back to Section 3.1, we can see that the longest-hitter on the PGA Tour during the years 2010-2018, Rory McIlroy, had a lower percentage of fairways found than Brian Gay, the shortest-hitter.

One last interesting find from this study can be seen in Figure 3 where we visualise *avgDist* over the years 2010-2018. From this, we can see that there is a clear upward trend in average driving distance off the tee with a difference of 9.12 yards over the 9 years. As we mentioned earlier, it is likely that the increase of the use of technology in the world of golf is a key reason for this increment in average distance.

6 Annex

6.1 Importing data

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as sc
from scipy.stats import pearsonr
from scipy.stats import spearmanr

names = ['name',
         'rounds',
         'fPerc',
         'year',
         'avgDist',
         'gir',
         'avgPutt',
         'avgScram',
         'avgSco',
         'points',
         'wins',
         'top10',
         'avgSGPutts',
         'avgSGTot',
         'SG:OTT',
         'SG:APR',
         'SG:ARG',
         'money']

myAtt = ['fPerc','avgDist','gir','avgSco', 'SG:APR']
data = pd.read_csv("pgaTourData.csv")
```

6.2 Missing data & outliers

```
for att in myAtt:
    if data[att].dtype == "int64" or data[att].dtype == "float64":
        print('Analysis of ' + att)
        print("Equal to zero")
        print((data[att] == 0).sum())
        print("Less than zero")
        print((data[att] < 0).sum())
        print("Standard deviation")
        print(data[att].std())
        print("Mean")
        print(data[att].mean())
        print("Null values")
```

```
        print(data[att].isnull().sum())
        print('-------------------------')
```

6.3 Cleaning Data

```
#Cleaning the data
#Only the players with wins have data in the wins columns
data['wins'].fillna(0, inplace=True)
data.dropna(subset=myAtt,inplace=True)
```

6.4 Data description

```
print(data[myAtt].skew())
print(data[myAtt].kurtosis())
data[myAtt].describe()
```

6.5 Comparing longest and shortest-hitter

```
minMaxDist = data[(data['avgDist'] == data['avgDist'].max()) |
                  (data['avgDist'] == data['avgDist'].min())]
#Pretty prints
minMaxDist
```

6.6 Histograms

```
plt.rc('font', size=12)
data[myAtt].hist(figsize=(10,10))
plt.show()
```

6.7 Density plots

```
plt.rc('font', size =20)
data[myAtt].plot(kind='density', subplots=True, layout=(5,2), sharex=False,figsize
plt.show()
```

6.8 Scatter plots of 5 attributes

```
plt.rc('font', size =20)
pd.plotting.scatter_matrix(data[myAtt], figsize=(30,30), alpha=0.5)
```

6.9 Average distance over the years

```
avgDist = data.groupby('year')['avgDist'].mean()
print(avgDist)
plt.rc('font', size =12)
plt.plot(avgDist)
plt.show()
```

6.10 Hypothesis & Correlation tests for *avgDist* & *SG:APR*

```
x = data['avgDist']
y = data['SG:APR']
stat, p = spearmanr(x, y)
# stat, p = pearsonr(x, y)
print('stat=', stat)
print('p=', p)
#Change
if p > 0.05:
 print('Probably independent')
else:
 print('Proabably dependent')

print('T-test: p = '
      + str(sc.ttest_ind(x.values, y.values, equal_var=False)[1]))
```

6.11 Hypothesis & Correlation tests for *avgDist* & *fPerc*

```
x = data['avgDist']
y = data['fPerc']
stat, p = spearmanr(x, y)
# stat, p = pearsonr(x, y)
print('stat=', stat)
print('p=', p)
if p > 0.05:
 print('Probably independent')
else:
 print('Proabably dependent')

print('T-test: p = '
      + str(sc.ttest_ind(x.values, y.values, equal_var=False)[1]))
```

6.12 Correlation Table

```
myAtt = ['avgDist','SG:APR', 'fPerc']
print("Correlation Table")
corr = data[myAtt].corr()
corr.style.background_gradient().set_precision(2)
```

6.13 Scatter plots of 3 attributes

```
plt.rc('font', size =20)
pd.plotting.scatter_matrix(data[myAtt], figsize=(30,30), alpha=0.5)
plt.show()
```

References

[1] jmpark746. Kaggle pga tour data. https://www.kaggle.com/jmpark746/pga-tour-data-2010-2018.