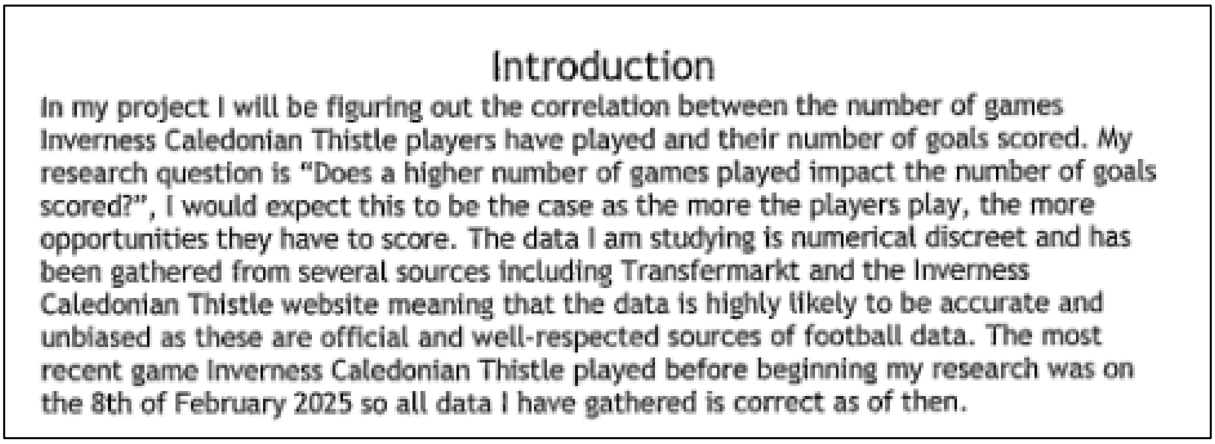
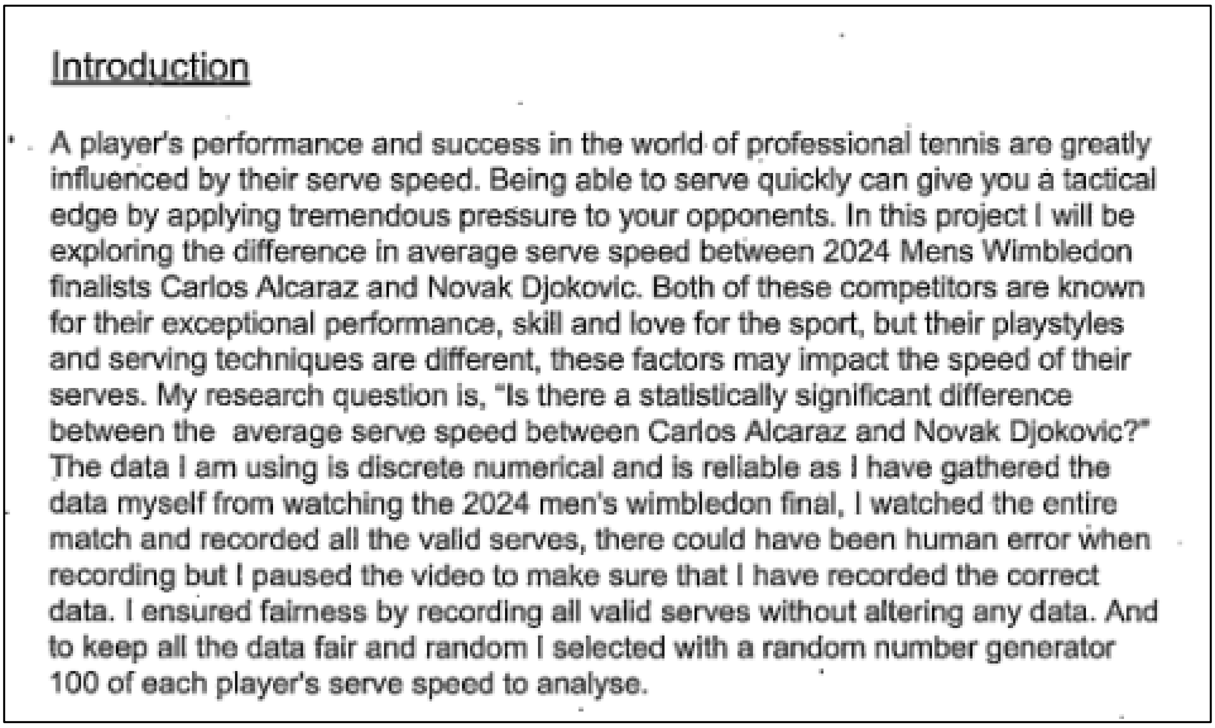
Candidate 1 evidence



Introduction

In my project I will be figuring out the correlation between the number of games Inverness Caledonian Thistle players have played and their number of goals scored. My research question is "Does a higher number of games played impact the number of goals scored?", I would expect this to be the case as the more the players play, the more opportunities they have to score. The data I am studying is numerical discreet and has been gathered from several sources including Transfermarkt and the Inverness Caledonian Thistle website meaning that the data is highly likely to be accurate and unbiased as these are official and well-respected sources of football data. The most recent game Inverness Caledonian Thistle played before beginning my research was on the 8th of February 2025 so all data I have gathered is correct as of then.

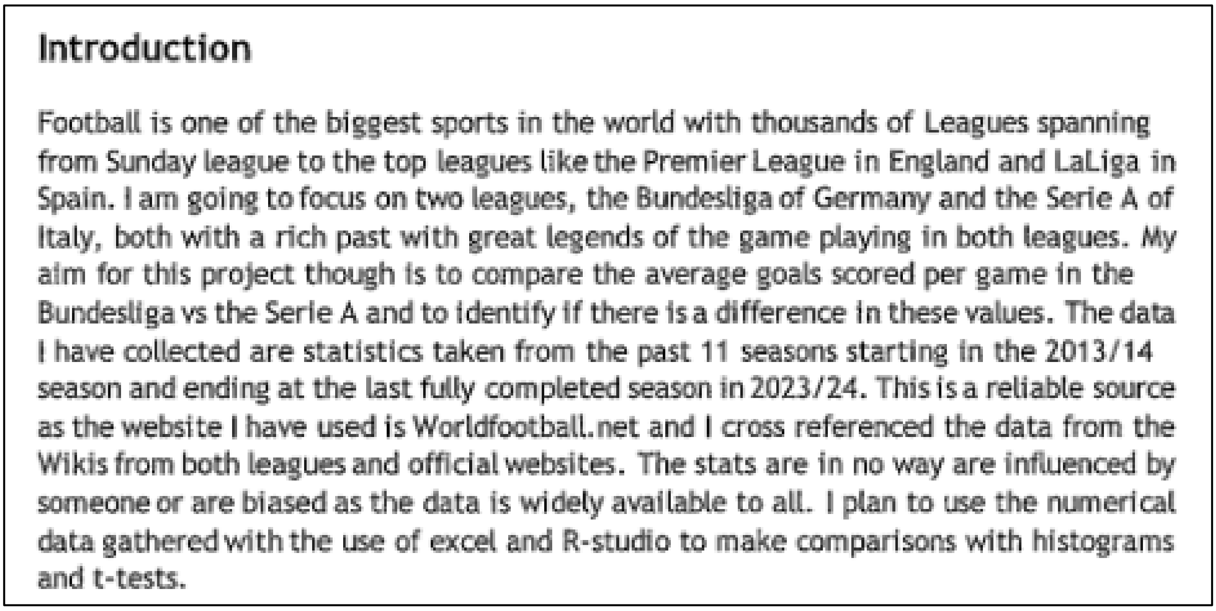
Candidate 2 evidence



Introduction

A player's performance and success in the world of professional tennis are greatly influenced by their serve speed. Being able to serve quickly can give you a tactical edge by applying tremendous pressure to your opponents. In this project I will be exploring the difference in average serve speed between 2024 Mens Wimbledon finalists Carlos Alcaraz and Novak Djokovic. Both of these competitors are known for their exceptional performance, skill and love for the sport, but their playstyles and serving techniques are different, these factors may impact the speed of their serves. My research question is, "Is there a statistically significant difference between the average serve speed between Carlos Alcaraz and Novak Djokovic?" The data I am using is discrete numerical and is reliable as I have gathered the data myself from watching the 2024 men's wimbledon final, I watched the entire match and recorded all the valid serves, there could have been human error when recording but I paused the video to make sure that I have recorded the correct data. I ensured fairness by recording all valid serves without altering any data. And to keep all the data fair and random I selected with a random number generator 100 of each player's serve speed to analyse.

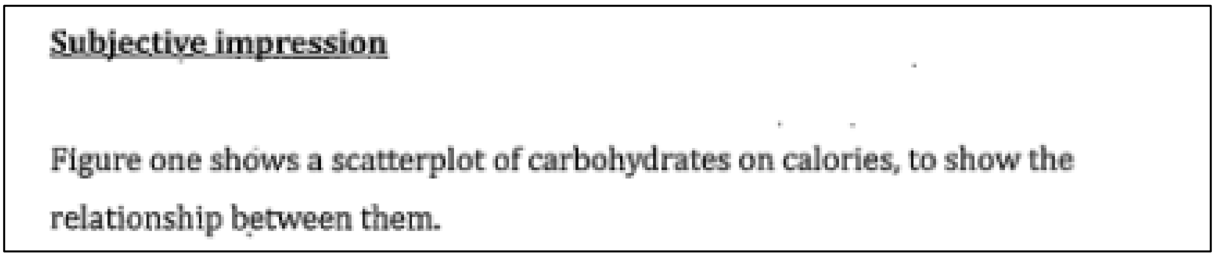
Candidate 3 evidence



Introduction

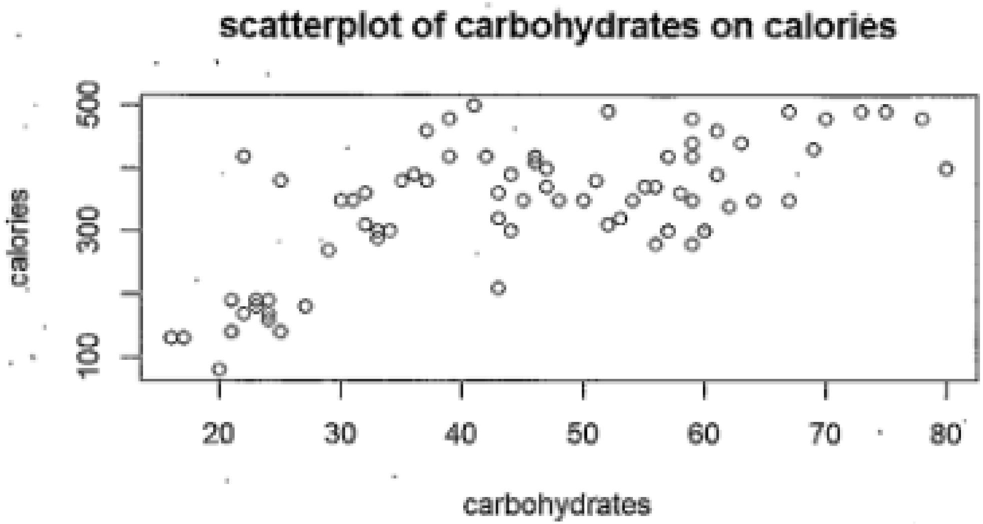
Football is one of the biggest sports in the world with thousands of Leagues spanning from Sunday league to the top leagues like the Premier League in England and LaLiga in Spain. I am going to focus on two leagues, the Bundesliga of Germany and the Serie A of Italy, both with a rich past with great legends of the game playing in both leagues. My aim for this project though is to compare the average goals scored per game in the Bundesliga vs the Serie A and to identify if there is a difference in these values. The data I have collected are statistics taken from the past 11 seasons starting in the 2013/14 season and ending at the last fully completed season in 2023/24. This is a reliable source as the website I have used is Worldfootball.net and I cross referenced the data from the Wikis from both leagues and official websites. The stats are in no way are influenced by someone or are biased as the data is widely available to all. I plan to use the numerical data gathered with the use of excel and R-studio to make comparisons with histograms and t-tests.

Candidate 4 evidence



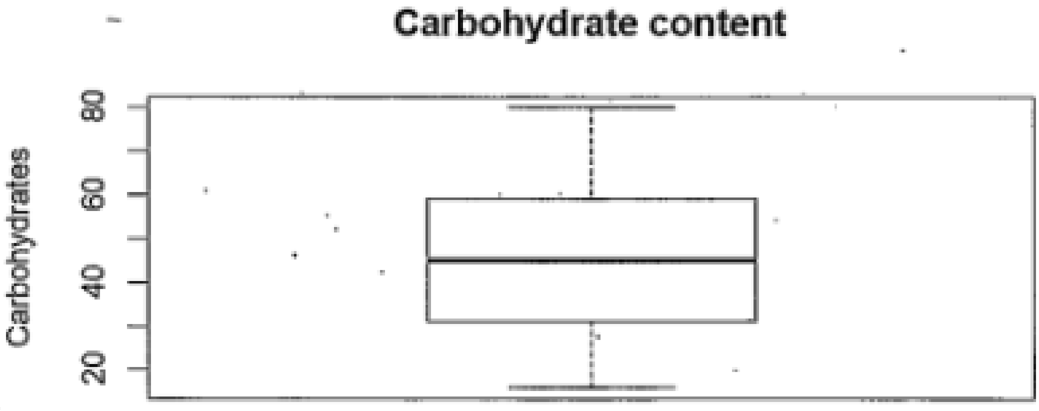
Subjective impression

Figure one shows a scatterplot of carbohydrates on calories, to show the relationship between them.

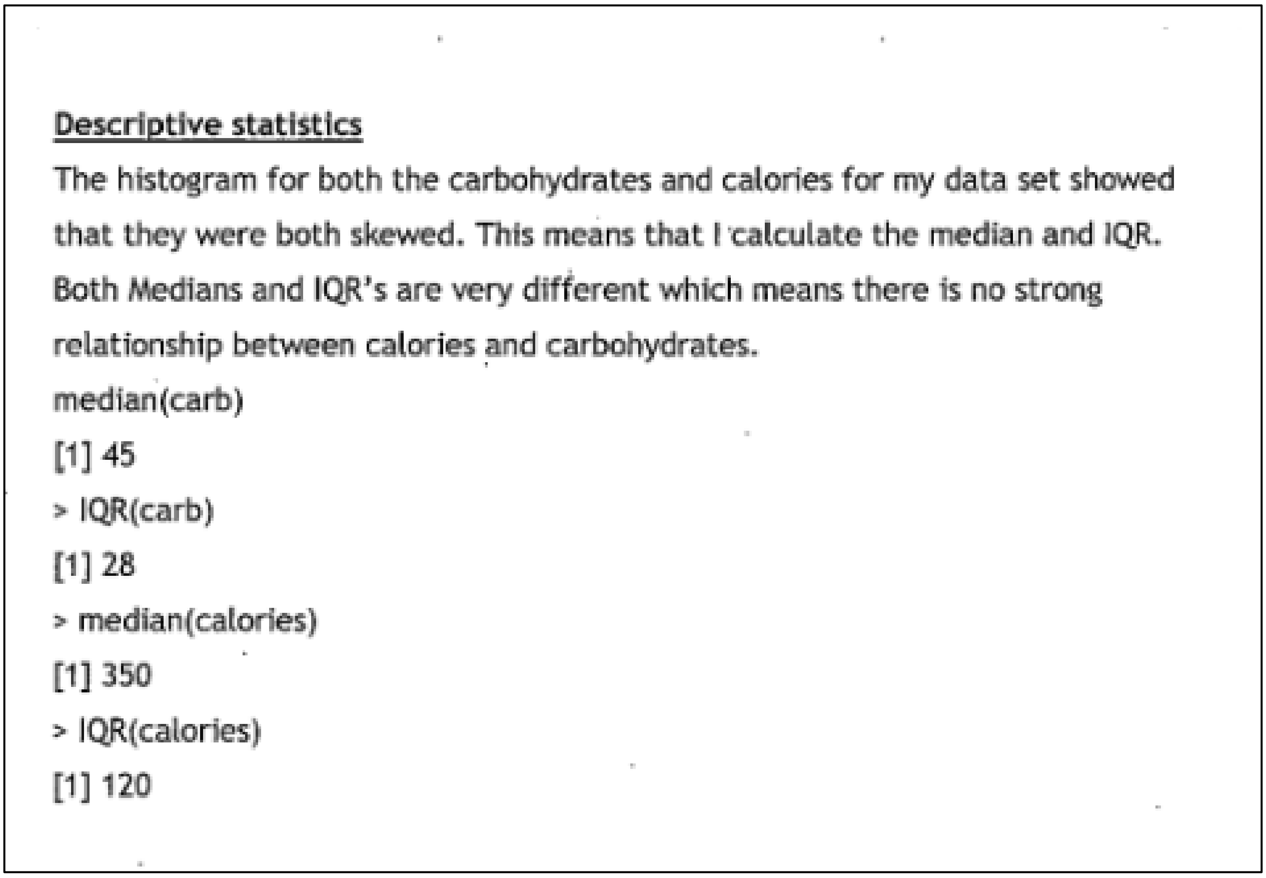


There is a positive relationship between carbohydrates and calories. The graph shows that on average the higher carbohydrate content the higher level of calories in Starbucks products.

Figure two shows a boxplot of carbohydrate content in Starbucks products.



We can see from the graph the carbohydrate content has an average of 45g.



Descriptive statistics

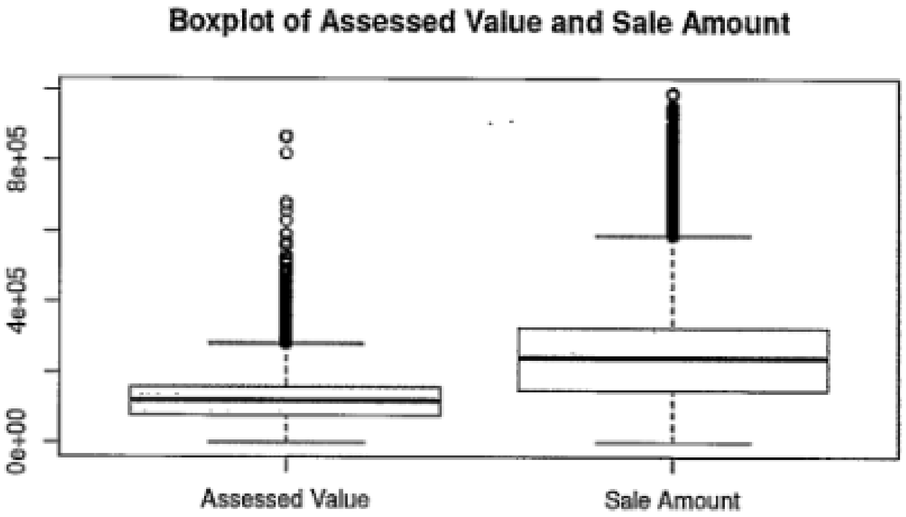
The histogram for both the carbohydrates and calories for my data set showed that they were both skewed. This means that I calculate the median and IQR. Both Medians and IQR's are very different which means there is no strong relationship between calories and carbohydrates.

```
median(carb)
[1] 45
> IQR(carb)
[1] 28
> median(calories)
[1] 350
> IQR(calories)
[1] 120
```

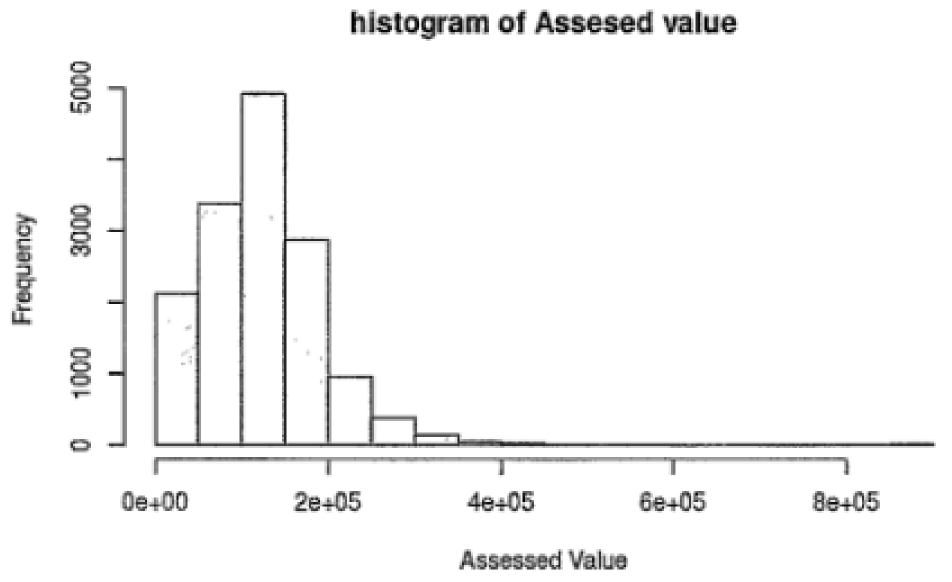
Candidate 5 evidence

Subjective impression

Boxplot



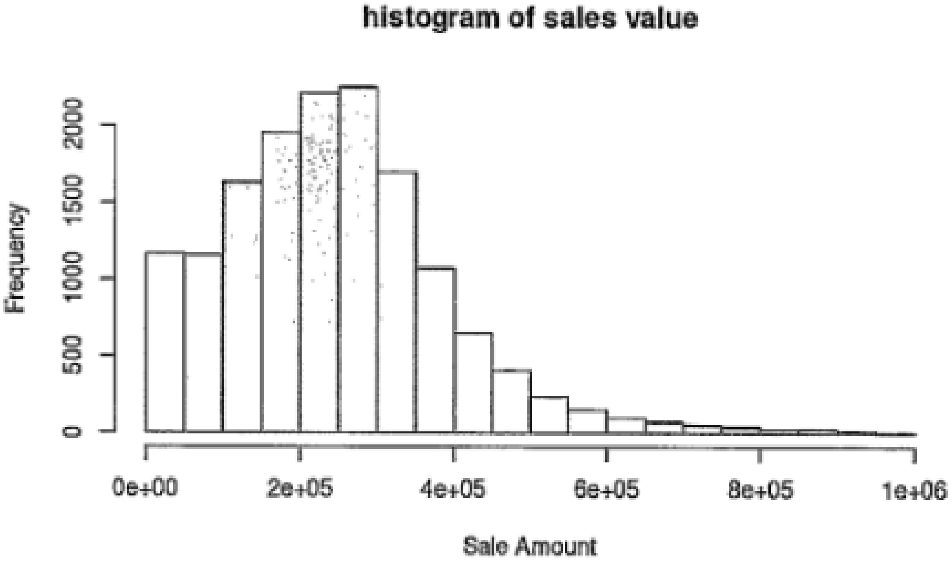
I have used a boxplot. It allows me to compare the average,it also allows me to tell the distribution of data and if it's varied.



histogram of Assesed value

I have used a histogram because it shows me if the data is normally distributed or not.

Histogram



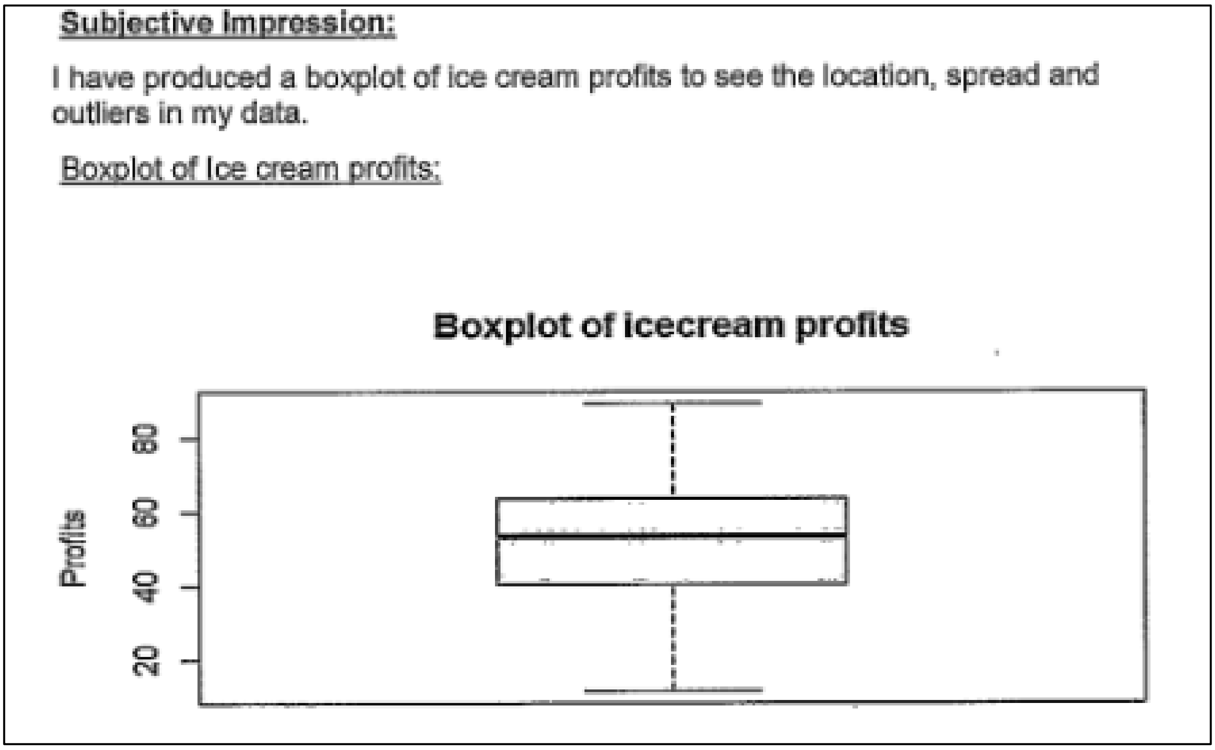
I have used a histogram because it shows me if the data is normally distributed or not.

Descriptive Statistics

Assessed value
Median = 120360
Interquartile range = 81720

Sale value
Median = 240000
Interquartile range = 176500

Candidate 6 evidence



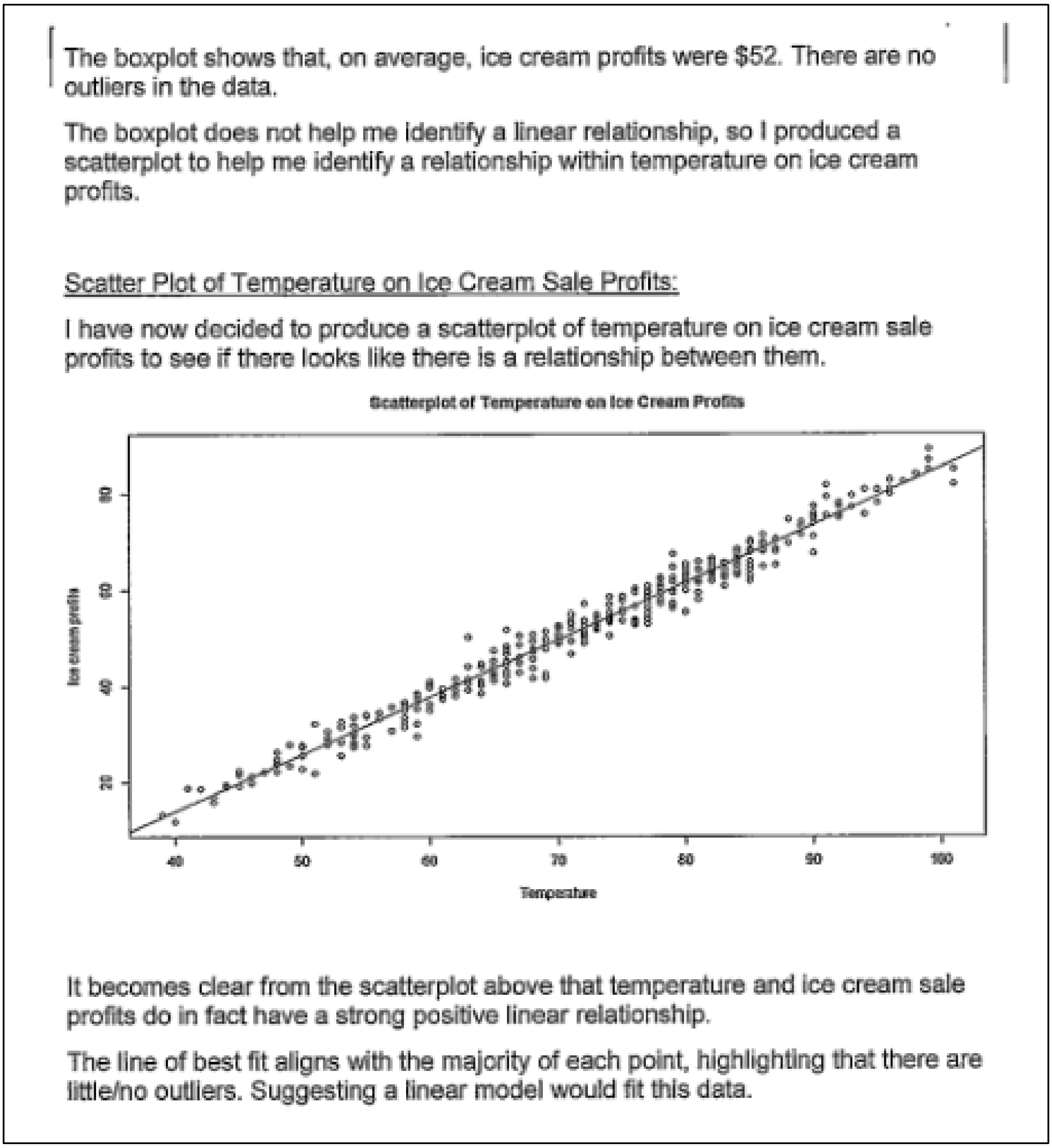
Subjective Impression:

I have produced a boxplot of ice cream profits to see the location, spread and outliers in my data.

Boxplot of Ice cream profits:

Boxplot of icecream profits

Profits
20
40
60
80



The boxplot shows that, on average, ice cream profits were $52. There are no outliers in the data.

The boxplot does not help me identify a linear relationship, so I produced a scatterplot to help me identify a relationship within temperature on ice cream profits.

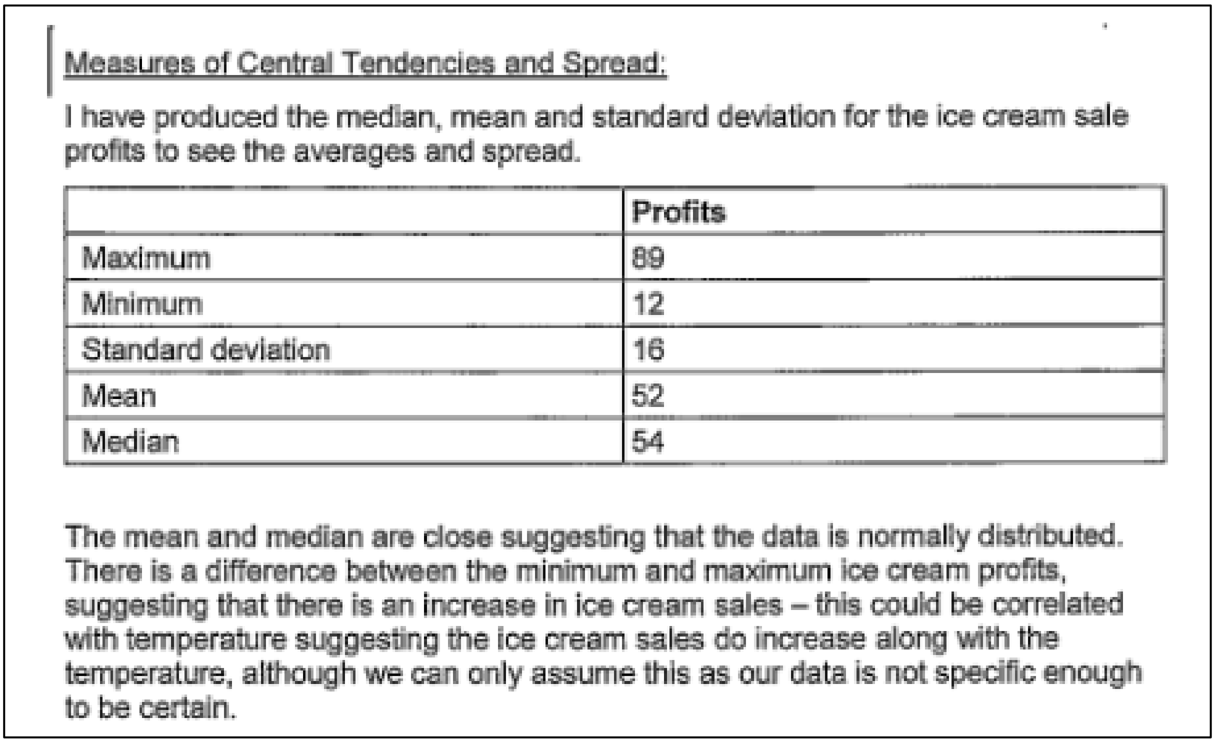
Scatter Plot of Temperature on Ice Cream Sale Profits:

I have now decided to produce a scatterplot of temperature on ice cream sale profits to see if there looks like there is a relationship between them.

Scatterplot of Temperature on Ice Cream Profits

It becomes clear from the scatterplot above that temperature and ice cream sale profits do in fact have a strong positive linear relationship.

The line of best fit aligns with the majority of each point, highlighting that there are little/no outliers. Suggesting a linear model would fit this data.



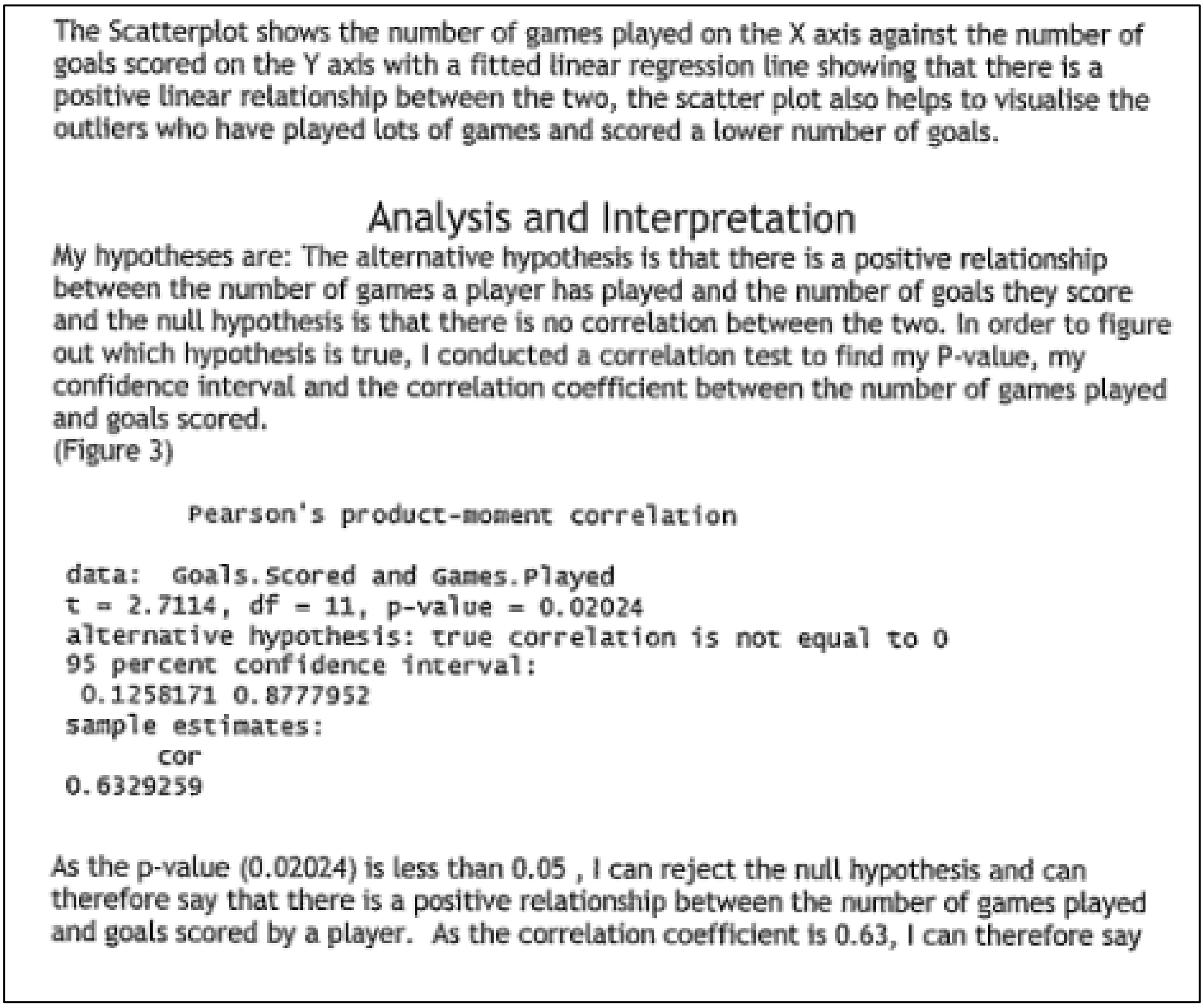
Measures of Central Tendencies and Spread:

I have produced the median, mean and standard deviation for the ice cream sale profits to see the averages and spread.

	Profits
Maximum	89
Minimum	12
Standard deviation	16
Mean	52
Median	54

The mean and median are close suggesting that the data is normally distributed. There is a difference between the minimum and maximum ice cream profits, suggesting that there is an increase in ice cream sales – this could be correlated with temperature suggesting the ice cream sales do increase along with the temperature, although we can only assume this as our data is not specific enough to be certain.

Candidate 7 evidence



The Scatterplot shows the number of games played on the X axis against the number of goals scored on the Y axis with a fitted linear regression line showing that there is a positive linear relationship between the two, the scatter plot also helps to visualise the outliers who have played lots of games and scored a lower number of goals.

Analysis and Interpretation

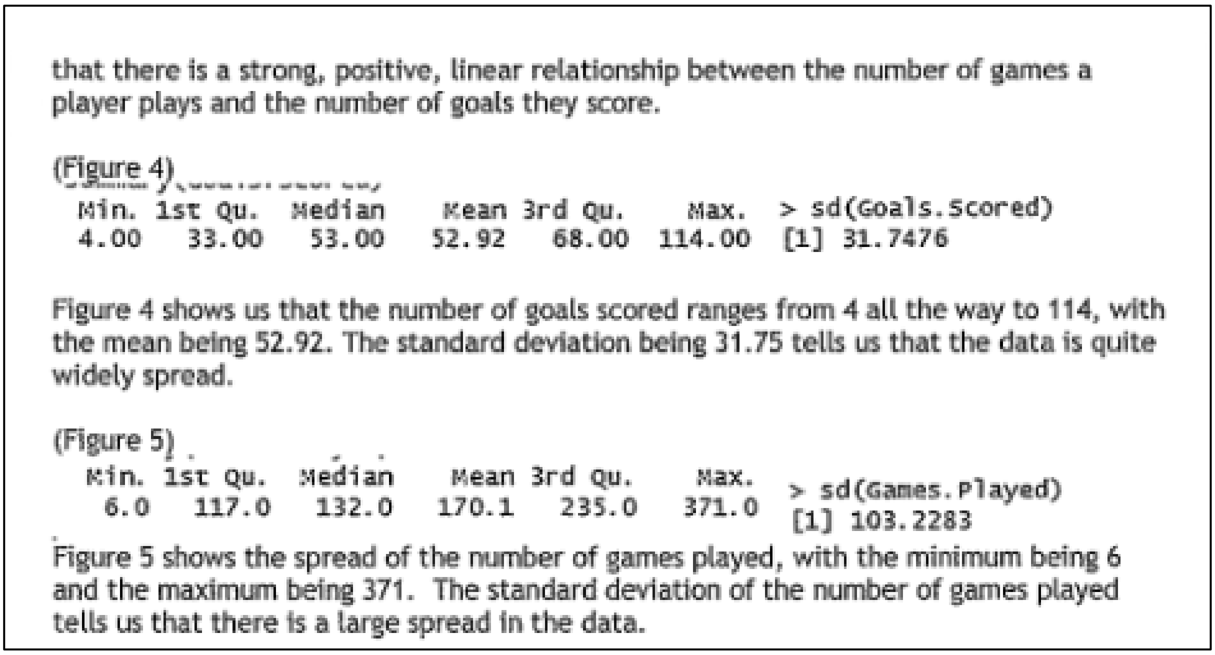
My hypotheses are: The alternative hypothesis is that there is a positive relationship between the number of games a player has played and the number of goals they score and the null hypothesis is that there is no correlation between the two. In order to figure out which hypothesis is true, I conducted a correlation test to find my P-value, my confidence interval and the correlation coefficient between the number of games played and goals scored.

(Figure 3)

```
        Pearson's product-moment correlation

data:  Goals.Scored and Games.Played
t = 2.7114, df = 11, p-value = 0.02024
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1258171 0.8777952
sample estimates:
      cor
0.6329259
```

As the p-value (0.02024) is less than 0.05 , I can reject the null hypothesis and can therefore say that there is a positive relationship between the number of games played and goals scored by a player. As the correlation coefficient is 0.63, I can therefore say that there is a strong, positive, linear relationship between the number of games a player plays and the number of goals they score.



(Figure 4)

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  > sd(Goals.Scored)
 4.00   33.00   53.00   52.92   68.00  114.00  [1] 31.7476
```

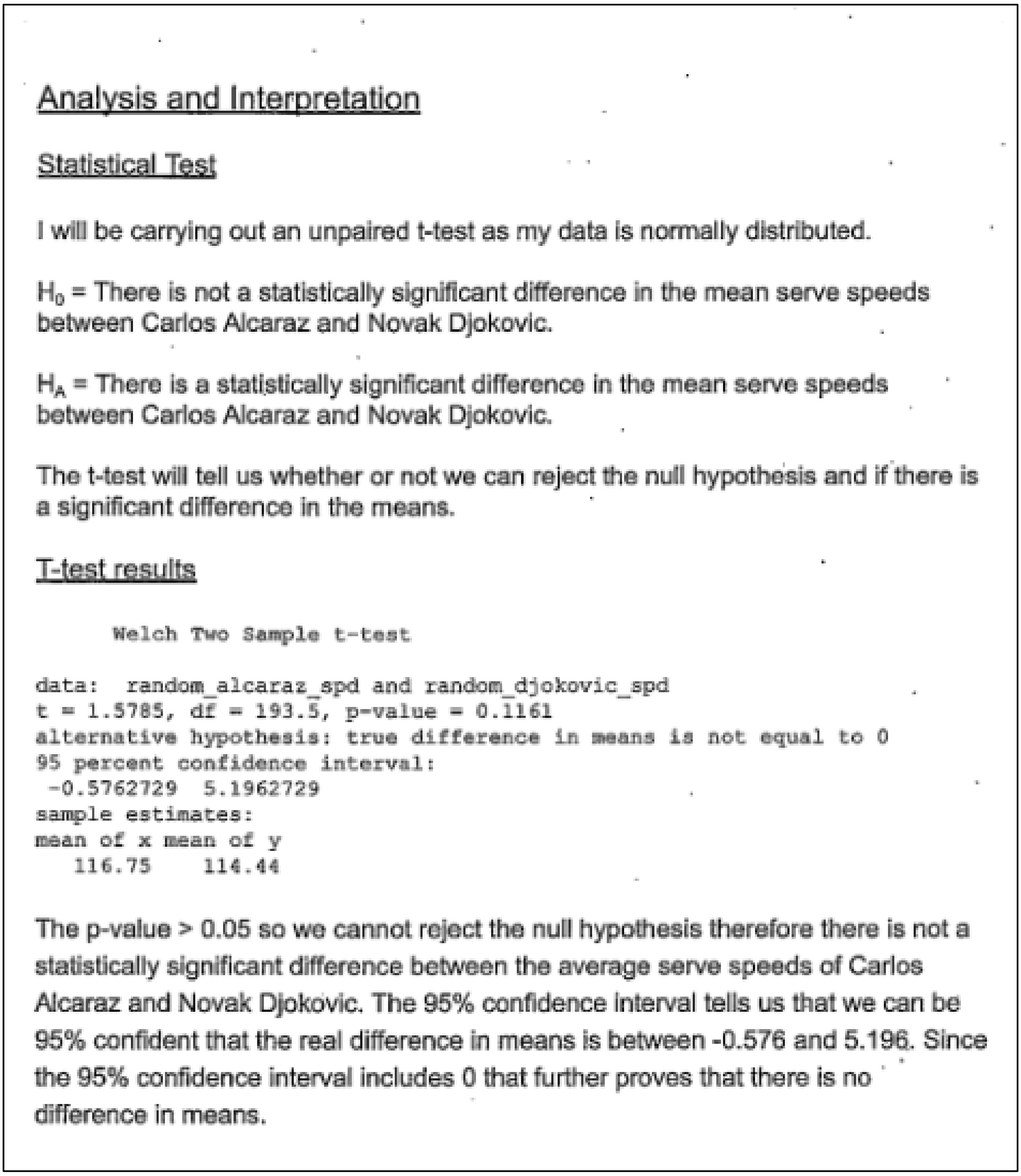
Figure 4 shows us that the number of goals scored ranges from 4 all the way to 114, with the mean being 52.92. The standard deviation being 31.75 tells us that the data is quite widely spread.

(Figure 5)

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.0   117.0   132.0   170.1   235.0   371.0  > sd(Games.Played)
                                               [1] 103.2283
```

Figure 5 shows the spread of the number of games played, with the minimum being 6 and the maximum being 371. The standard deviation of the number of games played tells us that there is a large spread in the data.

Candidate 8 evidence



Analysis and Interpretation

Statistical Test

I will be carrying out an unpaired t-test as my data is normally distributed.

H_0 = There is not a statistically significant difference in the mean serve speeds between Carlos Alcaraz and Novak Djokovic.

H_A = There is a statistically significant difference in the mean serve speeds between Carlos Alcaraz and Novak Djokovic.

The t-test will tell us whether or not we can reject the null hypothesis and if there is a significant difference in the means.

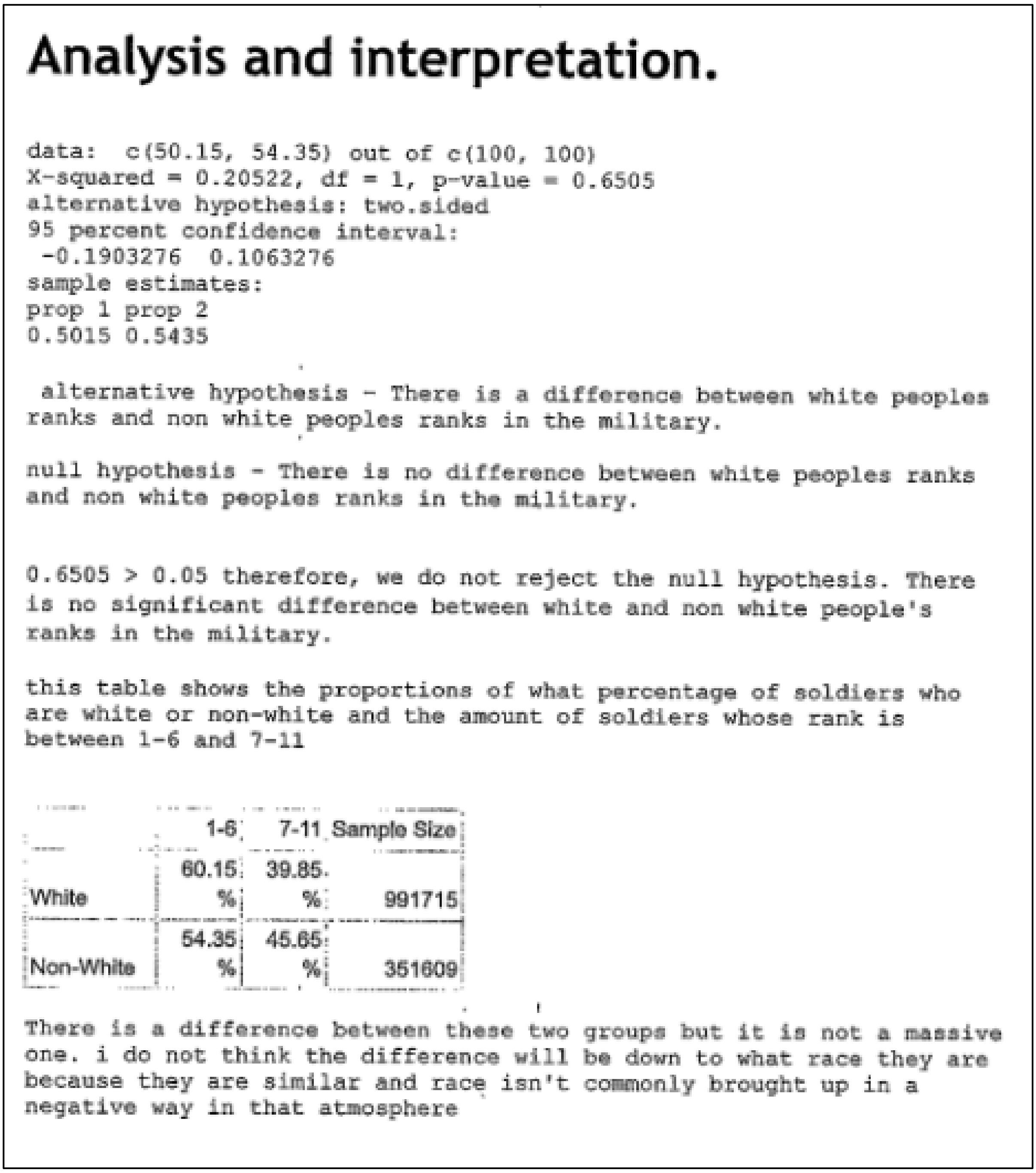
T-test results

```
      Welch Two Sample t-test

data:  random_alcaraz_spd and random_djokovic_spd
t = 1.5785, df = 193.5, p-value = 0.1161
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5762729  5.1962729
sample estimates:
mean of x mean of y
   116.75    114.44
```

The p-value > 0.05 so we cannot reject the null hypothesis therefore there is not a statistically significant difference between the average serve speeds of Carlos Alcaraz and Novak Djokovic. The 95% confidence interval tells us that we can be 95% confident that the real difference in means is between -0.576 and 5.196. Since the 95% confidence interval includes 0 that further proves that there is no difference in means.

Candidate 9 evidence



Analysis and interpretation.

```
data:  c(50.15, 54.35) out of c(100, 100)
X-squared = 0.20522, df = 1, p-value = 0.6505
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1903276  0.1063276
sample estimates:
prop 1 prop 2
0.5015 0.5435
```

alternative hypothesis - There is a difference between white peoples ranks and non white peoples ranks in the military.

null hypothesis - There is no difference between white peoples ranks and non white peoples ranks in the military.

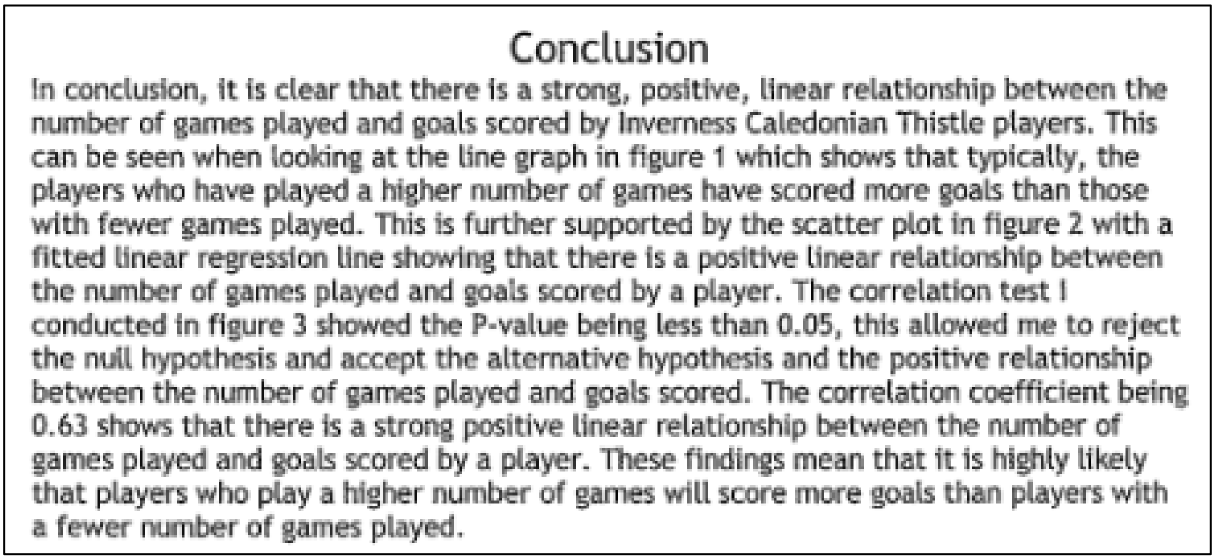
0.6505 > 0.05 therefore, we do not reject the null hypothesis. There is no significant difference between white and non white people's ranks in the military.

this table shows the proportions of what percentage of soldiers who are white or non-white and the amount of soldiers whose rank is between 1-6 and 7-11

	1-6	7-11	Sample Size
White	60.15 %	39.85 %	991715
Non-White	54.35 %	45.65 %	351609

There is a difference between these two groups but it is not a massive one. i do not think the difference will be down to what race they are because they are similar and race isn't commonly brought up in a negative way in that atmosphere

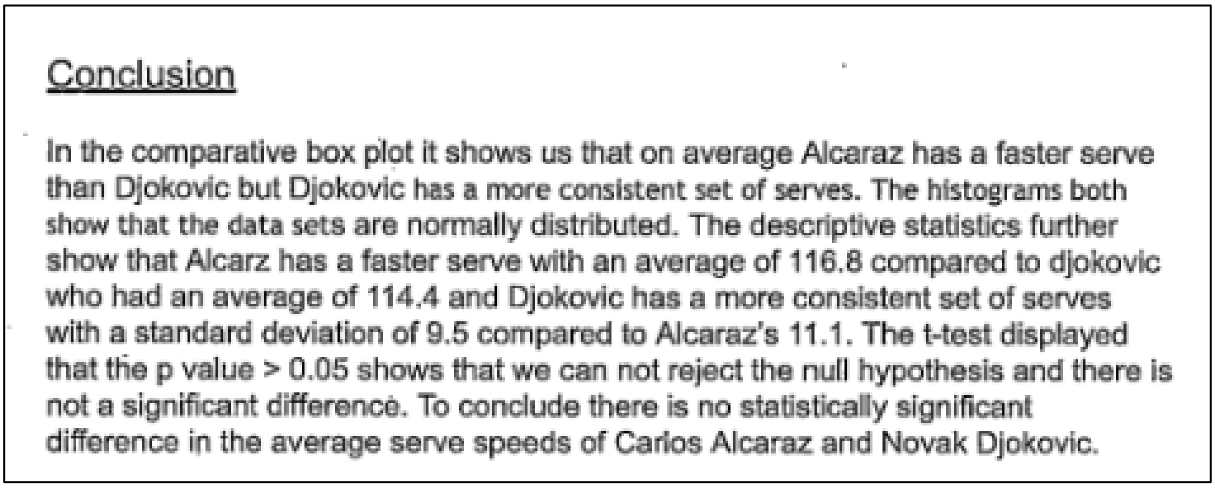
Candidate 10 evidence



Conclusion

In conclusion, it is clear that there is a strong, positive, linear relationship between the number of games played and goals scored by Inverness Caledonian Thistle players. This can be seen when looking at the line graph in figure 1 which shows that typically, the players who have played a higher number of games have scored more goals than those with fewer games played. This is further supported by the scatter plot in figure 2 with a fitted linear regression line showing that there is a positive linear relationship between the number of games played and goals scored by a player. The correlation test I conducted in figure 3 showed the P-value being less than 0.05, this allowed me to reject the null hypothesis and accept the alternative hypothesis and the positive relationship between the number of games played and goals scored. The correlation coefficient being 0.63 shows that there is a strong positive linear relationship between the number of games played and goals scored by a player. These findings mean that it is highly likely that players who play a higher number of games will score more goals than players with a fewer number of games played.

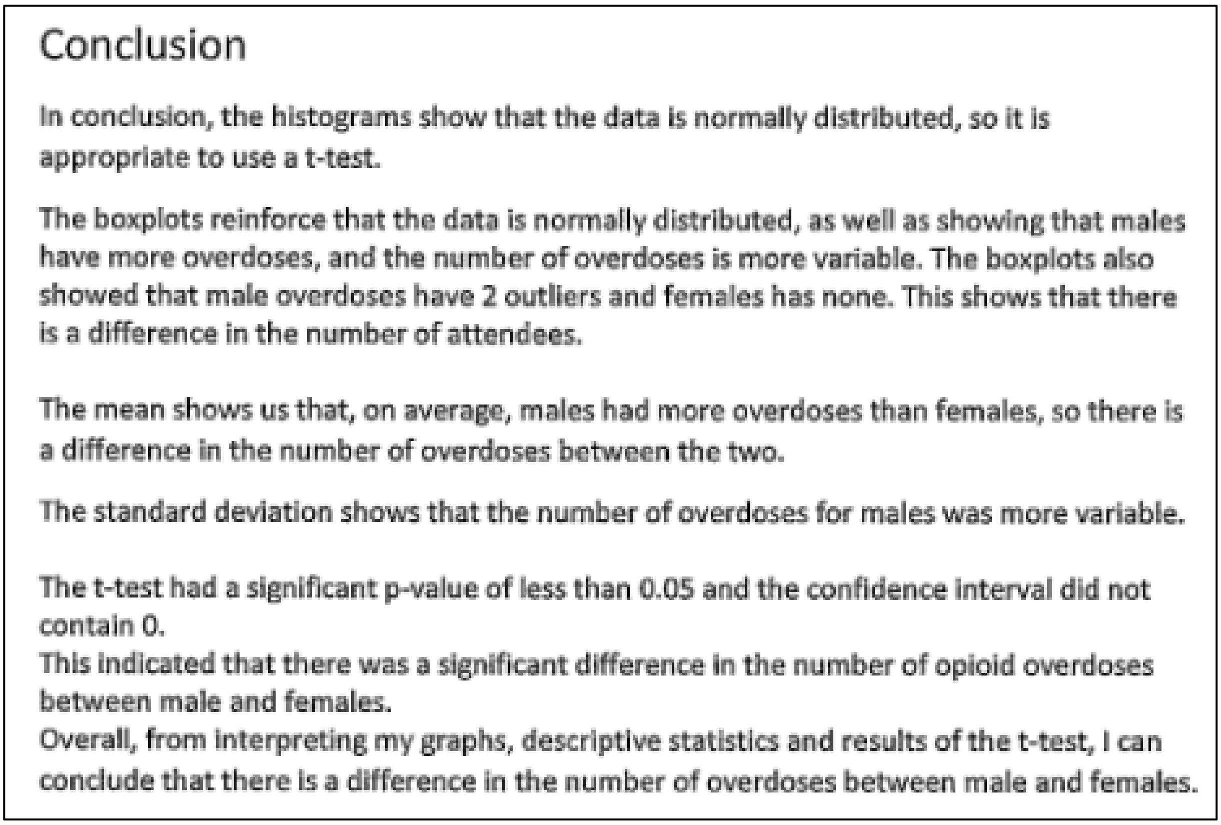
Candidate 11 evidence



Conclusion

In the comparative box plot it shows us that on average Alcaraz has a faster serve than Djokovic but Djokovic has a more consistent set of serves. The histograms both show that the data sets are normally distributed. The descriptive statistics further show that Alcarz has a faster serve with an average of 116.8 compared to djokovic who had an average of 114.4 and Djokovic has a more consistent set of serves with a standard deviation of 9.5 compared to Alcaraz's 11.1. The t-test displayed that the p value > 0.05 shows that we can not reject the null hypothesis and there is not a significant difference. To conclude there is no statistically significant difference in the average serve speeds of Carlos Alcaraz and Novak Djokovic.

Candidate 12 evidence



Conclusion

In conclusion, the histograms show that the data is normally distributed, so it is appropriate to use a t-test.

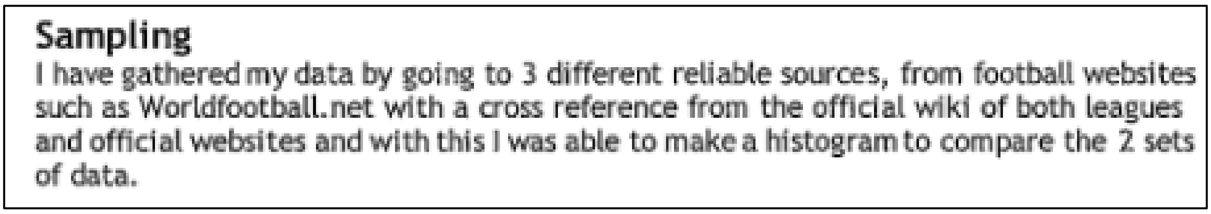
The boxplots reinforce that the data is normally distributed, as well as showing that males have more overdoses, and the number of overdoses is more variable. The boxplots also showed that male overdoses have 2 outliers and females has none. This shows that there is a difference in the number of attendees.

The mean shows us that, on average, males had more overdoses than females, so there is a difference in the number of overdoses between the two.

The standard deviation shows that the number of overdoses for males was more variable.

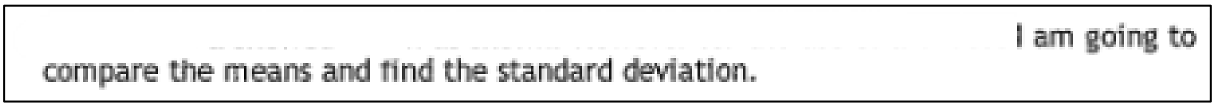
The t-test had a significant p-value of less than 0.05 and the confidence interval did not contain 0.
This indicated that there was a significant difference in the number of opioid overdoses between male and females.
Overall, from interpreting my graphs, descriptive statistics and results of the t-test, I can conclude that there is a difference in the number of overdoses between male and females.

Candidate 13 evidence



Sampling

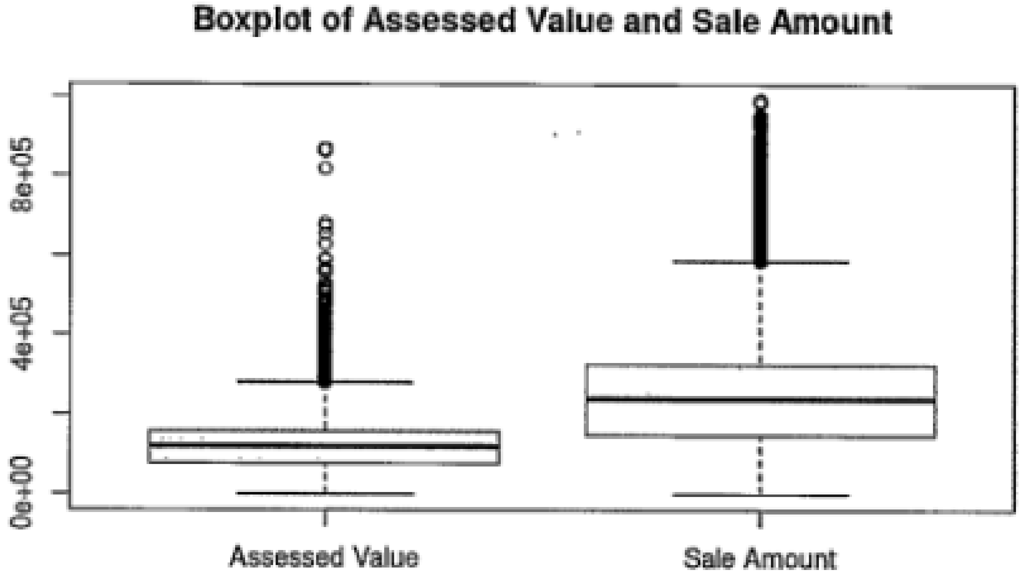
I have gathered my data by going to 3 different reliable sources, from football websites such as Worldfootball.net with a cross reference from the official wiki of both leagues and official websites and with this I was able to make a histogram to compare the 2 sets of data.



I am going to compare the means and find the standard deviation.

Candidate 14 evidence

Boxplot



I have used a boxplot. It allows me to compare the average,it also allows me to tell the distribution of data and if it's varied.

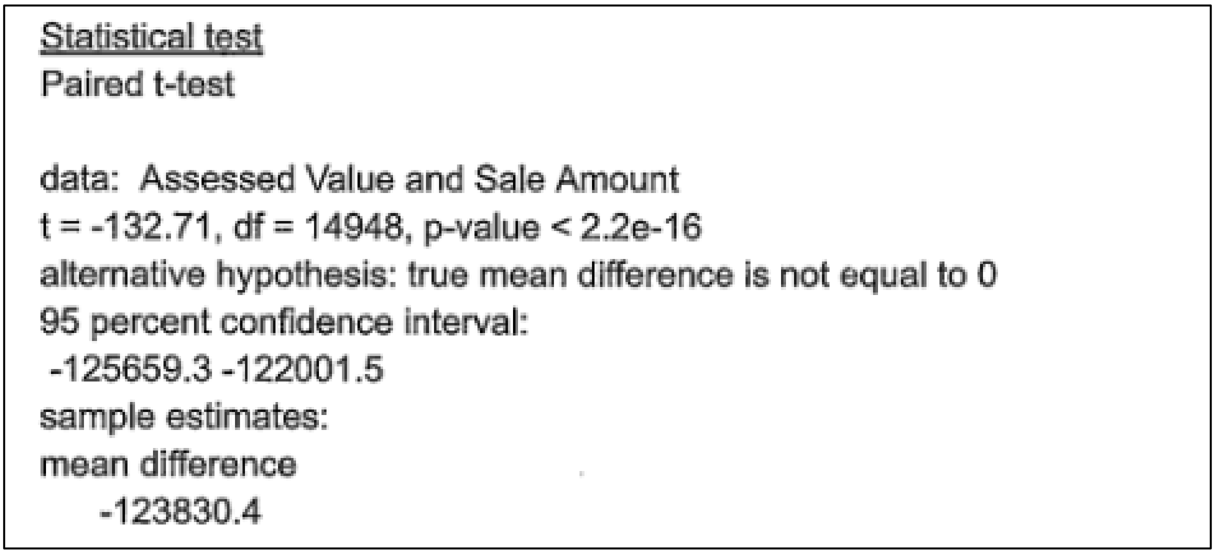
Descriptive Statistics

Assessed value

Median = 120360
Interquartile range = 81720

Sale value

Median = 240000
Interquartile range = 176500



Statistical test

Paired t-test

```
data:  Assessed Value and Sale Amount
t = -132.71, df = 14948, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -125659.3 -122001.5
sample estimates:
mean difference
      -123830.4
```