

Candidate evidence

Introduction

Candidate 1

Introduction

In this project I will be exploring the correlation between the quality of education and the number of new books published per 1 million people. My research question is, "Is there a correlation between the quality of education received in a country and the number of new books published? I got this data from Our World in data, they are a charity organisation based in the UK, this makes the data useful because it comes from a non-profit organisation, meaning that the data is less likely to be biased. The most recent numerical data on how many books were published per 1 million people came from 2009, this made both the data from how many books were published and the mean of the pisa test score performance less useful as the data is 14 years old. Due to the low amount of countries who were involved in the most recent pisa test, the data used is less likely to be omitting any key facts, making it less biased.

Introduction

Candidate 2

Introduction

Scotland's educational qualifications for maths for students aged 14-16 was named Standard Grade, this was introduced in 1986 and phased out in August of 2013 and was replaced by the Scottish Qualifications Authority (SQA). Standard grade maths covered topics such as geometry, trigonometry, algebra, percentages and statistics. In 2013, National 5 Maths replaced Standard Grade however similar topics were still covered. In this project I will investigate 'Since the year 2000, has there been a difference in attainment of Level 5 SCQF Maths between males and females until the year 2022?' The data I have gathered is numerical and is reliable as the information has been sourced directly from the official website for the Scottish Qualifications Authority. The data is fair and unbiased as it is the official results attained by students after the completion of their exam or appeal and has been marked by professional SQA markers by following an agreed code of practice. I will produce scatter graphs and bar charts to examine the attainment for both genders across the different years. I will compare and analyse my data to determine if there are any difference between males and females. I predict that males will have a higher-level attainment than females as during my years at school, statistically on average males achieved better grades in their maths tests compared to females within my school.

Introduction

Candidate 3

Introduction

Women's charities are reporting a high level of domestic abuse rates, so this makes this an important social problem. This report will analyse the domestic abuse rates throughout the ages of women.

My research question is to determine if there is a relationship between the number of females suffering domestic abuse and age.

The data I am studying is numerical.

Introduction and Conclusion

Candidate 4

Introduction

Price vs weight is a valid and accurate comparison as when a vehicles weight increases it is likely that the price will also increase this project will seek to find out whether this is true or false and find the evidence for either conclusion.

For this project I have used a scatter graph as it is the best and clearest way to display the data within a graph and shows the clear increase of price when weight is increased I have included a scatter graph to show the increase of price when weight is increased.

Subjective impression

Candidate 5

Subjective impression

Initially a bar chart on snowfall each year was constructed (figure1.)

Figure 1

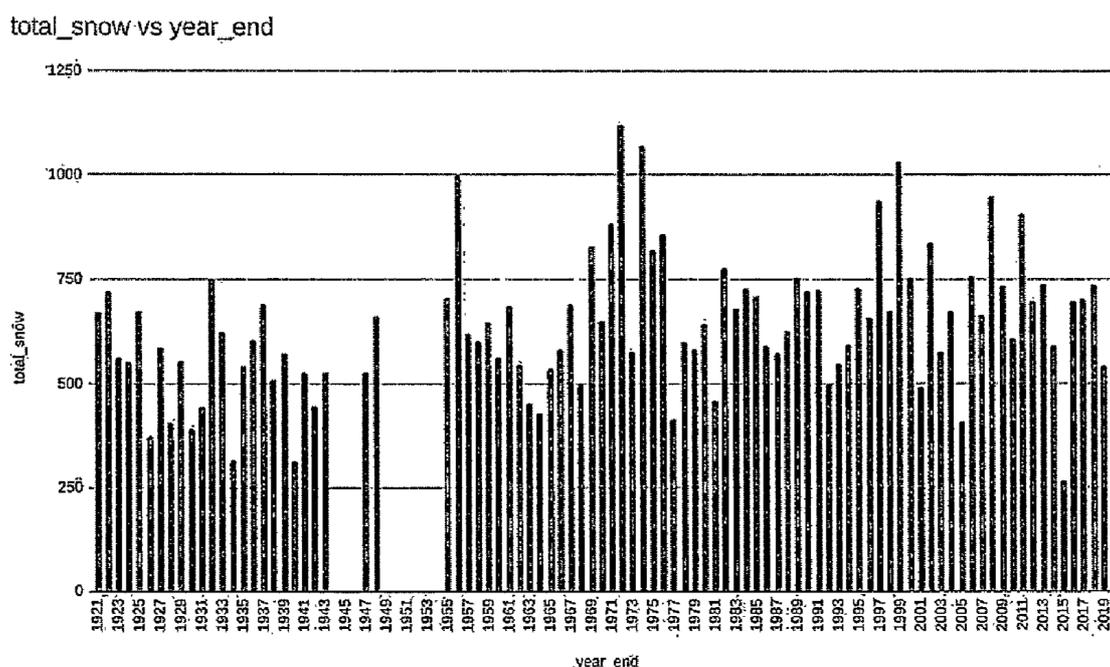
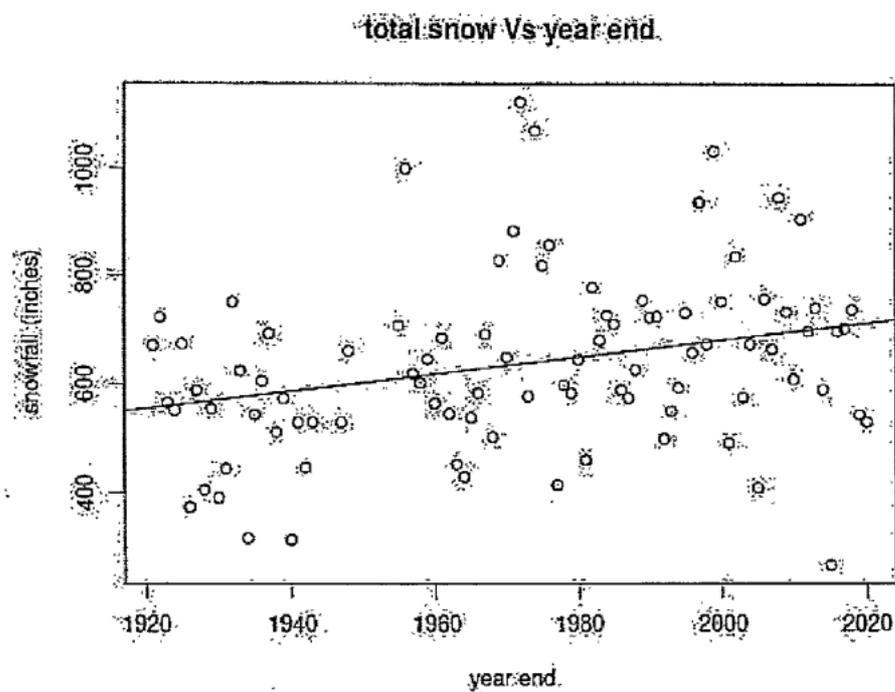


Figure 1 shows a variable picture over 100 years of data and indicates a linear relationship may not exist in this data. Also from 1939-1945 the years weren't recorded because of World War II. From looking at this data set there is no increase in snow over the 100 year time period. A scatter graph was also constructed with a linear regression shown in (figure 2.)

Figure 2

Figure 2 shows a slight increase trend over 100 years. There are outliers in this data as 1975-1976 is about 1122 inches. Cor: 0.2866893 this shows us that this data has a weak linear relationship.



The mean and standard deviation was calculated for the average snowfall at Mount Rainier for over 100 years. (figure3)

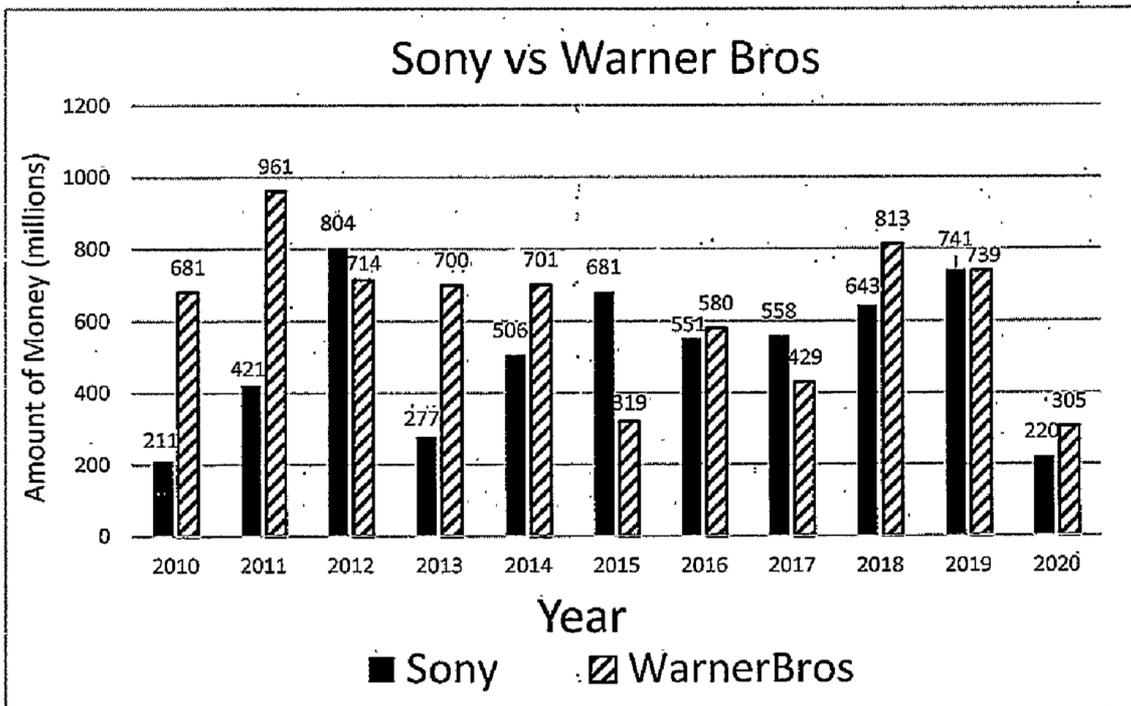
Table 1

Mean (snowfall)	638.1044
Standard deviation (snowfall)	163.807

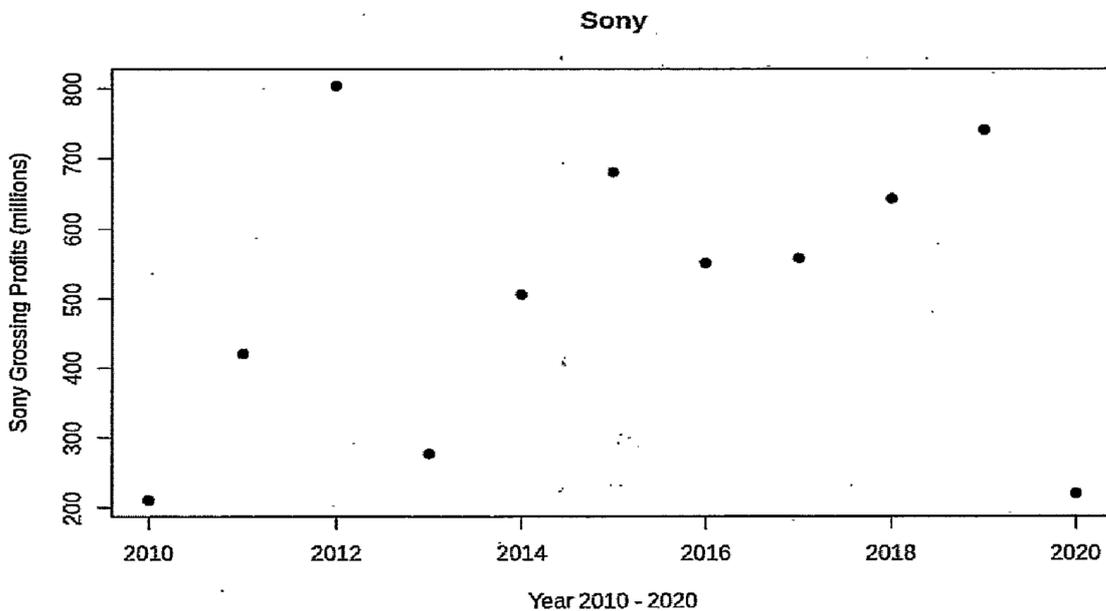
Table 1 tells us that on average over a 91-year period there were 638 inches of snowfall on Mount Rainier, with an expectancy that the snowfall will be between 635 inches to 645 inches. The high standard deviation shows that it varies each year.

Subjective impression

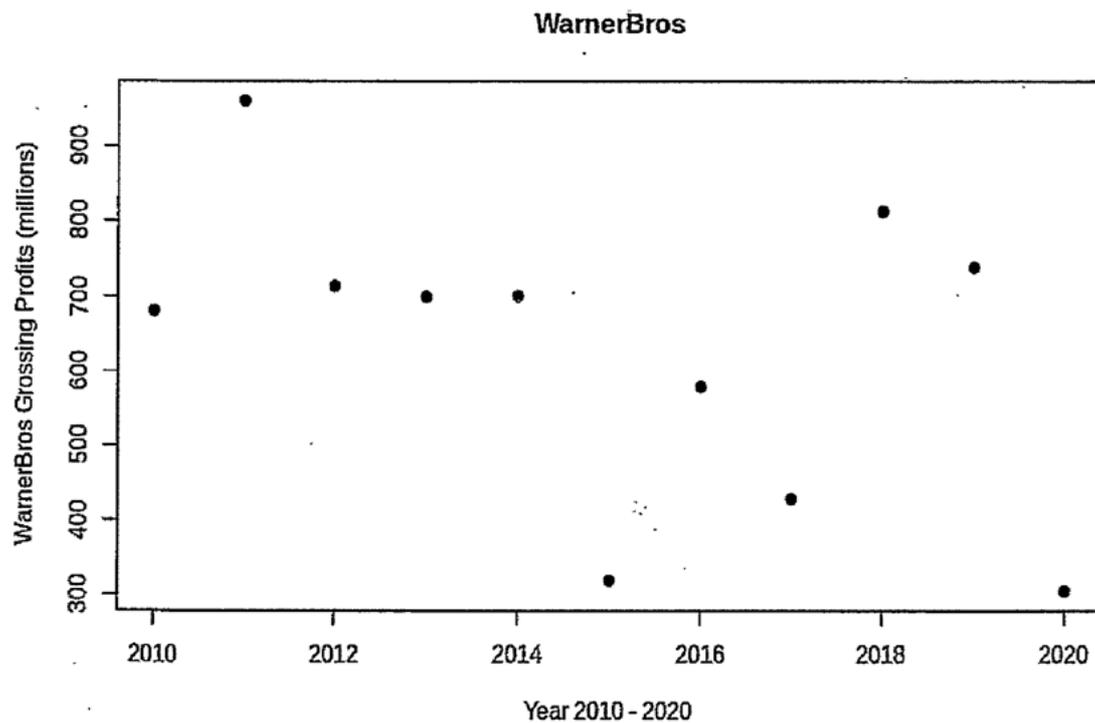
Candidate 6



The bargraph shows the amount of money Sony and Warner Bros made from 2010 to 2020. You can see from the graph that although it is not consistent, WarnerBros has made more money in the period of 10 years, 7 times.



The scatterplot shows the company Sony's Grossing Profits against the years 2010 - 2020. You can see that the data is not consistent due to the different years that the company made the best profits. The mean number of profit that Sony has made is \$520.27 million between 2010 and 2020. The standard deviation is 206.50.



This scatterplot shows Warner Bros grossing profit from each year ranging from 2010 - 2020. As you can see the data is not consistent here because in different years the company has made more or less profits. The mean number of profit that Warner Bros has made is \$631.1 million between 2010 and 2020. The standard deviation is 204.88.

Subjective impression

Candidate 7

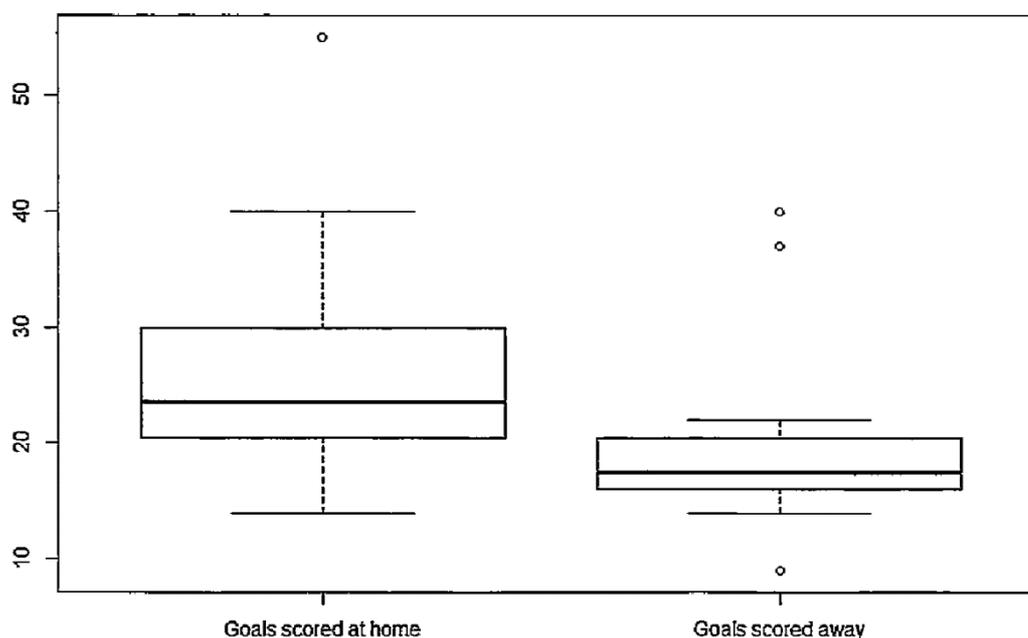
Graphical displays

I will insert and analyse a few graphical displays including boxplots and histograms. By using these graphical displays, I will be able to make a visual comparison between them and it is an easy way to look at datasets. I will look at the mean and standard deviation and I will then conduct a t-test for my datasets and will be able to check if there is a significant difference in the amount of goals scored at home and the amount of goals scored away by looking at the p-value and 95% confidence interval.

Boxplot

A boxplot is a graphical display that shows us the distribution of data and can allow us to make comparisons between datasets. They give 5 key bits of data, the minimum value, the lower quartile, the median, the upper quartile and the maximum value. They are very effective as they allow us to make an easy comparison by giving us a visual summary. An outlier is when a value is outside the overall distribution pattern and is significantly different to the other values which can affect the results.

Boxplot for goals scored at home and goals scored away in the SPFL 2021-2022 season

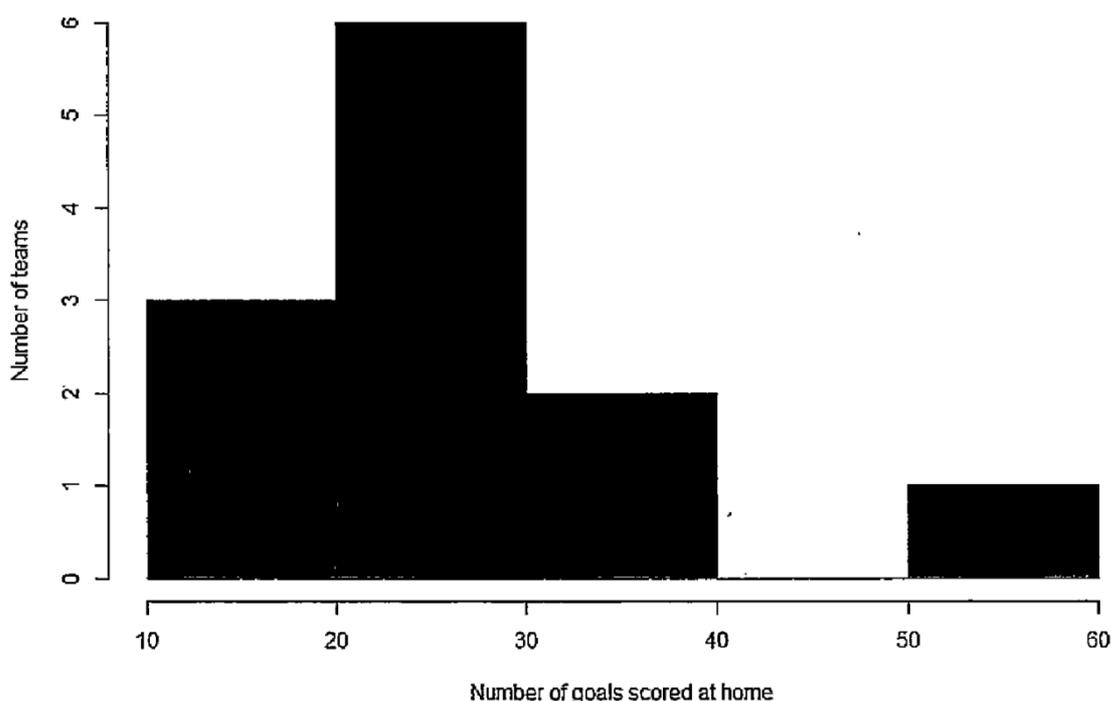


We can clearly see from the boxplot just by looking at it, that there are more goals scored at home as the median lies between 20 and 30 whereas the median for the away goals scored lies between 10 and 20 and the maximum value for goals at home is 40 compared to goals away which is only just above 20, while the minimum values are roughly the same and there isn't much difference in them. The goals scored at home is a more spread dataset with the minimum value around 15 and the maximum value is 40, the goals scored away dataset is less spread as the minimum value is 15 and the maximum value is 20. The boxplot identified values in both datasets which are far from the distribution pattern and therefore they are outliers which affect the results.

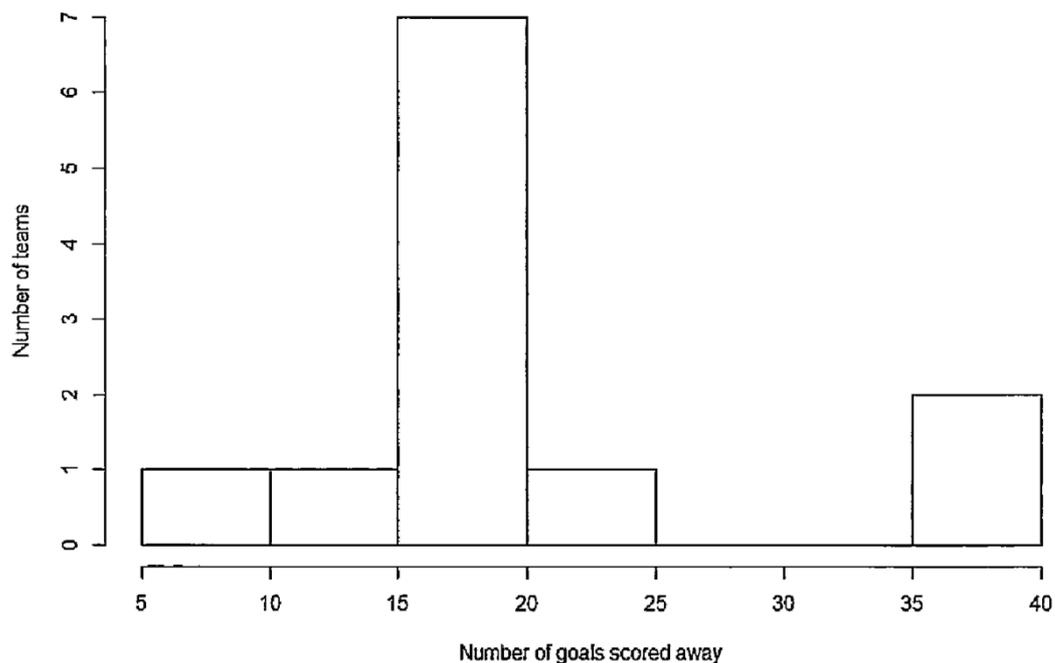
Histogram

A histogram is a graphical display that shows the frequency distribution of a dataset and allows us to make comparisons between the datasets. It shows how often a value occurs in a dataset and is another very easy graphical display to just look at and make visual comparisons. I have changed the colour of the histograms to make it easier to compare and see the difference between the 2 datasets.

Histogram for Goals scored at home in the SPFL 2021-2022 season



Histogram for Goals scored away in the SPFL 2021-2022 season



The histogram for goals scored at home is a positive skewed histogram as the peak of the data is on the left hand side and it gradually gets lower as it moves to the right and tails down. We also see there is one team which scored in between 50 and 60 goals at home. The data on the goals scored at home is more spread than the goals scored away meaning there is more of a difference in the amount of goals scored at home whereas the goals scored away are largely grouped in between the values of 15-20 goals which is less spread but it is not clear to see a significant difference from looking at the histograms.

Analysis and interpretation

Candidate 8

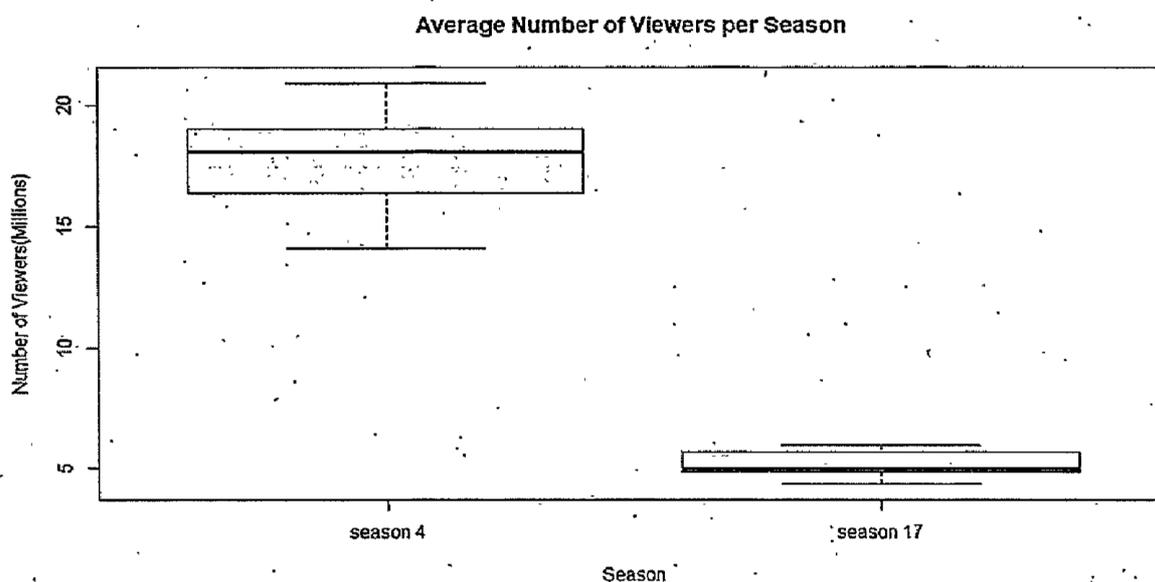
Graphical Displays

To form an initial impression of the data I will produce graphical displays such as boxplots and histograms.

Comparative boxplot

(Figure 1)

I have constructed a comparative boxplot for the average number of viewers for each season. This demonstrates the 5 key data points, the minimum, the lower quartile, the median, the upper quartile, the maximum value and the spread of this data. This will allow me to analyse the average values and the spread of the data.

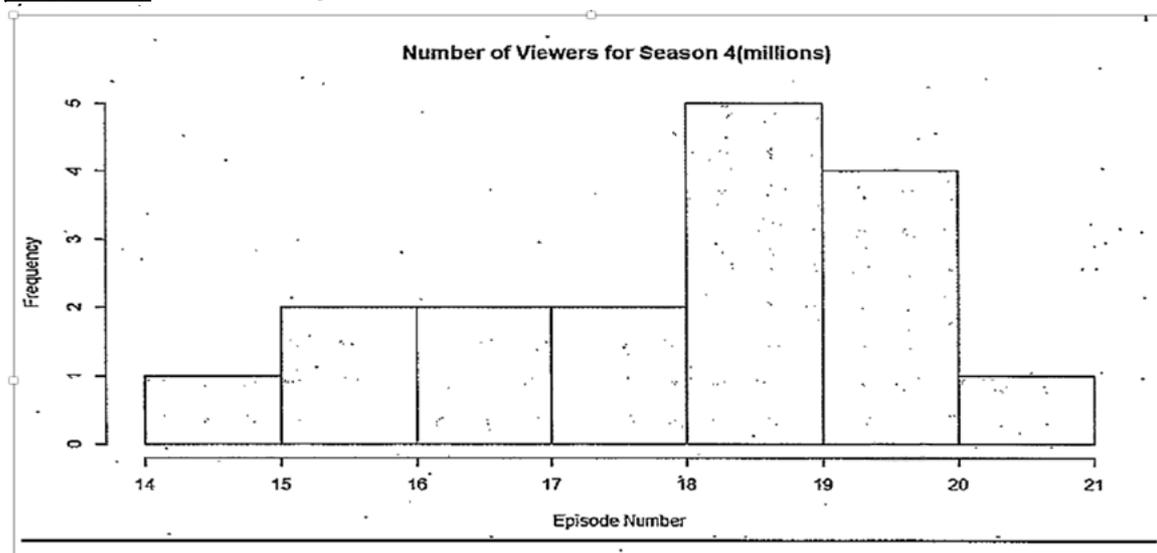


The comparative boxplot allows me to see that on average, the number of viewers for season 4 was significantly higher than the number of viewers for season 17. However the number of viewers for season 17 was more consistent than the number of viewers for season 4. We know this because the spread of the boxplot for season 4 is wider than the spread of the boxplot for season 17. To allow me to carry out further analysis, I will now test my data for normality using histograms.

Histograms

The histograms below allow me to see if the data for seasons 4 and 17 follow a normal distribution.

(Figure 2)



(Figure 3)

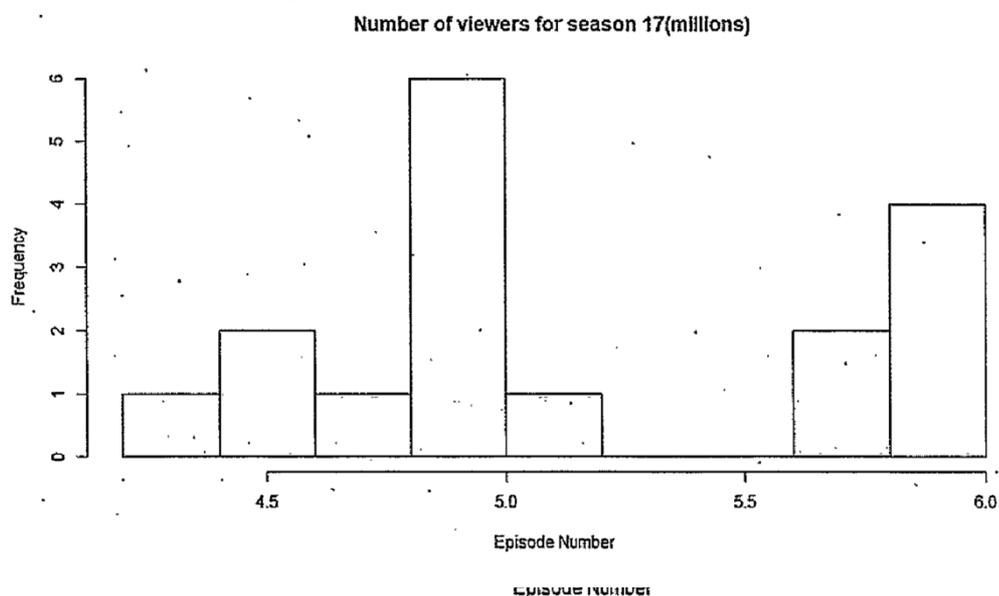


Figure 3 also allows us to see that the data produced a bell shaped curve. Although the curve is not as distinct as in figure 2, we can still assume that the data follows a normal distribution. Now that I have assumed normality, I am able to continue on with different statistical tests.

Descriptive Statistics

Statistics Table

The table below will allow me to analyse different statistics involving the data.

(Figure 4)

Statistics	Season 4	Season 17
Minimum	14.11	4.330
1 st Quartile	16.37	4.810
Median	18.09	4.980
Mean	17.77	5.168
3 rd Quartile	19.04	5.690
Maximum	20.93	5.960
SD	1.78019	0.5454188

Figure 4 further reiterates what the boxplot demonstrated and shows that season 4 had on average, a significantly higher amount of viewers than season 17 since the mean value for season 4 was higher than the mean value for season 17 (since $17.77 > 5.17$). However it also again shows that the number of viewers for season 17 was more consistent than the number of viewers for season 4 as the standard deviation value for season 17 is less than the standard deviation value for season 4 (since $0.55 < 1.78$). I have now carried out t tests to allow me to see if there is a statistically significant difference between the number of viewers for season 4 and 17.

Statistical Analysis

T.Test

I then carried out T test's to obtain the p value and the confidence interval for the data involving the number of viewers for season 4 and season 17.

(Figure 5)

My hypothesis are as follows:

H ₀	There is not a statistically significant difference between the number of viewers for season 4 and season 17.
H ₁	There is a statistically significant difference between the number of viewers for season 4 and season 17.

(Figure 6)

The results from my t-test are as follows:

Confidence Interval	P-value
(11.65890 ,13.54934)	2.2x10 ⁻¹⁶

The T.Test I carried out allowed me to compare the number of viewers between season 4 and season 17. The p-value was 2.2×10^{-16} which is less than 0.05 and therefore we can reject the null hypothesis and so we are able to say that there is a statistically significant difference between the number of viewers for season 4 and season 17. The confidence interval then tells us that we are 95% confident that there is a difference of between 11.66 and 13.55 million viewers for season 4 and season 17.

Analysis and interpretation

Candidate 9

Analysis and Interpretation

The mean and standard deviation of the number of delays caused by weather conditions at Los Angeles International Airport and New York's JFK Airport were calculated. (Figure 3)

Figure 3:

Mean (New York JFK)	72.52318
Standard Deviation (New York JFK)	47.03273

Mean (Los Angeles LAX)	96.49007
Standard Deviation (Los Angeles LAX)	44.69689

Figure 3 tells us that on average, Los Angeles' mean number of delays due to weather conditions is higher than that of New York JFK ($96.49007 > 72.52318$).

This shows that New York has fewer delays caused by weather conditions than Los Angeles. There is a slight difference in the number of delays between New York (JFK) and Los Angeles. Overall, the spread is slightly higher in Los Angeles than in New York ($47.03273 > 44.69689$). This shows that Los Angeles has more delays than New York.

The comparative boxplot (Figure 2) highlights a great difference between the New York JFK and Los Angeles airports. I am going to assume that both sets of data are normally distributed, and I will now run a Two Sample t-test to determine if there is a statistically significant difference between the mean number of delays caused by weather conditions at New York JFK and Los Angeles LAX airports. Our null hypothesis would state that there is no difference between the mean number of delays caused by weather conditions between New York JFK and Los Angeles airports whereas our alternative hypothesis would state that there is a difference between the mean number of delays caused by weather conditions between New York JFK and Los Angeles airports.

```
Welch Two Sample t-test
data: JFK.NUMBER.OF.DELAYS and LAX.NUMBER.OF.DELAYS
t = -4.5873, df = 301.18, p-value = 6.597e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -34.41790 -13.75315
sample estimates:
mean of x mean of y
 72.30921  96.39474
```

As our p-value is below 0.05 (**6.597 e-06**), we can reject the null hypothesis. Therefore, we can say that there is a statistically significant difference between the number of delays caused by weather conditions at New York JFK and Los Angeles airports.

We can say with 95% confidence that the data will lie between -34.41790 and -13.75315, and we can say that there is a statistically significant difference in the number of delays caused by weather conditions between New York JFK and Los Angeles airports.

Analysis and interpretation

Candidate 10

```
3:1 (Top Level) ↕
Console Terminal Jobs
R 4.1.3 ~/
i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this messa
> View(Aidan_Project_Aidan_Project)
> attach(Aidan_Project_Aidan_Project)
> cor.test(fat,cal)

Pearson's product-moment correlation

data: fat and cal
t = 0.58425, df = 17, p-value = 0.5667
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3352749  0.5588935
sample estimates:
 cor
0.1403

> summary(fat)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.300  3.350  3.600  4.868  5.100 13.100
> |
```

Conclusion

Candidate 11

Conclusion

In conclusion, there is a clear link between the rates of deaths from opioid Overdoses and the years progressing, which results in more deaths from the drugs happening from ODs every year since each year they are getting more common to obtain. As seen from the bar graph, the rate is increasing at a rapid rate

Conclusion

Candidate 12

Conclusion

In conclusion, when looking at the boxplots on the number of required knee replacement surgeries for males and females they suggest that there is a difference due to the median of the number of females being higher than the males as well as the maximum value also being higher for females. The bar charts also suggest that there is a difference as in the bar chart for the required knee replacements for females the maximum value reaches around 880 surgeries for 75-79 year olds and for the same age range in males, there was only around 760 surgeries, suggesting a difference. The descriptive statistics also suggest that there is a difference between male and females as on average the mean for females surgeries is larger meaning there are more surgeries for females as well as the standard deviation being higher showing the data for the number of female surgeries is also more varied. However, with the confidence interval being 95% this shows that we could be 95% confident that there was no significant difference between the received knee replacement surgeries for males and females. Finally, the two sample t-test gave strong evidence to suggest that there isn't a statistically significant difference between the number of males and females that received knee replacement surgeries, as we do not reject the null hypothesis that the difference between the estimated mean and the expected mean is zero. This is due to the p-value being larger than the significance level meaning it's not rejected. Therefore due to the t-test we can determine that although there is a difference between males and females receiving knee replacement surgeries, there isn't a large or significant difference between females and males requiring knee replacement surgeries.