Applications of Mathematics (Higher): project

Candidate evidence

Candidate 1 evidence

Introduction

This report is an investigation into violent assault rates in Alabama. This report will analyse violent assault rates in Alabama between 1960-79, and 2000-19, and will determine if there is a statistically significant difference between the two data sets.

Is there are a statistically significant difference between the violent assault rate in Alabama per 100,000 people during time period 1960-1979 and 2000-2019.

The data being used is numerical continuous as it has been measured over a period of time.

The data is sourced from the CORGIS dataset project, which has compiled the data from an official website of the Federal Bureau of Investigation (FBI), reporting Unified Crime Statistics. This ensures the reliability of the data as the FBI receives these statistics from 18,000 US law enforcement agencies of all levels.

.

The data is robust and reliable as the FBI is a public body and likely to be transparent. There may be bias present as the FBI does not directly gather this data, it is instead self-reported by state, local and tribal law enforcement agencies, who may be inclined to lie to benefit their own interests.

Candidate 2 evidence

Introduction

BMW and Porsche are two of the biggest car manufacturers in the world, they are both Germany based with Porsche being based in Stuttgart and BMW in Munich, both manufacturers have SUV's, Performance/Track and luxury cars. In this project I aim to see if the car manufacture has a relationship with fuel economy on highways. The data being used is numerical (mpg), categorical (manufacturer). The data being used was extracted from study conducted by an assistant professor by the name of Austin Cory Bart of the University of Delaware., I know this data is reliable as it has been collected by an assistant professor.
The null hypothesis for this project is; there is no correlation between car manufactures and fuel economy (mpg) on the highway.
The alterative hypothesis for this project is; there is a correlation between car manufacturers and fuel economy (mpg) on the highway.

Candidate 3 evidence

Introduction

my question Is you have a hgher chance of you dieing id you where younger on the titanic

The ship Titanic sank in 1912 with the loss of most of its passengers. Details can be obtained on 1309 passengers and crew on board the ship Titanic. The main use of this data set is Chi-squared and logistic regression with survival as the key dependent variable. Summary statistics for the categorical variables can be demonstrated and the cost of the ticket (fare) is very skewed so it can be used to demonstrate skewed data and differences between means and medians etc.
The titanic data has also been linked to numerous articles in the press including this one:
Australasia

Candidate 4 evidence

Subjective impression

Figure 1

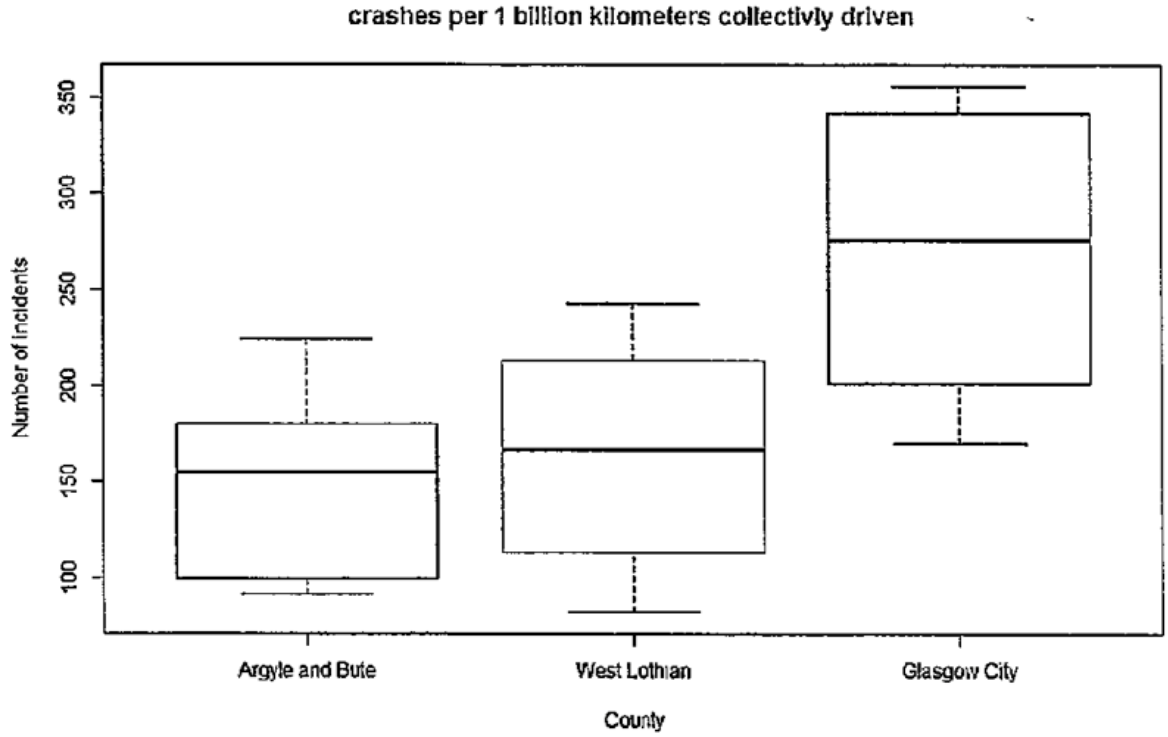


Figure 1 shows how many crashes happen per 1 billion miles driven collectively by the public. We can see that Glasgow City has more collisions than Argyle and Bute and West Lothian. While Argyle and Bute and West Lothian have a similar number.

Figure 2

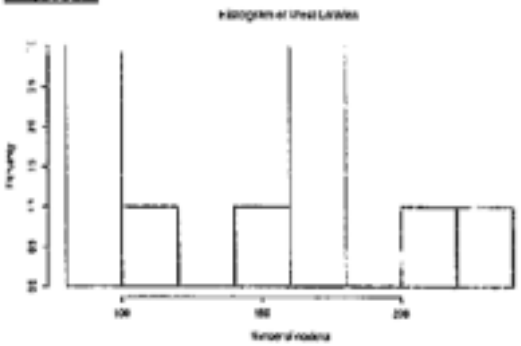


Figure 3

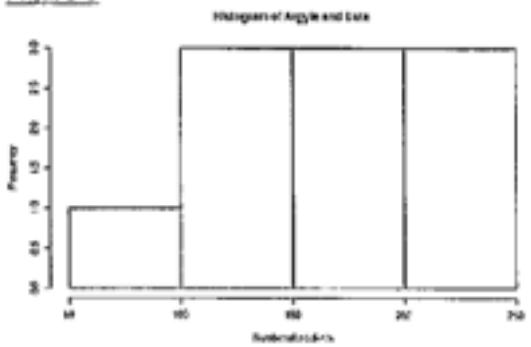
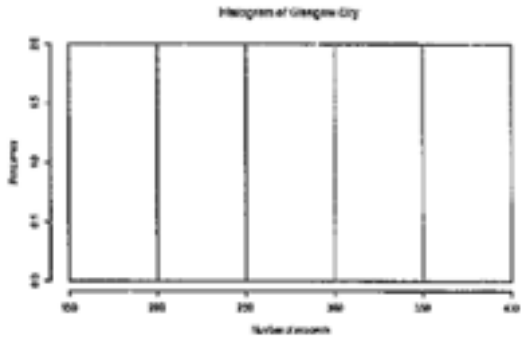


Figure 4



These charts show me the distribution of the data. The data is not normally distributed.

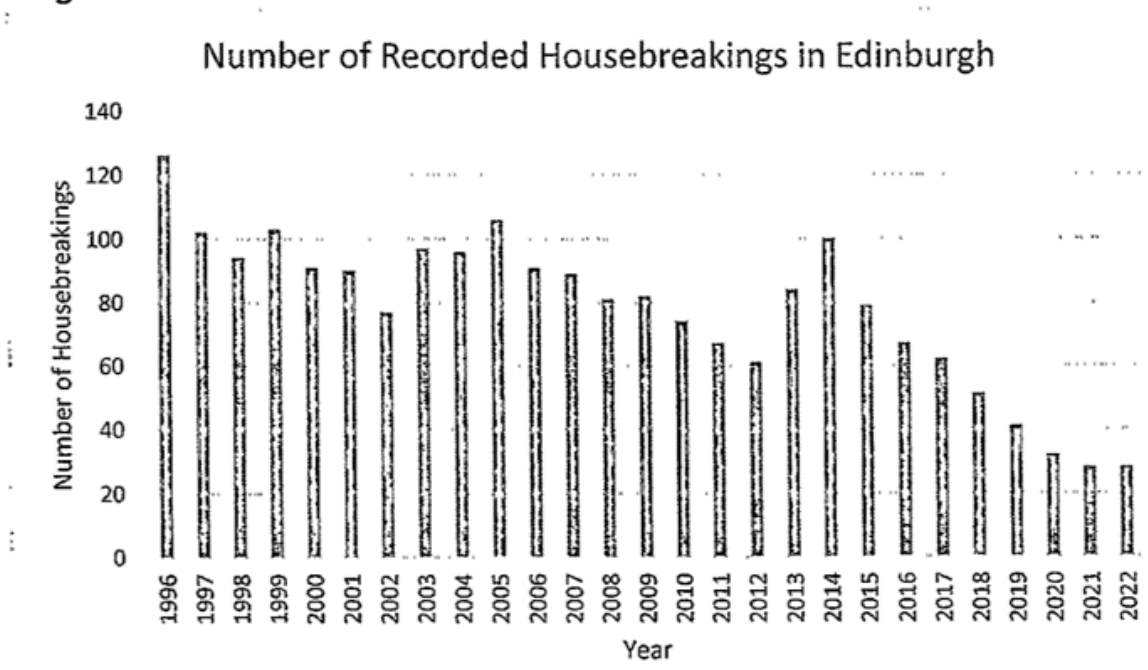
	median	Q1	Q3	Interquartile range
West Lothian	155	101.8	179.5	77.7
Glasgow City	276	208.8	333.8	125
Argyle and Bute	167	121	206	85

Despite the data being not normally distributed I will still run a t.test since I don't know how to run another test.

Candidate 5 evidence

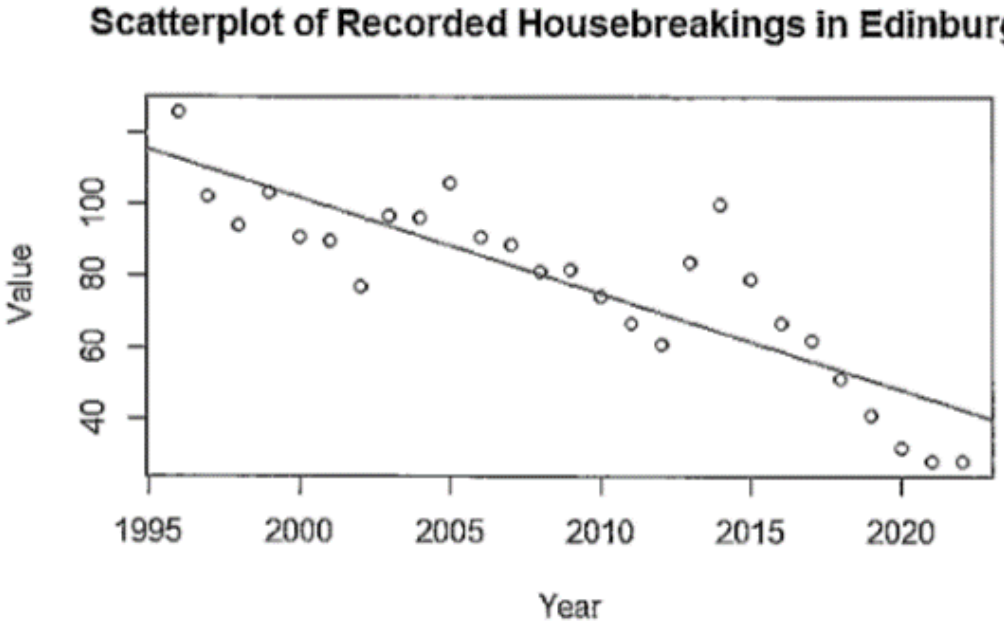
Subjective impression

Diagram 1



In diagram 1, I constructed a bar chart which shows a good variable picture over the 26 years of data. I can also indicate from this bar chart that there is a strong decreasing trend from 2014 onwards. This shows that a linear relationship may exist in this data.

Diagram 2



In diagram 2, I constructed a scatterplot to find if there is a linear relationship over the 26 year time period from 1996-2022. This clearly shows that a linear relationship may exist as there is a downwards trend in the data from about 2015 onwards. It also shows that house breaking crimes in Edinburgh were at their peak in 1996 and have gradually decreased overtime.

Descriptive statistics

I have calculated the mean and standard deviation of the number of recorded house breakings in Edinburgh over the 26 year period shown below.

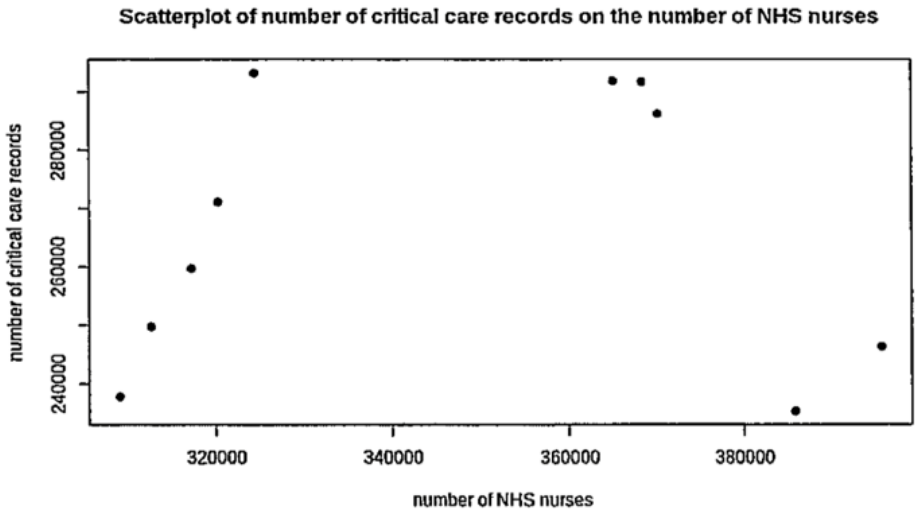
Mean (Housebreakings)	77.74074
Standard Deviation (Housebreakings)	25.11143

Over the 26 year time period it shows that the mean was 77.74074. The standard deviation was 25.11143 which is high, so this indicates that the number of housebreakings over the years were very varied.

Candidate 6 evidence

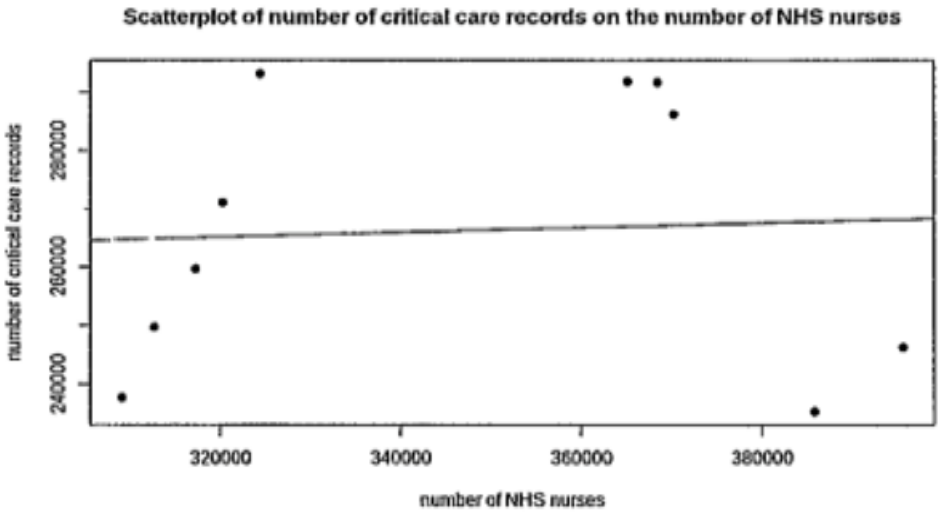
Subjective impression

Scatterplot



I have used a scatterplot to compare the number of critical care records on the number of NHS nurses. I have used this to give an initial idea about the association between my two sets of data. From this I have gained a rough idea that there is a week linear relationship between my two sets of data

Line graph



I have used a line graph to further add to my idea of the association between my two points given to me by my scatter graph. From this line graph we can clearly see there is a week linear relationship between the number of critical care records and number of NHS nurses, backing up our original assumption and suggesting there is little correlation between the two.

Descriptive Statistics	number of critical care records	number of NHS nurses
standard deviation	2342210.62%	3320547.86%
mean	26626720.00%	34691460.00%

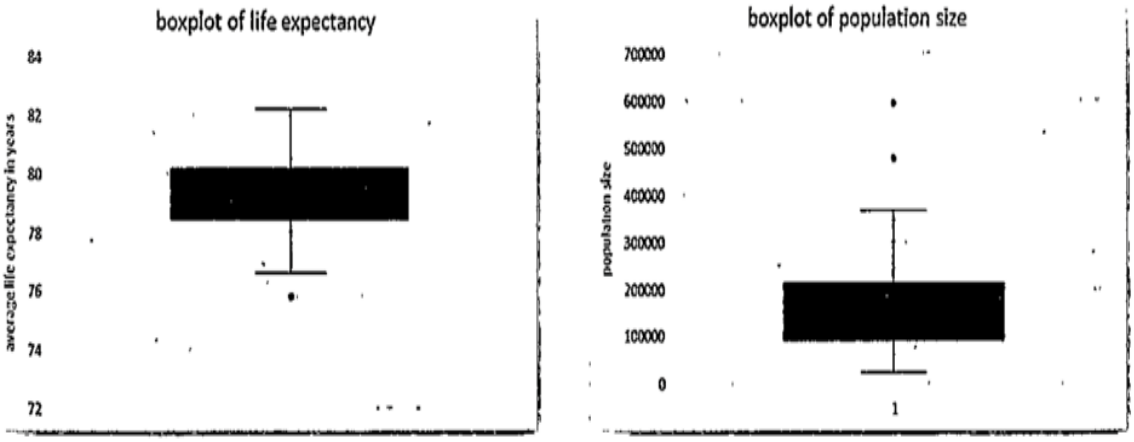
The descriptive statistics show that over the 10 years the average percentage of number of NHS nurses is greater than the number of critical care records, and the high standard deviation of number of NHS nurses suggests to us that the percentage of number of NHS nurses varied largely.

Candidate 7 evidence

Subjective impression, analysis and interpretation

I will begin by producing boxplots as it is a good visual test to check the distribution of the data and check for outliers.

Figure 1 Boxplots of Average Life Expectancy and Population size

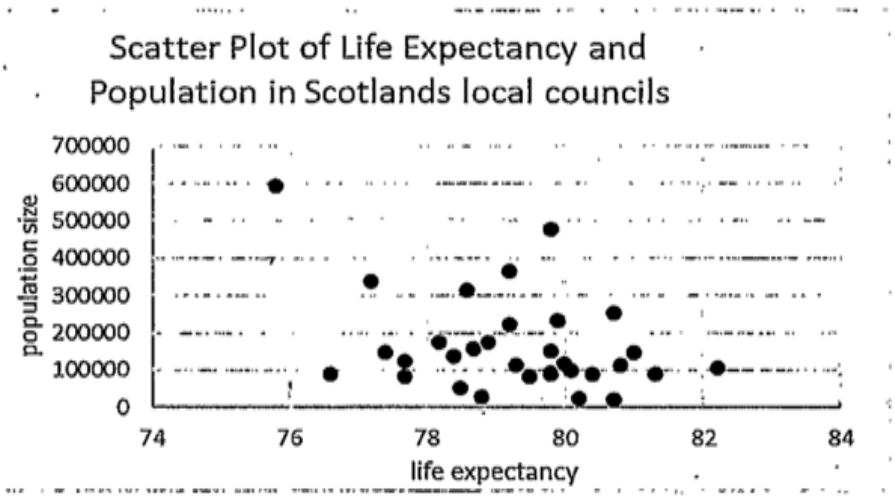


Since there are outliers my sample is skewed, however I am going to assume the population and life expectancy are normally distributed. The boxplot of life expectancy we can see has 1 outlier which is Glasgow city council and the boxplot on population size has 2 outliers which are Glasgow city council and Edinburgh city council.

<u>Analysis and interpretation</u>

I will now create a scatterplot to visually check for a relationship between population size and life expectancy.

Figure 2



In figure 2 we can see that there isn't much of a trend between the population size and life expectancy. However, I will check this further using a correlation test.
I calculated the mean and standard deviation (figure 3) for the life expectancy and the population of the local councils in Scotland.

Figure 3

Mean (life expectancy)	79.25625
Standard Deviation (life expectancy)	1.398423
Mean (Population)	165481.3438
Standard Deviation (Population)	127626.6925

Figure 3 tells us that on average life expectancy in the local councils of Scotland was 79 and the average population of those local councils is 165481 (rounded to nearest whole number). The large standard deviation in population size shows that the population varies greatly from council to council. I no longer need the mean and standard deviation for further calculations.

I have done a correlation test on the data that has been presented in the charts above which shows us the correlation coefficient of 0.01386422. There is a very weak positive relationship between the Population Size and the Life Expectancy in Local Councils in Scotland.

Figure 4

```
        Pearson's product-moment correlation

data:  Average.LE and Population.Size
t = 0.078435, df = 32, p-value = 0.938
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3258291  0.3503872
sample estimates:
       cor
0.01386422
```

The P-value is 0.938 this is more than 0.05 which means that there is no significant linear relationship between life expectancy and population size in Scotland.
Although there is no linear relationship between my two variables I will produce a linear model on the data which could be used for predictions if there was.

Linear Model

I will now compute a linear model that could be used to predict the life expectancy given the population size
the equation of the linear regression model can be calculated as: life expectancy= -0.000004005 (population size) +79.92

Figure 5

```
Call:
lm(formula = Average.LE ~ Population.Size)

Coefficients:
     (Intercept)  Population.Size
       7.992e+01        -4.005e-06
```

Figure 6

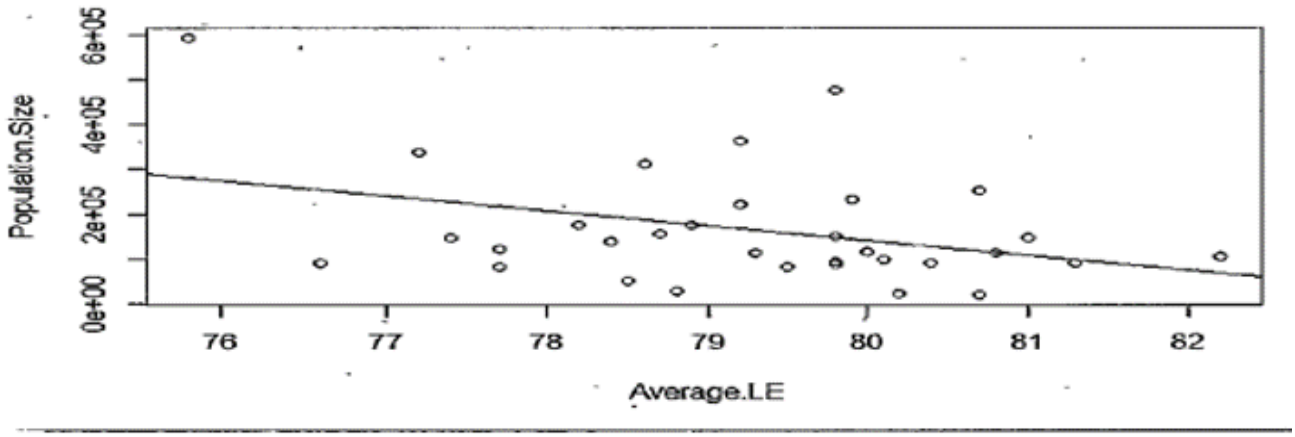


Figure 6 shows there is a decrease in trend in the population and life expectancy of the local councils in Scotland showing linear relationships. I carried out a linear regression for the whole data set.

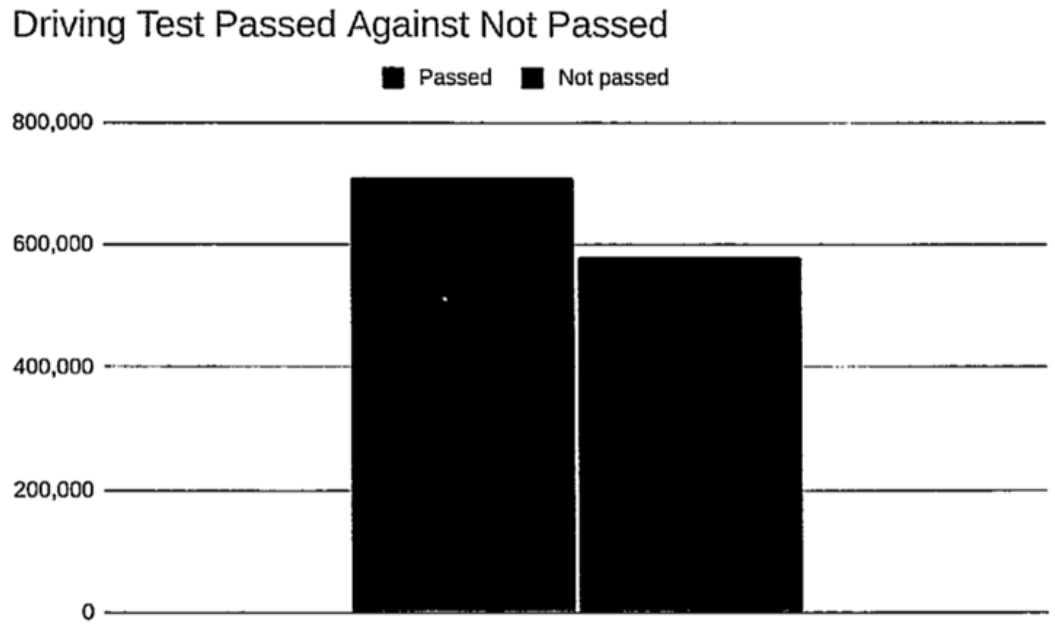
```
Call:
lm(formula = Average.LE ~ Population.Size)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9557 -1.0108  0.2037  0.8484  2.7016

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.992e+01  3.891e-01 205.379   <2e-16 ***
Population.Size -4.005e-06  1.862e-06  -2.151   0.0396 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.344 on 30 degrees of freedom
Multiple R-squared:  0.1336,    Adjusted R-squared:  0.1047
F-statistic: 4.627 on 1 and 30 DF,  p-value: 0.03965
```

The R squared value is 0.1336, this means that 13.36% of our data can be explained by our model. This shows us that the model does not fit out data well and can't be confidently used to predict the life expectancy given the population size of a local council.

Candidate 8 evidence

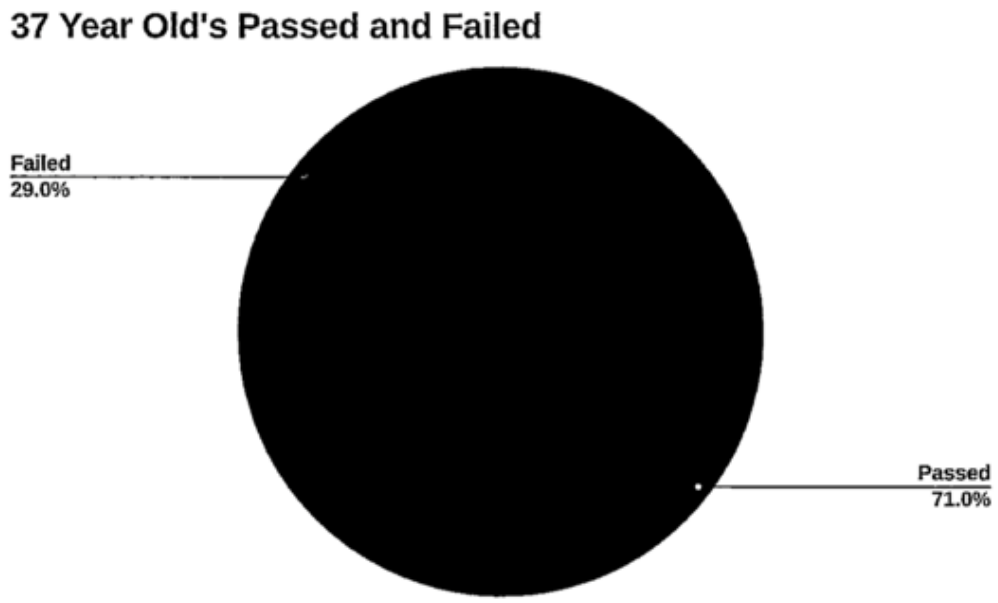
Subjective impression, analysis and interpretation

Subjective Impression
Figure 1:



This bar chart is helpful because it suggests that 17-year-olds have a higher pass rate than 37-year-olds. This is shown by 129798 more 17-year-olds passed than 37-year-olds this suggests that the younger you are the higher the rate you will have to pass is more likely.

Figure 2:



This pie chart shows the pass rate in % of 37-year-olds when sitting a driving test. This shows that 71% of people pass and 29% of 37-year-olds fail. This supports my theory that the younger you are, the higher your chance of passing your driving test.

	Passed	Failed	Total
Seventeen	711135	581337	1292472
Thirty-seven	47237	19250	66487

From this data, we can see that the total number of driving tests conducted for the age group seventeen is significantly higher compared to the age group thirty-seven. Additionally, the number of tests passed for the age group seventeen is substantially higher than the number of tests passed for the age group thirty-seven.

Analysis and Interpretation:

I will be conducting a z-test to determine whether there is a difference between you being older and taking a driving test or being younger and taking a driving

test. I have chosen a z-test because my data is categorical and compares the proportions of two different age groups.

Null hypothesis: there is no difference in age and passing a driving test between 17 and 37-year-olds.

Alternative hypothesis: there is a difference in age and passing a driving test between 17 and 37-year-olds.

2-sample test for equality of proportions with continuity correction

data: c(711135, 4737) out of c(1292472, 66487)
X-squared = 58191, df = 1, p-value < 2.2e-16
alternative hypothesis: two. sided
95 percent confidence interval:
0.4768230 0.4811091
sample estimates:
prop 1 prop 2
0.55021308 0.07124701

p-value = 0.00000000000000022 As the p-value is less than 0.05, therefore, we reject the null hypothesis.

95 percent confidence interval: 0.4768230 0.4811091
We can be 95% confident that the driving test pass rate and age lie between 0.4768230 and 0.4811091.

Candidate 9 evidence

Conclusion

Conclusion

To conclude my study I can look back and tests and graphs produced. The two histograms show that more types of red meat have a higher protein content then then majority of white meat types. On the other side, we see there are more white meats that have a low protein than red meats. This shows there is a difference in protein levels between red and white meats. The descriptive statistics give us a clear number to prove a difference in protein levels. There is a median average of 28.57 grams of protein for red meats and only 26.74 for white meats. This gives us a clear number to conclude the question of whether there is a difference. This pairs with the IQR where we can see red meats are more varied at 4.07 compared to 3.01 for white. By using the Mann Whitney test to find the p value of 1.428e-05 it tells us to reject the null hypothesis, again showing there is a difference. The 95% confidence interval found between -3.892662 and -1.236489 is more numerical proof of a difference as we see two number that are not 0 showing that the space in-between the two is the difference. By taking all of these points into consideration from visuals to units, I can answer the research question. 'Is there a difference between the amounts of protein found in red meats and white meats?' yes, there is a difference between amounts of protein in red and white meats as on average red meats have a higher protein count then white meats. Despite this, they are both nutritious forms of protein to help with muscle growth and to help with fulfilling a balanced diet.

Candidate 10 evidence

Conclusion

In conclusion, the research question investigated in this project was whether there is a relationship between antenatal care and stillbirth rates from countries across the world in the period from 2011-2021.

When looking at the original scatter graph (Figure 1) there was an indication of a relationship, this was then further proved using a correlation coefficient and hypothesis testing. In doing this the R studio output showed that there was a moderate correlation and there was enough evidence to reject the null hypothesis which is 'There is no relationship correlation between Stillbirth rate and Antenatal care'.

The null hypothesis was rejected because of the P value being less than 0.05 and the 95% confidence interval not covering 0.

As shown by the histograms (Figure 2 and 3) both sets of data are skewed which then showed that median and IQR were the appropriate descriptive statistics to use. These numbers were used to measure the spread and location of the data sets.

Overall this project has shown that sufficient Antenatal care, at least 4 visits does have an impact on the Stillbirth Rate of Countries across the world for 2011-2021.