Statistics Final Project Report

Ian Smith
School of Computing
National College of Ireland
Dublin, Ireland
x19178816@student.ncirl.ie

Abstract

Different types of statistical analysis were carried out for this project which were Logistic Regression, Kruskal-Wallis test, Independent Samples t-test and a Chi-squared test for Independence. The logistic regression model classified life expectancy into two categories with an accuracy of 88.24%. The Kruskal-Wallis showed that there was a statistically significant difference in life expectancy between different continents (p-value = 1.76×10^{-6}). The independent samples t-test showed that there is a significant difference between the mean GPA of Male and Female students (p-value = 0.004). The p-value from the Pearson Chi-Square test was 0.389 which showed that there is not a significant association between a student's gender and them having a child or not.

I. Introduction

Three pieces of statistical analysis were carried out as part of this project. These involved developing a logistic regression model, performing a Kruskal-Wallis ranked sum test, an independent samples t-test and a Chi-square test for independence.

A. Logistic Regression

The aim of the logistic regression is to develop a model that can classify life expectancy into two categories which are 'High' and 'Low'. Different features such as adult mortality rate, alcohol consumers and child mortality rate are used as predictors.

B. Kruskal-Wallis Test

A Kruskal-Wallis ranked sum test was carried out in order to determine if there was a significant difference in life expectancy between 4 different continents which were South America, Asia, Africa and Europe.

C. Fundamentals of Statistics

In this section an independent samples t-test and a chi-squared test for independence were carried out. The t-test was used to see if there was a significant difference between the mean GPA of students based on their gender. The chi-squared test was used to see if there was an association between a student's gender and them having a child or not.

II. Methodology & Dataset Description

A. Logistic Regression

The dataset that was used for the logistic regression model was obtained from the World Health Organisation's online database (www.who.int/gho/database). There were 6 variables in total which were labelled life expectancy, healthcare expenditure, childhood mortality rate, adult mortality rate, drinking rate and alcohol consumption.

Methodology:

- A total of 119 countries were selected from the WHO's website. This was carried out by downloading the relevant variables as csv files, importing them in RStudio and merging into one large dataframe.
- In order to convert the dependent variable life expectancy into a form suitable for regression, the median life expectancy of all the countries was calculated and determined to be 78.5 years. Any countries that were greater than or equal to that value were classified as 'high' (1) and under that value were classified as 'low' (0).
- After checking for collinearity it was found that adult mortality rate and child mortality rate were very highly correlated (0.93). The decision was taken to drop child mortality rate.
- The analysis began by fitting the model with all of the independent variables. Any variables that did not make a significant contribution to the model (p>0.05) were removed. The final model contained two predictors which were named adult mortality rate and alcohol consumers.
- Values were selected from 2016 for uniformity.

Dataset description:

Table I: Final dataset for the logistic regression model

Variable Name	Type	Class	Unit of Measurement
Life Expectancy	Numeric	Dichotomous Variable	Years
Adult Mortality Rate	Numeric	Continuous	Prob. of dying between 15 & 60 years per 1000 pop
Alcohol consumers	Numeric	Continuous	Alcohol consumers over the past 12 months (%)

B. Kruskal-Wallis Test

The dataset used for the Kruskal-Wallis Test was obtained on the World Health Organisation's online database (www.who.int/database). The unprocessed dataset contained 3 columns which were Country, Year and Life expectancy (in years).

Methodology:

- A total of 40 countries were selected from the WHO's website. 10 countries from each of the 4 continents (Europe, Asia, South America and Africa) were selected.
- A new column called continent was created using Excel which specified the continent in which each country resided. The final csv file was imported into R.
- The final dataframe consisted of two columns which were life expectancy in years and continent.

Dataset description:

Table II: Final dataset for the Kruskal-Wallis test

Variable Name	Type	Class	Unit of Measurement
Life Expectancy	Numeric	Continuous Variable	Years
Continent	Factor	Categorical	Continent the country is located in

C. Fundamentals of Statistics

The dataset used for the final section was the "CollegeStudentData.sav" file. The data contains information about a survey conducted at a college about various characteristics of students. It contains 50 rows and 18 variables. The software used to import the dataset and perform the analysis was carried out in SPSS.

III. Model Results & Evaluation

A. Logistic Regression

In this section, various assumptions for the logistic regression will be discussed and the results of the model output will be interpreted.

```
Call:
glm(formula = Life_binary ~ . - healthcare_exp - drinking_services -
    cmortality_rate, family = binomial(link = "logit"), data = df_factor)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.14122  -0.10507   0.05581   0.24257   1.84063

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       5.17177    1.55460   3.327 0.000879 ***
adult_mort_rate  -0.08132    0.01986  -4.094 4.23e-05 ***
alcohol_consumers  0.05780    0.01888   3.062 0.002198 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 164.893  on 118  degrees of freedom
Residual deviance:  53.474  on 116  degrees of freedom
AIC: 59.474

Number of Fisher Scoring iterations: 8
```

Figure 1: Output of the logistic regression model.

1) Assumptions: There are certain assumptions of a logistic regression model that must be met.

Sample Size: To check that the sample size is large enough to perform logistic regression, the formula N > 50+ 8(m) (where N is the sample size and m is the number of predictors) will be used. For this project the sample size (119) should be 50 + 8 times the number of predictors (5).

$$119 > 50 + 8(5)$$
$$119 > 98$$

From the above test, it can be concluded that the sample size is sufficiently large enough to perform logistic regression analysis.

Multicollinearity: In order to check for multicollinearity a correlation matrix for all the variables was created. Any variables that showed correlation over .65 were deemed to be too strongly correlated and were dropped.

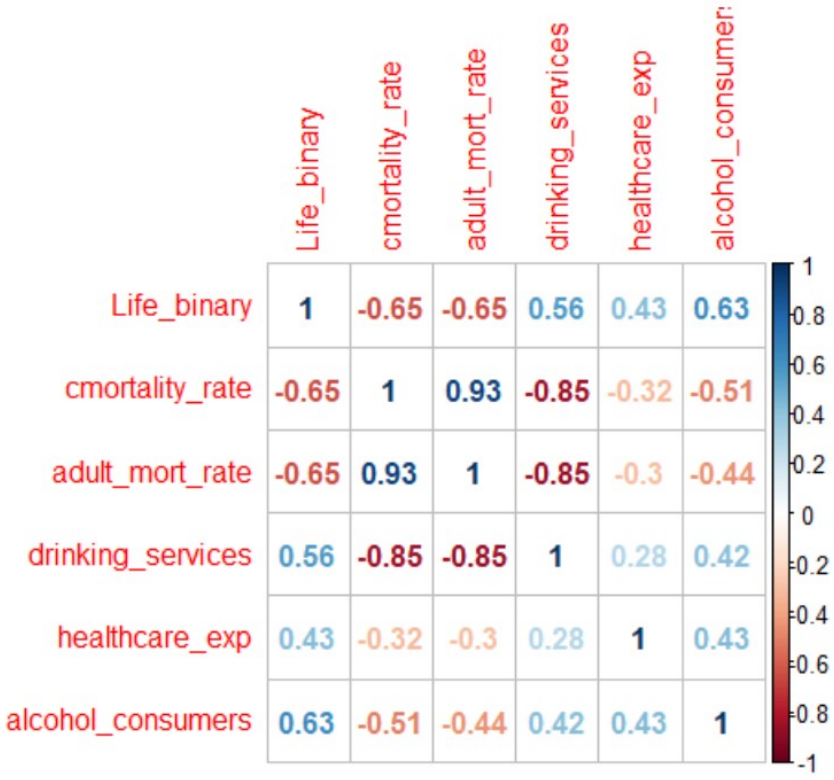

Figure 2: Correlation matrix for the variables in the logistic regression model.

The variables adult mortality rate and child mortality rate were highly correlated so the decision was taken to drop child mortality rate.

Pseudo R-squared and the Hosmer & Lemshow Test:

- The values of the Cox & Snell and Nagelkerke R square are analogous to the R^2 found in multiple regression. The values obtained for the Cox & Snell and Nagelkerke were 0.607 and 0.81 respectively. This means that the amount of behaviour explained by the independent variables in the model varies between 60.7% and 80.1%.
- The Hosmer & Lemshow test gives an indication of how good the model fit is.

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  df1$Life_binary, fitted(logfit7)
X-squared = 5.7502, df = 8, p-value = 0.6752
```

Figure 3: The Hosmer & Lemshow test for the logistic regression model.

The p-values of 0.6752 shows that there is not enough evidence at the 5% level of significant to suggest that the logistic regression model has poor fit. As a result, we fail to reject the null hypothesis and conclude that the model has a good fit.

2) Interpreting the Result: To interpret the result different evaluation metrics such as the classification table (a.k.a confusion matrix) will be used.

Confusion matrix: The confusion matrix helps to describe how well a classification model performs.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 51  7
         1  7 54

               Accuracy : 0.8824
                 95% CI : (0.8105, 0.9342)
    No Information Rate : 0.5126
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7646

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8852
            Specificity : 0.8793
         Pos Pred Value : 0.8852
         Neg Pred Value : 0.8793
             Prevalence : 0.5126
         Detection Rate : 0.4538
   Detection Prevalence : 0.5126
      Balanced Accuracy : 0.8823

       'Positive' Class : 1
```

Figure 4: The confusion matrix for the logistic regression model with additional evaluation metrics.

The confusion matrix shows that the model performed very well when classifying life expectancy by high or low. There were 51 true positives and 54 true negatives. There were only 7 false positives and false negatives. The percentage of 'Low' life expectancy classified as correct was 87.93% and the percentage of 'High' life expectancy classified as correct was 88.52%. The overall accuracy was 88.24%.

Wald Test:

The Wald Test is used to test the statistical significance of each coefficient in the model. The output in R is given below.

```
Wald test for adult_mort_rate
 in glm(formula = Life_binary ~ . - healthcare_exp - drinking_services -
    cmortality_rate, family = binomial(link = "logit"),
    data = df_factor)
F =  16.76479  on  1  and  116  df: p= 7.8581e-05
```

Figure 5: The Wald test for the adult mortality rate variable.

```
Wald test for alcohol_consumers
 in glm(formula = Life_binary ~ . - healthcare_exp - drinking_services -
    cmortality_rate, family = binomial(link = "logit"),
    data = df_factor)
F =  9.37661  on  1  and  116  df: p= 0.0027319
```

Figure 6: The Wald test for the alcohol consumption variable.

The Wald test shows that at the 5% level of significance both of the variables contribute significantly to the model. This shows that the two independent variables are indeed predictors of life expectancy. Also, the 95% confidence interval for exp(β) in fig. 7 shows that we are 95% confident that the adult mortality rate will have an odds ratio between 0.8807 and 0.9530. Likewise, for alcohol consumers the odds will fall between 1.0243 and 1.1051.

Logistic Equation & Odds Ratio:

The coefficients in the equation had the following values:

- β_0 (constant) = 5.1718
- β_1 = -0.0813
- β_2 = 0.0578

Using the coefficients the logistic regression equation can be written as follows:

$$p(X) = \frac{e^{5.1718-0.0813X1+0.0578X2}}{1+e^{5.1718-0.0813X1+0.0578X2}} \tag{1}$$

where X1 is the value for adult mortality rate, X2 is the value for alcohol consumption and p(X) is the probability of being assigned to high or low life expectancy. The coefficients were estimated using the maximum likelihood method.

Maximum Likelihood Estimates

Parameter	DF	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1	5.1718	1.5546	3.3268	9e-04
adult_mort_rate	1	-0.0813	0.0199	-4.0945	0.0000
alcohol_consumers	1	0.0578	0.0189	3.0621	0.0022

Odds Ratio Estimates

Effects	Estimate	95% Wald Conf. Limit	
adult_mort_rate	0.9219	0.8807	0.9530
alcohol_consumers	1.0595	1.0243	1.1051

Figure 7: Maximum Likelihood and Odds Ratio estimates for the logistic regression model.

The 'Estimate' column represents the odds ratio of a country having high or low life expectancy. An odds ratio of 0.9219 means that for every 1 unit increase in the life expectancy of a country the odds in favor of life expectancy being in the high category decreases by a factor of 0.9219, all things being equal. An odds ratio of 1.0595 means that for every one unit increase in life expectancy the odds in favor or life expectancy being in the high category increases by a factor of 1.0595, holding all other variables constant.

3) Result Statement: A logistic regression model was created using R to predict the life expectancy category of 119 different countries based on adult mortality rate and alcohol consumption. The assumptions of logistic regression were met and the model explained variations in the life expectancy category between 60.7% and 80.1%. The two predictors were determined to be statistically significant. The life expectancy category was classified with an accuracy of 88.24%.

B. Kruskal-Wallis Test

The first test that was tried out was a one-way ANOVA. Not all of the assumptions for the ANOVA were met which lead to the Kruskal-Wallis test being implemented. Also, the results of the Kruskal-Wallis ranked sum test will be given and interpreted. Before checking any of the assumptions some summary statistics were calculated and the data was visualised using a box-plot.

Continent <fctr>	count <int>	mean <dbl>	sd <dbl>	median <dbl>	IQR <dbl>
Africa	10	66.95	7.857799	67.85	11.125
Asia	10	73.27	5.486357	73.30	8.400
Europe	10	83.99	1.175160	83.75	1.525
South America	10	78.75	2.352894	78.90	1.550

Figure 8: Summary statistics for the data.

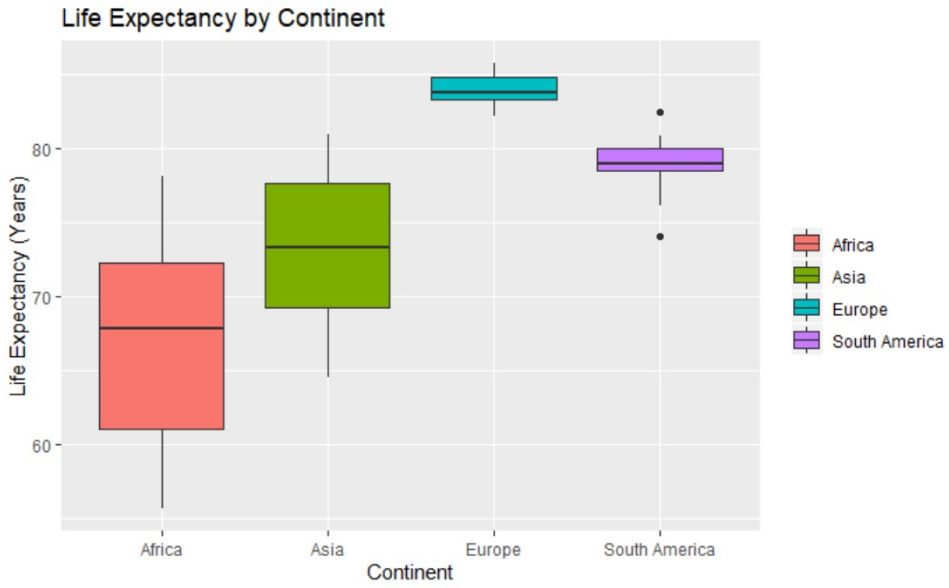

Figure 9: The life expectancy distribution for each continent.

1) ANOVA Assumptions: The assumptions for the ANOVA test are given below.

Continuous Scale:

The dependent variable must be measured using a continuous scale. Table II shows that the life expectancy variable meets this assumption.

Normal Distribution:

There is an assumption made that the populations from which the samples are taken are normally distributed.A Q-Q plot can show if this assumption is met.

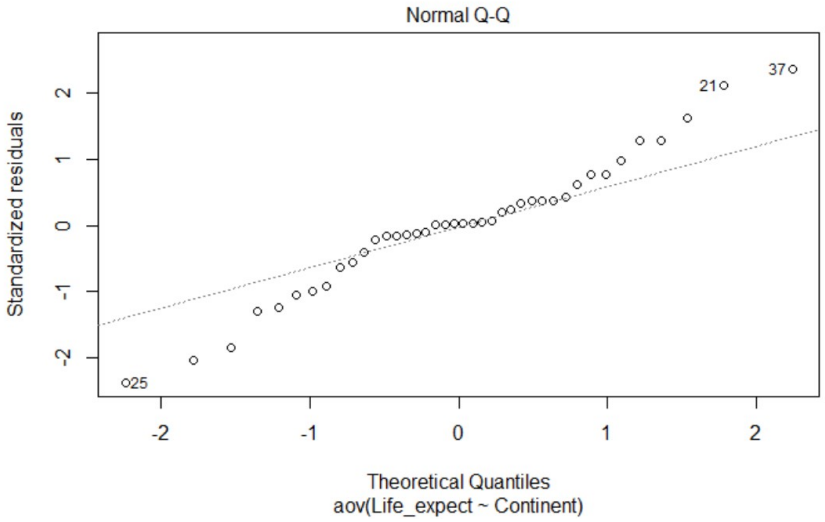

Figure 10: Q-Q plot for one-way ANOVA

The above Q-Q plot shows that the distribution is not normally especially at the beginning and end of the scale. The Shapiro-Wilk test can also be used to check for normality.

```
        Shapiro-Wilk normality test

data:  life_exp$Life_expect
W = 0.91243, p-value = 0.004486
```

Figure 11: Shapiro-Wilk test for normality

The Shapiro-Wilk test for normality gives a p-value of 0.004486 which is less than the critical value of 0.05. Therefore, we can conclude that the data is not normally distributed and the normality assumption is not met. The one-way ANOVA should not be used and the Kruskal-Wallis test should be used instead. ***Homogeneity of Variances:***
Levene's test was used to determine if this assumption was broken or not.

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value    Pr(>F)
group  3  9.7671 7.469e-05 ***
      36
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 12: Levene's test for homogeneity of variances.

The p-value of 7.469×10^{-5} is less than the significance level of 0.05. This means that there is evidence to suggest that the variance across groups is statistically significant.
Independence & Random Sampling
The observations are obtained independently and are sampled randomly from the population. The normality and homogeneity of variances assumptions were broken which means that an ANOVA is not suitable for this particular dataset. Instead, a Kruskal-Wallis ranked sum test will be used as it is a non-parametric test (doesn't make any assumptions about the population) and can be used when the assumption of normality fails. Welch's F-statistic would not be suitable here as the assumption of normality is not met.

2) Interpretation of Results: The output of the Kruskal-Wallis test is given below.

```
        Kruskal-Wallis rank sum test

data:  Life_expect by Continent
Kruskal-Wallis chi-squared = 29.498, df = 3, p-value = 1.76e-06
```

Figure 13: Output of the Kruskal-Wallis test.

The output of the Kruskal-Wallis test gives a p-value of 1.76×10^{-6}. As this is less than the significance level of 0.05, we conclude that there are statistically significant differences in life expectancy by continent.
Due to the fact that the variances were unequal, a Games-Howell post-hoc test was carried out.

	diff <dbl>	ci.lo <dbl>	ci.hi <dbl>	t <dbl>	df <dbl>	p <chr>
Asia-Africa	6.32	-2.35	14.99	2.09	16.09	.200
Europe-Africa	17.04	9.26	24.82	6.78	9.40	<.001
South America-Africa	11.80	3.95	19.65	4.55	10.60	.004
Europe-Asia	10.72	5.27	16.17	6.04	9.82	.001
South America-Asia	5.48	-0.11	11.07	2.90	12.20	.055
South America-Europe	-5.24	-7.68	-2.80	6.30	13.23	<.001

Figure 14: The Games-Howell post-hoc test

3) Result Summary: A Kruskal-Wallis ranked sum test was conducted to compare the life expectancy between different continents. The four continents were Europe, Asia, South America and Africa and there were 10 countries within each group. There was a statistically significant difference at the p<0.05 for the four different continents: $\chi^2(3) = 29.498, p = 1.76 \times 10^{-6}$.

Post-hoc comparisons using the Games-Howell test indicated that the mean score for Europe (M=83.99, SD = 1.18) was significantly different from Asia (M=73.27,SD = 5.49), Africa (M=66.95, SD = 7.86) and South America (M=78.75, SD=2.35). The mean score for South America did differ significantly from Africa. The mean score for Asia did not differ from Africa, nor did the mean score for South America differ from Asia.

C. Fundamentals of Statistics

Two different techniques were used in this section which were an independent samples t-test and a chi-square test for independence. The software used to carry out both tests was SPSS.

1) Independent Samples t-test: The independent samples t-test was used to see if there was a statistically significant difference in the mean GPA between Male and Female students in the sample.

Hypotheses:

$H_0 : \mu_{Male} - \mu_{Female} = 0$
$H_1 : \mu_{Male} - \mu_{Female} \neq 0$
where μ_{Male} and μ_{Female} are the mean GPA values for male and female students respectively.

Significance Level:

The significance level that will be used for this test is $\alpha = 0.05$.

Variables:

The test variable is "currgpa" (Current GPA of the students) and the grouping variable is "gender" (Male and Female).

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
student's current gpa * gender of student	50	100.0%	0	0.0%	50	100.0%

Figure 15: Case processing summary for the independent samples t-test. All of the records in the sample were used for this test.

Report

student's current gpa

gender of student	Mean	N	Std. Deviation
males	3.023	26	.3983
females	3.333	24	.3171
Total	3.172	50	.3907

Figure 16: Mean and standard deviation for current GPA factored by gender.

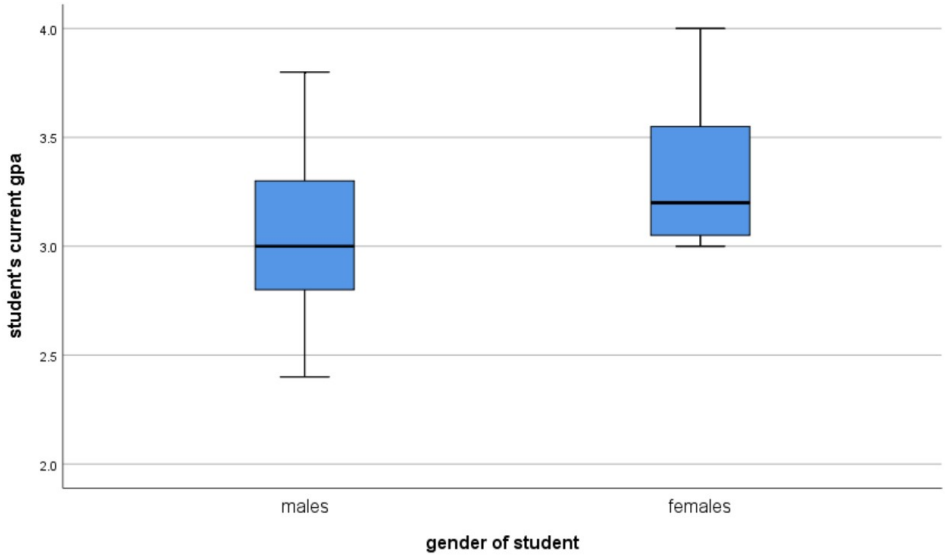

Figure 17: Box-plot showing the distribution of current GPA by gender.

After stating the null and alternate hypothesis, selecting the significance level and variables to be used the test was carried out. The results of it are given below.

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
student's current gpa	Equal variances assumed	.370	.546	-3.030	48	.004	-.3103	.1024	-.5161	-.1044
	Equal variances not assumed			-3.058	47.023	.004	-.3103	.1015	-.5143	-.1062

Figure 18: The results of the independent sample t test.

Result Summary:

The p-value of Levene's test is 0.546. Therefore, we fail to reject the null hypothesis of Levene's test and conclude that the variance of mean GPA between males and females is not significantly different. The t-test gave a p-value of 0.004 which is less than our significance level $\alpha = 0.05$. We can reject the null hypothesis and conclude that the mean GPA between male and female students is significantly different. The 95% confidence interval from our test is [-0.5161,-1.044].

Result Statement:

There was a statistically significant difference in current GPA values between Male and Female students ($t_{48} = -3.030$, p=0.004).

2) Chi-squared test for independence: The Chi-square test for independence was used to see if there is am association between a student' gender and if they have a child or not. ***Hypotheses:***

H_0 : There is no relationship between a student's gender and if they have a child or not.

H_1 : There is a relationship between a student's gender and if they have a child or not.

Significance Level:
The significance level used for this test is $\alpha = 0.05$.
Variables:
The variables used in this test were "gender" and "does subject have children."

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
does subject have children * gender of student	50	100.0%	0	0.0%	50	100.0%

Figure 19: Case processing summary for the chi-squared test. All of the records were used for this test.

does subject have children * gender of student Crosstabulation

			gender of student		
			males	females	Total
does subject have children	no	Count	14	10	24
		Expected Count	12.5	11.5	24.0
		% of Total	28.0%	20.0%	48.0%
		Residual	1.5	-1.5	
	yes	Count	12	14	26
		Expected Count	13.5	12.5	26.0
		% of Total	24.0%	28.0%	52.0%
		Residual	-1.5	1.5	
Total		Count	26	24	50
		Expected Count	26.0	24.0	50.0
		% of Total	52.0%	48.0%	100.0%

Figure 20: Crosstabulation table. With the expected count values shown, we can confirm that all cells have an expected value greater than 5, satisfying one of the assumptions for the chi-squared test for independence.

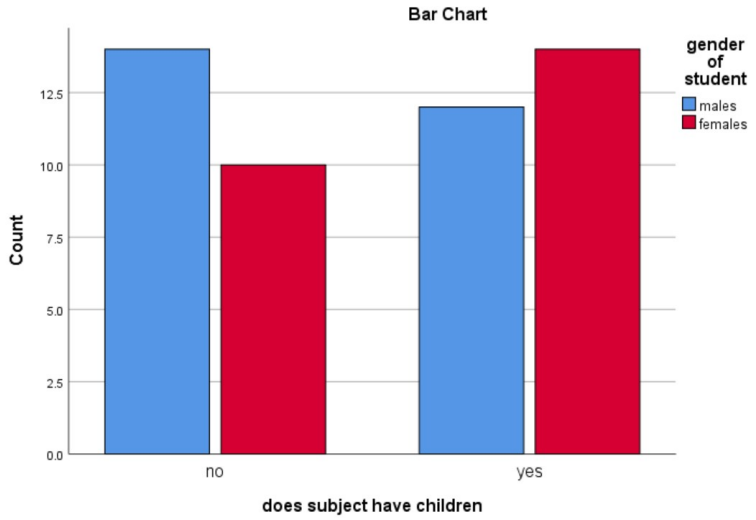

Figure 21: Bar chart breaking down a student's gender and if they have children or not.

After stating the null and alternative hypothesis, selecting the level of significance and variables to be used the test was carried out. The results of it are given below.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.742[a]	1	.389		
Continuity Correction[b]	.334	1	.563		
Likelihood Ratio	.744	1	.388		
Fisher's Exact Test				.413	.282
Linear-by-Linear Association	.727	1	.394		
N of Valid Cases	50				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 11.52.

b. Computed only for a 2x2 table

Figure 22: Result table of the chi-square test for independence.

Result Summary:
The value of the test statistic is .742. No cell had an expected count less than 5, so the expected cell count assumption was met. There is one degree of freedom. The corresponding p-value of the test statistic (0.389) is greater than the significance level, so we fail to reject the null hypothesis and conclude that there is no association between a student's gender and them having any children.
Result Statement:
There was not a statistically significant association between a student's gender and them having any children ($\chi^2(1) = 0.742, p = 0.389$).

IV. Conclusion

In this project, we have demonstrated different statistical techniques and ensured that the proper assumptions were met for each them. The logistic regression model classified life expectancy into two categories with a high degree of accuracy (88.24%). The Kruskal-Wallis test found that there was a statistically significant difference between life expectancy across the four continents examined. The independent samples t-test found a significant difference in GPA between male and female students. Finally, the chi-squared test for independence found no association between a student's gender and them having any children or not.