National College of Ireland

STATISTICS FOR DATA ANALYTICS

Submitted by
Abdul Azadh Abdul Saleem - x18203621
MSc in Data Analytics – JAN/Batch B - 2020/21

PART A
BINARY LOGISTIC REGRESSION

Binary Logistic Regression is popularly used as a Classification Algorithm, categorized under the Supervised learning technique. This regression technique is used to predict the probability of the occurrence of the observation values into one of the two categories of the dichotomous Dependent variable (i.e. two dependent values).

OBJECTIVE:

This analysis is mainly focused on classifying and predicting whether the victims of road accidents are severely injured or not, in 176 countries around the globe.

DATASET:

This Road Accident dataset is obtained by merging six different datasets from World Health Organization’s official website, as various factors leading to accident. This dataset has information of 180 countries. Road Traffic death, Urban Speed Limit, Blood Alcohol Concentration and Road Safety Strategy are considered as the Predictor variables in this analysis and the Severity of the injury of the accident victim is considered as the Dependent variable.

Data pre-processing and transformation has been done with the help of R language. Datasets were downloaded in .csv format, which are then imported in R language. With the help of DPLYR package the values in the datasets were merged with inner join method, with Country column as key column. Data cleaning has been done by dropping the Null values with DROP() function. Since the raw data has the dependent variable to be continuous value, it is converted into categorical value ‘1’ and ‘0’, in which 1 denotes the accident victim is not severely injured and 0 denotes severely injured victim of the accident. This process dataset is saved in .csv format which the fed into IBM SPSS software for the purpose of analysis. Since the values of the categorical Independent variables Blood Alcohol Concentration (BAC) and Road Safety Strategy were in string format i.e. “Yes” or “No”, they have been transformed into ‘1’ and ‘2’ respectively for the analysis.

Link of data source:
Independent variable:

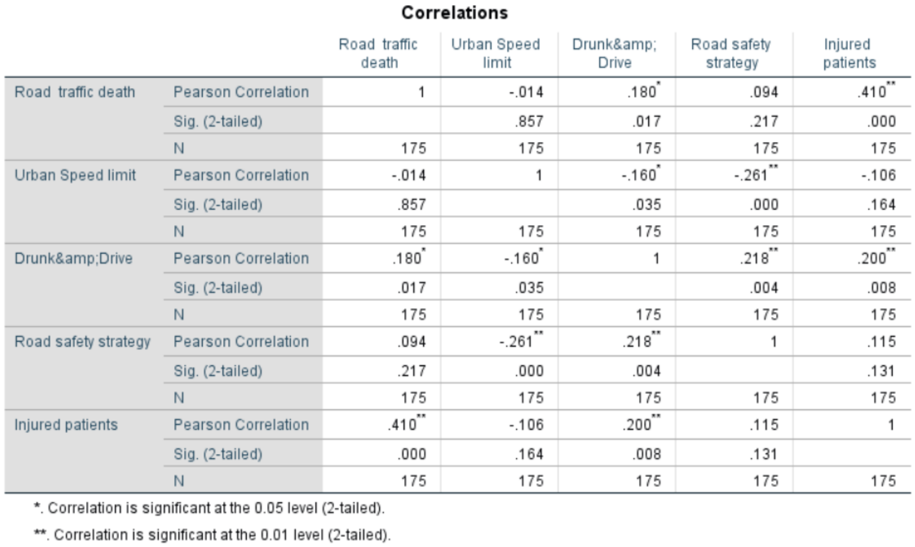
- Road Traffic death - https://apps.who.int/gho/data/node.main.A997?lang=en
- Urban Speed Limit - https://apps.who.int/gho/data/node.main.A1007?lang=en
- Blood Alcohol Concentration - https://apps.who.int/gho/data/node.main.A1002?lang=en
- Road Safety Strategy - https://apps.who.int/gho/data/node.main.A1012?lang=en

Dependent variable:

- Severity of Injury - https://apps.who.int/gho/data/node.main.A1019?lang=en

ASSUMPTIONS:

1. **Dependent variable is dichotomous –** This dependent variable is converted into categorical variable ‘0’ and ‘1’, which satisfies this assumption.
2. **Collinearity and Multi-Collinearity –** Collinearity and Multi-Collinearity has been checked with the help of Correlation matrix table provided below.



Correlations

		Road traffic death	Urban Speed limit	Drunk&amp; Drive	Road safety strategy	Injured patients
Road traffic death	Pearson Correlation	1	-.014	.180*	.094	.410**
	Sig. (2-tailed)		.857	.017	.217	.000
	N	175	175	175	175	175
Urban Speed limit	Pearson Correlation	-.014	1	-.160*	-.261**	-.106
	Sig. (2-tailed)	.857		.035	.000	.164
	N	175	175	175	175	175
Drunk&amp;Drive	Pearson Correlation	.180*	-.160*	1	.218**	.200**
	Sig. (2-tailed)	.017	.035		.004	.008
	N	175	175	175	175	175
Road safety strategy	Pearson Correlation	.094	-.261**	.218**	1	.115
	Sig. (2-tailed)	.217	.000	.004		.131
	N	175	175	175	175	175
Injured patients	Pearson Correlation	.410**	-.106	.200**	.115	1
	Sig. (2-tailed)	.000	.164	.008	.131	
	N	175	175	175	175	175

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

It has been observed that Dependent variable and Independent variables are less correlated, which satisfies the condition on Collinearity and Independent variables have Pearson Correlation less than 0.7 within themselves, this satisfies the condition of Multi-Collinearity.

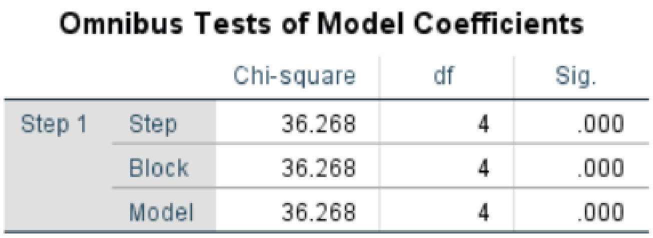
3. **Sample Size:** Logistic Regression needs large number of records with high amount of values to classify the Output. Taking 176 records in this analysis satisfies this assumption.
4. **Outliers:** This dataset has Outliers which influences less to classify the Dependent variable; hence this assumption is verified.
5. **Goodness-of-fit:** This dataset is analyzed and has found to be having Goodness-of-fit.

ANALYSIS OF LOGISTIC REGRESSION MODEL:

This analysis has been conducted in the IBM SPSS Statistics software. Here the variables RoadTrafficDeath, UrbanSpeedLimit, BAC and Roadsafetystrategy has been provided as the independent variable and SeverityofInjury has been provided as Dependent categorical variable. Then under the ‘Statistics and Plot’ category in the ‘options’ feature, the Classification plots, Hosmer-Lemeshow goodness-of-fit, Casewise listing of residuals has been chosen and CI for exp(B) is kept in 95%.

Omnibus Test:

This test has been conducted to check the performance of the model. Significance level here is observed as 0.00 which is less 0.05, shows that the variables are statistically significant. Result of Omnibus Test of Model Coefficients is provided below.

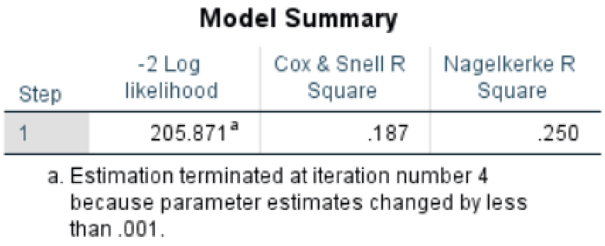


Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	36.268	4	.000
	Block	36.268	4	.000
	Model	36.268	4	.000

Model Summary:

This Model Summary table helps to find the value of distraction in dependent variable in predicting the output. This can be done with the help of Cox & Snell R Square and Nagelkarke R Square values, here the values are observed as 0.187 and 0.250 respectively. This output has been taken to prove that the predicted value has distraction somewhere between 18% to 25% from the actual value.



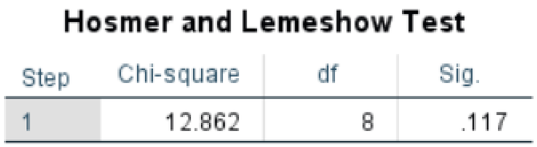
Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	205.871[a]	.187	.250

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test:

This Hosmer and Lemeshow table help to prove that the assumption of Goodness-of-fit has been satisfied. The Significance value observed in the analysis must be more than 0.05, which is 0.117 here. This explains the presence of correlation between predictor variable and dependent variable.

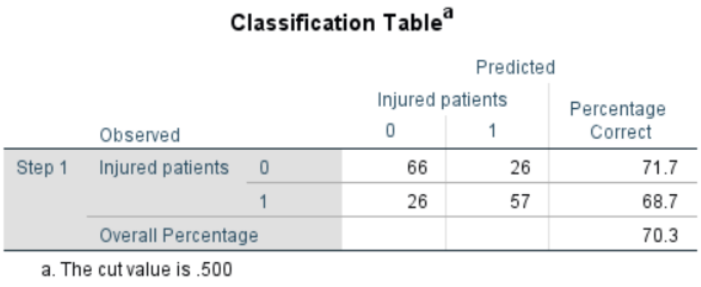


Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	12.862	8	.117

Classification Table:

This table shown in Classification table is Confusion Matrix, which is used to check for the accuracy of the output, which is here 70.3%. Model Specificity observed here is 71.7 and Model Sensitivity is 68.7.

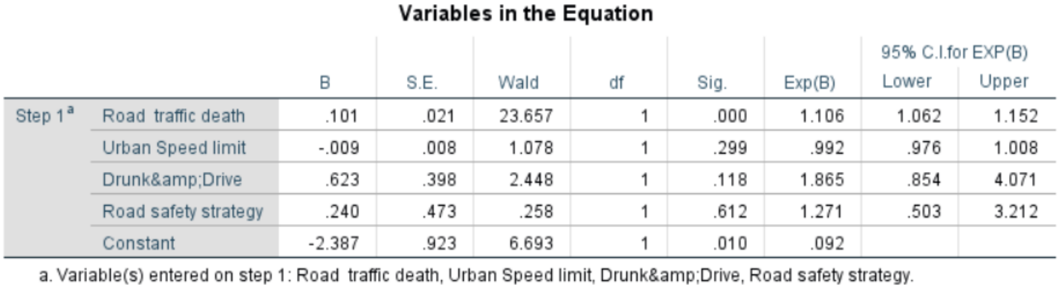


Classification Table[a]

	Observed		Predicted		
			Injured patients		Percentage Correct
			0	1	
Step 1	Injured patients	0	66	26	71.7
		1	26	57	68.7
	Overall Percentage				70.3

a. The cut value is .500

VARIABLES IN THE EQUATION:



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Road traffic death	.101	.021	23.657	1	.000	1.106	1.062	1.152
	Urban Speed limit	-.009	.008	1.078	1	.299	.992	.976	1.008
	Drunk&Drive	.623	.398	2.448	1	.118	1.865	.854	4.071
	Road safety strategy	.240	.473	.258	1	.612	1.271	.503	3.212
	Constant	-2.387	.923	6.693	1	.010	.092		

a. Variable(s) entered on step 1: Road traffic death, Urban Speed limit, Drunk&amp;Drive, Road safety strategy.

The B value in this table denotes the contribution of the independent variable in predicting the value of output variable. In this case, the per unit change in Predictor variable Drunk&Drive (BAC) will increase 0.623 log odds of Output variable.

log(p/1-p) = b0 + b1*x1 + b2*x2 + b3*x3 + b3*x3

Substituting the values of B in this equation, to derive the equation for this Logistic Regression as shown below.

log(p/1-p) = -2.387 + 0.101*RoadTrafficDeath -0.009*UrbanSpeedLimit + 0.623*BAC + 0.240*RoadSafetyStrategy

The independent variables UrbanSpeedlimit, Drunk&Drive and RoadsafetyStrategy has significance value more than 0.05 which specifies that these variables are less contributing for the prediction of output, while Road traffic Death high contributes for the prediction. Exp(B) is the exponential of the Coefficients, this provides us the odd’s ratio of the predictor. In this analysis, the odds of victim are drunk and getting severely injured is 1.865 higher than the opposite.

CONCLUSION:

This study has been conducted to analyze the Severity of road accidents of the victims from different countries. This resulted that this Binary logistic model could analyze with accuracy of 70.3% of this case across various countries.

PART B
TWO WAY ANOVA

Analysis of Variance between two groups, also known as ANOVA. Two-way ANOVA helps to identify the mean difference between the groups of two independent variable and the main purpose of this model is to understand the interaction between the dependent variable and independent variables.

DATASET DESCRIPTION:

The dataset utilized for this analysis has been extracted from the World Health Organization official site. The dataset used here is obtained by combining the data obtained three related dataset which has information from Global Information System on Alcohol and Health. This data has information collected from 195 countries around the globe about their National policy on Alcohol beverage, Awareness activities conducted on youth about alcohol consumption and percentage of youth under 15 to 19 years of age, addicted to the habit of Alcohol consumption.

Datasets were preprocessed and joined with the inner join method, keeping country as the primary key. In the resulting dataset obtained, further processing of removing the observations with missing values is done. Finally, the 156 observations were resulted on the above process, which is then further utilized for this analysis. For the process of Two-way ANOVA, the independent variable ‘Legal’ which represents the whether the country has legally defined National policy for Alcohol beverage. The presence and absence of this National policy is specified here as ‘Yes’ or ‘No’ which has been encoded into ‘1’ and ‘0’, respectively. Likewise, the other independent variable ‘awareness’ denotes the awareness conducted for the youth of the country about alcohol consumption, encoding them with ‘1’ and ‘0’ where ‘1’ denotes that the awareness activity has been conducted and ‘0’ denotes the no awareness activity has been conducted. The independent variable considered here are categorical variables, which has two groups each. On the other hand, dependent variable is a continuous variable which has data about the percentage of youth with alcohol addiction.

Source of the dataset has been provided below,
Independent variable:

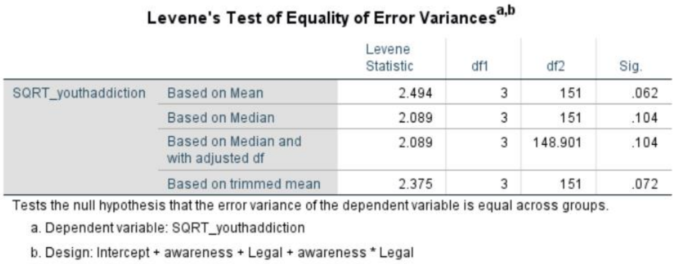
- Alcoholic beverage legally defined by country – https://apps.who.int/gho/data/node.main.A1123?lang=en
- Youth drinking awareness activities – https://apps.who.int/gho/data/node.main.A1198?lang=en

Dependent variable:

- 15-19 years old, current drinkers (%) – https://apps.who.int/gho/data/node.main.A1214?lang=en

VERIFICATION OF ASSUMPTION:

- Dependent variable considered in this analysis is Continuous variable.
- Two independent variables, Legal and awareness considered here has two categorical groups each.
- Independence of Observations is satisfied, there is no relationship between values in the different groups. This assumption has been checked with Independent sample T test.
- No Significant outliers, no influence of outliers on the output has been observed.
- Verification of normal distribution of dependent variable for every group in independent variable, has been conducted with Shapiro-Wilk test and resulted that significance value related from all the group observations are more than 0.05. In order to attain the normal distribution, the dependent variable has been transformed using Square root transformation and the resulting observation has been provided for Shapiro-Wilk test.
- Homogeneity of variances for each combination of two groups in independent variable, will help to identify whether our analysis is correct or not, in other words Goodness of fit. Levene's test has been used in order to check this assumption, whose result has been shown below in figure.



Levene's Test of Equality of Error Variances[a,b]

		Levene Statistic	df1	df2	Sig.
SQRT_youthaddiction	Based on Mean	2.494	3	151	.062
	Based on Median	2.089	3	151	.104
	Based on Median and with adjusted df	2.089	3	148.901	.104
	Based on trimmed mean	2.375	3	151	.072

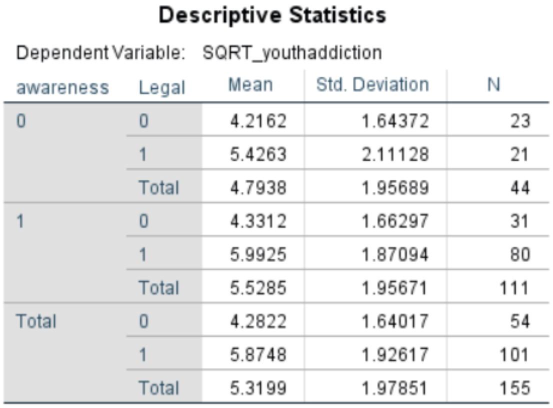
Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: SQRT_youthaddiction

b. Design: Intercept + awareness + Legal + awareness * Legal

Critical value here is F(3,151) = 2.494 and p-value is 0.62 which is greater than 0.05, shows that there is equal error variance of target variable across other factors. This shows that the assumption of Homogeneity of variance is satisfied.

INTERPRETATION OF RESULT:



Descriptive Statistics

Dependent Variable: SQRT_youthaddiction

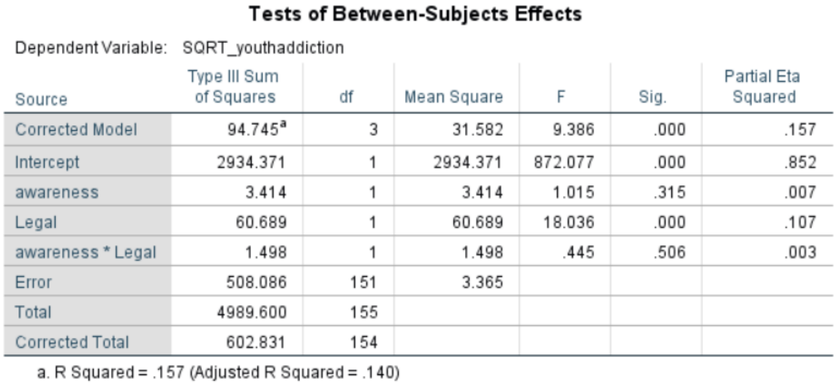
awareness	Legal	Mean	Std. Deviation	N
0	0	4.2162	1.64372	23
	1	5.4263	2.11128	21
	Total	4.7938	1.95689	44
1	0	4.3312	1.66297	31
	1	5.9925	1.87094	80
	Total	5.5285	1.95671	111
Total	0	4.2822	1.64017	54
	1	5.8748	1.92617	101
	Total	5.3199	1.97851	155

Descriptive Statistics:

From the Descriptive Statistics table, we could observe that the Mean of absence of awareness (5.43) against the presence of legal national policy for alcohol consumption is significantly influencing the percentage of youth addiction to alcohol in the country. Likewise, the mean of presence of awareness (5.99) against the presence of legal national policy has significant influence on the percentage of youth addiction to alcohol. And finally, the presence and absence of awareness among youth in the country along with Legal National policy for Alcohol consumption highly influences the percentage of youth alcohol consumers in the country, with mean value of 5.87.

Between-Subjects Effects Tests:

This table helps us to get the details about the significance effect of both the individual independent variable and their interaction on the target variable.



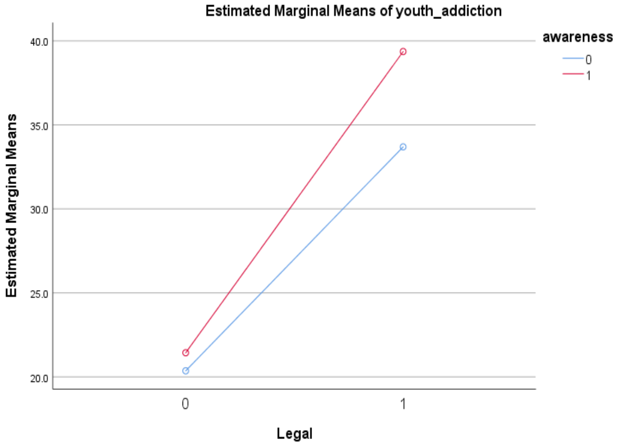
Tests of Between-Subjects Effects

Dependent Variable: SQRT_youthaddiction

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	94.745[a]	3	31.582	9.386	.000	.157
Intercept	2934.371	1	2934.371	872.077	.000	.852
awareness	3.414	1	3.414	1.015	.315	.007
Legal	60.689	1	60.689	18.036	.000	.107
awareness * Legal	1.498	1	1.498	.445	.506	.003
Error	508.086	151	3.365			
Total	4989.600	155				
Corrected Total	602.831	154				

a. R Squared = .157 (Adjusted R Squared = .140)

As per the result obtained in the above table, it is the Legal F(1,155) = 18.04 has the significant effect on the percentage of youth addiction to alcohol, with p-Value less than 0.05. Whereas, the other variable and the interaction of both the independent variables, does not have significant effect on influencing the dependent value. This ANOVA model can explain around 14% of variable on target value, which is also called as Accuracy.



This graph helps to get the information that the Legal National policy on alcohol and Awareness among the youth about the alcohol consumption does not have significant interaction to present the dependent value.

Since there only two groups in each of the independent variable, it is not possible to present the Multiple comparison table, which is Post Hoc (Tukey) test for this data.

CONCLUSION:

The two-way ANOVA was conducted to examine whether there is any effect of the Legal National policy for Alcohol beverage and Awareness activity conducted on youth for Alcohol consumption influences the percentage of youth consuming alcohol in the country. This analysis results that there is no significant interaction between these factors on the percentage of youth consuming alcohol F(1,155)= 0.445, p-Value= 0.506.

PART C

INDEPENDENT-SAMPLES T TEST

The main purpose of conducting Independent-Samples T Test is to identify if there is any significant difference between two unrelated group. In this method the mean of unrelated groups is compared on the continuous dependent variable.

OBJECTIVE:

The purpose of this analysis is to compare the mean height of the student of different ages. The age group of students considered for this test is less than 22 and 22-29.

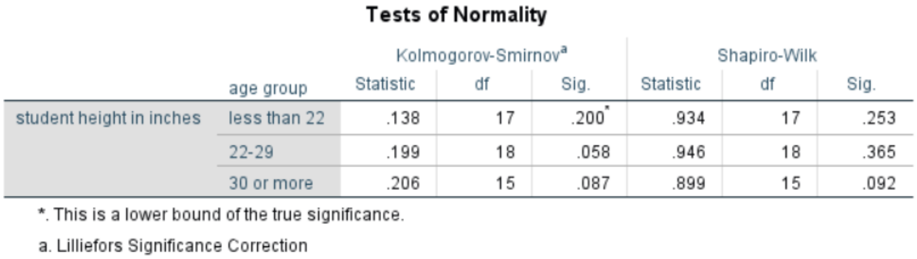
HYPOTHESIS:

Null Hypothesis (H_0) – The mean height of two different age group of students are equal (i.e. no statistical difference between the variables)

Alternative Hypothesis (H_1) – The mean height of two different group are unequal (i.e. there is statistical difference between the variables)

VERIFICATION OF ASSUMPTION:

- Dependent variable is measured on a Continuous scale.
- Independent variable has two categorical groups.
- Normality of dependent variable:



Tests of Normality

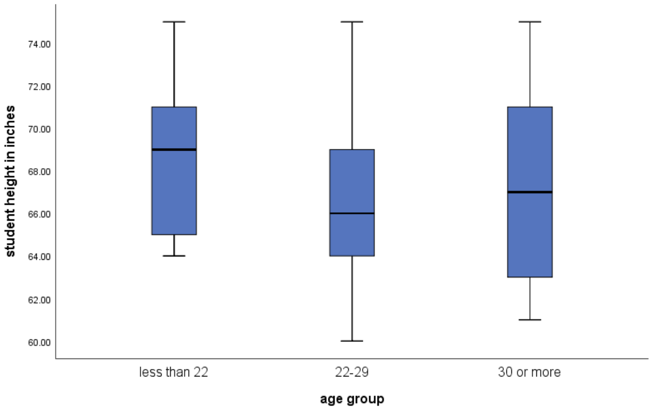
		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	age group	Statistic	df	Sig.	Statistic	df	Sig.
student height in inches	less than 22	.138	17	.200[*]	.934	17	.253
	22-29	.199	18	.058	.946	18	.365
	30 or more	.206	15	.087	.899	15	.092

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

This test of Normality of the dependent variable with respect to different groups is analyzed with the help of Shapiro-Wilk test. Here, the significance value we got for each group is more that 0.05, which results the presence of Normality in the dependent variable with respect to each individual group in independent variable.

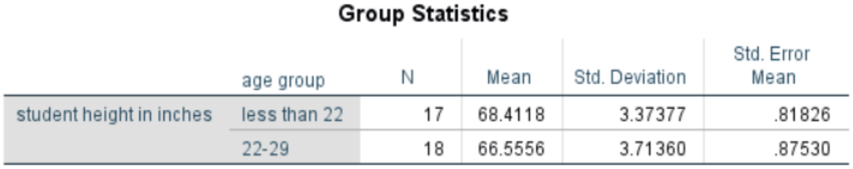
- No significant outliers, they dependent variable does not have any outlier that could affect the analysis.



INTERPRETATION OF RESULT:

After conducting the Independent-Samples T test for the variables considered for this analysis, the following two tables of Group Statistics and Independent sample T test were obtained which are studied below.

Group Statistics:



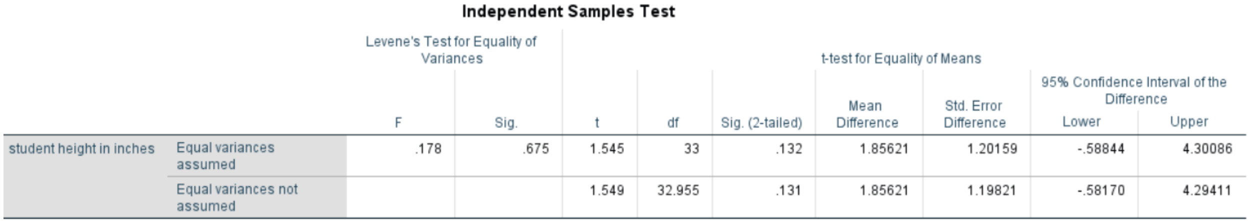
Group Statistics

	age group	N	Mean	Std. Deviation	Std. Error Mean
student height in inches	less than 22	17	68.4118	3.37377	.81826
	22-29	18	66.5556	3.71360	.87530

This table results that there is a slight difference between the sample mean of both individual age groups, where the mean value of group 'less than 22' age stands higher than the other age group '22-29'. And 'N' shows the count of observations considered from each group, which has a difference of 1.

Independent Samples Test:

The primary purpose of this table is to help us to verify whether our assumption is correct or not, with the help of Significance value obtained. In addition, to this result of Levene's test helps to verify the assumption of Homogeneity of variance.



Independent Samples Test

		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 height in inches	Equal variances assumed	.178	.675	1.545	33	.132	1.85621	1.20159	-.58844	4.30086
	Equal variances not assumed			1.549	32.955	.131	1.85621	1.19821	-.58170	4.29411

The result of Levene's test provided for two conditions, one for Equal variance assumed and other for Equal variance not assumed. The significance value obtained for this Levene's test is 0.675 which is more than 0.05, this shows that there is equal variance of error of dependent variable. This test also proves the Goodness-of-fit of the model, which is satisfied here.

The significance value of 2 tailed test is 0.132 which is greater than 0.05 alpha value, which infers that the **Null Hypothesis is not rejected**.

CONCLUSION:

The Independent Samples T test has been conducted on the data of student, to check whether there is any difference between the heights of students of two different age group considered in the analysis. And it has been observed that there is **no statistically significant difference** between the heights of two age group of students.

CHI-SQUARE TEST OF INDEPENDENCE

Chi-Square test of Independence, also called as Pearson's chi-square test, helps to identify the relationship between two variables which are Categorical.

OBJECTIVE:

This purpose of this analysis is to identify whether the age of the student has a relation with watching sports on television.

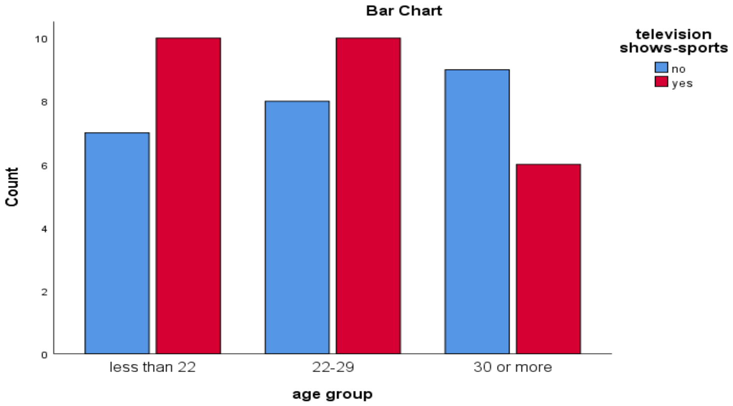
HYPOTHESIS:

Null Hypothesis (H_0) – The age of student does not have any relationship with watching sports on television.

Alternate Hypothesis (H_1) – The age of the students has effect on watching sports on television.

VERIFICATION OF ASSUMPTIONS:

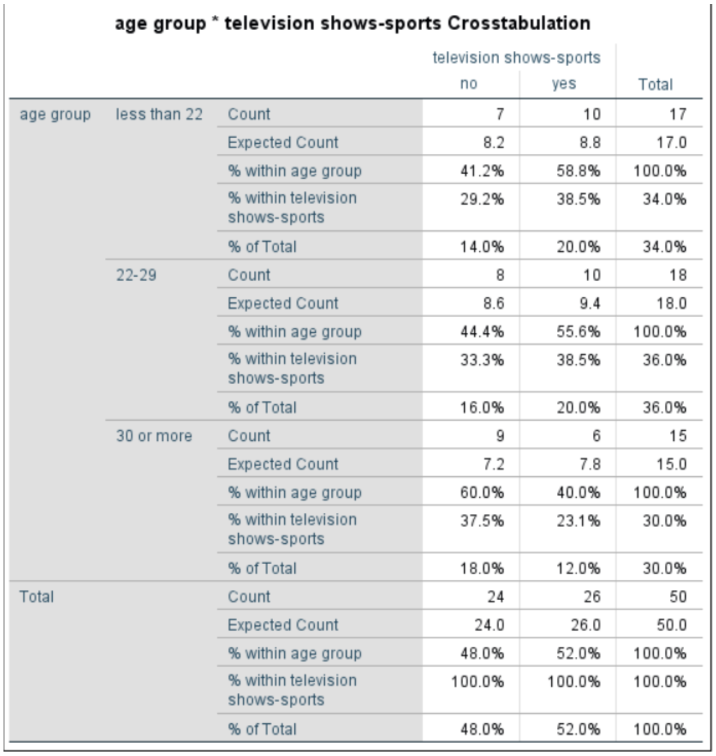
- Variables considered in this analysis are measured in Ordinal or Nominal level.
- Variables has two or more independent groups, individually.
- The Bar chart shown below, helps to visually read the data. It provides information about one categorical with respect to the frequency of occurrences of the other categorical variable.



INTERPRETATION OF THE RESULT:

Cross Tabulation:

Cross tabulation stands as the important aspect of this interpretation. This table has the information like the observer count and expected count of each categorical variable on each categories of other variable. Cross tabulation table for this analysis is shown below.



age group * television shows-sports Crosstabulation

			television shows-sports		
			no	yes	Total
age group	less than 22	Count	7	10	17
		Expected Count	8.2	8.8	17.0
		% within age group	41.2%	58.8%	100.0%
		% within television shows-sports	29.2%	38.5%	34.0%
		% of Total	14.0%	20.0%	34.0%
	22-29	Count	8	10	18
		Expected Count	8.6	9.4	18.0
		% within age group	44.4%	55.6%	100.0%
		% within television shows-sports	33.3%	38.5%	36.0%
		% of Total	16.0%	20.0%	36.0%
	30 or more	Count	9	6	15
		Expected Count	7.2	7.8	15.0
		% within age group	60.0%	40.0%	100.0%
		% within television shows-sports	37.5%	23.1%	30.0%
		% of Total	18.0%	12.0%	30.0%
Total		Count	24	26	50
		Expected Count	24.0	26.0	50.0
		% within age group	48.0%	52.0%	100.0%
		% within television shows-sports	100.0%	100.0%	100.0%
		% of Total	48.0%	52.0%	100.0%

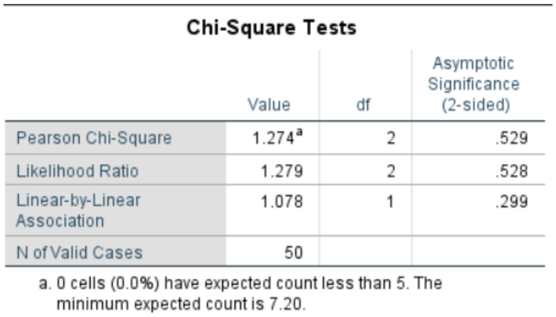
- Most of the student who has their age less than 22 and 22-29, watch sports on television.

- While the count of student who has the age of 30 or more do not watch sports on television and higher than the count of students of same age watching sports on television.

- Overall, the observed count of people of all age, watching sports on television is slightly higher than count of people of all age who do not watch sports on television. And Expected count obtained is also equal to the observed count, altogether.

Pearson's Chi-Square test:

The result obtained from Pearson's Chi-square test helps to verify the assumption of this analysis. The Pearson's chi square significance value obtained in this analysis is more that 0.05 alpha values which denotes that the Null hypothesis assumed in this analysis is not rejected, this value is shown in the figure below.



Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	1.274[a]	2	.529
Likelihood Ratio	1.279	2	.528
Linear-by-Linear Association	1.078	1	.299
N of Valid Cases	50		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 7.20.

The Pearson's Chi-square test has provided us the value of ***$\chi^2(1) = 1.274$***, with the significance value of p-Value = 0.529 which is greater than 0.05, signifies that the Null hypothesis assumption is correct.

CONCLUSION:

From this study conducted to find the relationship between the age groups of students and watching sports on television, it is concluded that the age of students **does not have any effect** on watching sports on television.