DATA ANALYSIS PROJECT

Data analysis is the process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusions and supporting decision-making.
In this SPSS project, we'll be inspecting different research questions and as well as ways to explore and analyze them.

SCOPE OF THE PROJECT

1. TREATING MISSING VALUES : Exploring the missing data and finding out means to treat them.
2. FERTILIZER ANALYSIS : Investigating the effects of various fertilizers on parsley plant to know which fertilizer is the best and if there's any significant difference between them.
3. DEPRESSION MEDICATION ANALYSIS : Investigating whether various medicines listed for depression is actually significant in mean Beck's Depression Inventory.
4. HEALTH CARE COST ANALYSIS : Surveying how patient attribute such as sex, age, drinking, smoking and exercise can predict health care cost.
5. PSYCHOLOGICAL TEST ON CHILDREN : Running a correlation analysis on psychological test on 128 children ranging from 12 – 14 years old to determine the relationship between them.

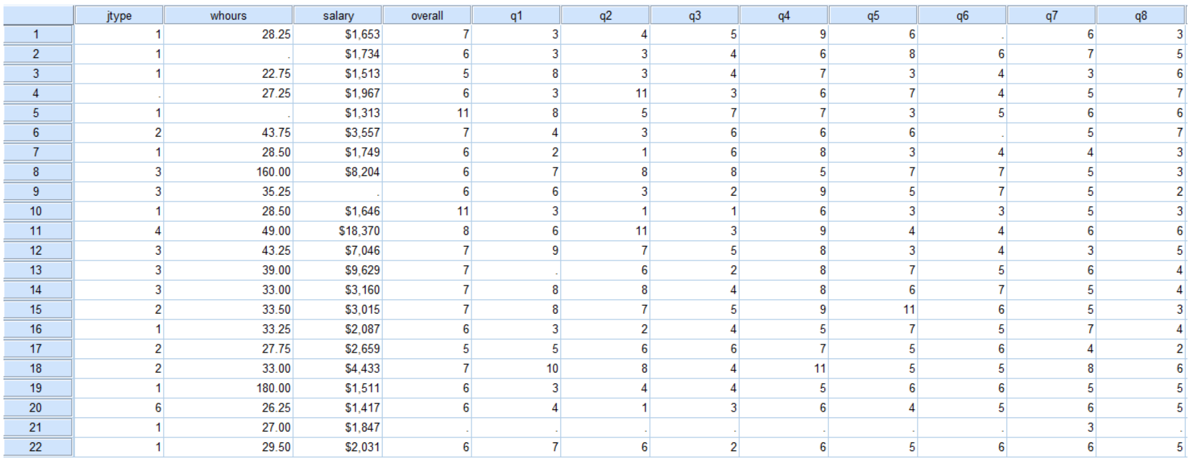
OBJECTIVES

To show/address particular research questions using appropriate statistical techniques and interpret results of different analysis and each meaningful conclusions and also to show realistic applications and practical suggestion of the findings and to reflect on the learning skills/expertise.

IMPORTANCE OF THE PROJECT

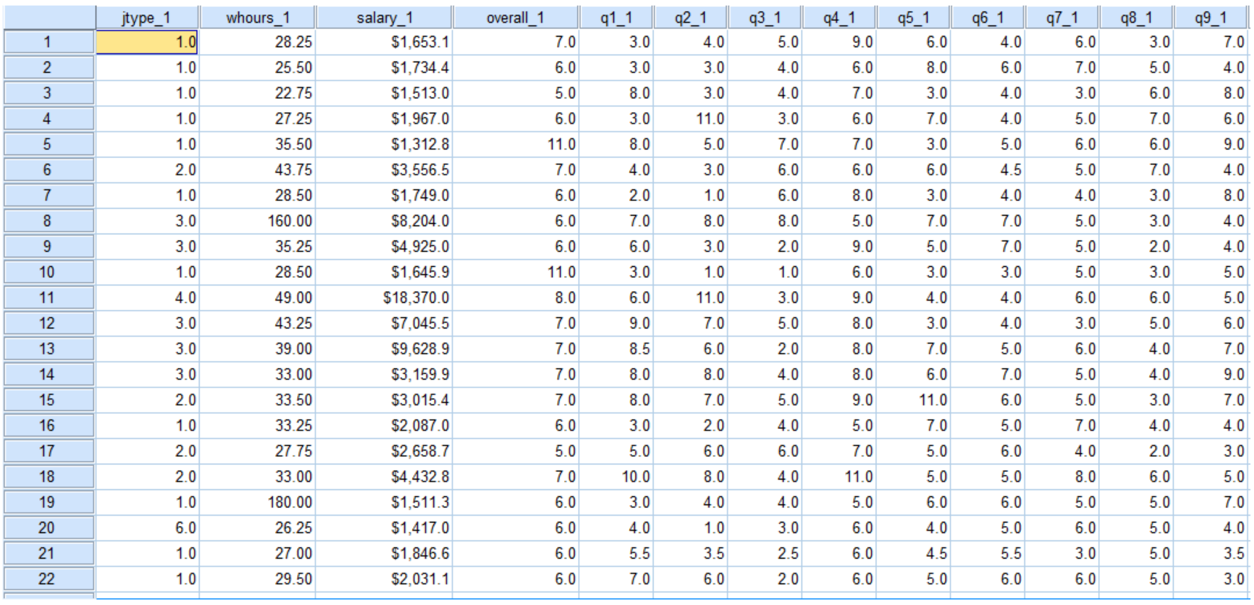
This project emphasizes on the significance of statistical analysis in decision making, problem solving and policy formulation across different domains. It also emphasizes on the importance of data quality, handling missing values and understanding data distribution to extract meaningful insights.

Q1.Treating Missing values : Here, we treated the missing values through the 'Missing value' option using Linear Interpolation method to complete our dataset and go ahead with our analysis.



	jtype	whours	salary	overall	q1	q2	q3	q4	q5	q6	q7	q8
1	1	28.25	$1,653	7	3	4	5	9	6	.	6	3
2	1	.	$1,734	6	3	3	4	6	8	6	7	5
3	1	22.75	$1,513	5	8	3	4	7	3	4	3	6
4	.	27.25	$1,967	6	3	11	3	6	7	4	5	7
5	1	.	$1,313	11	8	5	7	7	3	5	6	6
6	2	43.75	$3,557	7	4	3	6	6	6	.	5	7
7	1	28.50	$1,749	6	2	1	6	8	3	4	4	3
8	3	160.00	$8,204	6	7	8	8	5	7	7	5	3
9	3	35.25	.	6	6	3	2	9	5	7	5	2
10	1	28.50	$1,646	11	3	1	1	6	3	3	5	3
11	4	49.00	$18,370	8	6	11	3	9	4	4	6	6
12	3	43.25	$7,046	7	9	7	5	8	3	4	3	5
13	3	39.00	$9,629	7	.	6	2	8	7	5	6	4
14	3	33.00	$3,160	7	8	8	4	8	6	7	5	4
15	2	33.50	$3,015	7	8	7	5	9	11	6	5	3
16	1	33.25	$2,087	6	3	2	4	5	7	5	7	4
17	2	27.75	$2,659	5	5	6	6	7	5	6	4	2
18	2	33.00	$4,433	7	10	8	4	11	5	5	8	6
19	1	180.00	$1,511	6	3	4	4	5	6	6	5	5
20	6	26.25	$1,417	6	4	1	3	6	4	5	6	5
21	1	27.00	$1,847	.	.	.	.	.	.	.	3	.
22	1	29.50	$2,031	6	7	6	2	6	5	6	6	5

Here, if you look closely at our dataset you'll see that there are missing values in our dataset and we'll go ahead to fill the missing values.



	jtype_1	whours_1	salary_1	overall_1	q1_1	q2_1	q3_1	q4_1	q5_1	q6_1	q7_1	q8_1	q9_1
1	1.0	28.25	$1,653.1	7.0	3.0	4.0	5.0	9.0	6.0	4.0	6.0	3.0	7.0
2	1.0	25.50	$1,734.4	6.0	3.0	3.0	4.0	6.0	8.0	6.0	7.0	5.0	4.0
3	1.0	22.75	$1,513.0	5.0	8.0	3.0	4.0	7.0	3.0	4.0	3.0	6.0	8.0
4	1.0	27.25	$1,967.0	6.0	3.0	11.0	3.0	6.0	7.0	4.0	5.0	7.0	6.0
5	1.0	35.50	$1,312.8	11.0	8.0	5.0	7.0	7.0	3.0	5.0	6.0	6.0	9.0
6	2.0	43.75	$3,556.5	7.0	4.0	3.0	6.0	6.0	6.0	4.5	5.0	7.0	4.0
7	1.0	28.50	$1,749.0	6.0	2.0	1.0	6.0	8.0	3.0	4.0	4.0	3.0	8.0
8	3.0	160.00	$8,204.0	6.0	7.0	8.0	8.0	5.0	7.0	7.0	5.0	3.0	4.0
9	3.0	35.25	$4,925.0	6.0	6.0	3.0	2.0	9.0	5.0	7.0	5.0	2.0	4.0
10	1.0	28.50	$1,645.9	11.0	3.0	1.0	1.0	6.0	3.0	3.0	5.0	3.0	5.0
11	4.0	49.00	$18,370.0	8.0	6.0	11.0	3.0	9.0	4.0	4.0	6.0	6.0	5.0
12	3.0	43.25	$7,045.5	7.0	9.0	7.0	5.0	8.0	3.0	4.0	3.0	5.0	6.0
13	3.0	39.00	$9,628.9	7.0	8.5	6.0	2.0	8.0	7.0	5.0	6.0	4.0	7.0
14	3.0	33.00	$3,159.9	7.0	8.0	8.0	4.0	8.0	6.0	7.0	5.0	4.0	9.0
15	2.0	33.50	$3,015.4	7.0	8.0	7.0	5.0	9.0	11.0	6.0	5.0	3.0	7.0
16	1.0	33.25	$2,087.0	6.0	3.0	2.0	4.0	5.0	7.0	5.0	7.0	4.0	4.0
17	2.0	27.75	$2,658.7	5.0	5.0	6.0	6.0	7.0	5.0	6.0	4.0	2.0	3.0
18	2.0	33.00	$4,432.8	7.0	10.0	8.0	4.0	11.0	5.0	5.0	8.0	6.0	5.0
19	1.0	180.00	$1,511.3	6.0	3.0	4.0	4.0	5.0	6.0	6.0	5.0	5.0	7.0
20	6.0	26.25	$1,417.0	6.0	4.0	1.0	3.0	6.0	4.0	5.0	6.0	5.0	4.0
21	1.0	27.00	$1,846.6	6.0	5.5	3.5	2.5	6.0	4.5	5.5	3.0	5.0	3.5
22	1.0	29.50	$2,031.1	6.0	7.0	6.0	2.0	6.0	5.0	6.0	6.0	5.0	3.0

Now, we've filled in the missing values just with the same method. Fixing missing data was vital because it could make our results wrong and mess up our conclusions. It keeps our work strong and our project high quality.

Q2 . A farmer wants to know which fertilizer is best for his parsley plants. So he tries different fertilizers on different plants and weighs these plants after 6 weeks.

Firstly, we check the descriptive statistics by comparing the means of each fertilizers separately.

Report

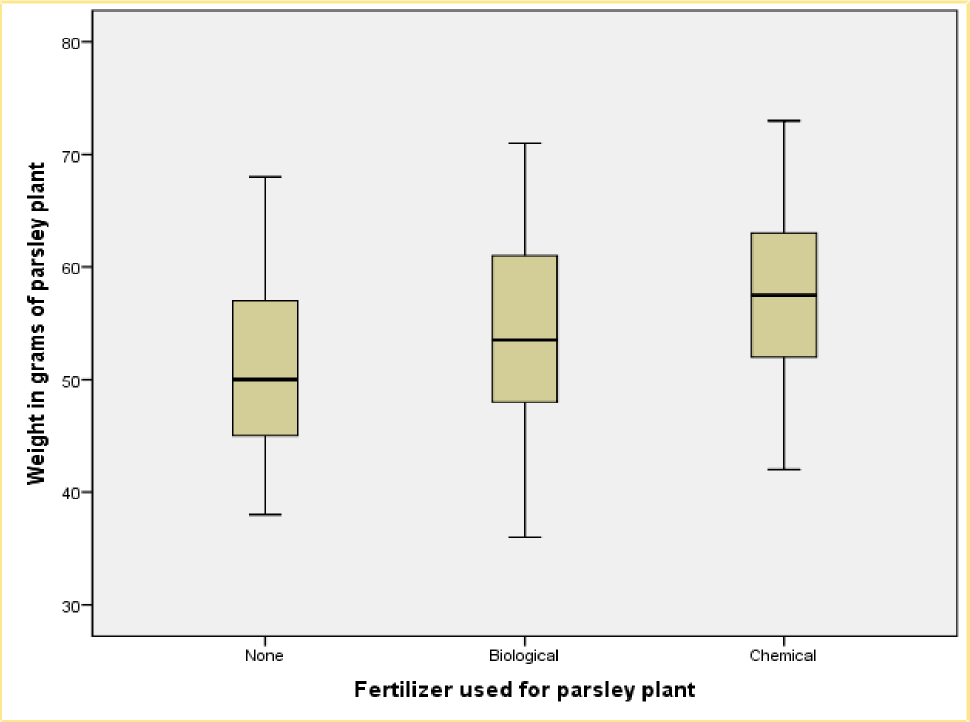
Weight in grams of parsley plant

Fertilizer used for parsley plant	Mean	N	Std. Deviation
None	51.20	30	7.836
Biological	53.63	30	8.857
Chemical	56.97	30	7.854
Total	53.93	90	8.445

In this table above, we have 30 samples of fertilizers each. The chemical fertilizer has the highest mean weight of 56.97 which is almost 57 grams and a standard deviation of 7.854, the biological has a mean of 53.63 and a standard deviation of 8.857(has the highest standard deviation because of the spread or dispersion of data points within the dataset). The None has the lowest mean which is 51.20 and a standard deviation of 7.836.

Next, we explore our data using histogram

Weight in grams of parsley plant



From the boxplots, we see that there no outliers and the distribution are roughly symmetric and the center of the distributions don’t appear to be hugely different. The median weight for the None fertilizer used is slightly lower than the median weight of the biological and chemical. Also the chemical weight tend to be slightly higher than the None and biological.

Before we run our analysis we must make sure our data meet the following assumptions;

- Dependent variable must be continuous
- Independent variable must be categorical
- Independent observations
- Random sample of data
- Data must be normally distributed
- Homogeneity of variance
- No outliers

This data met the assumptions of the one-way ANOVA test and the hypothesis states that;

Null hypothesis: All population means are equal

Alternate hypothesis: At least one of the population mean is not equal to the others.

Now let’s run our analysis,

ANOVA

Weight in grams of parsley plant

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	502.867	2	251.433	3.743	.028
Within Groups	5844.733	87	67.181		
Total	6347.600	89			

Above is the result of our analysis, the data was subjected to a One way ANOVA test at 95% confidence interval with all assumptions being met. The test found that p-value is 0.028 which is less than 0.05 level of significance, so we reject the null hypothesis that all population means are equal.
Again, we concluded that the weight of the plants is significantly different for at least one of the fertilizer group(p<0.028).

Q3. Very basically, 100 participants suffering from depression were divided into 4 groups of 25 each. Each group was given a different medicine. After 4 weeks, participants filled out the BDI, short for Beck's depression inventory. Our main research question is; Did our different medicines result in different mean BDI scores? A secondary question is whether the BDI scores are related to gender in any way.

Here, we want to check the relationship between BDI and medicine.

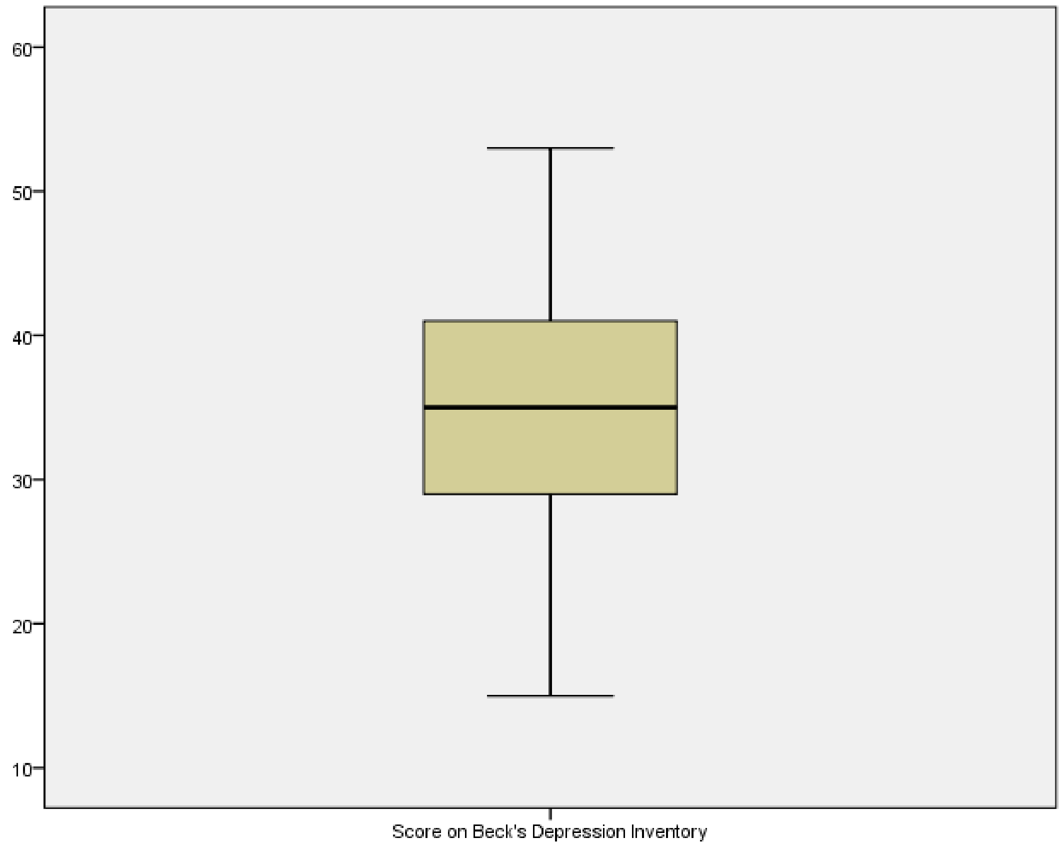
Firstly, we check the descriptive statistics of the various medicine which is shown below;

Descriptives

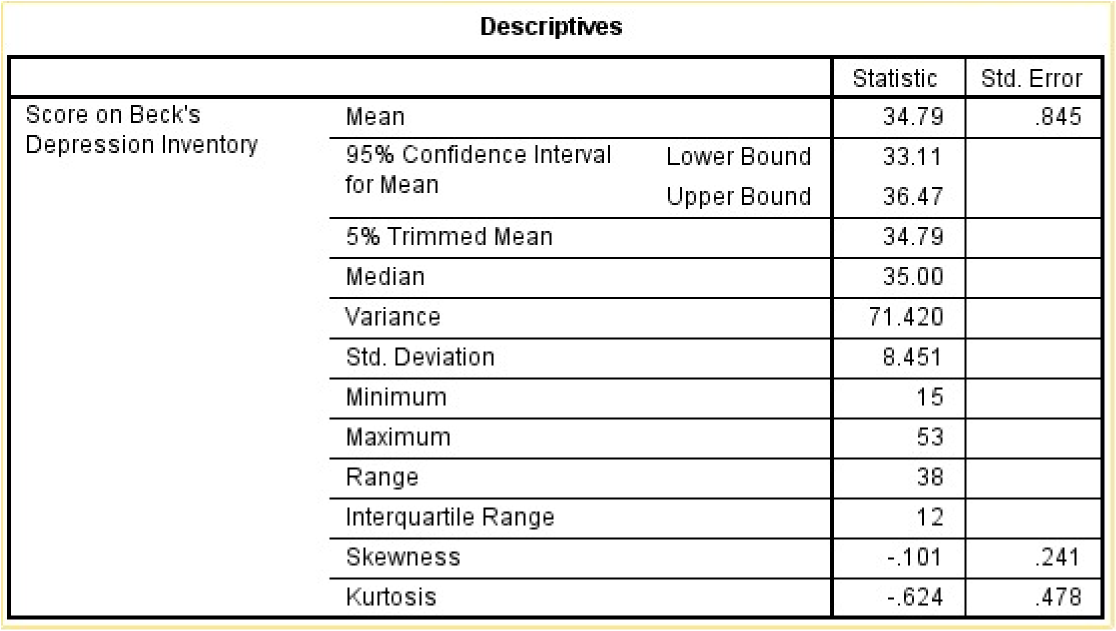
Score on Beck's Depression Inventory

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
None	25	41.72	6.328	1.266	39.11	44.33	27	51
Placebo	25	35.04	7.056	1.411	32.13	37.95	25	53
Homeopathic	25	35.52	6.634	1.327	32.78	38.26	19	47
Pharmaceutical	25	26.88	6.704	1.341	24.11	29.65	15	43
Total	100	34.79	8.451	.845	33.11	36.47	15	53

From the above table, None has a mean of 41.72 which is the highest mean and a standard deviation of 6.328, Placebo has a mean of 35.04 and a standard deviation of 7.056 which is the highest (has more spread of data), Homeopathic has a mean of 35.52 and standard deviation of 6.634 and lastly, pharmaceutical has a mean of 26.88 which seems to be the lowest and standard deviation of 6.704.



From the boxplot above, the data looks slightly negatively skewed and has no outlier.



Descriptives

			Statistic	Std. Error
Score on Beck's Depression Inventory	Mean		34.79	.845
	95% Confidence Interval for Mean	Lower Bound	33.11	
		Upper Bound	36.47	
	5% Trimmed Mean		34.79	
	Median		35.00	
	Variance		71.420	
	Std. Deviation		8.451	
	Minimum		15	
	Maximum		53	
	Range		38	
	Interquartile Range		12	
	Skewness		-.101	.241
	Kurtosis		-.624	.478

Looking at the descriptive table above, the data has a skewness of -.101 which indicates that the distribution of data is slightly negatively skewed. In a negatively skewed distribution, the tail on the left side (the lower values) is longer or more stretched out than the tail on the right side (the higher value). The kurtosis has a value of -0.624 which indicates that the distribution of data has negative kurtosis. Negative kurtosis means that the distribution has higher tails and is less peaked or leptokurtic.

Now let’s run our analysis,

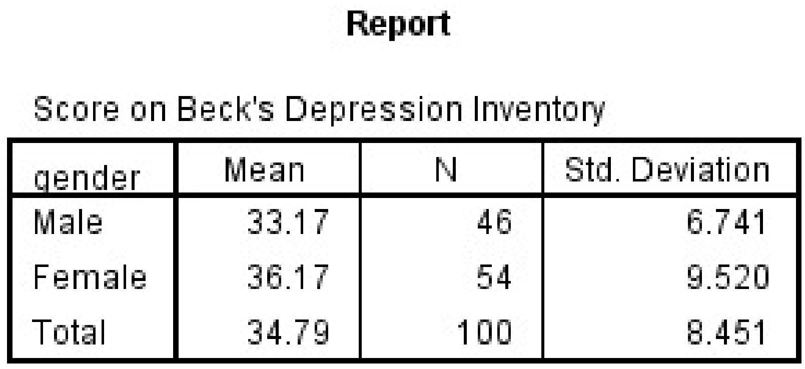
ANOVA

Score on Beck's Depression Inventory

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2779.710	3	926.570	20.730	.000
Within Groups	4290.880	96	44.697		
Total	7070.590	99			

Here, the data was subjected to a one-way ANOVA test at 95% confidence interval with all assumptions being met to know the relationship between BDI and the various medicine. The test found that p-value is 0.000 which is less than 0.05 level of significance showing that there’s a statistical relationship between BDI and one of the medicines.

Now let’s check the relationship between BDI and gender,



Report

Score on Beck's Depression Inventory

gender	Mean	N	Std. Deviation
Male	33.17	46	6.741
Female	36.17	54	9.520
Total	34.79	100	8.451

From the table above, the male has a mean of 33.17 which is the lowest and a standard deviation of 6.741 while the female has a mean of 36.17 which is the highest and a standard deviation of 9.520 due to the large spread of data.

Let’s analyze our data and check the relationship between gender and BDI.

ANOVA

Score on Beck's Depression Inventory

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	222.481	1	222.481	3.184	.077
Within Groups	6848.109	98	69.879		
Total	7070.590	99			

The test found that p-value is 0.077 which is greater than 0.05 level of significance showing that there’s no statistical relationship between BDI and any of the gender.

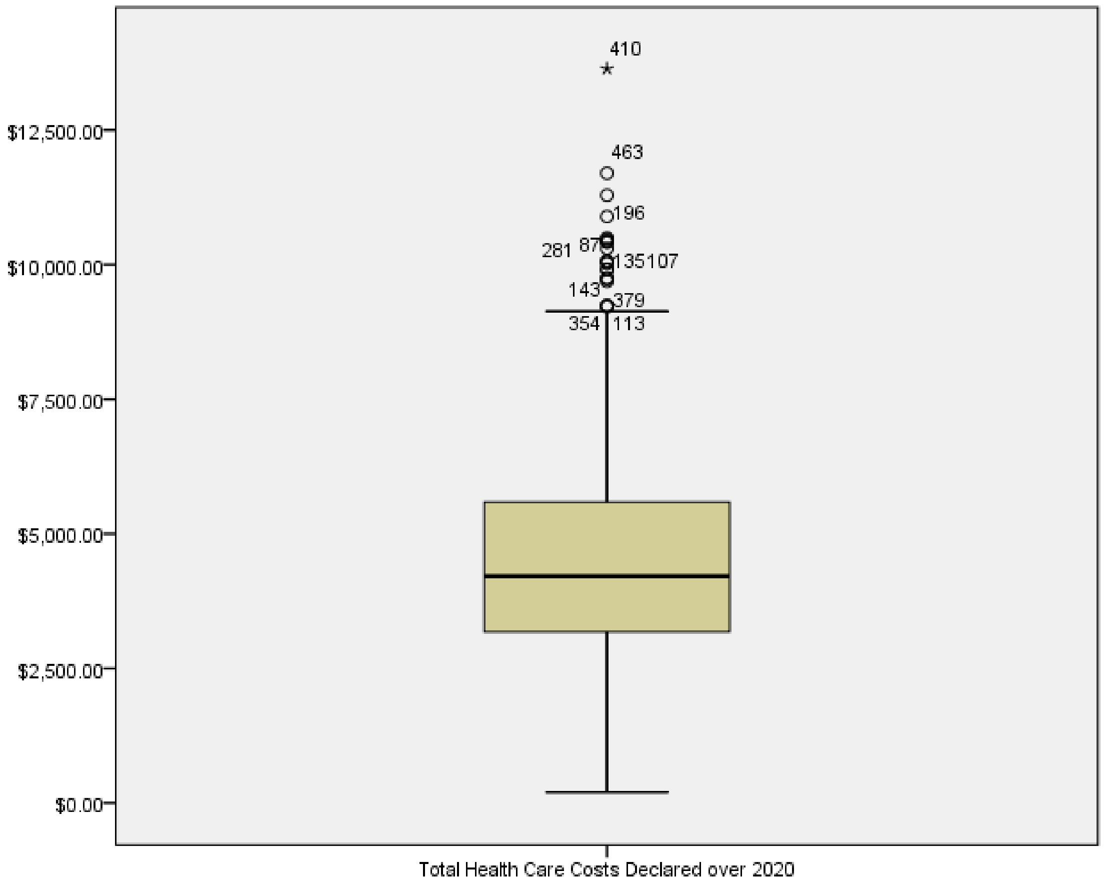
Q4. A scientist wants to know if and how health care costs can be predicted from several patient characteristics which are sex, age, drinking, smoking and exercise.

Here, the dependent variable is health care costs (In US dollars). Then the independent variables are sex, age, drinking, smoking and exercise.

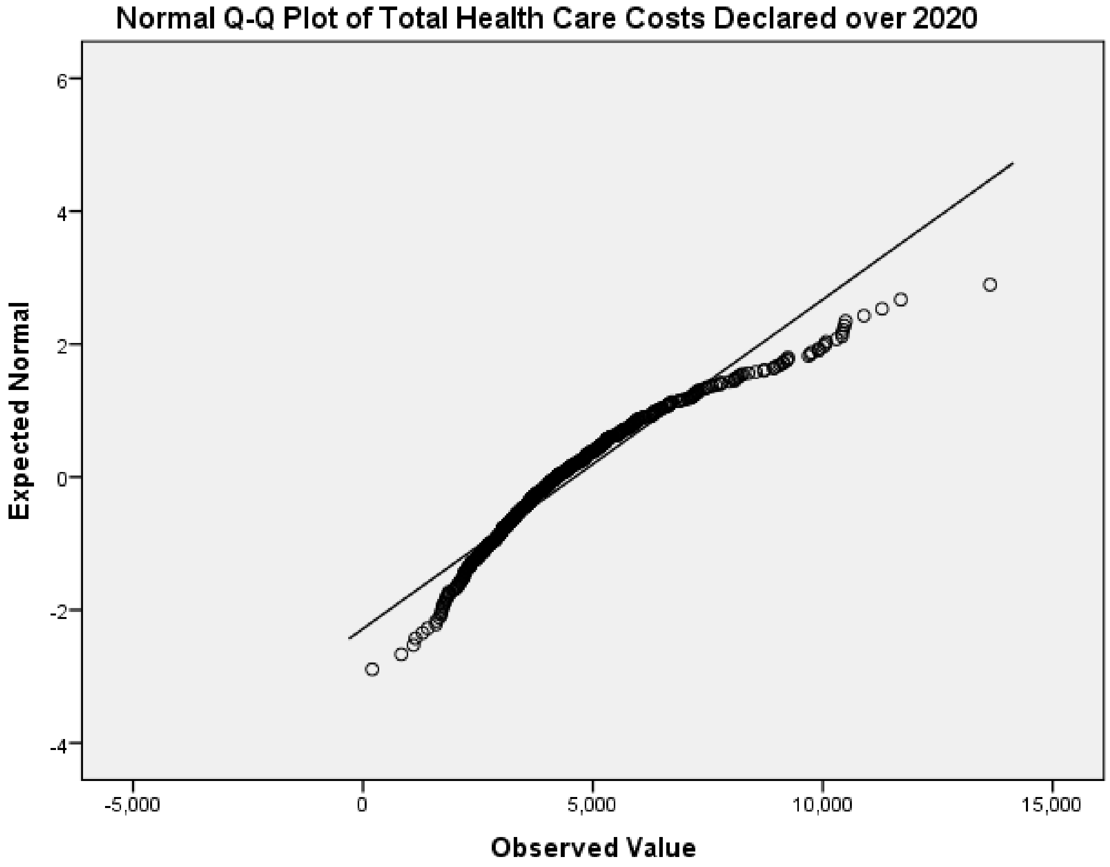
DATA CHECKS FOR REGRESSION

- The dependent variable is quantitative
- The independent observation is categorical
- There are enough sample sizes

Firstly, we have to explore our data to check for outliers and to know if the frequency distribution looks plausible.



From the box plot above, we can see that the data is roughly symmetric with outliers (the tail on the right is longer or stretched than the tail on the left).



From the Q-Q plot above, it shows that our data looks slightly non-normal because the data didn't really fall greatly on the line and has outliers and they are legitimate data points and they represent genuine extreme values. Therefore, transforming them or removing them may not be appropriate unless they are outrightly skewed.

Now let's run a multiple Linear Regression to analyze our data;

MULTIPLE REGRESSION ASSUMPTIONS;

- Independent observations
- Homoscedasticity
- Linearity

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.629[a]	.395	.390	$1,577.10344	.395	67.889	5	519	.000

a. Predictors: (Constant), Average Consumption of Alcoholic Beverages per Week, Sex, Age at Survey Completion (Years), Average Hours of Exercise per Week, Average Consumption of Cigarettes per Day

- The R shows the multiple correlation coefficients. It shows the pearson correlation between the actual scores and those predicted by our regression model. R= 0.629 shows a positive strong correlation.

- R-square shows the goodness of the fit which indicates that the models explains 39.5% of the variation in the dependent variable.
- Adjusted R square takes into account the number of independent variables in the model. An adjusted R-squared value of 0.390 indicates that approximately 39.0% of the variability in the dependent variable is explained by the independent variables in a regression model.
-

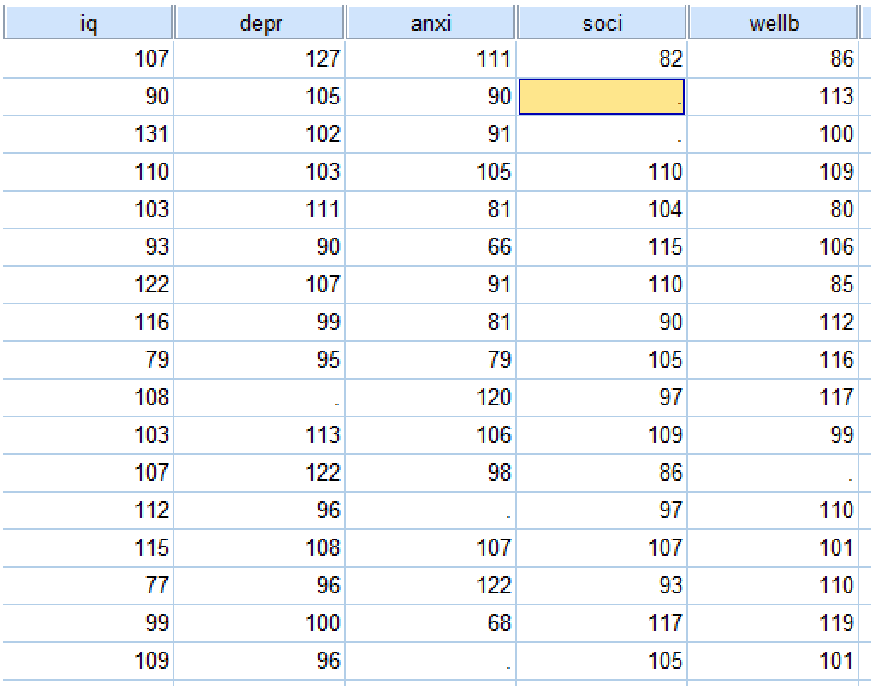
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-3263.586	1059.163		-3.081	.002	-5344.360	-1182.812
	Sex	509.271	139.565	.126	3.649	.000	235.089	783.454
	Age at Survey Completion (Years)	114.658	12.452	.322	9.208	.000	90.196	139.121
	Average Hours of Exercise per Week	-271.270	36.300	-.281	-7.473	.000	-342.584	-199.956
	Average Consumption of Cigarettes per Day	139.414	17.384	.311	8.020	.000	105.263	173.565
	Average Consumption of Alcoholic Beverages per Week	50.386	10.275	.192	4.904	.000	30.201	70.572

a. Dependent Variable: Total Health Care Costs Declared over 2020

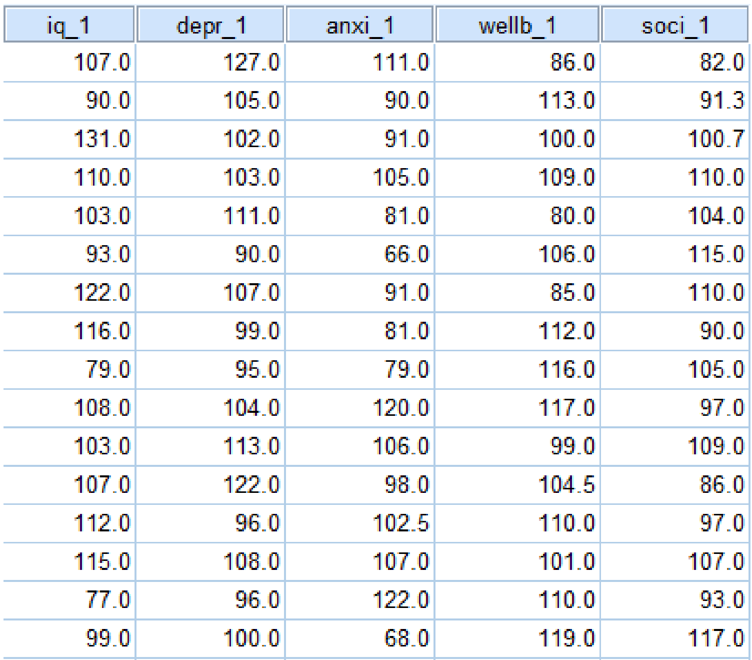
B-coefficient(unstandardized coefficient) : Each b-coefficient indicates the average increase in costs associated with a 1-unit increase in a predictor, that is to say that 1-year increase in age results in $114.7 increase in costs and an hour increase in exercise per week results with a -$271.3 increase in yearly health costs. Also, a 1-unit increase in sex is associated with an average $509.3 increase in costs.

Q5. psychological test data on 128 children between 12 and 14 years old. Run a correlation analysis on the variables.



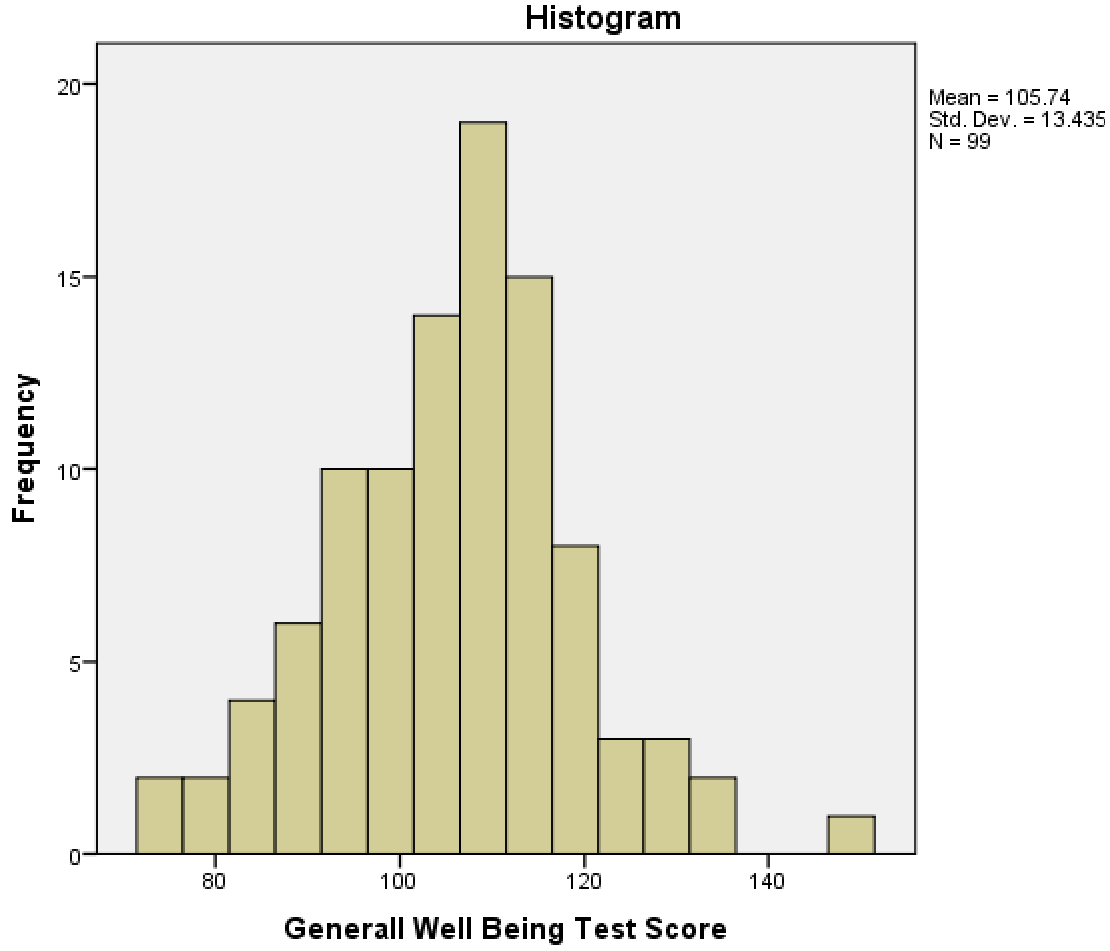
iq	depr	anxi	soci	wellb
107	127	111	82	86
90	105	90	.	113
131	102	91	.	100
110	103	105	110	109
103	111	81	104	80
93	90	66	115	106
122	107	91	110	85
116	99	81	90	112
79	95	79	105	116
108	.	120	97	117
103	113	106	109	99
107	122	98	86	.
112	96	.	97	110
115	108	107	107	101
77	96	122	93	110
99	100	68	117	119
109	96	.	105	101

From the above dataset, we noticed we have some missing values and our data won't look reasonable without treating the missing values. So we start by treating the missing values.



iq_1	depr_1	anxi_1	wellb_1	soci_1
107.0	127.0	111.0	86.0	82.0
90.0	105.0	90.0	113.0	91.3
131.0	102.0	91.0	100.0	100.7
110.0	103.0	105.0	109.0	110.0
103.0	111.0	81.0	80.0	104.0
93.0	90.0	66.0	106.0	115.0
122.0	107.0	91.0	85.0	110.0
116.0	99.0	81.0	112.0	90.0
79.0	95.0	79.0	116.0	105.0
108.0	104.0	120.0	117.0	97.0
103.0	113.0	106.0	99.0	109.0
107.0	122.0	98.0	104.5	86.0
112.0	96.0	102.5	110.0	97.0
115.0	108.0	107.0	101.0	107.0
77.0	96.0	122.0	110.0	93.0
99.0	100.0	68.0	119.0	117.0

From the dataset above it shows that the missing value has been treated so we'll run a histogram plot to check if our data looks plausible.



From our histogram above it shows that our data looks plausible/symmetric and has outliers and they are legitimate data points and the represent genuine extreme values.

Now let's run a pearson correlation analysis;

DATA CHECKS FOR PEARSON CORRELATION

- Two or more continuos variable
- Bivariate normality
- Random sample of data from the population

Correlations

		LINT(iq)	LINT(depr)	LINT(anxi)	LINT(wellb)	LINT(soci)
LINT(iq)	Pearson Correlation	1	.143	.126	-.079	-.150
	Sig. (2-tailed)		.108	.155	.373	.092
	N	128	128	128	128	128
LINT(depr)	Pearson Correlation	.143	1	.274**	-.781**	-.318**
	Sig. (2-tailed)	.108		.002	.000	.000
	N	128	128	128	128	128
LINT(anxi)	Pearson Correlation	.126	.274**	1	-.268**	-.454**
	Sig. (2-tailed)	.155	.002		.002	.000
	N	128	128	128	128	128
LINT(wellb)	Pearson Correlation	-.079	-.781**	-.268**	1	.332**
	Sig. (2-tailed)	.373	.000	.002		.000
	N	128	128	128	128	128
LINT(soci)	Pearson Correlation	-.150	-.318**	-.454**	.332**	1
	Sig. (2-tailed)	.092	.000	.000	.000	
	N	128	128	128	128	128

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

From the correlation table above, the correlation of any variable and itself is 1 and the number of missing observations for each variable is N= 128. From the result above , the strongest correlation is between depression and overall well-being r = 0.781 and is statistically significant at 0.01 level of significance ($p=0.00<0.05$).