Cognixia

A Collabera LEARNING SOLUTIONS COMPANY

Cancer Incentive

Project Theme : A medical Drugs company have been paying huge amount of incentives for Sales representatives across 6 different locations for dealing customers /patients having cancer. So ,we need to analyze huge volumes of data ,dig deeper into it and gain insights out of the data for suggesting a “Safe maximum” percentile amount that can be paid to every sales rep. based on their achievement in targets.

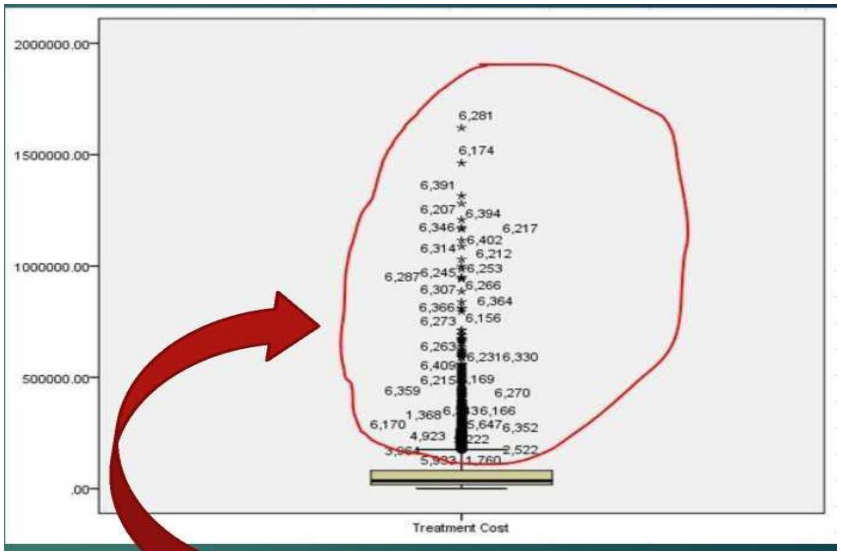
Tools Used : SPSS ,R and Excel
Visualizations : Box and Whisker Plots,Q-Q Plot and Histogram .
Stages Involved : Data Gathering , Cleaning (detecting outliers and removing them) ,analyzing and suggesting safe maximum amount .

Exploratory Data Analysis Methods : Five Number summary Analysis, Descriptive Statistics and Calculation of z-scores for removing outliers ,Cross Tabulations .

tive Statistics of complete Data Set , Outliers Identification using B ox -Plot

Descriptives

			Statistic	St cl. Error
Cost	Mean		70188.0511	1299.58372
	95% Confidence Interval for Mean	Lower Bound	67640.4332	
		Upper Bound	72735.6690	
	5% Trimmed Mean		54720.4513	
	Median		34829.0000	
	Variance		10837785796.777	
	Std. Deviation		104104.68672	
	Minimum		150.00	
	Maximum		1.62E+06	
	Range		1619968.00	
	Interquartile Range		63511.50	
	Skewness		4.906	.031
	Kurtosis		39.544	.061



Observe Skewness and kurtosis (Not in

Outliers(Extreme Values are present)

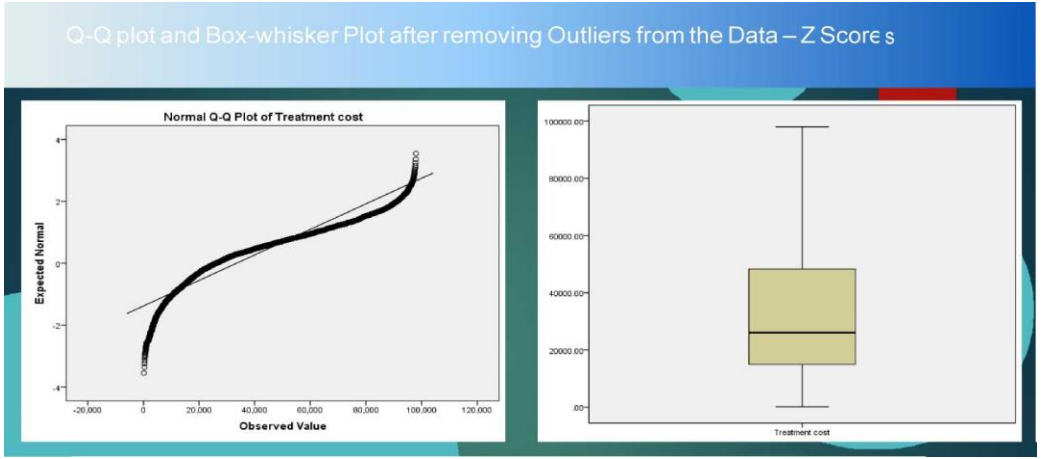
As we identified that there are set of extreme Values(Bad Data), our task is to remove them by not considering them for Analysis . So, We will do an Outlier treatment using two Different methods .

Method 1 : Calculate Z-scores until the Extreme values are ‘0’.
Method 2: Calculate Lower and Upper Median Values and remove data points that doesn’t fall under this range.

Z-Score for a Particular data point = (Data Value - Mean)/ Standard Deviation .

Z-score range is -3 to +3 . If the value is less than -3 or greater than +3 , we will not consider those data values for analysis.

After doing an iteration process of removing outliers, we will consider remaining data for Analysis.



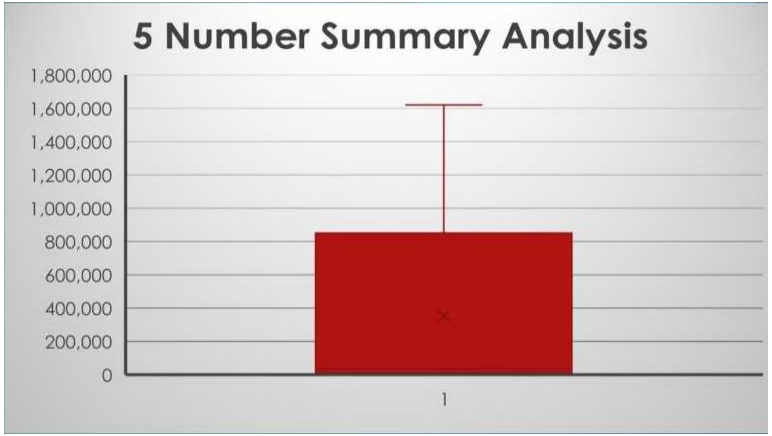
				Missing		No. of Extremes[a]	
	N	Mean	Std. Deviation	Count	Percent	Low	High
VAR00021	5121	33484.5310	24248.57220	1296	20.2	0	0

-So , from above table and charts, we can Say that Extreme values(Low and High) have been removed and data is Positively skewed ,i.e. because of long tail on right hand side shown in Box Plot. 20% of values are not considered for Analysis.

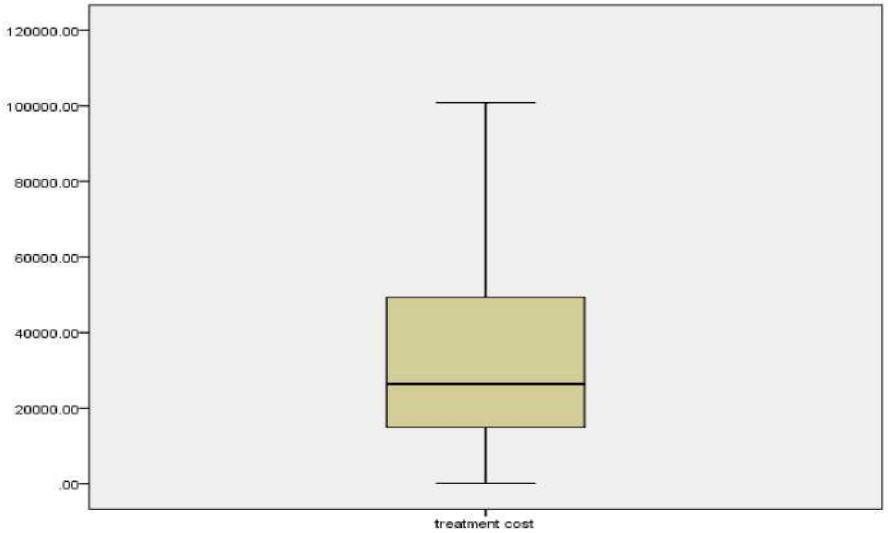
This method tells us whether selected sample of data is normally distributed based on Five Numbers shown below (Q0,Q1,Q2,Q3 and Q4)

5 Number summary Analysis

Q0(Minimum Value in the range)	150
Q1(25 %)	17306
Q2(Median)	34,829
Q3(75%)	80809
Q4(Maximum value in the range)	1620118
IQR(Q3-Q1)	63503
LOWER MEDIAN OUTLIER	-77948.5
UPPER MEDIAN OUTLIER	176063.5



- Because of Extreme Values, Graph looks to be asymmetric .



-Extreme Values have been removed and it is clearly visible in box plot that there were no Outliers . However, Data is positively or right skewed because of extreme values on right hand side of Median .

Formulae for identifying Lower and Upper Median Values

Lower Range: Q1- 1.5* IQR
Upper Range : Q3+1.5*IQR

Univariate Statistics

			Missing		No. ofExtremes[a]
N	Mean	Std. Deviation	Count Percent	Low	High
VAR00035 5165	34047.4145	24897.20002	1242 19.4	0	0

Insights :

- The 75 th percentile value is 49377 and 80% Value is which makes a difference of 7000 .
- Like Wise , the difference between 95th and 90th percentile values is 11779 which was a sudden increase .
- By this we can say that 90 % of values are lesser than 74448 .
- Hence, its suggestible to keep the treatment cost at 90 percent

Conclusion : This inferences were drawn based out of the calcu

the output based on the number of data points that are considered. However, for data to be normally distributed ,removal of extreme values in important .

Statistics	
treatment cost	
N Valid	5165
Missing	1242
Mean	34047.4145
Median	26419.0000
Mode	20000.00[a]
Skewness	.899
Std. Error of Skewness	.034
Kurtosis	-.157
Std. Error of Kurtosis	.068
Minimum	150.00
Maximum	100869.00
Percentiles 25	14984.5000
50	26419.0000
75	49377.5000
80	56386.8000
85	64879.5000
90	74448.8000
95	86227.9000

For more information or to set up an appointment, kindly contact us today.