Cancer data: Location wise Treatment targets

Project description, Tools and Techniques

Data Inputs & Tools

- Description :
  - Understand the historical data and fix the targets For Disease wise treatment cost in Location wise
- Data Inputs : Bits Data.xlsx
  - Provided data having Disease_Name and Locations wise treatment cost
- Tools used : SPSS, Excel and R

Techniques

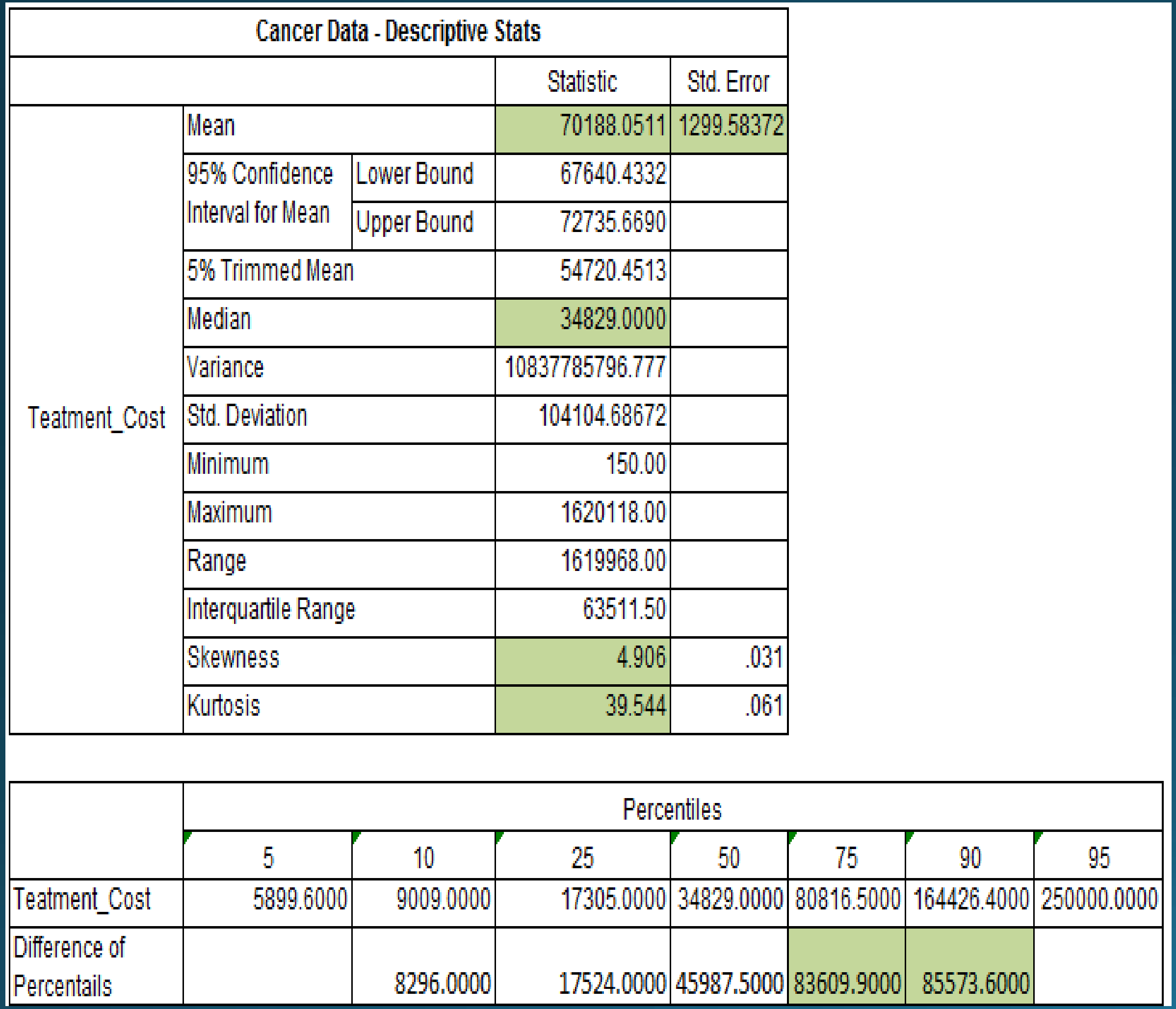
Exploratory data analysis

- Histogram
- Box and Wisher plots
- Q-Q plot

Descriptive Stats

- Central Tendency – Identify the Patron of input data
  - Mean, Median, Mode, Min, Max
- Dispersions- Identify the Quality of input data
  - Standard Deviation, Kurtosis, Skewness, Range and Standard Error
- Data Normalization & identify the outliers in data
  - Z-Score, 5 number analysis

Cancer data- Descriptive Stats



Cancer Data - Descriptive Stats

			Statistic	Std. Error
Teatment_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		1620118.00	
	Range		1619968.00	
	Interquartile Range		63511.50	
	Skewness		4.906	.031
	Kurtosis		39.544	.061

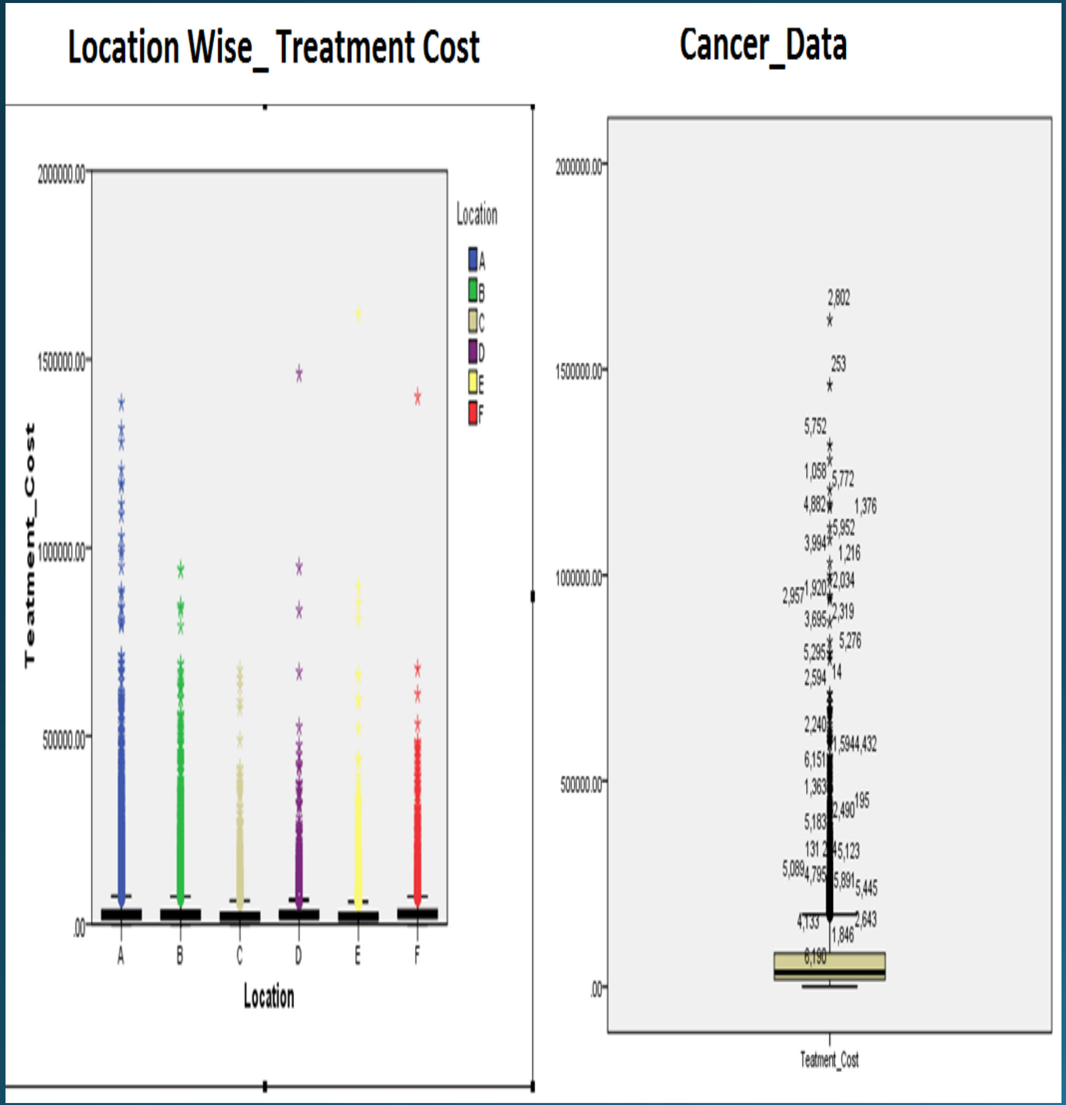
	Percentiles						
	5	10	25	50	75	90	95
Teatment_Cost	5899.6000	9009.0000	17305.0000	34829.0000	80816.5000	164426.4000	250000.0000
Difference of Percentails		8296.0000	17524.0000	45987.5000	83609.9000	85573.6000	

Insights: Descriptive Stats

- ❖ Mean and median are not similar and Standard error is high
- ❖ Skewness and Kurtosis are not under range
- ❖ Range of the data also to high Min (150)and Max (1620118)
- ❖ Almost 90% of treatment cost is below 9009.00
- ❖ 75th and 90th Percentiles difference is very high

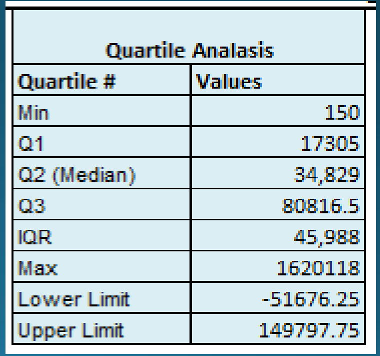
Note: Based on the descriptive stats, input data having outliers

Exploratory data analysis



Insights: Boxplot

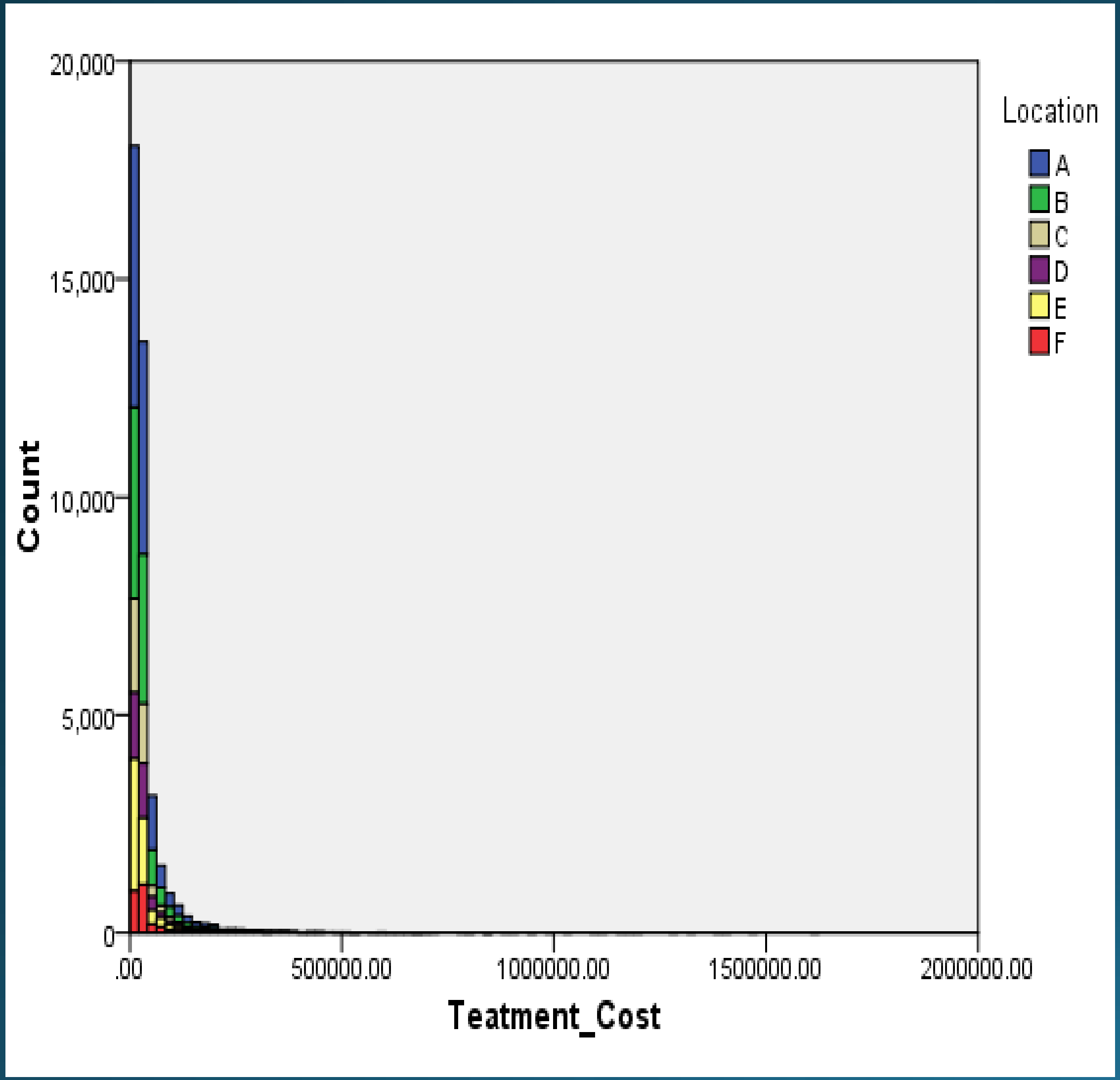
- ❖ Most of the data having Outliers
- ❖ **Less than Lower limit** and **gather than Upper limit** consider the outliers
- ❖ To identify the outliers we are using below formulas
    - ❖ **Lower Limit : Q1-1.5*IQR**
    - ❖ **Upper Limit : Q3+1.5*IQR**
- ❖ The magenda colour data is outliers in all locations cancer data
- ❖ Location "A" having more data and more outliers
- ❖ Location "C" having less data and Less outliers
- ❖ Below are the Quartile wise summary details



Quartile Analasis

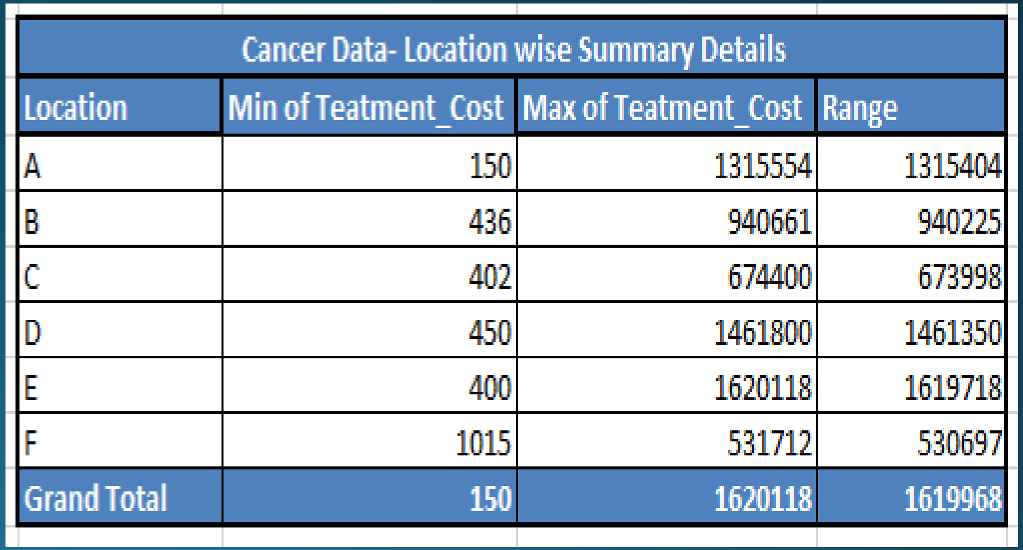
Quartile #	Values
Min	150
Q1	17305
Q2 (Median)	34,829
Q3	80816.5
IQR	45,988
Max	1620118
Lower Limit	-51676.25
Upper Limit	149797.75

Exploratory data analysis-Histogram



Insights: Histogram

- ❖ Data looks like Right skewed (Positive)
- ❖ Skewness range is -0.8 to +0.8 but input cancer data skewness is "4.90621764277889", data is not in Skewness range
- ❖ Cancer data location wise summary details.



Cancer Data- Location wise Summary Details			
Location	Min of Teatment_Cost	Max of Teatment_Cost	Range
A	150	1315554	1315404
B	436	940661	940225
C	402	674400	673998
D	450	1461800	1461350
E	400	1620118	1619718
F	1015	531712	530697
Grand Total	150	1620118	1619968

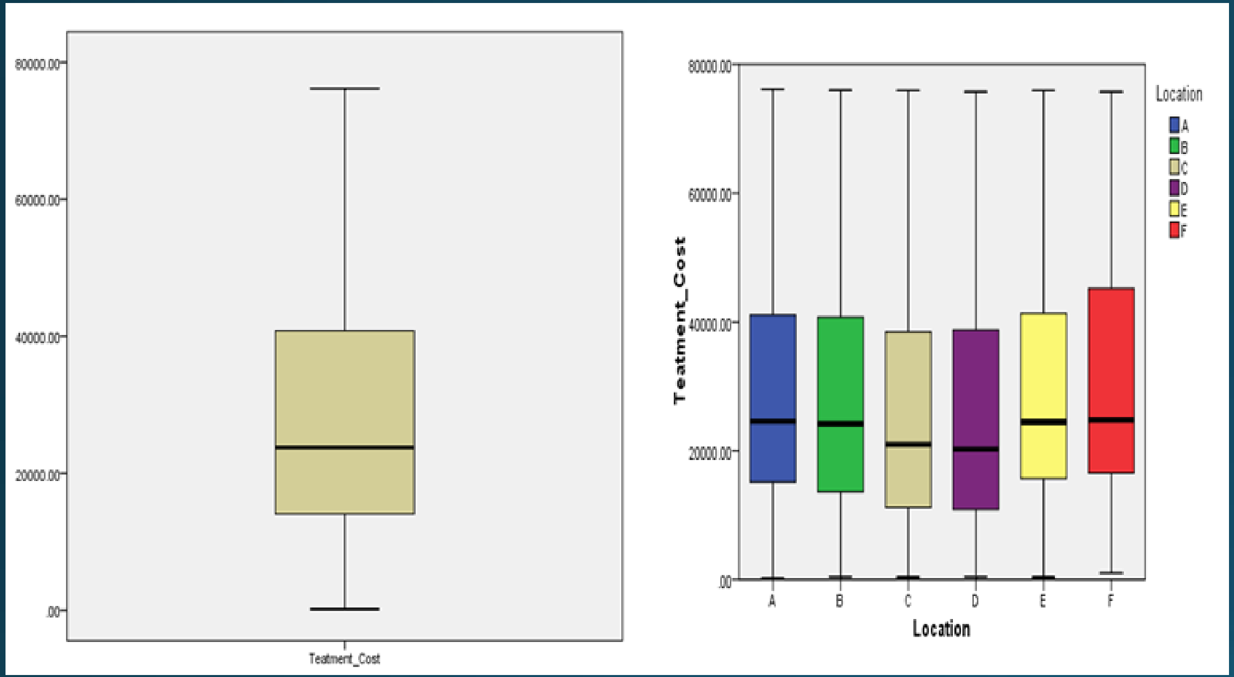
Removing the outliers

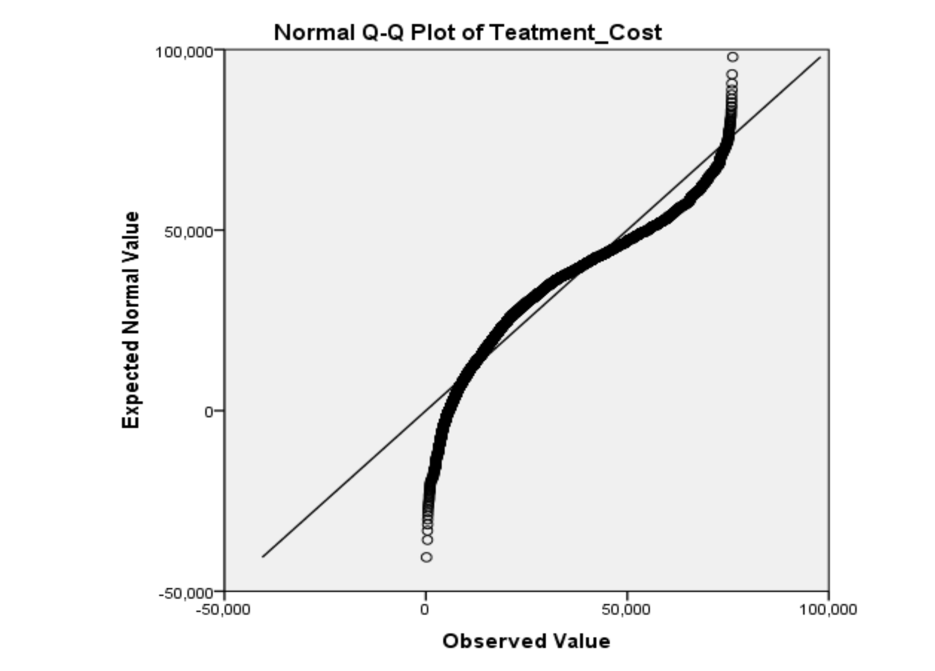
- As per the initial data understandings, some Bad data available. Our task to avoid the bad data using the outlier treatment
- To removing the outliers we are following 2 different methods
- Method (A) : using the Z-Score until removing the extreme values "Zero"
  - For Z-Score range we calculated -1.96 to +1.96 with 95% of Confidence interval,
  - If data is having less than -1.96 and grater than +1.96 will not consider data values for this analysis
  - For Z-Score we used the formula Z-Score=(data Value-mean/Stdv)
- Method (B) : Using the lower and Upper Quartile limits
  - Lower limit **"Q1-1.5*IQR"** and Upper limit **"Q3+1.5*IQR"**
  - After removing the outliers will start the analysis remaining data

Note:

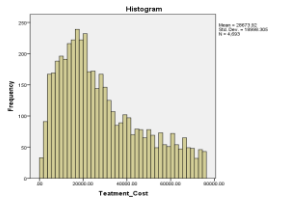
- ❖ For this data we are using Z-Score calculation using 95% of confidence interval
- ❖ When I am trying to remove the outliers using 5 number analysis, more data removing to control the skewness so we consider Z-Score technique to remove the outliers for this data

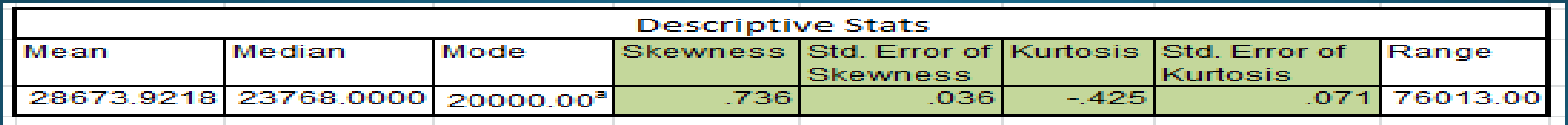
Box-whisker & Q-Q plot after removing Outliers using– Z Scores





- After removing the outliers, data is positively skewed and skewness value is “0.736” and skewness also under the range
- We consider the 95% confidence value to remove the outliers using the Z-Score
- Because of long tail on right hand side shown in Box Plot. 25% of values are not considered for Analysis.
- After removing the outliers below are the descriptive stats summary details





Descriptive Stats							
Mean	Median	Mode	Skewness	Std. Error of Skewness	Kurtosis	Std. Error of Kurtosis	Range
28673.9218	23768.0000	20000.00ª	.736	.036	-.425	.071	76013.00

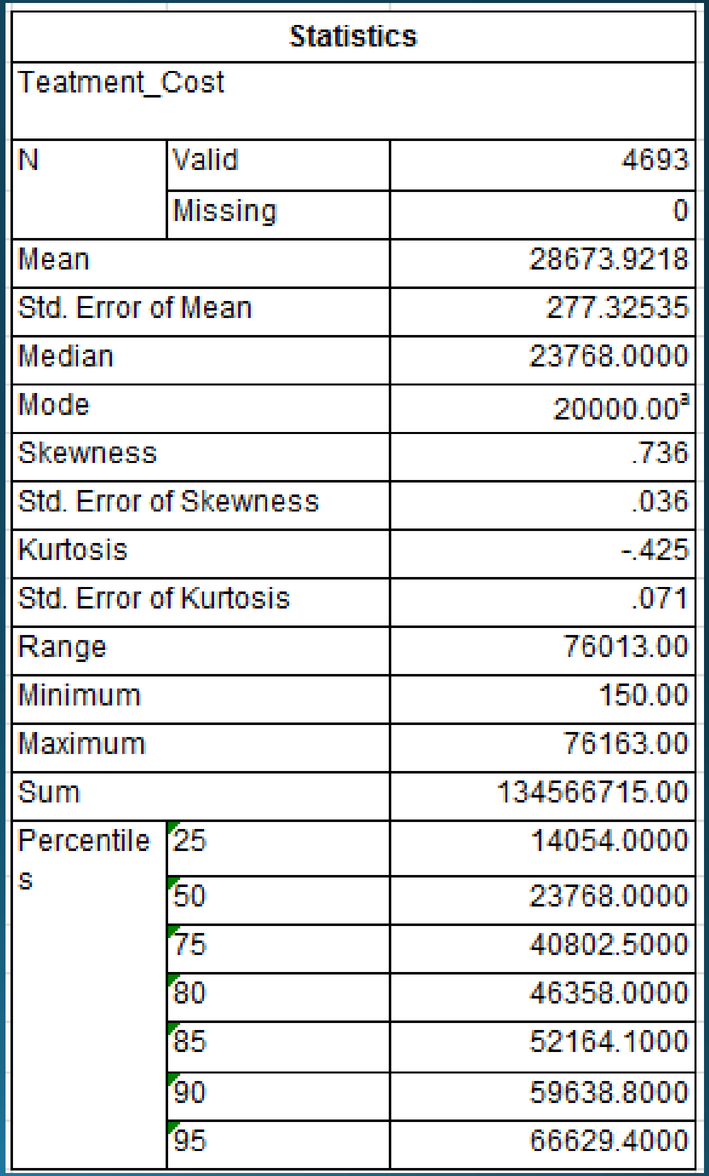
Cancer data Insights

- The different between 25th to 50th percentile values is 17034, which is sudden increased
- Same way when we compare the 50th and 75th percentile not that much difference.
- The difference between 75th and 95th percentiles values is to high, 95th percentiles cost is increased more than 50% of 75th Percentile
- The different between 75th to 85th and 85th to 95th values are not much different,
- Hence, its suggestible to keep the treatment cost at **85th percentile** .

Note :

This treatment cost we finalized based on the calculations done through Z scores using confidence interval is 95% to remove the outliers continues 6 times iterations to normalize the data

There might be difference if we change the confidence interval and numbers. However the current data is normally distribution after removing the outliers

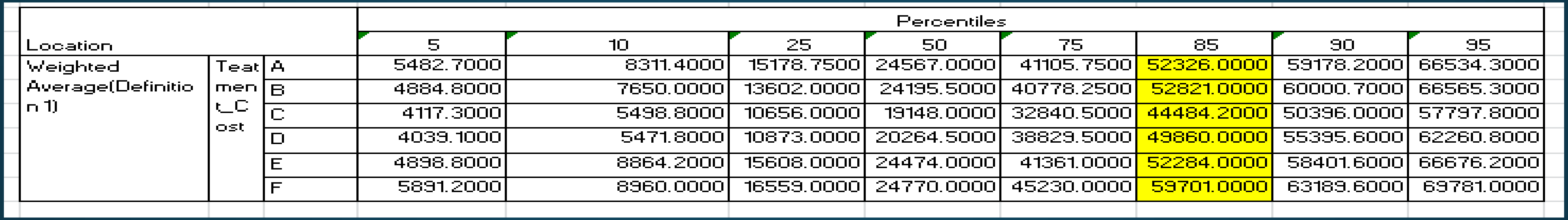


Statistics

Teatment_Cost

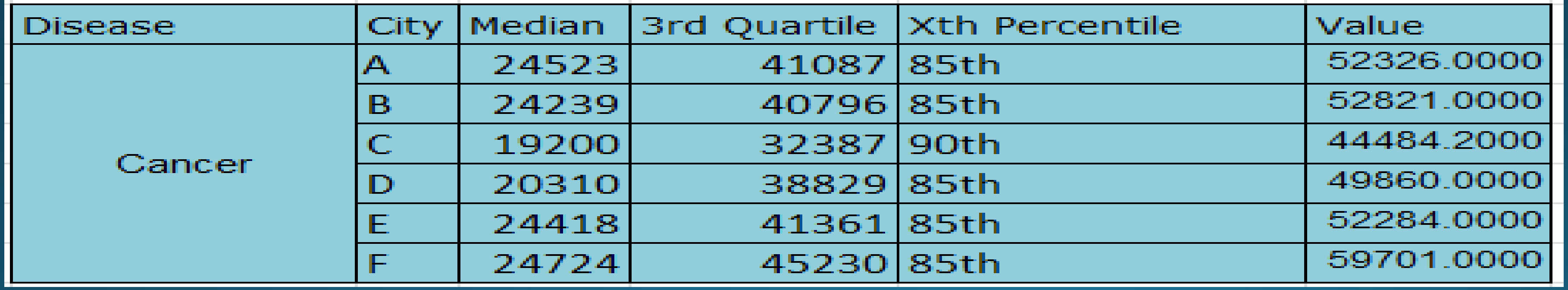
N	Valid	4693
	Missing	0
Mean		28673.9218
Std. Error of Mean		277.32535
Median		23768.0000
Mode		20000.00[a]
Skewness		.736
Std. Error of Skewness		.036
Kurtosis		-.425
Std. Error of Kurtosis		.071
Range		76013.00
Minimum		150.00
Maximum		76163.00
Sum		134566715.00
Percentiles	25	14054.0000
	50	23768.0000
	75	40802.5000
	80	46358.0000
	85	52164.1000
	90	59638.8000
	95	66629.4000

Cancer data Insights :Locations wise



Location			Percentiles							
			5	10	25	50	75	85	90	95
Weighted Average(Definition 1)	Teatment_Cost	A	5482.7000	8311.4000	15178.7500	24567.0000	41105.7500	52326.0000	59178.2000	66534.3000
		B	4884.8000	7650.0000	13602.0000	24195.5000	40778.2500	52821.0000	60000.7000	66565.3000
		C	4117.3000	5498.8000	10656.0000	19148.0000	32840.5000	44484.2000	50396.0000	57797.8000
		D	4039.1000	5471.8000	10873.0000	20264.5000	38829.5000	49860.0000	55395.6000	62260.8000
		E	4898.8000	8864.2000	15608.0000	24474.0000	41361.0000	52284.0000	58401.6000	66676.2000
		F	5891.2000	8960.0000	16559.0000	24770.0000	45230.0000	59701.0000	63189.6000	69781.0000

- ❖ As per the requirement we calculated Location wise cancer treatment cost.
- ❖The difference between 75th and 95th percentiles values is to high, 95th percentiles cost is increased more than 50% of 75th Percentile
- ❖The different between 75th to 85th and 85th to 95th values are not much different,
- ❖Hence, its suggestible to keep the treatment cost at **85th percentile for all Locations.**
- ❖ Please below table for reference purpose.



Disease	City	Median	3rd Quartile	Xth Percentile	Value
Cancer	A	24523	41087	85th	52326.0000
	B	24239	40796	85th	52821.0000
	C	19200	32387	90th	44484.2000
	D	20310	38829	85th	49860.0000
	E	24418	41361	85th	52284.0000
	F	24724	45230	85th	59701.0000