HOPE AI

IQR Analysis Document

Github link of Data:

https://github.com/JayachandraPrabha/3.Data-Science/blob/main/Placement.csv

1) **IQR: (Interquartile Range):**
    - To find the outliers present in the data which differs significantly from other observations.
    - Measures the spread of the middle half of the data.
    - It is the range for the middle 50% of the sample.
        - Formula: IQR = Q3(75%) - Q1(25%)

    Outliers arise due to,

    ➔ Changes in system behavior

    ➔ Fraudulent behavior

    ➔ Human error

    ➔ Instrument error or

    ➔ Simply through natural deviations in populations

2) **Why "1.5" in IQR method of outlier detection?**
    - It controls the sensitivity of the range and hence the decision rule.
    - A bigger scale would make the outliers to be considered as data points, while a smaller one would make some of the data points to be perceived as outliers
    - Lesser bund outliers:Q1-1.5*IQR
    - Greater bound outliers:Q3+1.5*IQR

Let us consider the given data, some of the basic informations about the given data are,

1) The overall shape of the dataset: **215 rows × 15 columns.**

2) The data basically deals with the survey of the candidates who have **placed and not placed** in the placement and the columns in the data are,
    - 'sl_no', 'gender', 'ssc_p', 'ssc_b', 'hsc_p', 'hsc_b', 'hsc_s', 'degree_p', 'degree_t', 'workex', 'etest_p', 'specialization', 'mba_p', 'status', 'salary'
    - In the above mentioned columns there were qualitative columns / variables (Categorical data) and quantitative columns / variables (Numerical data), hence both were separated.

3) Below is the obtained data from the dataset:

	sl_no	ssc_p	hsc_p	degree_p	etest_p	mba_p	\
Mean	108.0	67.303395	66.333163	66.370186	72.100558	62.278186	
Median	108.0	67.0	65.0	66.0	71.0	62.0	
Mode	1	62.0	63.0	65.0	60.0	56.7	
min	1.0	40.89	37.0	50.0	50.0	51.21	
Q1:25%	54.5	60.6	60.9	61.0	60.0	57.945	
Q2:50%	108.0	67.0	65.0	66.0	71.0	62.0	
Q3:75%	161.5	75.7	73.0	72.0	83.5	66.255	
99%	212.86	87.0	91.86	83.86	97.0	76.1142	
Q4:100%	215.0	89.4	97.7	91.0	98.0	77.89	
IQR	107.0	15.1	12.1	11.0	23.5	8.31	
1.5Rule	160.5	22.65	18.15	16.5	35.25	12.465	
Lesser	-106.0	37.95	42.75	44.5	24.75	45.48	
Greater	322.0	98.35	91.15	88.5	118.75	78.72	
Min	1	40.89	37.0	50.0	50.0	51.21	
Max	215	89.4	97.7	91.0	98.0	77.89	

	salary
Mean	288655.405405
Median	265000.0
Mode	300000.0
min	200000.0
Q1:25%	240000.0
Q2:50%	265000.0
Q3:75%	300000.0
99%	NaN
Q4:100%	940000.0
IQR	60000.0
1.5Rule	90000.0
Lesser	150000.0
Greater	390000.0
Min	200000.0
Max	940000.0

4) From the obtained insights, let us segregate the need informations:

colName	Min	Lesser outliers	Q1	Median (Q2)	Q3	Max	Greater outliers
ssc_p	40.89	37.95	60.6	67	75.7	89.4	98.35
hsc_p	**37**	42.75	60.9	65	73	**97.7**	91.15
degree_p	50	44.5	61	66	72	**91**	88.5
etest_p	50	24.75	60	71	83.5	98	118.75
mba_p	51.25	45.48	57.94	62	66.25	77.89	78.72
salary	2,00,000	1,50,000	2,40,000	2,65,000	3,00,000	**9,40,000**	3,90,000

Result:

Hence the outliers in the given data were as follows:

colName	Lesser bound outliers	Greater bound outliers
ssc_p	Nil	Nil
hsc_p	**37 < (42.75)**	**97.7 > (91.15)**
degree_p	Nil	**91> (88.5)**
etest_p	Nil	Nil
mba_p	Nil	Nil
salary	Nil	**9,40,000 > (3,90,000)**

- In hsc_p, **37** is the lesser bound outliers & **97.7** is the greater bound outliers
- In salary, **9,40,000** is the greater bound outliers
