Data activity 5: Health dataset

Data Activity 5

Using the **Health_Data**, please perform the following functions in R:

- Find out mean, median and mode of variables ***sbp, dbp and income.***
- Find out the five-figure summary of ***income*** variable and present it using a Boxplot.
- Run a suitable hypothesis test to see if there is any association between systolic blood pressure and presence and absence of peptic ulcer.

Remember to **record** your findings in your e-portfolio

Learning Outcomes

- Systematic understanding of the key mathematical and statistical concepts and techniques which underpin mechanisms in Data Science and AI.
- Apply mathematical and statistical methods in these fields to help in the decision-making process.

Statistical analysis:

Systolic blood pressure(sbp):

Minimum	91.0
1st quartile	114.0
Median	123.0
Mean	127.7
3rd quartile	141.8
maximum	195.0
Mode	12

Diastolic Blood pressure (dbp):

Minimum	60.0
1st quartile	74.0
Median	82.0
Mean	82.77
3rd quartile	90.0
maximum	115.0
Mode	13

Income:

Minimum	52933
1st quartile	68637
Median	86561
Mean	85194
3rd quartile	99696
maximum	117210
Mode	NA

Boxplot – income:

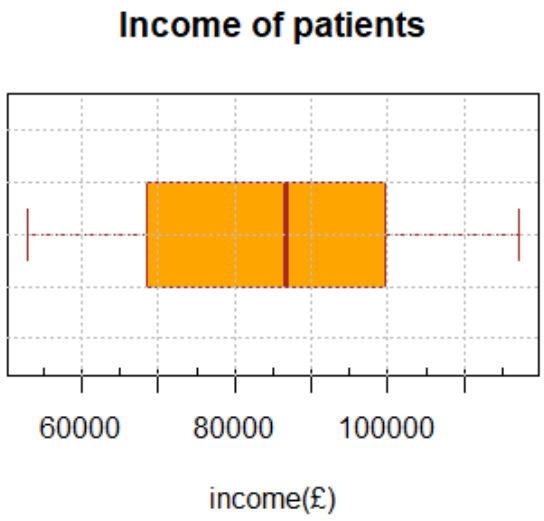

Figure 1: This box-plot demonstrates the distribution of income among patients.

Systolic blood pressure and the presence and absence of peptic ulcer:

Null hypothesis: there is not association between systolic blood pressure and the presence and absence of peptic ulcers.

Alternative hypothesis: There is an association between systolic blood pressure and the presence and absence of peptic ulcers.

Before the choice of the statistical test can be confirm, the nature of the data needs to be assessed.

Systolic blood pressure = quantitative data, independent variable

Peptic ulcer = categorical, dependent variable

It can be deducted from looking at the nature of the peptic ulcer dataset that there are two levels present, presence/absence. This deduces the choice of the statistical down to two: an independent sample t-test or a Mann-Whitney U test.

The preferred test can be decided by performing a normality test.

To identify the normality of the data, a histogram was drawn.

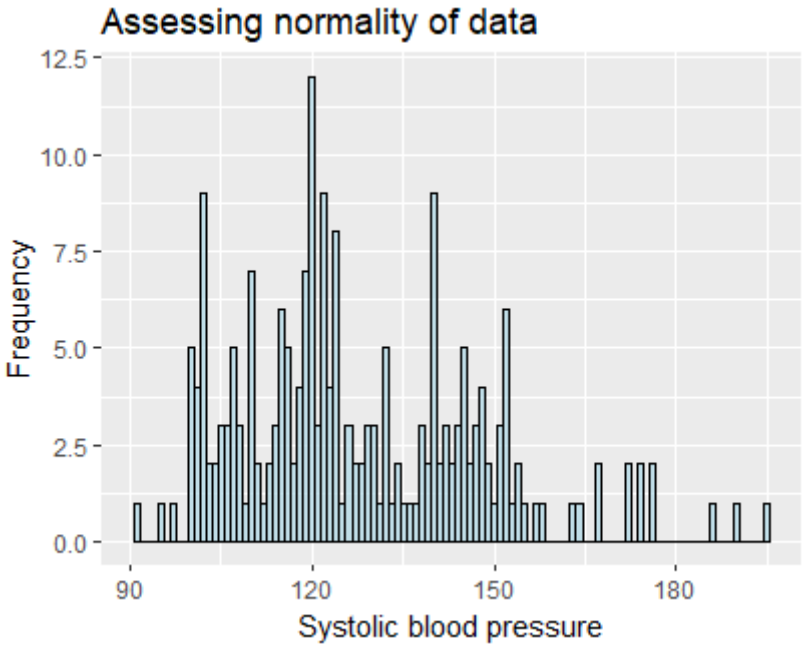

Figure 2: A histogram demonstrating the distribution of systolic blood pressure among the patients.

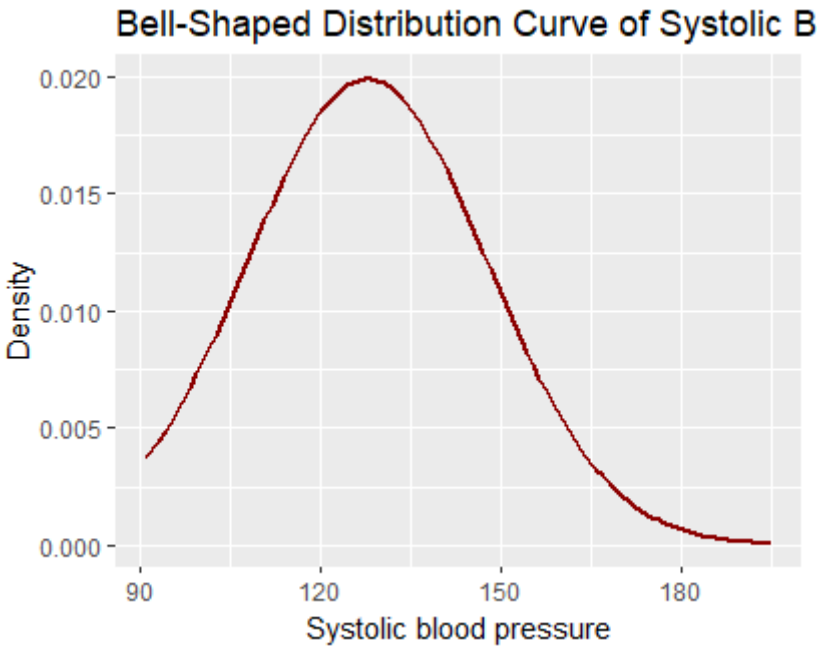

Figure 3: Bell-shaped curve demonstrating the distribution of systolic blood pressure of patients.

From Figures 2 and 3, it can be determined that there is a right-skewed distribution. This tells you that the data is non-normal.

Due to this, the preferred statistical test is a Mann-Whitney U test.

These are as follows:

W = 44100

P- value < 2.2e-16

Alternative hypothesis: true location shift is not equal to 0.

Test Statistic (W): The test statistic (W) is 44100. This is the Wilcoxon rank sum test statistic, which measures the sum of ranks assigned to the values in the two groups. It is used to assess whether the distributions of the two groups are the same.

P-value: The p-value is very small (p-value < 2.2e-16), essentially zero. This suggests strong evidence against the null hypothesis. In the context of the Mann-Whitney U test, a small p-value indicates a significant difference between the distributions of the two groups.

Alternative Hypothesis: The alternative hypothesis is given as "true location shift is not equal to 0." This is a general phrasing for the Mann-Whitney U test, indicating that the distributions of the two groups are not equal.

Interpretation:

With a very small p-value, the null hypothesis is rejected. This suggests that there is an association between systolic blood pressure and the presence/absence of a peptic ulcer.

From previous research conducted, it would be an accurate conclusion to draw. Research which has assessed the occurrence of peptic ulcers in men and their blood pressure (Segawa *et al.*, 1995) has found gastric ulcers in men were related to systolic and diastolic pressure. No relationship, however, was found for women.

References:

Segawa, K. et al. (1995) 'Incidence of peptic ulcer in men is inversely correlated with blood pressure: study in an apparently healthy Japanese population', The American Journal of Gastroenterology, 90(3), pp. 399–402. Available at: https://pubmed.ncbi.nlm.nih.gov/7872277/#:~:text=The%20incidence%20of%20duodenal%20ulcer (Accessed: 14 January 2024).

Data Activity 6

Find out the mean, median and mode of 'age' variable.

Age:

Minimum	6.00
1st quartile	21.00
Median	27.00
Mean	26.51
3rd quartile	32.00
maximum	45.00
Mode	26

Find out whether median diastolic blood pressure is same among diabetic and non-diabetic participants.

In the diabetes column no = 2, yes = 1.

Therefore the median for people with diabetes = 83dbp and those with no diabetes = 82dbp

Find out whether systolic BP is different across occupational group.

value label
1 GOVT JOB
2 PRIVATE JOB
3 BUSINESS
4 OTHERS

Occupation	**Systolic BP**
Gov job	129.3833
Private Job	126.3469
Business	127.8571
Others	127.0192