**Assignment1**

**Uchenna Ilodigwe**

**Big Data Solutions Architecture, Conestoga College**

**Prog8430-Data Analytics Mathematics & Algorithms**

**October 5th,2021**

**Question 1.1.a**

The table below shows the total income for each marital status category.

|  |  |  |
| --- | --- | --- |
|  | Marital\_Status | Income |
| 1 | divorced | 47305893 |
| 2 | married | 72416620 |
| 3 | never | 31039004 |
| 4 | widowed | 8788126 |

**Question 1.1b:** The highest total income is from the **MARRIED** marital status category

**Question 1.2.a:** The mean age of respondents born in Asia is 40.96

**Question 1.2.b:** The mean age of respondents born in Asia weighted by the number of children they have is 40.61

Question 1.3.a: The table below shows the mean score of the political awareness test per gender

|  |  |
| --- | --- |
| Gender\_UI | Score\_UI |
| female | 0.0068008 |
| male | 0.0007063 |

**Question 1.3.b:** The females have a higher score of 0.0068008

**Question 1.4**: The 34th and 63rd percentiles of percentage of time taken on test are 0.20 and 0.61 respectively.

Question 2.1

Piechart showing the total respondents in each political affiliation

Chart

Description automatically generated

Question 2.1b: The Political Affiliation contains the most respondents is the Conservative

Question 2.1c: The Political Affiliation has the fewest respondents is the( Other ) political affiliation

Question 2.2a:

The table shows the shows the percentage of respondents from each Region that are in the Treatment group.

|  |
| --- |
|  |

|  |  |
| --- | --- |
| nation\_UI | percentage\_UI |
| Asia | 0. 2349057 |
| Europe | 0. 1443396 |
| North America | 0. 5155660 |
| Southern | 0. 1051887 |

Question 2.2.b: The region that has the highest percentage of people in the Treatment group is NorthAmerica

1. Question 2.2.c: The region that has the lowest percentage of people in the Treatment group is the Southern region

**Question 2.3.a:**

Chart, bar chart

Description automatically generated

**Question 2.3.b:** The Region that has the lowest mean score is North America

* 1. Question 2.3.c. The Region that has the highest mean score is the Southern Region

Question 2.4.a:Histogram showing the distribution of percentage food per income

Chart, histogram

Description automatically generated

Question 2.4.b:

0-0.2 has the highest frequency.

Question 2.5.a

Chart, box and whisker chart

Description automatically generated

Question 2.5.b: According to the charts, the married has the highest average income because the box plot representing married martial has the highest top.

Question 2.5.c: The divorced marital status has the lowest average income because the bottom of the boxplot can be traced to the lowest income on the y (Income)axis

Question 2.5.d. The married marital status has the greatest variability in income because it has the largest difference between the top and bottom of the box plot

Question2.6.a: Histogram representing the distribution of household incomeChart, histogram

Description automatically generated

Question 2.6.b:Histogram representing the distribution of standardized score

Chart, histogram

Description automatically generated

Question 2.6.c:

Chart, scatter chart

Description automatically generated

Question 2.6.d: In conclusion, there is a relationship between score and income, because the x and the y co-ordinates seem to be increasing at about the same pace.

Question 2.6.e: Using Pearson's correlation , correlation coefficient = 0.4568418

Question3.1.a:

Chart, line chart

Description automatically generated

Question 3.1.b: The test to be conducted is the Shapiro-wilk normality test.

Question 3.1.c: Political Awareness test score is normally distributed because the p-value for the Shapiro-wilk test is 0.508 which is above 0.50 which provides marginally significant evidence of a normal distribution

Question 3.2.a:The hypothesis test used is the T-TEST.

Question 3.2.b: The reasons for the test are the following:

1. The data is conceptually independent

2. Political Awareness test score is normally distributed because the p-value for the shapiro-wilk test is 0.508 which is above 0.50 which provides marginally significant evidence of a normal distribution

3. having checked the variance of treatment and control group, the p.value is 0.4887 which is greater than 0.50 so there is no significant difference in variance of the two groups.

Since the three conditions are met for normality we will use the t-test test.

Question 3.2.c: From the t-test the p-value is greater than 0.05 i.e 0.1419 which gives strong evidence that the average test scores are significantly similar between the treatment and control groups.

Question 3.3.a Chart, box and whisker chart

Description automatically generated

The p- value 0.471 is greater than 0.05 and so there is strong evidence to suggest that there is no significant difference in the variance.

Question3.3.b Chart, box and whisker chart

Description automatically generated

The p-value is 2e-16 which is less than 0.05 so there is a strong evidence that shows that difference in variance.

2e-16