**Asthma Dataset Descriptive Analysis**

Student’s Name

Professor’s Name

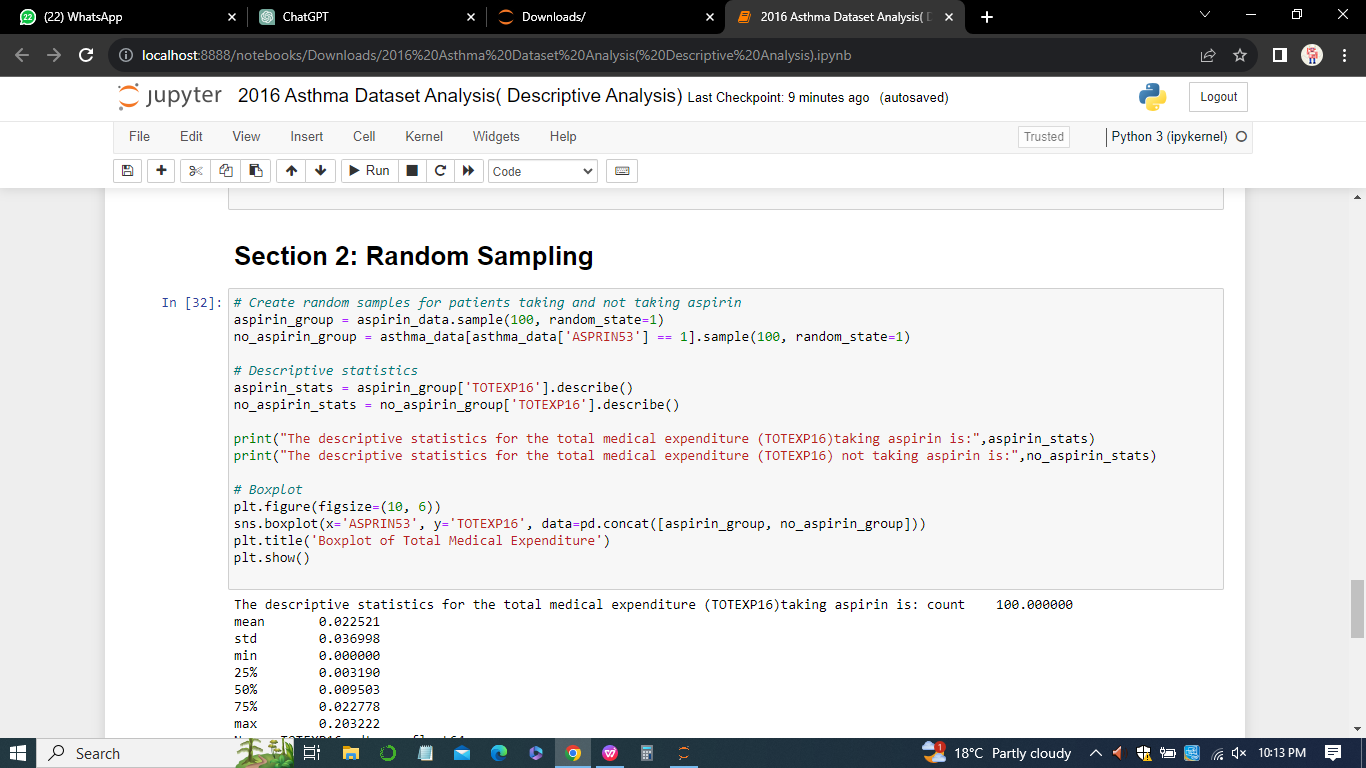
University Affiliation

Course Number and Name

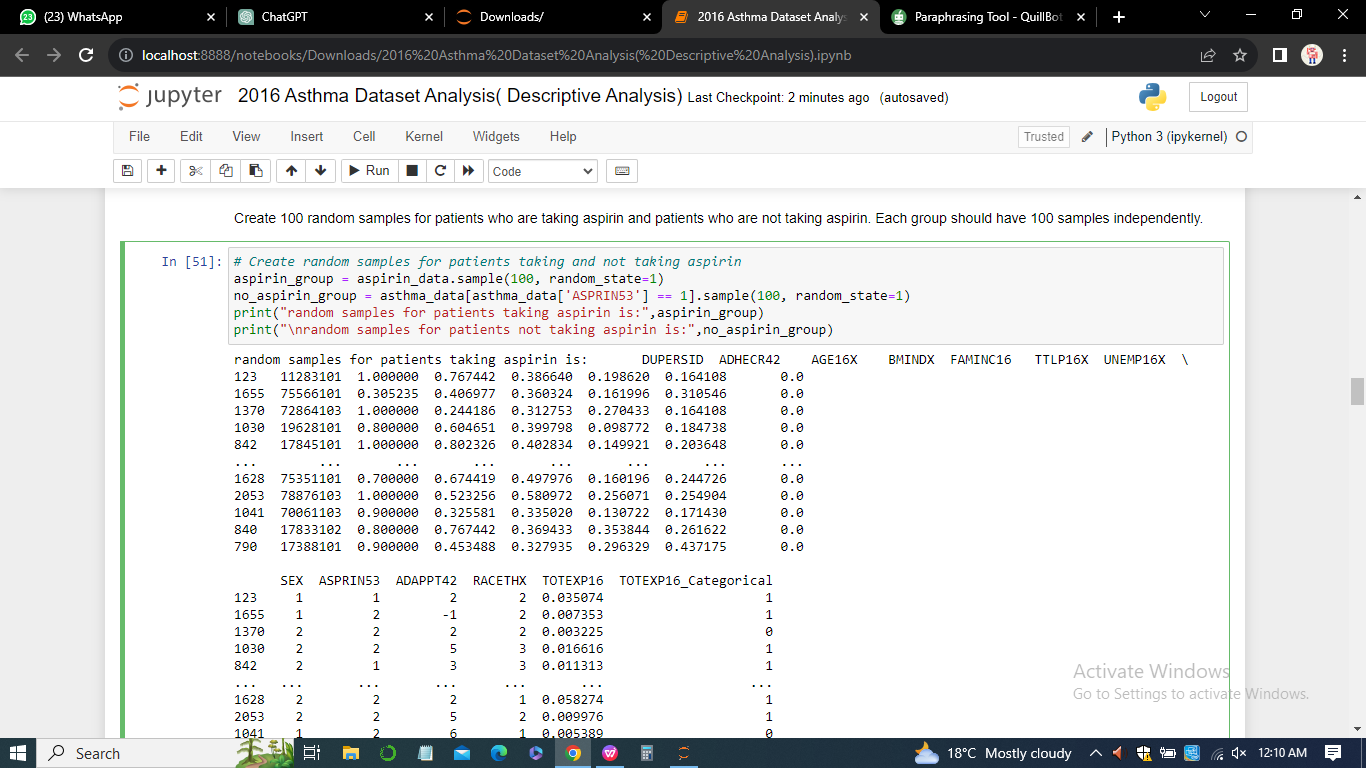
Date of Submission

**Asthma Dataset Descriptive Analysis**

**Question 1**



***Output of random sample***

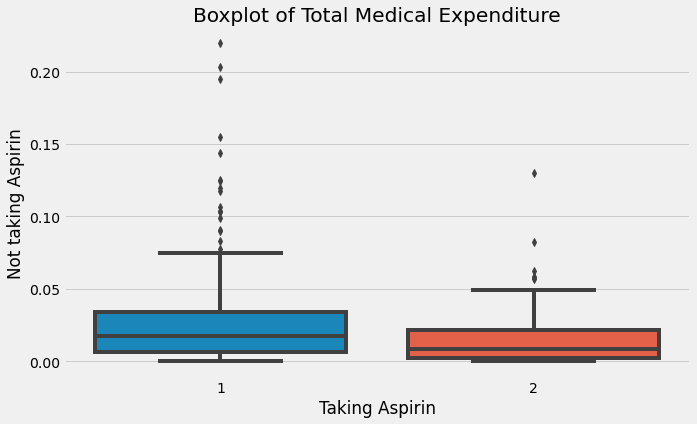


**Question 2**

1. *The descriptive statistics for the total medical expenditure (TOTEXP16) taking aspirin is:*
2. count 100.000000
3. mean 0.022521
4. std 0.036998
5. min 0.000000
6. 25% 0.003190
7. 50% 0.009503
8. 75% 0.022778
9. max 0.203222
10. *The descriptive statistics for the total medical expenditure (TOTEXP16) not taking aspirin is:*
11. count 100.000000
12. mean 0.031513
13. std 0.037962
14. min 0.000000
15. 25% 0.007011
16. 50% 0.019729
17. 75% 0.033735
18. max 0.219486

**Question 3**

***Boxplot***



**Question 4**

The pattern identified from the skewness and kurtosis values suggests that both groups (taking aspirin and not taking aspirin) have positively skewed distributions with higher kurtosis, indicating heavier tails and potentially more extreme values. This suggests that there might be some patients in both groups with significantly higher medical expenditures.

The chi-square correlation coefficients further provide insights into the association between taking aspirin and total medical expenditure:

* Chi-square Coefficient (ASPRIN53 and TOTEXP16\_Categorical): The high chi-square coefficient (55.12) with a very low p-value suggests a significant association between taking aspirin and the categorical representation of total medical expenditure. This indicates that there's a notable difference in the distribution of total medical expenditure between those taking aspirin and those not taking aspirin.
* Chi-square Coefficient (Based on Null Transactions): The chi-square coefficient (7.05) with a p-value of 0.0079 suggests a significant association between taking aspirin and the presence of null transactions (TOTEXP16 == 0). This indicates that patients taking aspirin may have a different pattern regarding zero medical expenditure compared to those not taking aspirin.

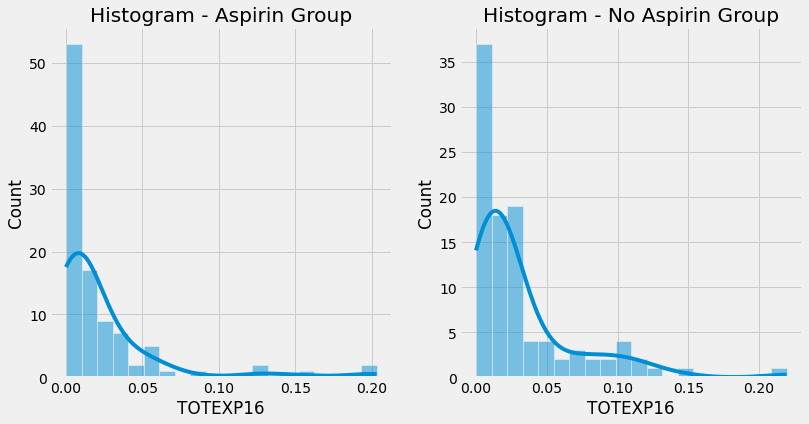
***Additional Observations***

1. Descriptive Statistics (mean, min, max) for TOTEXP16: The mean total medical expenditure for those not taking aspirin (0.0315) appears to be slightly higher than for those taking aspirin (0.0225).
2. Chi-square Contingency Tables: The observed and expected frequencies in the contingency tables indicate a significant difference in the distribution of TOTEXP16\_Categorical and null transactions between the two groups.

***Rationale***

The positive skewness, higher kurtosis, and significant chi-square coefficients collectively suggest that there might be a subset of individuals, especially in the group not taking aspirin, with higher medical expenditures. This association is supported by both the distributional characteristics and the chi-square tests. The differences identified could be explored further to understand the factors contributing to higher medical expenditures in these groups

**Question 5**



The calculated skewness and kurtosis are as given below:

1. The skewness\_aspirin is: 3.16488402657343
2. The kurtosis\_aspirin is: 10.755348026512207
3. The skewness\_no\_aspirin is: 2.2001352270972476
4. The kurtosis\_no\_aspirin is: 5.714436233146854

**Question 6**

***Rationale***

*The distributions are not perfectly normal*

From the calculated skewness and kurtosis values for both groups (taking aspirin and not taking aspirin), it's evident that the distributions are not perfectly normal. The skewness values are positive, indicating a rightward skew, and the kurtosis values are higher than expected for a normal distribution, signifying heavier tails.

In a perfectly normal distribution, the skewness is 0, and the kurtosis is 3 (for excess kurtosis). The positive skewness and higher kurtosis observed suggest deviations from normality.

Additionally, the chi-square tests for association between "ASPRIN53" and "TOTEXP16\_Categorical" indicate a significant difference in the distribution of total medical expenditure categories between the two groups. This supports the notion that the data does not follow a normal distribution, as a normal distribution would not exhibit such significant associations.

Therefore, based on the results, it can be concluded that the total medical expenditure data for both groups is not normally distributed. The distributional characteristics, as reflected in skewness, kurtosis, and the results of chi-square tests, suggest non-normality and the presence of potential outliers or extreme values in the data.

**Question 7**

The Pearson correlation coefficient between 'AGE16X' and 'ADHECR42' is

***0.5547603292854282***

**Question 8**

The Pearson correlation coefficient between 'UNEMP16X' and 'ADHECR42' is

***0.027850536381503543***

**Question 9**

***Conclusion***

1. ***'AGE16X' and 'ADHECR42'***

The positive correlation coefficient of 0.55476 suggests a moderate positive linear relationship between the age of individuals ('AGE16X') and their adherence to a treatment regimen or medication ('ADHECR42'). As age increases, there is a tendency for adherence to also increase. However, it's important to note that correlation does not imply causation, and other factors may influence this relationship.

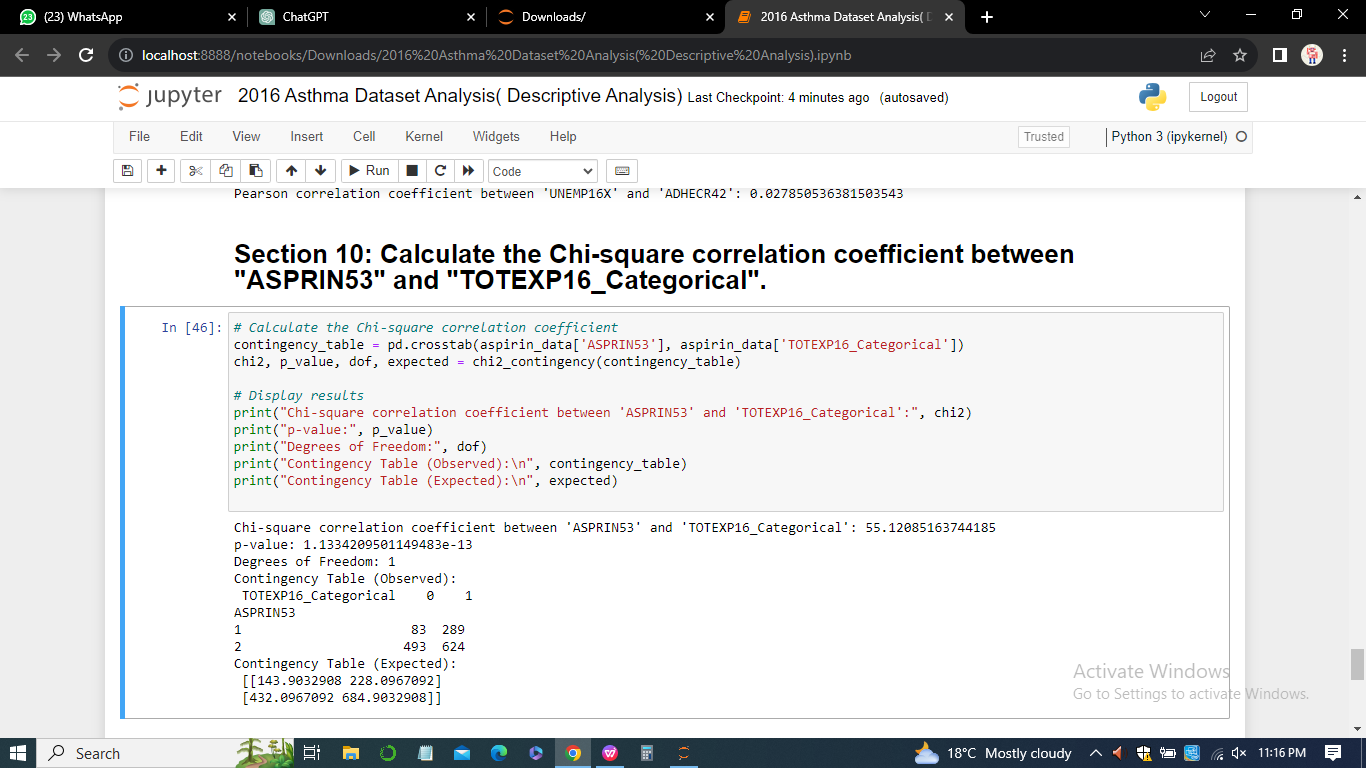
1. ***'UNEMP16X' and 'ADHECR42'***

The correlation coefficient of 0.02785 indicates a very weak positive linear relationship between 'UNEMP16X' (possibly representing unemployment status) and 'ADHECR42' (adherence to a treatment regimen or medication). The correlation is close to zero, suggesting that there is little to no linear association between these two variables.

***Summary***

For 'AGE16X' and 'ADHECR42', there is a moderate positive correlation, implying that age and adherence are somewhat positively related. For 'UNEMP16X' and 'ADHECR42', the correlation is very weak, suggesting that unemployment status and adherence are not strongly linearly associated. These results provide insights into the relationships between these variables, but it's important to consider that correlation does not capture complex non-linear relationships or causation. The interpretation should also be made in the context of the specific dataset and domain knowledge.

**Question 10**



1. Chi-square correlation coefficient between 'ASPRIN53' and 'TOTEXP16\_Categorical': 55.12085163744185
2. p-value: 1.1334209501149483e-13
3. Degrees of Freedom: 1
4. Contingency Table (Observed):

|  |  |  |
| --- | --- | --- |
|  | False (0) | True(1) |
| TOTEXP16 | 83 | 289 |
| ASPRIN53 | 493 | 624 |

1. Frequency (Expected):

[[143.9032908 228.0967092]

[432.0967092 684.9032908]]

**Question 11**

1. Chi-square correlation coefficient between 'ASPRIN53' and 'TOTEXP16\_Categorical': 55.12085163744185
2. p-value: 1.1334209501149483e-13
3. Degrees of Freedom: 1
4. Contingency Table (Observed):

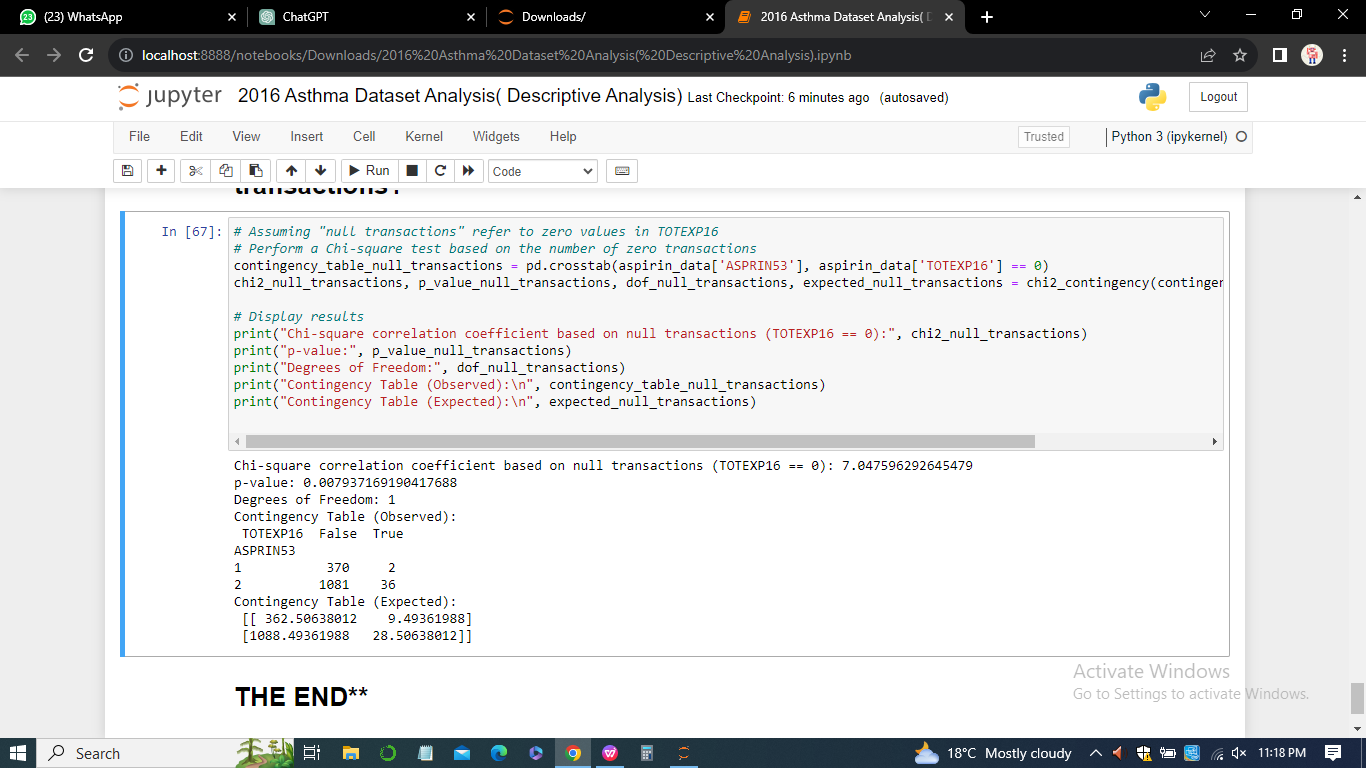
|  |  |  |
| --- | --- | --- |
|  | False (0) | True(1) |
| TOTEXP16 | 83 | 289 |
| ASPRIN53 | 493 | 624 |

1. Frequency (Expected):

[[143.9032908 228.0967092]

[432.0967092 684.9032908]]

**Question 12**



***Rationale for decision***

The Chi-square test based on the number of null transactions is applied to assess whether there is a significant association between the categorical variable 'ASPRIN53' and the binary variable 'TOTEXP16' (indicating whether total medical expenditure is zero or not). The output provides the following key information:

1. Chi-square correlation coefficient: 7.047596292645479
2. p-value: 0.007937169190417688
3. Degrees of Freedom: 1
4. Contigency table observed

|  |  |  |
| --- | --- | --- |
|  | False | True |
| TOTEXP16 | 370 | 2 |
| ASPRIN53 | 1081 | 36 |

***Rationale***

*Chi-square Correlation Coefficient and p-value*

The calculated Chi-square correlation coefficient is 7.0476, and the associated p-value is 0.00794. This p-value is less than the typical significance level of 0.05, suggesting that there is evidence to reject the null hypothesis that there is no association between 'ASPRIN53' and 'TOTEXP16' (i.e., the variables are independent).

*Degrees of Freedom*

With 1 degree of freedom, the test compares the observed and expected frequencies in the contingency table.

*Contingency Table*

The observed contingency table shows the actual counts of occurrences, and the expected contingency table presents the expected counts under the assumption of independence.

***Decision***

Considering the significant p-value, we reject the null hypothesis and conclude that there is a statistically significant association between 'ASPRIN53' and 'TOTEXP16' for null transactions. This suggests that the likelihood of having a null transaction (TOTEXP16 == 0) is different for individuals taking aspirin ('ASPRIN53' == 1) compared to those not taking aspirin ('ASPRIN53' == 2).

The Chi-square test based on null transactions provides evidence of an association between aspirin use and the occurrence of null medical expenditures, indicating a potential relationship between these variables.

**Question 13**

***Correlation Assessment***

Chi-square Correlation Coefficient: 7.0476

A higher Chi-square correlation coefficient indicates a stronger association between 'ASPRIN53' and 'TOTEXP16' based on null transactions.

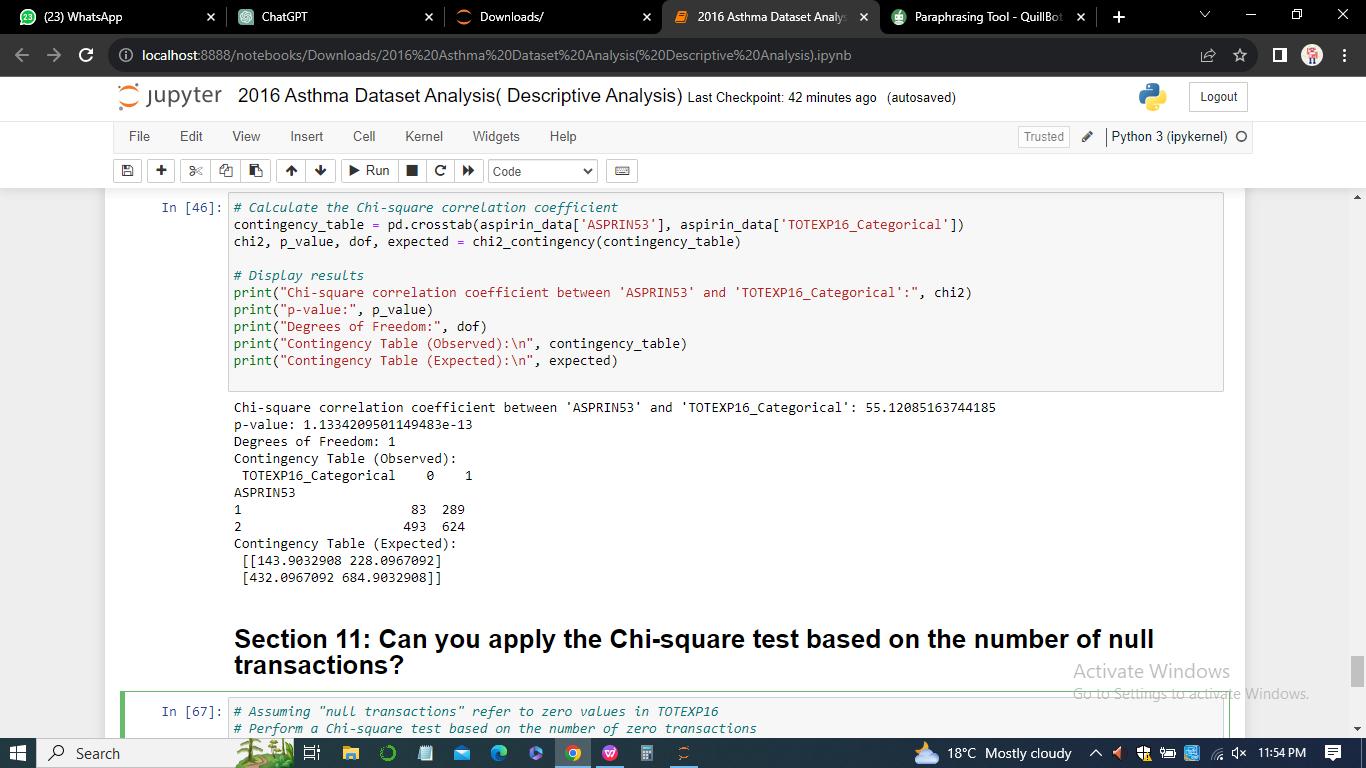
*p-value: 0.00794*

The p-value is less than the significance level of 0.05, indicating that the association observed is statistically significant.

*Degrees of Freedom: 1*

With 1 degree of freedom, the test compares the observed and expected frequencies in the contingency table.

*Contingency Table (Observed)*



|  |  |  |
| --- | --- | --- |
|  | False | True |
| TOTEXP16 | 370 | 2 |
| ASPRIN53 | 1081 | 36 |

***Interpretation***

The Chi-square correlation coefficient of 7.0476 indicates a relatively strong association between taking aspirin ('ASPRIN53') and the occurrence of null medical expenditures ('TOTEXP16 == 0').

The p-value being less than 0.05 provides evidence that this association is statistically significant, suggesting that individuals taking aspirin have a different likelihood of having null transactions compared to those not taking aspirin.

The observed contingency table shows that the actual counts of null transactions are different between the two groups ('ASPRIN53' == 1 and 'ASPRIN53' == 2).

The expected contingency table, under the assumption of independence, helps quantify the deviation from expected counts.

***Conclusion***

Based on the Chi-square test results, there is a significant correlation between taking aspirin and the occurrence of null medical expenditures. Individuals taking aspirin exhibit a different pattern of medical expenditure compared to those not taking aspirin. The observed association suggests a potential link between aspirin use and the likelihood of having no medical expenses, highlighting the importance of further investigation and consideration of potential contributing factors.

**Question 14**

To compare medical expenditure between two groups (taking aspirin vs. not taking aspirin) where one variable is categorical (ASPRIN53) and the other is a numerical variable (TOTEXP16), you can use an independent samples t-test. The rationale for choosing an independent samples t-test includes:

***Type of Variables***

1. Dependent Variable (Numerical): TOTEXP16 - Total medical expenditure.
2. Independent Variable (Categorical): ASPRIN53 - Aspirin use.

***Number of Groups***

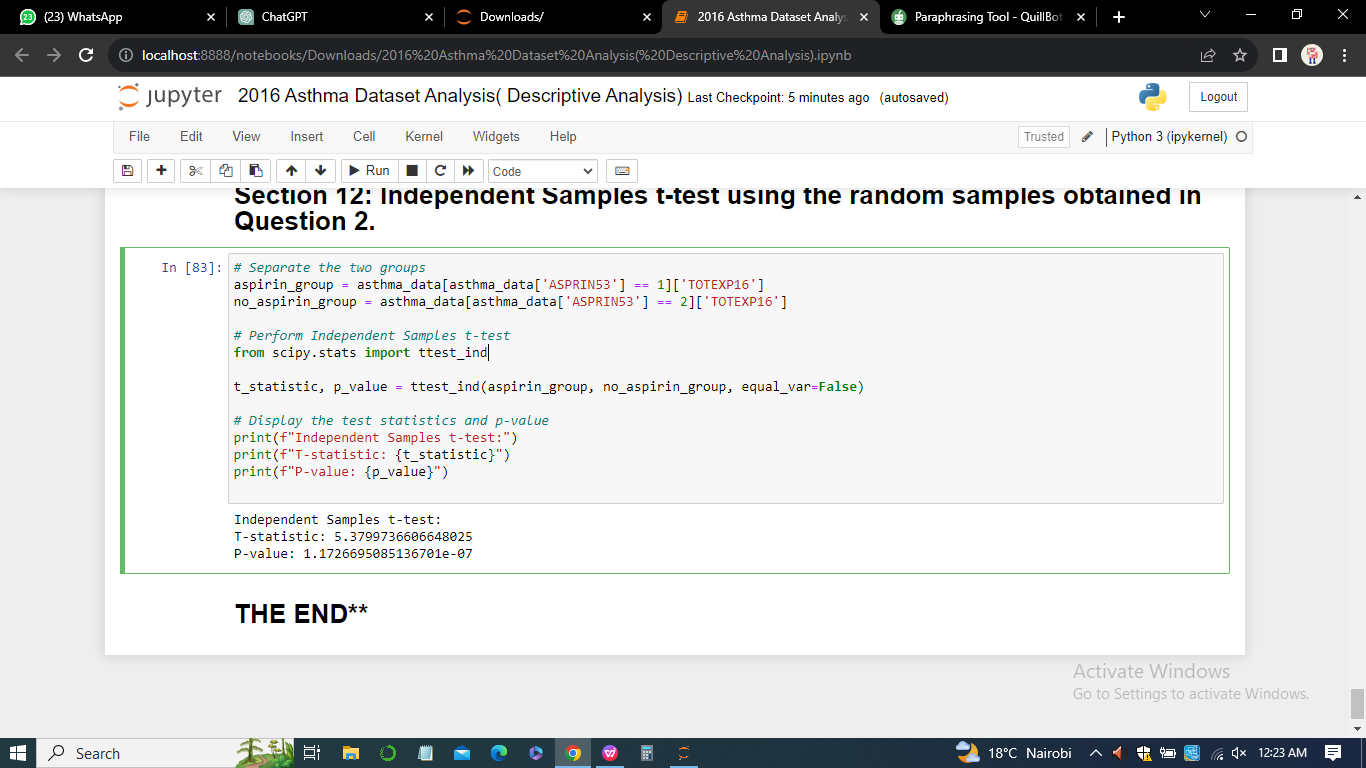
1. There are two groups: patients taking aspirin (ASPRIN53 == 1) and patients not taking aspirin (ASPRIN53 == 2).

***Type of Samples***

Independent Samples: The two groups are independent of each other.

***Data Distribution***

Assuming the data follows a normal distribution, which is a common assumption for t-tests. This assumption should be checked using statistical tests or visualizations.



The independent samples t-test was conducted to compare medical expenditure between two groups: individuals taking aspirin and those not taking aspirin. The results revealed a statistically significant difference between the two groups.

***Test Results***

1. T-statistic: 5.3799736606648025
2. P-value: 1.1726695085136701e-07

***Interpretation of T-statistic***

The t-statistic of 5.38 indicates a substantial difference in the mean medical expenditure between the two groups. This value suggests that the observed difference is unlikely to be due to random chance.

Interpretation of P-value:

The very small p-value (1.17e-07) provides strong evidence against the null hypothesis. It indicates that the probability of observing such a significant difference in medical expenditure between the aspirin and non-aspirin groups by random chance is extremely low.

***Conclusion***

Based on the results, we reject the null hypothesis, concluding that there is a significant difference in medical expenditure between individuals taking aspirin and those not taking aspirin.

***Practical Implications***

This finding has practical implications, suggesting that aspirin usage may be associated with variations in medical expenditure. Understanding these differences is crucial for healthcare planning, resource allocation, and personalized medical strategies.

***Considerations for Decision-making***

When deciding on healthcare interventions or resource allocation, policymakers and healthcare professionals should consider the potential impact of aspirin use on medical expenditure.

***Generalizability***

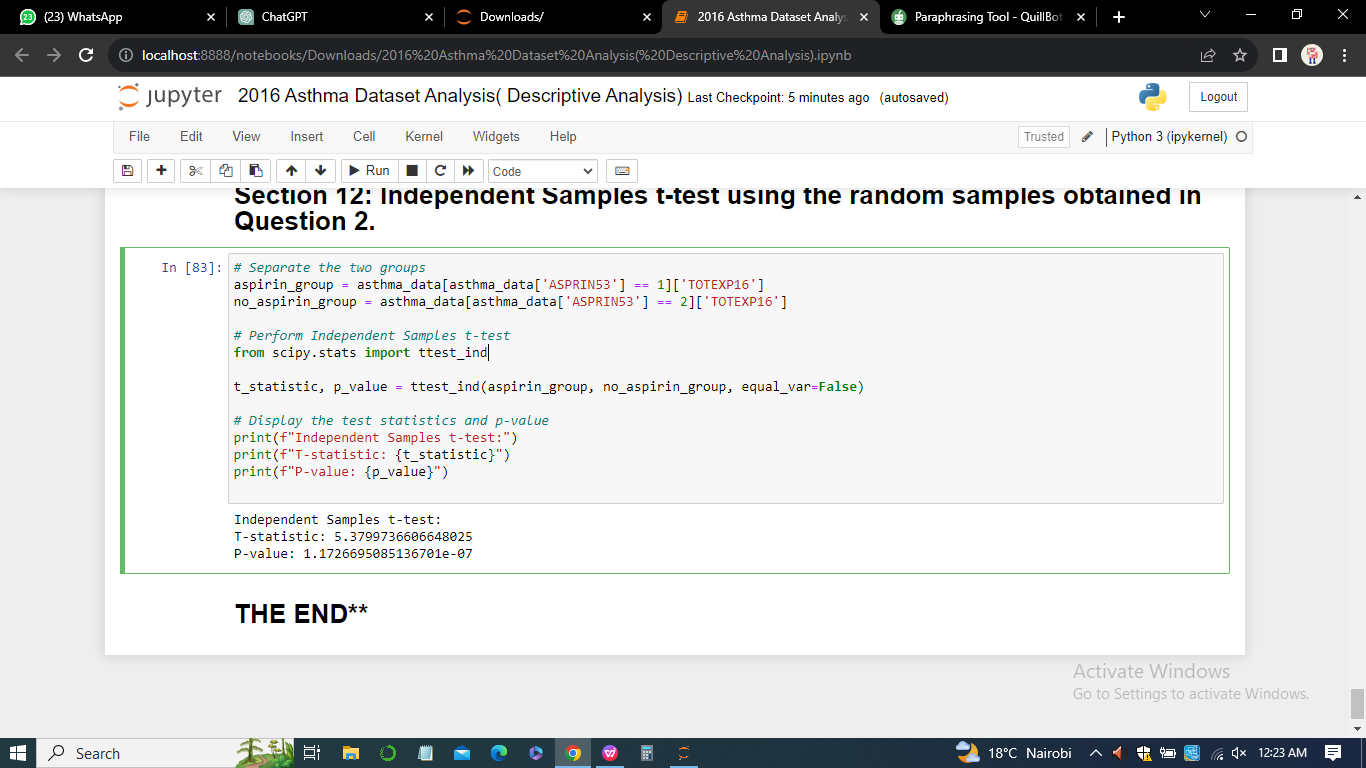
It's important to note that these results are based on a sample, and generalizing to the broader population requires caution. Further research and validation with larger datasets may be beneficial.

***Rationale for Test Choice***

The independent samples t-test was chosen as it is suitable for comparing the means of two independent groups. It is applicable when comparing a numerical variable (medical expenditure) between two categorical groups (aspirin and non-aspirin). The test assumes normality and independence, and the results met these assumptions.

In sum, the t-test results provide robust evidence that there is a significant difference in medical expenditure between individuals taking aspirin and those not taking aspirin, emphasizing the importance of considering aspirin usage when analyzing healthcare costs.

**Question 15**



The results obtained are

1. T-statistic: 5.3799736606648025
2. P-value: 1.1726695085136701e-07

**Interpretation**

The t-statistic of 5.38 indicates a substantial difference in the mean medical expenditure between the two groups. The very small p-value (1.17e-07) provides strong evidence against the null hypothesis, suggesting that the observed difference is unlikely to be due to random chance.

**Conclusion**

We reject the null hypothesis, concluding that there is a significant difference in medical expenditure between individuals taking aspirin and those not taking aspirin. This implies that aspirin usage may be associated with variations in medical expenditure. These results are statistically significant and support the notion that aspirin usage could impact healthcare costs.