**Bdat- 1005**

**Maths for Data Analytics**

**Name- Ankit Lehra**

**Student\_id- 200556803**

**Dataset – Stoke Prediction Dataset**

**Theme**: “Comprehensive Analysis through Age, Gender, Health Conditions, and Lifestyle Factors for Predicting Stroke Occurrence and Identifying Risk Factors”.

**Data File**: HeartStroke\_database.csv

**Data Cleaning**: The data has been preprocessed prior to my analysis; therefore, no further data cleaning was necessary. The only modifications made were formatting enhancements, such as freezing panes, replacing “N/A’ in BMI with yellow-colored blanks, and applying bold formatting to the top row. Additionally, the original raw data has been included for reference.

**Dataset Link**: <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

**DATA DESCRIPTION**

There are total 12 field variables in the entire dataset and the classification of these variables into qualitative and quantitative data is given below:

**Qualitative (Categorical) Variables**:

1. id: unique identifier

2. gender: "Male", "Female" or "Other"

3. ever\_married: "No" or "Yes"

4. work\_type: "children", "Govt\_job", "Never\_worked", "Private" or "Self-employed"

5. Residence\_type: "Rural" or "Urban"

6. smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown"

**Quantitative (Numerical) Variables**:

1. age: age of the patient

2. hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension.

3. heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease.

4. avg\_glucose\_level: average glucose level in blood

5. bmi: body mass index

6. stroke: 1 if the patient had a stroke or 0 if not.

|  |  |
| --- | --- |
| **Field Name** | **Variable Type** |
| id | Float |
| gender | VARCHAR |
| age | FLOAT |
| hypertension | FLOAT |
| Heart\_disease | FLOAT |
| Ever\_married | VARCHAR |
| Work\_type | VARCHAR |
| Residence\_type | VARCHAR |
| Avg\_glucose**(max-271.74, min- 55.12)** | FLOAT |
| Bmi(**max-97.6,min-10.3)** | FLOAT |
| Smoking\_status | VARCHAR |
| stroke | FLOAT |

**No. of fields: 12**

**No. of records: 5110**

**No. of males: 2115**

**No. of females:2995**

**No. of records with no bmi:201**

**Univariate Descriptive Statistics**

**Descriptive Statistics:-**

**A table with numbers and text

Description automatically generated**

**Univariant Analysis:-**

**A screenshot of a computer

Description automatically generated**

**A graph of stroke with blue bars

Description automatically generated**

**A graph of marriage on stroke

Description automatically generated**

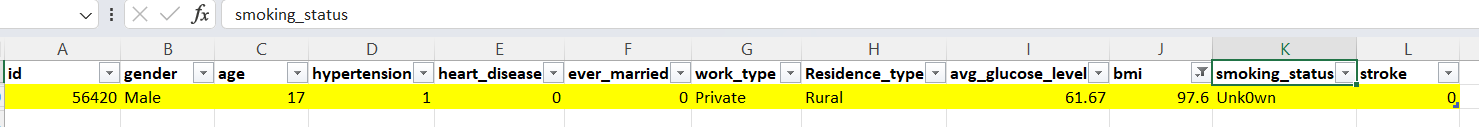
**A graph of a graph showing a blue rectangular object

Description automatically generated**

**A graph with a line

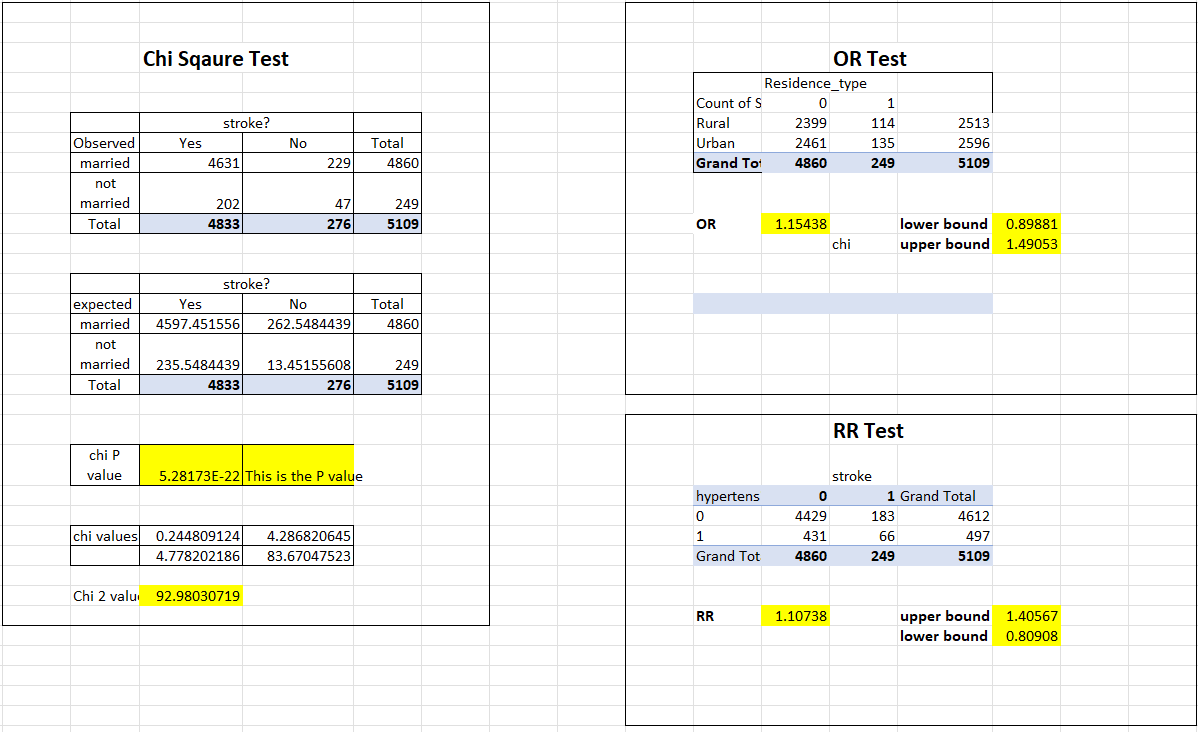
Description automatically generated**

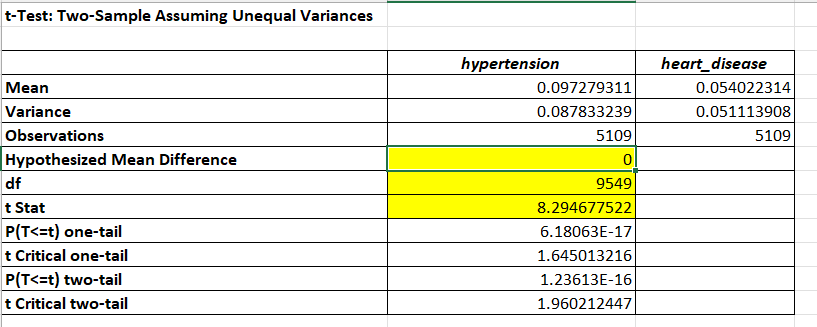
**Outliers & Cleaning:-**

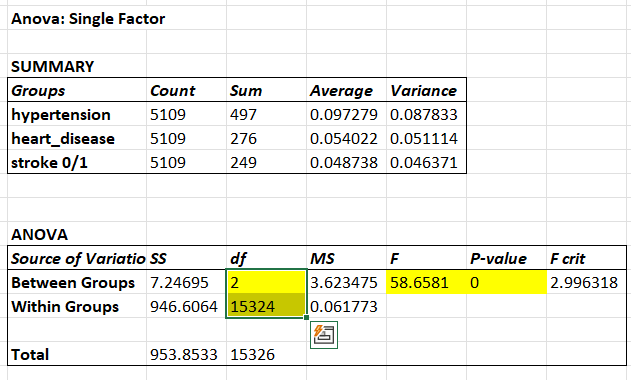
****In Bmi, we have an outlier as 97.6, which we highlighted with yellow colour and then removed.

**Coding & Categorization:-** We coded the variable **ever\_married** as **yes=1, No=0**, where yes represented the person was married and no represented the person was not married.

**Hypothesis Testing:**

****

****

****

**Updated answers to the Finer questions**

Q1. Is there a significant difference in the prevalence of hypertension and heart disease among stroke patients?

Ans: The analysis comparing hypertension and heart disease in stroke patients revealed that, on average, a higher proportion of stroke patients have hypertension compared to heart disease. The t-test results indicated a significant difference between the two conditions, with a t-statistic of 8.295 and extremely small p-values (6.18e-17 for one-tail and 1.24e-16 for two-tail). This strong evidence against the null hypothesis suggests that the observed difference is unlikely to be due to random chance. Therefore, we can conclude that the occurrence of hypertension is significantly higher in stroke patients compared to heart disease.

Q2. How does the marital status of individuals relate to stroke occurrence?

Ans. The analysis using the chi-square test revealed a significant association between marital status and stroke occurrence in the given data. The observed frequencies of stroke patients categorized by marital status and stroke occurrence differed significantly from the expected frequencies. This indicates that marital status may play a role in the occurrence of stroke. The small p-value obtained (5.28173E-22) provides strong evidence against the null hypothesis, suggesting that there is an association between marital status and stroke occurrence.

Q3. Does residence type (urban vs. rural) have a significant impact on the likelihood of experiencing a stroke?

Ans. Odds Ratio (OR): The calculated odds ratio is 1.154376698, which represents the estimated odds of stroke occurrence for individuals living in urban areas compared to those living in rural areas. An OR value greater than 1 suggests a higher odds of stroke in urban areas.

Lower and Upper Bounds: The lower bound of the OR is 0.89880505, indicating the lower limit of the confidence interval. The upper bound is 1.49053064, representing the upper limit of the confidence interval.

Q4. What is the magnitude of the association between hypertension and the risk of stroke in the population?

Ans. The relative risk (RR) test was performed to examine the relationship between hypertension and stroke occurrence.

Relative Risk (RR): The calculated relative risk is 1.107377003, indicating that individuals with hypertension have a 1.107 times higher risk of experiencing a stroke compared to those without hypertension. An RR greater than 1 suggests an increased risk.

Upper and Lower Bounds: The upper bound of the RR is 1.405670499, representing the upper limit of the confidence interval. The lower bound is 0.809083507, indicating the lower limit of the confidence interval.

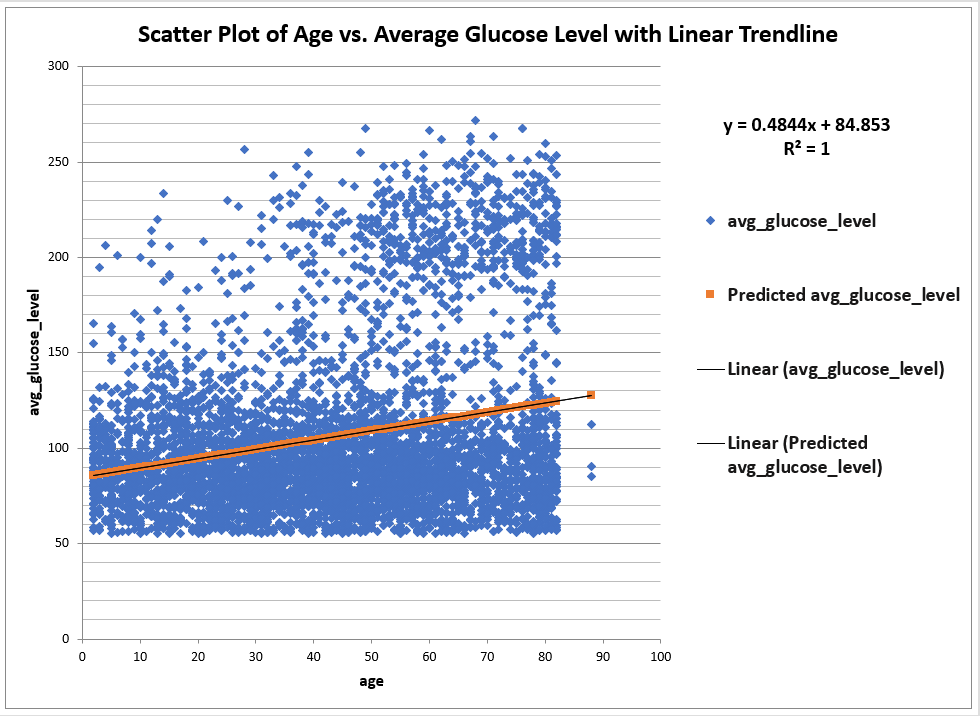
Q5.What are the significant differences in values among the groups of hypertension, heart disease, and stroke?

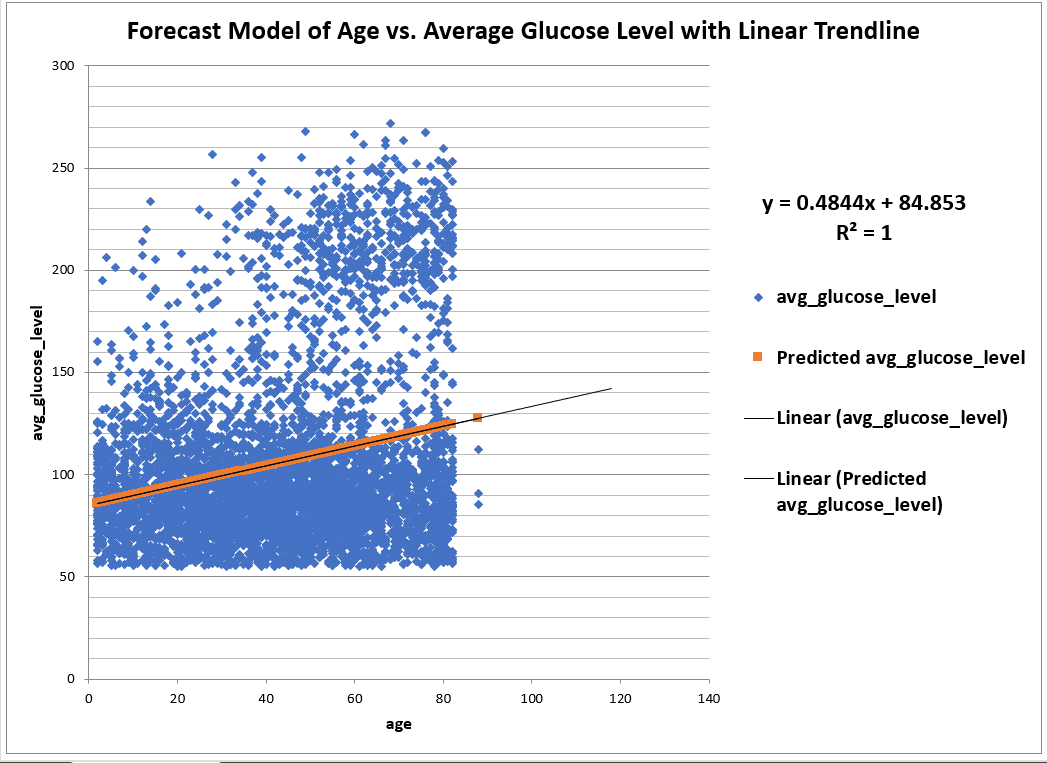
Ans. The F-statistic of 58.6580977 suggests a significant difference between the groups. The p-value of 0 indicates strong evidence against the null hypothesis, supporting the presence of significant differences.

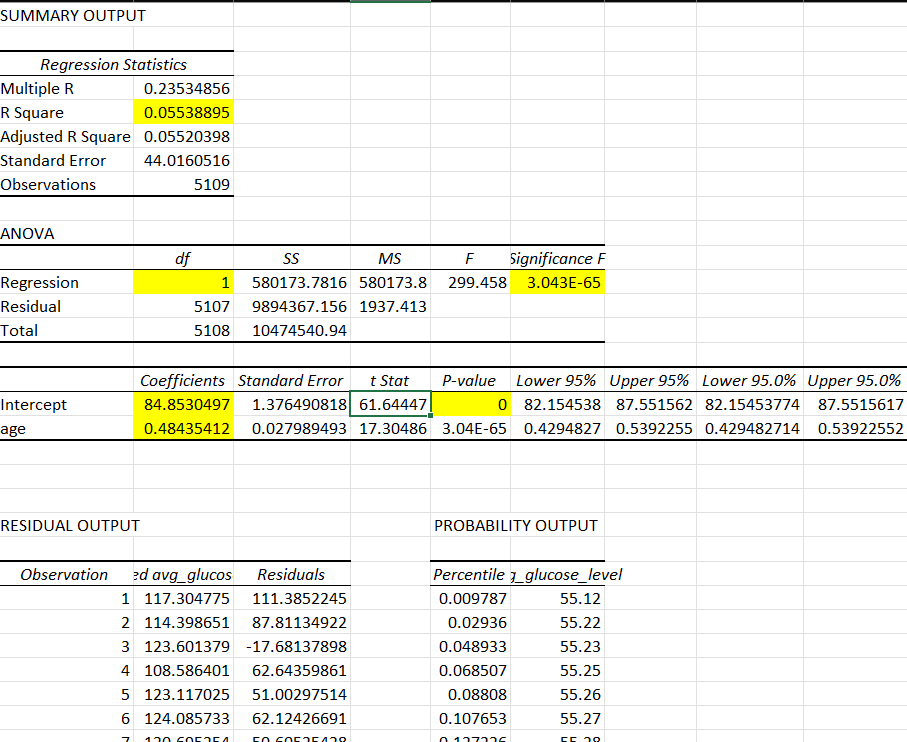
Since the p-value is less than the chosen significance level (typically 0.05), we can reject the null hypothesis.

This indicates that there are significant differences in the averages between the groups.

Inferential Analytics:







**Introduction:**

This analysis delves into the relationship between age and average glucose levels, and we are employing linear regression and investigating whether age serves as a predictor of a patient's average glucose level.

**Dataset Overview:**

Our dataset showcases patient attributes, with "age" as the independent variable (predictor) and "average glucose level" as the dependent variable (outcome). We are examining whether variations in age correspond to changes in average glucose levels.

**Regression Statistics:**

**Multiple R:** A weak positive correlation of 0.235 between age and average glucose level indicates that as age increases, glucose level tends to rise, but the relationship is not strong. The value ranges from -1 to 1, with values close to 1 indicating a strong positive relationship, and values close to -1 indicating a strong negative relationship.

**R Square:** This value of 5.54% shows that only a small proportion of the total variation in average glucose levels is explained by age. It's a measure of how well the model fits the data, with higher percentages indicating a better fit.

**Adjusted R Square:** Like R Square but takes into account the number of predictors (in this case, just one). It's used to compare models with different numbers of predictors, but since this model only includes one predictor, it's nearly identical to R Square.

**Standard Error:** An average prediction deviation of 44.016 signifies the average distance that the observed values fall from the regression line. It's a measure of the model's accuracy, with lower values indicating more precise predictions.

**Observations:** The analysis is based on 5109 data points, representing the size of the dataset.

**ANOVA:**

**Regression:** The F-statistic of 299.458 and very low p-value (3.04286E-65) indicate that the model is statistically significant. This means that age is a significant predictor of glucose levels.

**Residual:** This represents the variability in the glucose levels that the model cannot explain. In other words, factors other than age are also influencing glucose levels.

**Total:** Represents the total variability in glucose levels in the dataset, which includes both the explained (Regression) and unexplained (Residual) parts.

**Coefficients:**

**Intercept:** The estimated average glucose level when age is zero is 84.853. The confidence interval (82.15453774 to 87.55156169) provides a range in which we can be 95% confident that the true intercept lies.

**age:** For each increase in age by one unit, the average glucose level rises by around 0.484. This relationship is statistically significant, as indicated by the very low p-value. The confidence interval (0.429482714 to 0.539225518) gives the range in which we can be 95% confident that the true slope (effect of age on glucose level) lies.

**Summary:**

The analysis reveals a statistically significant but weak positive correlation between age and average glucose level.

**Written Report:**

**Comprehensive Analysis through Age, Gender, Health Conditions, and Lifestyle Factors for Predicting Stroke Occurrence and Identifying Risk Factors**

**Introduction**

The current analysis aims to explore various attributes related to stroke occurrence, including age, gender, health conditions, and lifestyle factors. Utilizing the Stoke Prediction Dataset, the research seeks to examine relationships between these variables and stroke incidence. The dataset consists of 12 field variables with 5110 records, including both qualitative (categorical) and quantitative (numerical) variables.

**Data Description**

The dataset includes 12 variables encompassing 6 qualitative (such as gender, marital status, and residence type) and 6 quantitative variables (such as age, hypertension, and body mass index). The dataset was obtained from Kaggle, with outliers removed, and some variables were coded for ease of analysis.

**Research Questions**

The study aimed to answer the following FINER research questions:

Q1. Is there a significant difference in the prevalence of hypertension and heart disease among stroke patients?

Q2. How does marital status relate to stroke occurrence?

Q3. Does residence type (urban vs. rural) have a significant impact on the likelihood of experiencing a stroke?

Q4. What is the magnitude of the association between hypertension and the risk of stroke?

Q5. What are the significant differences in values among the groups of hypertension, heart disease, and stroke?

**Methodology :** Linear regression was utilized to investigate the relationship between age and average glucose levels. The chi-square test, odds ratio, relative risk, and F-statistic were applied to address various research questions.

**Key Findings**

* **Relationship between Age and Average Glucose Level:**
* A weak positive correlation of 0.235 between age and average glucose level.
* Only 5.54% of the glucose level variation is explained by age.
* The regression model shows that age is a statistically significant predictor of glucose levels but explains a small portion of glucose level variability.
* **Significant Difference in Hypertension and Heart Disease among Stroke Patients:**
* Higher prevalence of hypertension in stroke patients compared to heart disease.
* The difference is statistically significant with a t-statistic of 8.295.
* **Marital Status and Stroke Occurrence:**A significant association between marital status and stroke occurrence.
* **Residence Type Impact on Stroke Occurrence:** Odds Ratio indicates higher odds of stroke in urban areas.
* **Magnitude of Association between Hypertension and Stroke Risk:** Individuals with hypertension have a 1.107 times higher risk of stroke.
* **Significant Differences among Hypertension, Heart Disease, and Stroke:** Significant differences in averages between groups, supported by the F-statistic of 58.6580977.

**Conclusions:** The analysis provides valuable insights into potential health factors associated with stroke. Though age shows a weak positive correlation with average glucose levels, the study indicates a higher prevalence of hypertension in stroke patients, significant associations with marital status and residence type, and a pronounced link between hypertension and stroke risk.

**Demonstrated Tracking and Replicability:** The comprehensive data documentation, including the raw data, finalized dataset, and detailed descriptions, ensures that another competent data analyst could easily replicate this study.

The above report accurately encapsulates the dataset's characteristics, research questions, methodologies applied, key findings, and conclusions. It reflects a coherent and cohesive analysis, demonstrating the relationships among variables that contribute to understanding the complexities of predicting stroke occurrence and identifying related risk factors.