EDA LOG			
ID	**Date**	**Description**	**Observations and Insights**
1	6/13/2025	Context of under 65 focus	1. Stroke cases make up **4.87%** of the original dataset. 2. Of the 249 total stroke cases, **159** (**≈64%**) are patients aged 65 and over—indicating that this age group heavily dominates stroke representation. dominantly represent stroke cases. 3.By excluding patients 65 and over, we reduce the **overwhelming effect** of age and allow other risk factors to emerge more clearly. 4. After removal, stroke cases under 65 will make up **2.20%** of the remaining dataset.
2	6/13/2025	Age column number summary analysis	1. After data filtering, minimum age is 0.08 and max is 64. 2. **No outlier** detected using 1.5 x IQR bounds.
3	6/13/2025	Average glucose level number summary analysis	1. Average glucose level shows large outliers over the 1.5 x IQR bound. 2. Upon research, glucose levels can be as extreme as over 600; thus, outliers were deemed **possibly valid** and not removed.
4	6/13/2025	Bmi column number summary analysis	1. The number summary for bmi shows large value **outliers** when using 1.5 x IQR upper bound. 2. Upon investigation, bmi of 97 or more is **rare** but possible. Thus, data with large values were **preserved**.
5	6/13/2025	Impute null bmi values	1. The bmi column contains large outliers, making the **mean** an **unsuitable** choice for imputation. The median value (**27.7**) was selected as it provides a more robust and less disruptive replacement for nulls. 2. After imputation, summary statistics such as mean and IQR shifted slightly, but not enough to compromise data integrity or require further adjustment.
6	6/13/2025	Bin Continuous Features	Based on CDC Standards, the age group was binned into: * children(0-17), young adult(18-24), adults(25-34), midlife adults(34-44), older adults(45-54), pre-seniors(55-64) Based on CDC standards, avg_glucose_level was binned into: * hypoglycemic(<70), normal (70-99), pre-diabetic (100-125), diabetic (126-199), and high diabetes (200+) Based on CDC standards, bmi was binned into: * underweight, normal weight, overweight, obesity class 1, obesity class 2, and obesity class 3
7	6/13/2025	Gender Distribution Analysis	1. Females represent **58.40%** of the under 65 population and accounts for **53.33%** of the stroke cases, the stroke rate is skewed because of the *difference in population representation* and not overall risk. 2. **Males have a higher stroke rate** compared to women within their respective groups (**2.47% vs 2.01%**). 3. These findings suggest that **men under 65** may *require more targeted early screening* compared to women.
8	6/13/2025	Age Group Distribution Analysis	1.The age distribution of patients under 65 is fairly even, except for **young adults (9.31%)**, who represent roughly half the size of other groups. 2. **In-group stroke risk increases steadily with age**, peaking at **7.05**% among pre-seniors. Children (**0.23%**) and young adults (**0%**) show very low stroke rates. 3. **Pre-seniors (55–64) account for 58.89% of all stroke cases under 65**, reinforcing the dominance of age as a stroke risk factor, even before age 65.

9	6/14/2025	Engineer Categorical Column Values	1. Added hypertension status, heart_disease_status, ever_married_status, and stroke_status columns. 2. Replaced 0/1 and no/yes values to more interpretable version (no hypertension/hypertension, no heart disease /heart disease, never married/married, no stroke/ had stroke.
10	6/14/2025	Work Type Distribution Analysis	1. Private job workers make up **59.04%** of the population. 2. Self-employed(**2.97%**) and government job (**3.04%**) workers have a **slightly higher** stroke rate compared to private company workers (**2.45%**). 3. Although private company workers account for 65.56% of stroke cases, this is likely **skewed by their large share** of the total population.
11	6/14/2025	Residence Type Distribution Analysis	1. The distribution for both categories are similar (50.37% vs 49.63%) 2. Urban residents (2.33%) have a slightly higher stroke rate than rural residents (2.07%). 3. Urban residents (53.33%) represent a higher stroke occurrence rate than rural residents (46.67%) overall. 4. The trend largely follows the population distribution where urban residents outnumber the rural residents.
12	6/14/2025	Marriage History Distribution Analysis	1. Married people (**59.14%**) account for 20% more than the never married people (**40.86%**). 2. Married people have a **3.36%** risk of having a stroke, *compared to other married people* , while people that are never married experience stroke at a rate of **0.54%** *compared to other never married people* . 3. Married individuals account for 90% of all stroke cases in the dataset.
13	6/14/2025	Smoking Status Distribution Analysis	1. Never smoked and unknown status patients represent roughly **69%** of the data. 2. Having a history of smoking (smokes, formerly smokes) showed a **higher stroke rate (3.79%** and **3.76%**). 3. Smoking or formerly smoked shows **almost double** the rate of stroke occurrence compared to the other groups. 4. Smokes and formerly smokes contributed roughly the **same number of stroke occurrences** while being roughly **half the size** of the other two groups.
14	6/14/2025	BMI Distribution Analysis	1. Patients within the normal and overweight categories make up **~55%** of the population. 2. Patients in the **normal weight** range are **7 to 8 times less likely** to experience stroke. 3. Patients that are **above normal weigh**t have a stroke rate of around **3%**. 4. Patients in the **overweight category** represent **42.22%** of the overall stroke cases while making up **28.47% of the population**.
15	6/15/2025	Average Glucose Level Distribution Analysis	1. Patients with **normal** average glucose level account for **48.19%** of the population and **36.67%** of the total stroke cases. The dominant representation of this group accounts for the higher stroke case occurrence. 2. Patients with **high diabetes** is the least represented at **5.59%** but it accounts for **18.89%** of stroke cases which is the second highest representation of stroke overall. Patients with high diabetes also shows an in group stroke rate of **7.46%,** which is **more than twice the rate of other groups**.
16	6/15/2025	Heart Disease Distribution Analysis	1. Patients with heart disease only represents **2.35%** of the population while accounting for **14.44%** of overall stroke cases. 2. Patients with heart disease have a stroke rate of **13.54%** or about 6x more than patients without heart disease **(1.93%**).

17	6/15/2025	Hypertension Distribution Analysis	1. Although patients with *hypertension* only make up **6.57%** of the population, they account for **17.78% of all stroke cases**. 2. Patients with hypertension experiences stroke at a rate of **5.97%**, which is **more than three times** higher than the **1.94%** stroke rate among patients without hypertension.
18	6/15/2025	Chi-Square Test	1. The **Chi-Square Test** is was used to evaluate the **statistical significance** of categorical features in relation to the target variable (stroke). 2. Both **gender and residence_type** had p-values that were **greater than the 0.05** significance threshold, indicating **no statistical significant association** between these two features and stroke in patients under 65. 3. As a result, gender and residence type will be **excluded from further analysis**.
19	6/16/2025	Mann-Whitney Test	1. Excel does not natively have a Mann-Whitney Test so **Python** was used to conduct this test. 2. Both **bmi** and **average glucose level** tested to have a p value that is significantly **lower than the 0.05** threshold, indicating that they are both **statistically significant predictors of stroke** for patients under the age of 65.
20	6/16/2025	Top In-Group Risk Factors	The **top 5** in group stroke risk factors are: 1. Heart Disease at **13.54%** 2. High Diabetes at 7.46% 3. Pre-Senior age at 7.05% 4. Hypertension at 5.97% 5. Smokes at 3.79%
21	6/16/2025	In Group Stroke Rate of Change	1. Heart disease has the highest in-group stroke rate at 13.54%. 2. The decline in rate is **drastic until the fifth factor** (smokes, **3.79%**) and then the decline rate slows down.
22	6/16/2025	Disproportionate Stroke Burden	The bar chart shows the following insights: 1. Pre-seniors account for **58.89%** of the stroke cases while only being **18.42%** of the population under 65. 2. Heart disease accounts for **14.44%** of the stroke cases while only being **2.35**% of the population.
23	6/16/2025	Age - Heart Disease Distribution	1. Pre-seniors (55-64) have the highest heart disease rate at 7.71% 2.There is a **drastic increase** in heart disease diagnosis starting with **midlife adults (0.58%)**, rising to **older adults (3.88%)**, and peaking with **pre-seniors (7.71%)**.
24	6/16/2025	Age - Hypertension Distribution	1. Pre-seniors have the highest rate of hypertension at **15.56%.** 2. The rate of hypertension diagnosis starts **increasing drastically in adulthood (25-34)**. 3. **Screening and preventive care for hypertension should begin in the mid-20s**, long before hypertension manifests.
25	6/16/2025	Age - Diabetes Distribution	1. Diabetes affects **1 in 4** (26.60%) **pre-seniors** (55-64). 2. **Drastic increase in diabetes rate** starts at **adult age** (25-34) at **9.38%** and **doubles** at older **adult stage** (45-54) at **20.30%**. 3. **Screening and preventive care plan** should be considered for patients **starting mid 20s**.
26	6/16/2025	Age - Smoking Distribution	1. **Stroke risk remains elevated even after quitting**, prevention must start early. Former smokers show **similar stroke risk** to current smokers, emphasizing that **not starting at all** is the most effective protection. 3. **Most smokers begin before age 35**. With smoking history **rising sharply between ages 18 and 34**, prevention campaigns must target **young adults and adolescents** before lifetime risk is locked in.