**Phase II: Data Exploration**

Prevalence of a stroke condition on patient population:

* Pie chart shows that 98% of patients are healthy and only 2% of patients have a stroke condition. Thus, this suggests need for adjustment on data imbalance upon training a ML model.

Count frequencies grouped by stroke condition on lifestyle and health indicator factors:

* **Gender**: stroke is not a gender specific condition. However, among patient populations with a stroke condition there are more female patients (55%) with a stroke condition than man (45%).
* **Marital Status**: interesting enough, there are more stroke patients with a married status (90%) than a single status (10%).
* **Residence Type**: residential area is not an important factor as almost equal proportion patients have stroke condition who reside in rural (49%) or urban area (51%).
* **Occupation Type**: work type is a quite interesting feature. As it demonstrates that there is high tendency for patients work in private sector (56%) or a self-employed (32%) have a stroke condition than patients work in government sector or children (i.e., kindergarten).
* **Smoking Status**: smoking status seems weekly associated with a stroke condition. As it shows that group of non-smoking patients (55%) have higher chances of having a stroke than group of smoking patients (45%).
* **Hypertension**: hypertension is not a significant determinant factor on stroke condition. It clearly shows that patients with no hypertension (74.5%) have more strokes than a group of patients with hypertension (25.5%).
* **Heart Disease**: heart disease is not a significant factor as well. As group of patients with no heart disease have a stroke (74.5%) than patients with heart disease condition (25.5%).

Distribution of stroke patient population: age, bmi and avg\_glucose\_level

* **Age**: a histogram shows that age distribution of patient population is non-uniformed. Also, it shows some tri-modal characteristics (i.e., two peaks on lower and upper end) suggesting that decent number of stroke patients from infants to child (i.e., less than a year to 5-year-old) and seniors (i.e., 80-year-old). However, most of patients are around mid-30s to 60s.
* **Bmi**: a histogram shows that quite a normally distributed patient population. Most patients bmi is centralized around range of 25 to 30.
* **Avg\_glucose\_level**: a histogram suggested quite a positive skewed (right) distribution of patient population for average glucose level. Many stroke patients are on lower end (i.e., 75 to 100) whereas some patients have extreme high level of average glucose than entire patient population.

Bar chart of patient population age on stroke condition by lifestyle and health indicator factors:

* Regardless of all bar charts are conditioned by different factors, in general patients with a stroke condition are older (i.e., aging) than normal patients. Mean age of stroke patients population is around age 65 to 70 except for very few population who are less than 10-year-old suggested by one of bar chart constructed by work types.

Correlation matrix plots on entire training set

* There are some interesting correlation patterns do exist among features. First, age and bmi show a positive moderate correlation (i.e., 0.39). Older patients tend to have higher bmi. Second, age and glucose show a positive weak correlation (i.e., 0.24). As patients become older, their average glucose level becomes higher. In addition, there is slightly stronger positive correlation with age and stroke condition (i.e., value of 0.16) in comparison to other numerical features (i.e., bmi and glucose level) correlated with stroke condition. Therefore, it is hard to make any generalization on trends or patterns about health monitoring features (i.e., age, bmi, glucose level) and stroke condition.

Experiment: down-sampling to conduct in depth exploratory analysis

* As discussed above, to confirm presence of whether any strong predictors (i.e., features) exist on train set, down-sampling was performed. In training set, there were 98% of non-stroke and only 2% of stroke cases present in total of 42512 observations. Thus, data imbalanced on label must be addressed for analyzing better insights. Thus, down-sampling was done to reduce number of non-stroke patient cases and balanced the proportion of non-stroke and stroke patient cases to be 1:1 ratio.

Correlation matrix plots on down-sampled training set

* As expected, the correlation between age vs. stroke becomes stronger with a value of 0.61. Also, avg\_glucose\_level vs. stroke shows higher correlation with a value of 0.24 than before (i.e., 0.08).
* Conversely, correlation among each numerical feature (i.e., age vs. bmi) reduced due to down-sampling which resulted in loss of signal from quite a lot of non-stroke patient cases.

Faceted scatter plots by demographic factors:

* **Age vs. bmi**: faceted scatter plot by gender suggests that both male and female have similar trends. It does show very weaker correlation trends whereas patients become older, their bmi increased.
* **Age vs. avg\_glucose\_level**: faceted scatter plot by gender suggests that both male and female have similar trends. It does show very weaker correlation trends whereas patients become older, their avg\_glucose\_level increased.