**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 (77.4%) than patients with heart disease condition (22.6%).

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 stroke 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.

Correlation matrix plot 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., 0.35). There is increased in bmi from younger age to mid age patient population. Second, age and glucose show a positive weak correlation (i.e., 0.24). Although trend is not so clear but as patients get older, their average glucose level becomes higher within certain age range. In addition, there is a weak 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) correlation against a 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.

Correlation matrix plot on stroke patients only

* Thus, author decided to split stroke patients from non-stroke patient population for gaining better insights on feature interaction. Hidden trends were revealed that. First, there is a weak negative correlation between age and bmi. Thus, for stroke patients as they become older, bmi is reduced. Also, a correlation between bmi and avg. glucose level increased barely minimum compared to entire population of patients.

Bar chart of mean age on stroke patients by lifestyle and health indicator factors:

* Regardless bar charts being conditioned by different factors, mean age of stroke patients was around mid-60. Except for occupation type, where mean age of patients who worked in children related sector was significantly low (i.e., mean age less than 10-year-old).

Age group stratification on stroke patient population

* Since author found the fact that “age” is the most important feature. For in depth analysis, age is stratified into different age group range pre-defined by author into following buckets: x<30, 30-40, 40-50, 50-60 etc. The following age group distribution was identified: 42% (age 70 to 80), 22% (age 60 to 70), 16% (age 50 to 60), 10% (age 80 to 90) and remaining 10%. Thus, almost 80% of stroke patients is distributed in age group of 50 to 80.

Experiment: down-sampling to conduct in depth exploratory analysis

* As discussed above, to build a model which can classify on stroke patients rather than non-stroke patients, resampling was performed. Originally, there were 98% of non-stroke and only 2% of stroke cases represent in 42,512 of total observations. Thus, down-sampling with randomized replacement was performed to reduce the sample size of non-stroke cases and adjust proportion of classes being 1:1 ratio by matching exact sample size (2,142) of stroke cases to non-stroke cases.

Faceted scatter plots by demographic factors on entire training set:

* **Age vs. bmi**: faceted scatter plot by a stroke condition suggests that there is a weak association of age and bmi. As patients get older, bmi is slightly increased which holds true on certain age range.
* **Age vs. avg\_glucose\_level**: faceted scatter plot by a stroke condition suggests there is a weak association of age and avg. glucose level. As patients get older, there is slight increased in avg. glucose level over time.