Introduction:

According to the CDC over 800,000 people in the United States have a stroke each year. Of these people just under a quarter have had a previous stroke, meaning that a majority of strokes occur in patients who may be unaware they are at risk. Strokes cost the economy an estimated 34 billion dollars so there is a large monetary incentive to prevent strokes before they occur. With this in mind, it would be valuable to have a statistical model that determines who is at risk of having a stroke.

Some potential clients for this model would be life and health insurance companies and government organizations. Life insurers could use the information to more accurately quantify the risk involved in their policies. If our model shows that someone is at risk for a stroke their life insurance policy will have an increased premium. Health insurance providers can use the model to offer preventive care treatments to those at risk which will save them money as opposed to expensive emergency medical services. Government organizations would be interested in the model to aid in their data collection and prediction processes. Additionally, they would be able to identify when their jurisdictions have higher levels of risk and allocate capital accordingly.

Using personal and medical data of over 20,000 people I believe we can construct a model that can accurately determine patients with a high risk of stroke. This data includes attributes such as gender, age, BMI, average glucose level, whether the subject has been married, presence of hypertension, presence of heart disease and a history of stroke. Using this information is ideal because these attributes are all for the most part, easily identifiable. Easily identifiable attributes are valuable because it increases the usefulness and scope of the model. The only attributes that even requires a doctor to diagnose is heart disease and hypertension. A person only needs to know their height and weight to find BMI and their are apps to find your average blood glucose level. This means that the model can be applied on virtually every person who has their yearly physical done.

Based on the amount of data and the attributes we have for each subject, it is possible to create a model that can help determine who is at risk of stroke. As previously stated this model would have large financial implications as well as being a useful tool for millions of people.While the data is relatively clean because it came from Kaggle we will still have to optimize it in order to achieve the best results. In addition to this we will run some exploratory analysis to see if any features are more predictive than we would expect.

Data Wrangling:

The data from this project came from Kaggle so it did not need as much work as other datasets. The data contained attributes such as an id number, gender, age, average glucose level, BMI, smoking status, type of work, type of residence (urban vs. rural), whether the subject was married, whether they suffered from hypertension and whether they had heart disease. A majority of the work was to optimize rather than clean the data but there were still a decent amount of missing values to deal with. In order to optimize the data, some variables were changed to binary variables. This will simplify future statistical analysis which is the goal of data optimization.

The first step and only real cleaning task was to decide what to do with the missing values for BMI and smoking status. For BMI, the obvious answer was to eliminate the missing values because replacing the missing values with a mean would have drastically misrepresented some cases. One possibility I considered was setting the BMI based on the correlation to average glucose level but with over 20 thousand non-null data points it seemed like an unnecessary risk because the variables share a weak and inconsistent correlation. As for the smoking status feature, further investigation revealed that about a quarter of the missing values were for subjects younger than 19. It is a safe assumption that most of the missing values for these individuals were non-smokers due to their age so anyone who had missing data under smoking status and was under 19 was considered a non-smoker. However, there remains a small percentage of teenage smokers in the dataset. One concerning element of this is that young people have very few strokes and having young smokers included in the dataset may hamper effective analysis and cause a model trained on this data to underestimate the dangers of smoking.. On the other hand, while children and young people are at low risk for stroke, there may be larger economic benefits to using preventative treatments earlier in life because it may help younger people avoid a lifetime of health complications. With this in mind, I decided to create two datasets. One that will exclude anyone under 19 and another that will not. I will examine and make my models based on both datasets and display the results separately. The models will most likely be similar in many ways but their differences will provide insights.

The major optimizations I made to the data were to convert some of the string columns into binary variables. This makes the data more optimal for graphing and statistical analysis. The variables I re-coded were, gender, residence type, smoking status, and ever married. To show what the one and zero would represent I changed the names of the variables to male, urban, has smoked and kept ever married the same. This was self explanatory for the most part, values that were once ‘male’, ‘urban’ and ‘yes’ were now ones and values like female, rural and no were now zeros. The only complication here was that smoking status/ has smoked had three values, ‘smokes,’ ‘formerly smoked,’ and ‘never smoked.’ To solve this a lambda function was implemented in order to encode a one if the subject formerly or currently smokes. This was prudent because smoking has lasting harmful effects. In the absence of a way to quantify the degree of their smoking this was the best course of action.

In addition to optimizing the data I created three new columns that I thought would be helpful in viewing the data graphically. These new columns were, BMI class, glucose class and risk factors. BMI class was created using a BMI chart. The four classes of BMI are underweight, normal, overweight and obese. They each respond to a different range of values and help humans to distinguish the effects of a BMI that is too high or too low. Next we have glucose class. There was not an easily available chart to distinguish glucose levels so I separated the subjects into 5 groups based on what range of percentile they were in. By doing this, we can now see what range of values are best for avoiding stroke. Next, risk factor was created in order to show how frightening the combination of heart disease and hypertension could be. To make this feature, I used a for loop which checked whether the patient suffered from hypertension heart disease, both or neither and saved that result in a pandas series. While these variables are not likely to help a machine learning model they are perfect for making interesting and eye catching visuals.

As previously mentioned we have two datasets, one where subjects are accepted regardless of age and one where only adults (over 18) are accepted. This will allow us to the effects of allowing children into the data. The main reason for this split is the possibility of dubious correlations that only occur in children with stroke, of which there are very few. With that being said, the additional dataset may also lead to some insights that would not be possible otherwise. At this time, an exclusively under 18 dataset is unnecessary and because the occurence of stroke is so low in children it is debatable if a model based on this dataset would even be useful.

EDA and Inference:

The goal of the exploratory analysis was to uncover the most interesting avenues for inferential analysis in addition to uncovering what features will most improve a future model. To this end, tools such as groupby objects, bar graphs, cumulative distribution functions (cdf), histograms decipher the data into a human readable format. Once in the more comprehensible format, we can analyze and determine how to proceed with our analysis.

Our first table groups the data by stroke victims and non victims. This table greatly colored this section of the analysis because it shows what factors are most prevalent in stroke victims. The main differences between victims and non victims in this table were age, presence of heart disease and presence of hypertension. None of these are particularly surprising because we know older people are more susceptible to all kinds of medical conditions and it stands to reason that someone who had heart issues would be more susceptible to a condition that stems from lack of blood in the brain. With the latter point in mind, we should examine the effects of our other features of not only likelihood for stroke but also for heart disease and hypertension. While heart disease has a slightly higher correlation with stroke both factors are important.

To show how risk increases with age, we grouped the subjects into 10 year age containers (20-29, etc). The goal was to examine how the risk of stroke changes as a patient gets older. Before the age of 30 there is nearly no risk of stroke in addition to less than one percent of patients who suffer from heart disease or hypertension, From that point on percentage of patients who suffer from stroke begin to increase rapidly to the point that patients in their fifties are 5 times more likely to suffer a stroke than those in their thirties. The rates of heart disease and hypertension spike similarly as age increases. This lends credence to our earlier statement that older people tend to be more susceptible to many medical conditions. Going forward age will certainly be an effective feature in a future model.

BMI and average glucose level are our only features that directly indicate bodily health, or at least they are supposed to. With this in mind, I was determined to gain useful insight from these variables. While their were not clear correlations between these variables and occurence of stroke, when we created a feature to group the patients by containers based on these variables, we were able to gain some insights. Interestingly, it appears people who are underweight are at a higher risk of stroke than those who are obese. This was a surprising finding and one of the only areas where the two datasets dramatically differ. The dataset that included under 18 year olds shows that underweight people are the least at risk for stroke. One reason for this is that BMI was most likely constructed with adults in mind so it is my suspicion that some children of healthy weight are marked as underweight. Additionally, there are only a few hundred underweight individuals in the over 18 dataset so any conclusions drawn from it may be questionable. In regards to glucose level, it appears that those in the highest bracket of average glucose level are the most at risk. When we plot a cdf of the average glucose levels of victims of stroke and non victims we see that about 40% of stroke victims have a blood glucose level of over 150 while less than 20% of non victims see these same levels. This means that stroke victims tend to have high levels of blood glucose.

Going into the analysis I thought gender and smoking would be somewhat intertwined. I held the (apparently incorrect) belief that men smoked more than women. Our dataset does not show that but it does show that men are more at risk due to a higher prevalence of heart disease. Smoking on the other hand, had less clear implications. In the adults only dataset, it appeared smoking did not have much of an effect and even showed smokers where at less risk for stroke than non smokers. The dataset including under 18 year olds contradicts this. It shows smokers are indeed at an increased risk. This should come as no surprise because a large majority of under 18 year olds have probably not smoked and not had a stroke. Smoking is obviously bad for your health but the contradiction between datasets definitely brings in questions about whether or not smoking will be a useful feature for a machine learning model.

In addition to graphical methods we used permutation sampling to see what effects were random noise and which were real. Unsurprisingly, the effects of hypertension and heart disease are far beyond randomness and significant in real terms with victims being 3-4 times as likely to have a stroke as opposed to those who do not have either condition. The effect of gender appears to be small but also statistically significant. Also unsurprisingly, place of residence appears to be superficial when predicting a stroke. Smoking once again is the most interesting result here because the datasets once again disagree. The adults only dataset shows that smoking’s positive effect is insignificant while the other dataset shows that the negative effects of smoking are significant. I am inclined to believe the latter because it conforms to what I already think but I will be keeping a close eye on this moving forward.