Final Project Report

1. Problem Selection

The problem we would like to solve is to predict strokes based on multiple individual factors. The factors that we will consider will be gender, age, heart disease, marital status, glucose levels, body mass index, smoking status, and hypertension. In addition, we are interested in determining which predictor variables have a large impact on whether or not someone gets a stroke.

We are going to solve this problem by using classification and clustering techniques. Classification techniques will include decision trees and k-nearest neighbors. Clustering techniques will include hierarchical clustering (simple and complete linkage) and k-means clustering. We believe clustering and classification are the most appropriate methods for solving our problem because they divide observations into different classes based on predictor variables. In our case, we want to predict whether a person has had a stroke based on key health indicators.

2. Data Collection

The dataset that we will be using is Stroke Prediction Dataset from kaggle. The dataset includes the stroke prediction based on gender, age, hypertension, heart disease, marital status, work type, residence type, Glucose level, and BMI. We'll be using this dataset to predict which of these factors have the biggest impact on someone getting stroke. The dataset will require preparation since there are some uneed attributes and incorrect formatting of data. That is, we will remove unique id and residence type since they seem irrelevant to what we require. In addition, we will make changes to gender, marital status, and smoking status to numerical observations because currently they are in string format.
Stroke Prediction Dataset: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

3. Data Preparation

To clean the data and transform the data into an appropriate format for data analysis, we decided to drop the columns “id”, “work_type”, and “Residence_Type” from the dataset, since we thought we wouldn't need this information to predict stroke. Moreover, we decided to change the “gender” column using dummy variables method by making it ‘1’ for Male and ‘0’ for Female. We also did the same thing for the “ever_married” column, by making it ‘1’ for married and ‘0’ for never married. Then, for the “smoking_status” column there were four possible options, ‘formerly smoked’, ‘never smoked’, ‘smokes’ or ‘Unknown’, so for this one, we decided to make a separate column for "formerly smoked", "never smoked", and "smokes". Then for each of them we used a dummy variable method where ‘1’ would mean true and ‘0’ would mean false. In cases where they all would be ‘0’ indicating false, then it would mean that the result is unknown. Finally, after looking at the data, we found that there were null values for about 100 entries for the "bmi" column. Since the dataset had over 5000 rows, we decided to remove those 100 rows, since removing 100 rows shouldn't affect our result.

4. Data Exploration

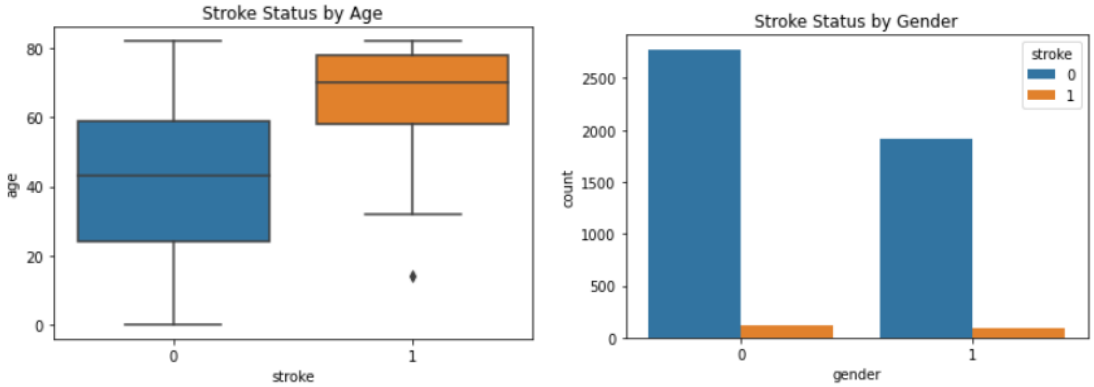


Figure 1

Figure 1 illustrates, in our opinion, the most important variables. After carefully examining age, we noticed that, generally, people who had a stroke are older. That is, those people are usually in the range of 60 years old to 80 years old. The median is 67 years of age for people who have had a stroke which is unsurprising since our predisposition was that stroke would most likely affect older aged people. Furthermore, gender was another important variable that we anticipated. In figure 1, male is indicated by the value 1 and female is represented by the value 0. There is a slight increase in females who have had a stroke compared to males who have had a stroke.

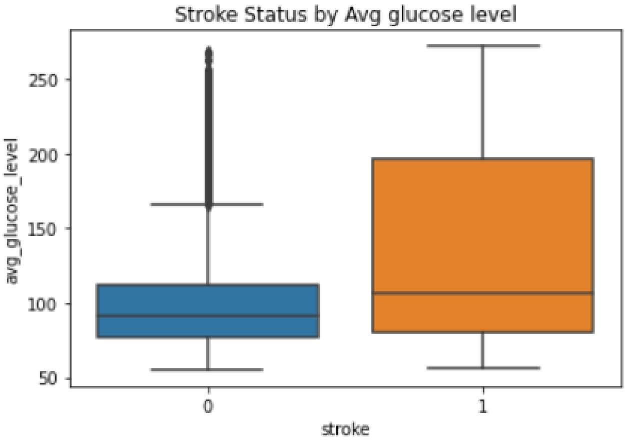


Figure 2

Figure 2 displays the boxplots we got from plotting the average glucose levels against whether someone had a stroke. The value of 0 on the x-axis indicates those that never had a stroke where a value of 1 indicates those who have. The data shows that there is a broader range in average glucose levels in those that did have a stroke compared to those that didn't. The interquartile range of those that did get strokes roughly ranges from 75 to 200, whereas for those that didn't have a stroke, that range was from ~75 to 110. This implies that strokes are more common as one's average glucose level rises. We also observe that amongst those who didn't have a stroke, those with glucose levels above 170ish were considered outliers. This differs greatly from the class of people who've had a stroke - the 4th quartile was around 290.

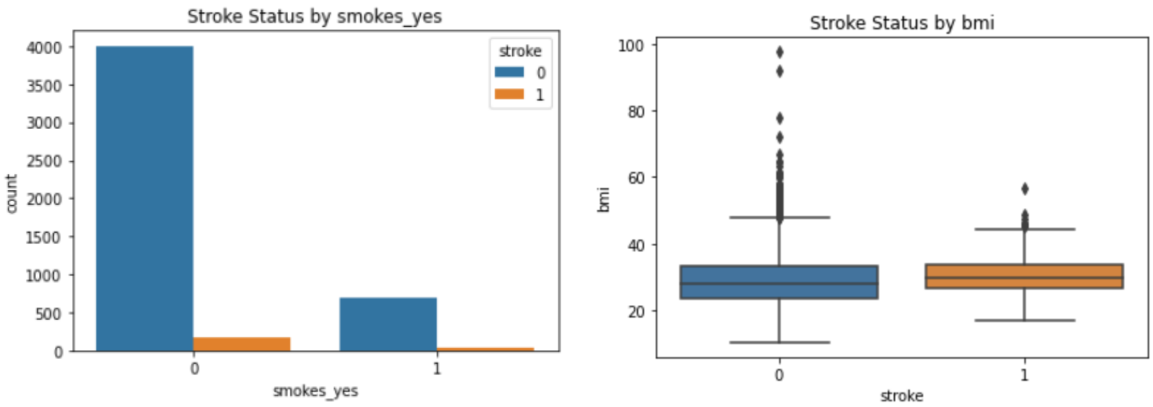


Figure 3

Figure 3 illustrates the most surprising data exploration we found. We initially assumed that BMI and currently smoking status would largely affect stroke status. However, the data showed that current smokers had less strokes than current nonsmokers. This was a surprise to us since it seemed counterintuitive. We decided that this may be due to the fact that the data is imbalanced and there could be a small number of observations of people who are smokers and have had strokes. Similarly, BMI levels indicated that there is only a slight increase in strokes for people who have high BMI. Again, this caught us by surprise since people who are overweight generally get strokes. Interestingly, there are quite a few outliers for non-stroke observations that have had no stroke which is counterintuitive. Again, since this dataset is imperfect and there is a data imbalance, we reasoned that the observations were not enough for people with high BMI and had a stroke.

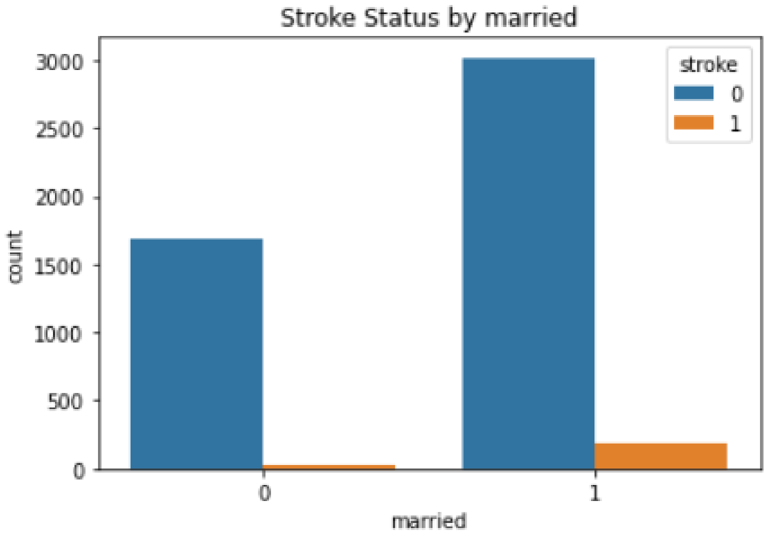


Figure 4

Figure 4 shows the count plot examining the relationship between marriage status and stroke status. The value of 0 on the x-axis corresponds to those who are not married while a value of 1 corresponds to those who are. Although the raw number of those who have had a stroke is generally small, when comparing proportions we noticed that those who were married were at a slightly greater chance of getting a stroke. This was interesting to us as we believed biomarkers would be the best indicators of one's chance of stroke.

5. Data Modeling

When modeling out data, we encountered a difficult issue which was that our data was very imbalanced for our stroke class. That is, we had 4700 no stroke observations and only 209 stroke observations. Since the not stroke was the minority class and it important when modeling the data, we decided to oversample the minority class, no stroke. There are different approaches to oversampling, but we chose SMOTE (Synthetic Minority Oversampling Technique). This method of oversampling would be a good alternative since it synthesizes and produces new observations from the existing minority class. Since there were 4700 no stroke observations and only 209 stroke observations, SMOTE would synthesize 4,491 stroke observations to be added to our existing data. The advantage of oversampling the data is that our models can make useful predictions and we would not lose the majority of our observations if instead we chose to perform undersampling. The disadvantage of this oversampling the data is that we have a lot of synthesized stroke observations which are not as ideal as real observations. Now, here is our data modeling results based upon our resampled SMOTE data.

For our data modeling, we decided to use classification and clustering techniques. For classification we used decision trees and k-nearest neighbors. We created 2 variations of Decision Tree models - one using entropy as the criterion and another using the gini index. For k-nearest neighbors, we created 2 variations as well. The first variation used 4 as the value of n_neighbors, and the second variation used 2.

Best Decision Tree Model:

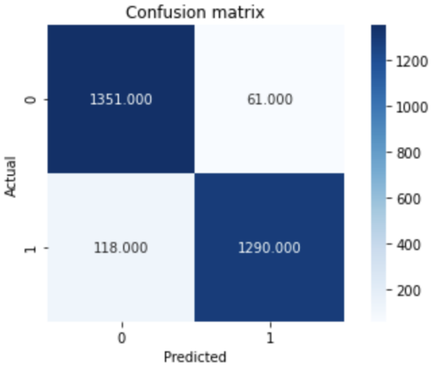


Figure 5

Figure 5 shows the confusion matrix produced by our decision tree model with entropy and whose predictor variables included marriage status, gender and age. The confusion matrix indicates that the majority of observations were indeed correctly predicted. The model had an accuracy of 93% which means the predicted observations were classified correctly 93% of the time. The error calculated to 0.06 which is very low. The F-1 score was 0.93 which was the highest among all models in both classification and clustering. Overall, the statistics indicate that the model is a good fit.

Best K-Nearest Neighbors Model:

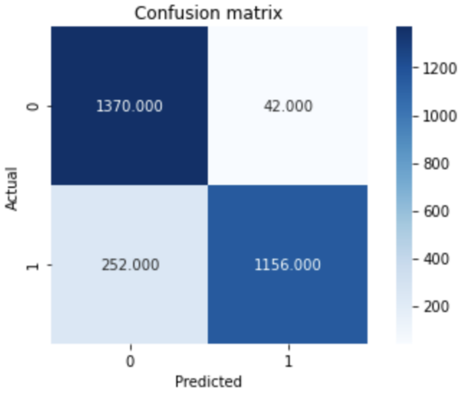


Figure 6

Figure 6 shows the confusion matrix produced by the k-nearest neighbors model with a n_neighbors value of 2. As with the best performing decision tree, this model's predictor variables were marriage status, gender and age. The confusion matrix shows the model did a better job at predicting class 0 observations (those who did not have a stroke). The model had an accuracy of 89%, an error of 10% and an F-1 score of 88%. Overall, this was a very well performing model.

Our clustering methods included hierarchical clustering and k-means clustering. One variation of hierarchical clustering utilized complete linkage while a second utilized single linkage. However, both hierarchical models used euclidean distance. For our k-means models, one variation had 10 as the number of neighbors while the second iteration used 20 neighbors for classification.

Best Hierarchical Model:

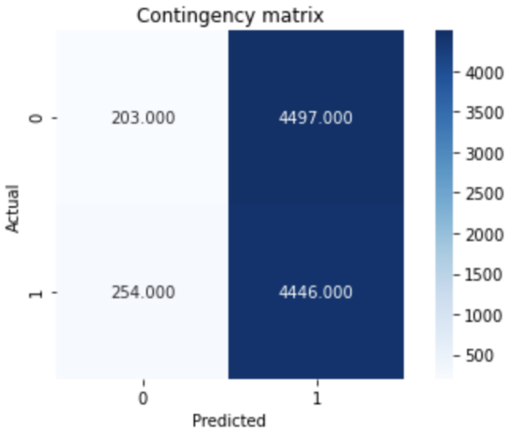
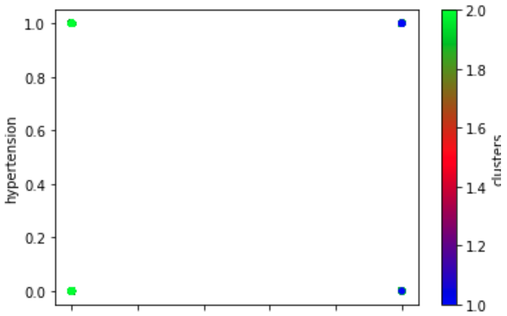
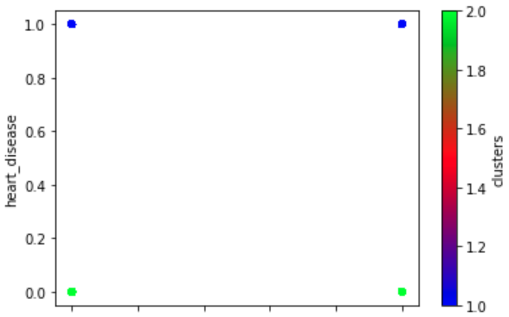


Figure 7

For our best hierarchical model, we actually had two models that gave the same exact results. The one common denominator between the two were their predictor variables: current smoker, hypertension and heart disease. One variation of the model used complete linkage while the other used single linkage. Figure 7 shows the contingency matrix produced by both. As the matrix shows, the models have a high recall when predicting Class 1 (those who've had a stroke), but a poor precision. This is due to the fact that the models had a large prediction bias in favor of Class 1 - it predicted 8,943 observations to belong to Class 1 while only predicting 457 observations for Class 0. The models did very poorly when predicting Class 0 with poor precision and recall. The silhouette coefficient for both models was 0.79. As predicted from the matrix, our adjusted rand index had a very small value of 9.8x10^-5.





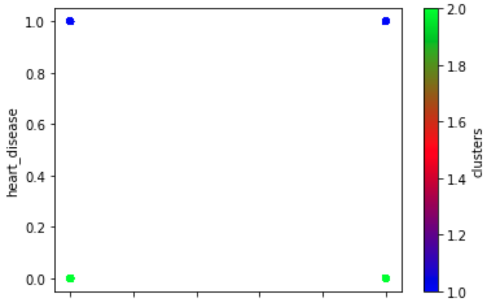


Figure 8

Figure 8 shows the scatter plots of our predictor variables. In clockwise order, the first plot shows the relationship between current smoker status and hypertension. We see the clusters are differentiated based on whether someone is currently smoking. The second plot shows the relationship between current smoker status and heart disease. This plot differentiates clusters based on whether someone has heart disease. The final plot shows the relationship between hypertension and heart disease. The clusters here are also differentiated by whether someone has heart disease. Since these predictor variables all produce binary values, we only see 4 points on each plot. In reality, there are more than 4 observations - they are just hidden perfectly under each other.

Best K-Means Model:

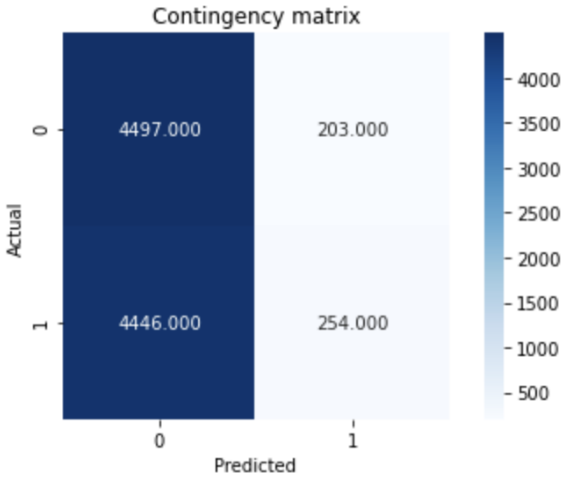


Figure 9

As with our best hierarchical model, we also had two best-performing variations that produced the same results. The first iteration had 10 initial centroids while the second iteration had 20 initial centroids. Both variations had 2 as the value of n_clusters since we wanted to differentiate clusters based on whether someone had a stroke or not. Interestingly, our K-means models produced the reverse results of those produced by our best performing hierarchical models. Rather than have a large prediction bias in favor of Class 1, the k-means models had a large bias in favor of Class 0. It predicted 8,943 observations to be those who never had a stroke, and 457 as those who did have a stroke.The models predicted Class 0 with high recall but poor precision. The models did very poorly when predicting Class 1 with poor precision and recall. The silhouette coefficient for both models was 0.79. As predicted from the matrix, our adjusted rand index had a very small value of 9.8x10^-5.

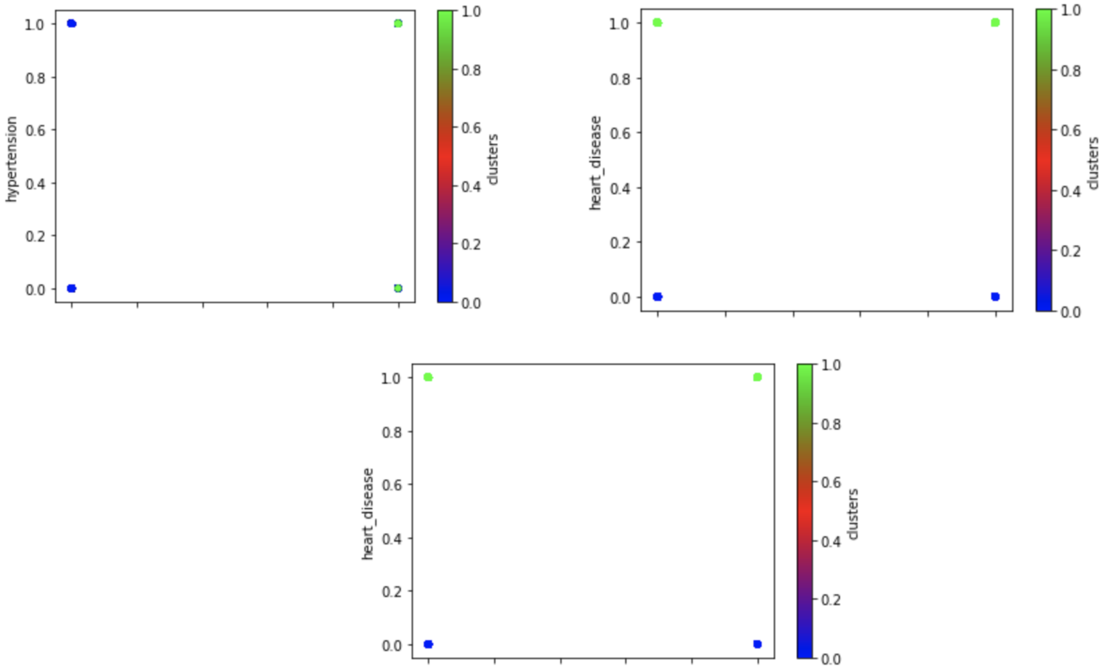


Figure 10

Figure 10 shows the scatter plots of our predictor variables, which are identical to the plots in Figure 8. In clockwise order, the first plot shows the relationship between current smoker status and hypertension. We see the clusters are differentiated based on whether someone is currently smoking. The second plot shows the relationship between current smoker status and heart disease. This plot differentiates clusters based on whether someone has heart disease. The final plot shows the relationship between hypertension and heart disease. The clusters here are also differentiated by whether someone has heart disease. Since these predictor variables all produce binary values, we only see 4 points on each plot. In reality, there are more than 4 observations - they are just hidden perfectly under each other.

Overall, our best performing model was our decision tree model that utilized entropy and whose predictor variables included marriage status, gender and age. We explored these variables further to see their individual impact on stroke risk.

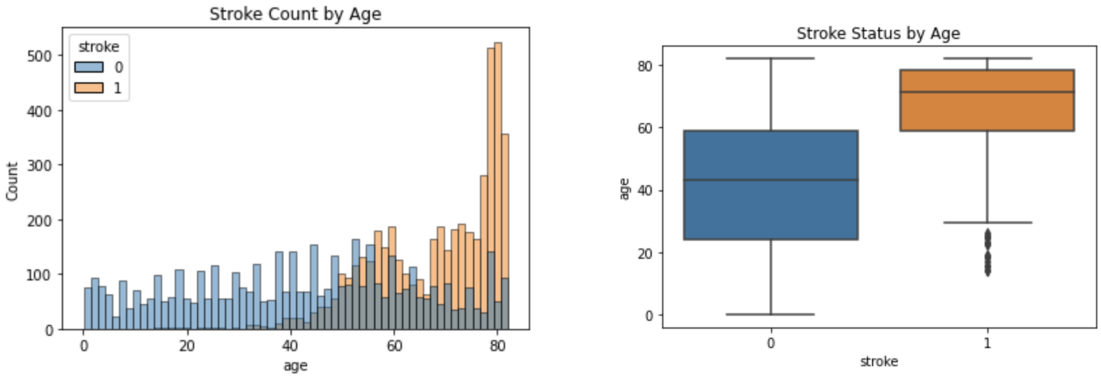


Figure 11

Figure 11 further expounds on the relationship between age and stroke risk. As we can see from the histogram and box plot, the risk of stroke greatly increases in the mid 40s and is common until the early 80s. Interestingly, in the box plot we can see that the interquartile range of those that do get a stroke (Class 1) begins at the end of Class 0's interquartile range. This happens at around 60 years of age. We also see that those that get a stroke between their mid-teens and mid-twenties are considered outliers.

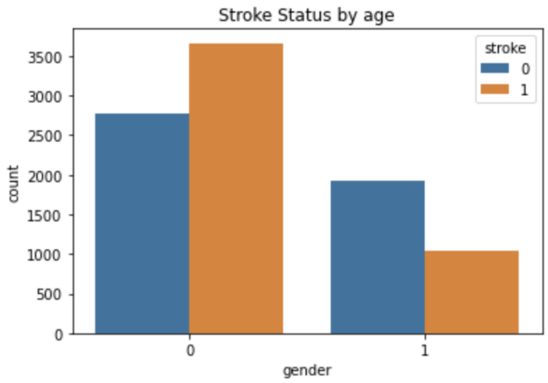


Figure 12

Figure 12 further explores the relationship between gender and stroke risk. In this bar plot, a gender value of 0 represents a female and value of 1 represents a male. The data shows that amongst those that do have a stroke, there are a greater number of female observations. This suggests that women are at a higher risk of having a stroke.

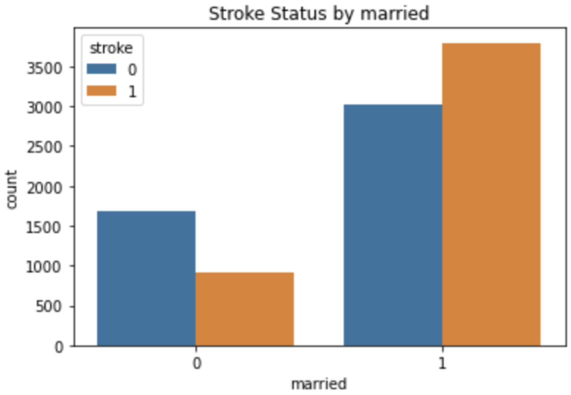


Figure 13

Figure 13 shows the relationship between marital status and risk of stroke. A married value of 0 indicates the person is not married while a value of 1 indicates they are. The plot shows that amongst those who are not married, the majority have not had a stroke. In the case of those who are married, there is a sizable increase in the number of people who have had a stroke previously versus those who haven't had a stroke. We can infer that in addition to biomarkers, one's environment and social status can influence their risk of stroke.