ST635

Intermediate Statistical Modeling for Business

Spring 2020

Stroke and Health Status - Project Report

Group Member

Sagar Bansal

Haochen Jiang

Liuqing Wang

Snow Wang

Table of Contents

Abstract

With the improvements of living standards, people nowadays pay more attention to their health. This paper talks about stroke, one of the biggest health problems in the U.S., focusing on the pre-existing health factors that will potentially rise one's risk of getting a stroke. We investigated relationships between stroke and other factors including age, hypertension, heart disease, average glucose level, body mass index, and smoking status. With the logistic regression model we fitted, it is clear that within the six factors, body mass index is the only factor that has little association with whether a person will get a stroke or not. The cluster analysis model showed that age and average glucose level are important factors that helped conclude the four different groups of people.

Section 1 - Introduction

Stroke, as known as a cerebrovascular disease, occurs when there is not enough blood flowing to the brain due to sudden blockage of arteries. It usually happens suddenly and can cause high mortality and disability rates. Some common symptoms of stroke include difficulty of speaking and walking, difficulty of understanding, and lack of coordination. In addition, there are three major types of stroke, transient ischemic attack, ischemic stroke, and hemorrhagic stroke. Ischemic stroke is the most common one and it counts nearly 87% of strokes in total. The results of stroke may vary from location of stroke in the brain and the degree of brain injuries.

According to the *Center for Disease Control and Prevention* (CDC), in the United States, there are more than 795,000 cases of stroke each year and almost 18% of them result in death. In fact, stroke is the fifth primary cause of death in the United States. On a global scale, each year, there are 15 million people who have strokes. There is no doubt that it must be taken seriously and definitely requires our immediate actions.

Our team is determined to take a deeper look at this disease. We will examine the underlying relationships and structures that stroke has with different factors including age, hypertension, heart diseases, average glucose level, body mass index, and smoking status. Then, we will interpret the relationships and perform statistical inferences on the importance of these variables. Additionally, we would like to find relationships between observations and see if we can identify meaningful groups for the observations.

In Section 2, we will explore more data characteristics on our dataset and identify if there is any obvious relationship between variables. In Section 3, we will first introduce a supervised learning technique, logistic regression analysis, and an unsupervised learning technique, cluster analysis, along with our interpretation and validation. Finally, we will conclude the overarching results in the Summary section.

Section 2 - Data Characteristics

We found the dataset on Kaggle.com, a public domain where datasets can be copied, modified, distributed, and performed work for any purposes without permission. The original dataset was published by Mckinsey & Company for one of their case studies. It is an observational study where the data were collected without any experimental manipulation and the researcher had no control over the variables. The study covers data for a large group of people in a period of time indicating it is cross-sectional instead of longitudinal where the researcher observes only one person but in many different time periods.

The primary interest of this project is to find the relationships between pre-existing health conditions and stroke, and therefore only 7 out of the 12 variables from the original dataset were chosen (see 1.1 in Appendix). The response variable is a binary showing whether or not a person has a stroke. Stroke equals to 1 means a person has stroke, whereas stroke equals to 0 means a person doesn't have a stroke. There are 6 predictors chosen including 3 quantitative variables - age, average glucose level, and body mass index, and 3 qualitative variables - hypertension, heart disease, and smoking status. (Note, we will use bmi in the rest of the report to replace body mass index). Variables like marriage status and work type were limited. The original dataset has 43400 observations. After cleaning out observations with null values, we have a total of 29072 observations to study with. Within these observations, 548 of them have had stroke experiences (also refers to whether people had strokes or not), which is only 1.88% of the sample size, and therefore an imbalance problem is observed. However, we believe the imbalance is a true representation of the population considering there's only a very small amount of people in the real world that are bothered with this disease. Table 1 provides detailed statistics for our dataset.

Table 1 Definitions of Variables and Summary Statistics

Qualitative Variables

Variable	Description	Percent of Sample Data
Stroke	1 if an observation has had a stroke experience	1.88%
	0 if an observation has never had an stroke experience	98.11%
HBP	YES if an observation has hypertension	11.15%
	NO if an observation does not have hypertension	88.85%
	***Hypertension is normally High Blood Pressure**	
HD	YES if an observation has heart disease	5.21%
	NO if an observation does not have heart disease	94.78%
	***Heart Disease describe a range of diseases that affect your heart**	
SMOKING_STATUS	NEVER SMOKED if an observation has never smoked	54.16%
	FORMERLY SMOKED if an observation smoked before but has quited now	24.42%
	SMOKES if an observation is currently smoking	21.42%

Quantatitive Variables

Variable	Description	Mean of Sample Data	Minimum of Sample Data	Maximum of Sample Data
AGE	Age of an observation	47.67	10	82
AVG_GLUCOSE_LEVEL	Average glucose level of an observation	106.40	55.01	291.05
	***Glucose level is normally Blood Sugar Level**			
BMI	Body Mass Index of an observation	30.05	10.1	92

From an initial investigation, we observed several relationships between the variables. As presented in Figure 1, there appear to be a relationship between age and stroke experience. We see that people who are older tend to have strokes. Also, Figure 2 shows that there may be a relationship between average glucose level and stroke experience, and the two can be correlated with each other.

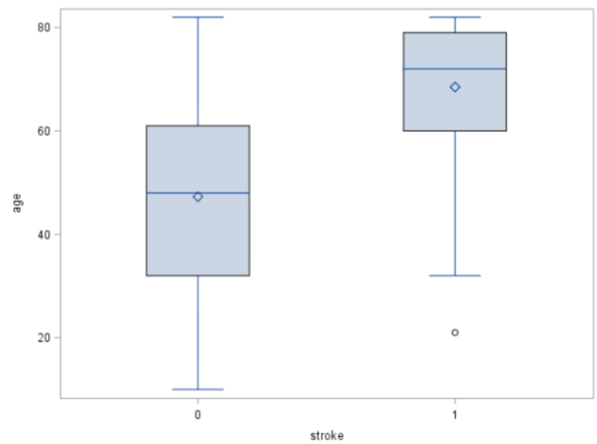

Figure 1 Boxplot of Age by Stroke Experience

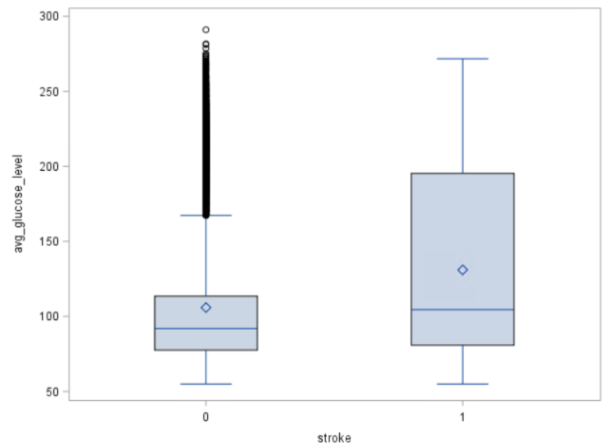

Figure 2 Boxplot of Average Glucose Level by Stroke Experience

Additionally, from Table 2, we see that only 1.51% of people who don't have hypertension had a stroke, whereas 4.91% of people who have hypertension had a stroke. As shown in Table 3, only 1.55% of people who don't have heart disease had a stroke, whereas 7.98% of people who have heart disease had a stroke. Therefore, we believe there might be some underlying relationships between hypertension and stroke, as well as heart disease and stroke. Nevertheless, it seems that bmi and smoking status do not have obvious relationships with stroke experience, but we would like to confirm it through our models, referring to Appendix 1.2 and 1.3 for the boxplot and frequency table.

Table 2 Stroke Experience by Hypertension

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of stroke by HBP			
		HBP		
	stroke	no	yes	Total
	0	25442 87.51 89.20 98.49	3082 10.60 10.80 95.09	28524 98.12
	1	389 1.34 70.99 1.51	159 0.55 29.01 4.91	548 1.88
	Total	25831 88.85	3241 11.15	29072 100.00
	Frequency Missing = 15			

Table 3 Stroke Experience by Heart Disease

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of stroke by HD			
		HD		
	stroke	no	yes	Total
	0	27129 93.32 95.11 98.45	1395 4.80 4.89 92.02	28524 98.12
	1	427 1.47 77.92 1.55	121 0.42 22.08 7.98	548 1.88
	Total	27556 94.79	1516 5.21	29072 100.00
	Frequency Missing = 15			

Section 3 - In Depth Analysis

This section will further examine the patterns observed in Section 2 using a supervised learning technique which is conducted with a clear target, and an unsupervised learning technique, which is conducted without a specific target. We will address the model selections, and then focus on the statement of the models and their interpretations.

Section 3.1: Logistic Regression Model

Model Selection

Our variable of interest in the dataset is binary, meaning the values are either 1 or 0. The goal is to find the relationships between predictors and the response as well as the coefficient interpretations, then make useful inference statements. Therefore, we recommend addressing the supervised learning analysis by fitting a logistic regression model with all six predictors mentioned in Section 2. The other supervised learning technique, which is classification tree, does not provide coefficients making it less ideal for our specific purposes.

Model Justifications

Our hypothesis model with assumptions can be found in the A2.1 of Appendix. While conducting the logistic regression model, we used maximum likelihood estimation to fit the model and estimate coefficients.

First of all, in order to evaluate model utility, we checked misclassification rate, receiver operating characteristics curve (ROC), and area under the curve (AUC). From the output SAS generated reported in Figure 3, we found that the model is useful as the ROC curve is always above the diagonal line and Area Under the Curve value is 0.8372 which is greater than 0.5. In addition, the misclassification rate of the logistic regression model is 25.8% at the 0.020 Prob Level that is considerably good, as illustrated in Table 4. Full Classification Table output can be found in Appendix 2.2. The corresponding sensitivity and specificity values are 77.9% and 74.1% respectively. Sensitivity is slightly higher than specificity indicating that we prefer to have better labeling of people who may experience stroke. Overall, the model is considered as useful.

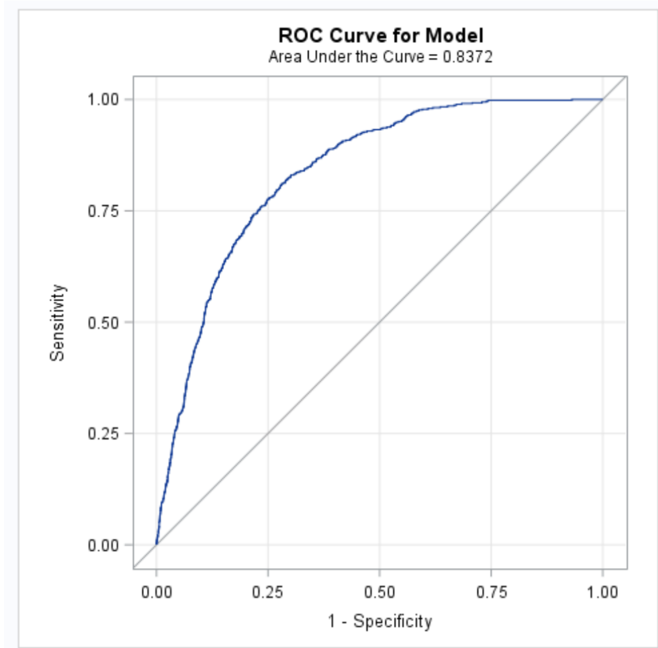

Figure 3 Receiver Operating Characteristic Curve for Logistic Model

Table 4 Partial Classification Table

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	548	0	28524	0	1.9	100.0	0.0	98.1	.
0.020	427	21149	7375	121	74.2	77.9	74.1	94.5	0.6
0.040	329	24529	3995	219	85.5	60.0	86.0	92.4	0.9
0.060	222	26294	2230	326	91.2	40.5	92.2	90.9	1.2

The misclassification rate is calculated as 100%-74.2%=25.8%

Secondly, to evaluate the model's validity, we used the Hosmer and Lemeshow Goodness-of-Fit Test to test the null hypothesis that the model generally fits well. The results of the Hosmer and Lemeshow test are presented in the Table 5 below. The p-value is 0.2079, which is greater than 0.05 indicating that we fail to reject the null hypothesis. As a result, there is no evidence that the model does not fit well. Moreover, there is no time series structure or special order in this dataset so it's unnecessary to check the independent issue. However, in order to be more precise, we checked the residual vs. index plot by looking at Model and Outlier Diagnostic. We didn't notice any clear pattern in the Pearson Residuals vs Case Number plot shown in Figure 4, and it confirms the independence check made earlier. Therefore, we cannot turn down the model based on validity issues and the model is reasonable when assessing the relationships between predictors and response.

Table 5 Hosemer-Lemeshow test results

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
10.8921	8	0.2079

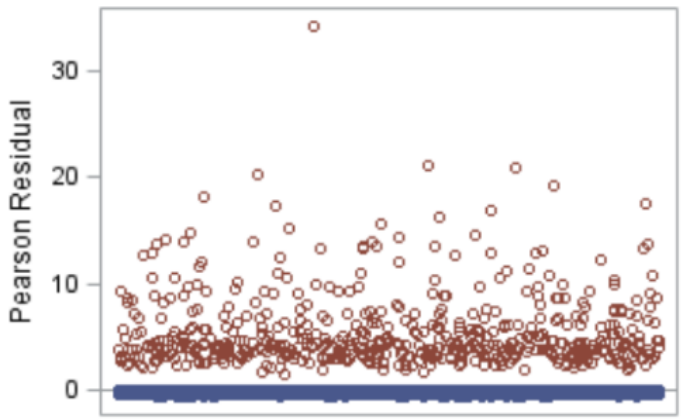

Figure 4 Pearson Chi-squared residual vs. index plot

Model Interpretation

In the following discussion, we will use odds of having a stroke to describe the probability that a person will experience a stroke or is more likely to experience a stroke than he/she will not experience a stroke. Table 6 shows the estimated coefficient and p-value of each variable. Using Wald Chi-squared test, we revealed that Age (0.0001), $\text{Hypertension}_{\text{No}}$ (0.0001), $\text{Heart Disease}_{\text{No}}$ (0.0001), Average Glucose Level (0.0001), $\text{Smoking Status}_{\text{Formerly\_Smoked}}$ (0.0335) and $\text{Smoking Status}_{\text{Never\_Smoked}}$ (0.0358) are statistically significant, whereas bmi (0.2722) does not significantly affect the odds of the response at 5% significance level. This result confirms our initial findings in Section 2 and provides evidence for our observation that bmi does not add much to the odds of having a stroke. As shown in the boxplot in Appendix A1, the bmi distribution for stroke and no stroke are very similar.

Table 6 Logistic Regression Model Estimates

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.3284	0.3936	346.7463	<.0001
age		1	0.0719	0.00371	374.7961	<.0001
HBP	no	1	-0.4384	0.1005	19.0161	<.0001
HBP	yes	0	0	.	.	.
HD	no	1	-0.6169	0.1125	30.0453	<.0001
HD	yes	0	0	.	.	.
avg_glucose_level		1	0.00384	0.000774	24.5504	<.0001
bmi		1	-0.00783	0.00713	1.2055	0.2722
smoking_status	formerly smoked	1	-0.2695	0.1268	4.5180	0.0335
smoking_status	never smoked	1	-0.2484	0.1183	4.4066	0.0358
smoking_status	smokes	0	0	.	.	.

In theory, the odds of a zero-year-old person with heart disease, hypertension, average glucose level value of 0 and BMI value of 0 who smokes will have a stroke is estimated to be approximately 0.0007. However, this is not practical since a person cannot have age, average glucose level and BMI values of zero. In terms of age, one year increase in age is associated with a 7.5% increase in the odds of having a stroke when all other variables remain fixed. It confirms that age is positively associated with strokes when average glucose level, bmi, smoking status, hypertension, and heart disease stays the same. In terms of average glucose level, a unit increase in average glucose level is associated with about 0.4% increase in the odds of having a stroke when all other variables remain fixed. It confirms that average glucose level is positively associated with strokes when bmi, smoking status, age, heart disease, and hypertension stay the same.

In terms of hypertension, the model estimates that people with no hypertension have about 35.5% lower odds of having a stroke than people with hypertension, when other variables are fixed. It confirms that people who have hypertension tend to have higher risk of getting strokes than people who don't have hypertension when average glucose level, bmi, smoking status, age, and heart disease stays the same. In terms of heart disease, the model estimates that people with no heart disease have about 46% lower odds of having a stroke than people with heart disease when other variables are fixed. It confirms that people who have heart disease tend to have higher risk of getting strokes than people who don't have heart disease when average glucose level, bmi, smoking status, age, and hypertension stays the same.

In terms of smoking status, the model estimates that people who formerly smoked have about 23.6% lower odds of having a stroke than people who currently smoke, when other variables are fixed. The model estimates that people who never smoke have 22% lower odds of having a stroke than people who smoke when other variables are fixed. This is an important discovery, because we didn't see obvious relationships between smoking status and stroke experience in our initial investigation. Moreover, it indicates that the impact of formerly smoking on the response is almost the same as never smoked when compared with smoking according to the model, when other variables remain fixed.

In conclusion, the logistic regression model concluded that age, smoking status, average glucose level, hypertension, and heart disease have significant impact on stroke and bmi does not. The result indicates that older people, people who currently smoke, people with hypertension or heart disease, and people with high average glucose levels should be aware of their risk of having a stroke, especially those with heart disease.

Section 3.2: Cluster Analysis

Model Selection

The goal for unsupervised learning analysis is to have a better understanding of different groups of people by finding overarching structures. We thereby recommend a cluster analysis using the quantitative variables including age, average glucose level, and bmi. The other unsupervised learning techniques like principal component analysis and common factor analysis are better for identifying relationships between variables but not ideal for our specific purpose.

Model Justifications

We conducted a hierarchical cluster analysis with average linkage criterion. Full SAS output can be found in Appendix. Since our dataset has about 30,000 observations and it's hard to identify jumps through the Dendrogram, as shown in Appendix Table A3.1, we only listed the last 20 clusters of Cluster History, shown in Appendix A3.2, and calculated the differences between each cluster. Based on the calculation shown in Appendix A3.3, we identified three clear jumps with the biggest differences, which are from 7 to 6 cluster configurations, 3 to 2 cluster configurations, and 2 to 1 cluster configurations. Hence, we have three candidate cluster configurations including 7 cluster configuration, 3 cluster configuration, and 2 cluster configuration.

Next, in order to find the best cluster configuration, we intended to label these three candidate cluster configurations and see if these labels are reasonable. The 3-cluster and 2-cluster configurations, as shown in Appendix 3.3, are almost the same except that the third cluster in the 3-cluster configurations consists of a single observation which is essentially an outlier. Clusters 1 and 2 in both cluster configurations have a wide range of values for the age, average glucose level, and bmi variables. Therefore, we were not satisfied with using these configurations as our final configuration. We believe there should be more groups of people in our dataset. By comparing means of different clusters in each candidate cluster configuration, 7 cluster configuration seems to be the most reasonable, which is shown in Table 7 below. More detailed outputs of each candidate cluster configuration and technical details can be found in Appendix.

Table 7 Cluster Configuration Means

The SAS System

The MEANS Procedure

CLUSTER	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
1	23527	age	23527	45.7979343	18.4257246	10.0000000	82.0000000
		avg_glucose_level	23527	87.2814082	17.0930001	55.0100000	135.9400000
		bmi	23527	29.4947592	6.9122666	10.1000000	72.2000000
2	1621	age	1621	44.5533621	20.3078260	10.0000000	82.0000000
		avg_glucose_level	1621	140.2167798	11.1836505	119.6500000	175.1000000
		bmi	1621	29.4806292	6.6278170	12.5000000	61.8000000
3	3875	age	3875	60.4356129	14.4589452	10.0000000	82.0000000
		avg_glucose_level	3875	206.5863381	23.2653195	148.6400000	272.8600000
		bmi	3875	33.5713806	7.7569170	15.0000000	80.1000000
4	45	age	45	39.9333333	8.3758310	19.0000000	52.0000000
		avg_glucose_level	45	255.9486667	12.8352948	228.2400000	281.5900000
		bmi	45	36.5911111	14.1449930	17.3000000	82.7000000
5	2	age	2	30.5000000	10.6066017	23.0000000	38.0000000
		avg_glucose_level	2	63.4650000	9.2843120	56.9000000	70.0300000
		bmi	2	85.0000000	9.8994949	78.0000000	92.0000000
6	1	age	1	80.0000000	.	80.0000000	80.0000000
		avg_glucose_level	1	291.0500000	.	291.0500000	291.0500000
		bmi	1	28.7000000	.	28.7000000	28.7000000
7	1	age	1	78.0000000	.	78.0000000	78.0000000
		avg_glucose_level	1	135.7300000	.	135.7300000	135.7300000
		bmi	1	89.0000000	.	89.0000000	89.0000000

Table 8 bmi Categories from Center for Disease Control and Prevention (CDC)

BMI	Considered
Below 18.5	Underweight
18.5 to 24.9	Healthy weight
25.0 to 29.9	Overweight
30 or higher	Obese

To better understand the 7 cluster configuration, we assumed age has two groups including younger people (10-50 years old) and older people (51-82 years old). Based on our observations on Table 7 above, the first cluster is composed of younger people who are around 46 years old, with slightly low average glucose level and overweight bmi. (Refer to Table 8 above for detailed bmi categories provided by CDC). The second cluster is composed of younger people who are around 45 years old, with medium average glucose level and overweight bmi. Compared to the first cluster, the only difference between them is that people in the second cluster have much higher average glucose levels. The third cluster is composed of older people who are around 60 years old, with high average glucose level and obese bmi. The fourth cluster is composed of younger people who are around 40 years old, with very high average glucose level and obese bmi. Cluster 5, cluster 6 and cluster 7 are hard to describe since they only have one or two observations. They did not fall into any of the main categories and could be outliers captured in separate clusters.

Based on our initial observations, we proposed four cluster labels. The first cluster represents younger people with comparatively healthy status. The second cluster represents younger people who struggle with their glucose level. The third cluster represents older people who struggle with healthy status. The fourth cluster represents younger people who have extremely unhealthy status.

To validate our selected cluster configuration, we used a qualitative variable, stroke, to check if the clusters have significant differences in the proportion of whether a person experienced a stroke or not. From Table 9, we can see that cluster 1 has 1.39 column percent of observations with a stroke, followed by 2.22 column percent in cluster 2 and 4.75 column percent in cluster 3. For cluster 4, the column percent of observations with a stroke is zero that is lowest among all four clusters (excluding clusters of outliers). The finding that younger people who have extremely unhealthy status have very low risk of having a stroke will need further investigation. Overall, this tells us that older people who struggle with healthy status have the highest risk of having a stroke, followed by the younger people who struggle with their glucose level, younger people with comparatively healthy status and younger people who have extremely unhealthy status.

Table 9 Frequency Table

The SAS System

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of stroke by CLUSTER								
		CLUSTER							
	stroke	1	2	3	4	5	6	7	Total
	0	23199 79.80 81.33 98.61	1585 5.45 5.56 97.78	3691 12.70 12.94 95.25	45 0.15 0.16 100.00	2 0.01 0.01 100.00	1 0.00 0.00 100.00	1 0.00 0.00 100.00	28524 98.12
	1	328 1.13 59.85 1.39	36 0.12 6.57 2.22	184 0.63 33.58 4.75	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	548 1.88
	Total	23527 80.93	1621 5.58	3875 13.33	45 0.15	2 0.01	1 0.00	1 0.00	29072 100.00

When we conducted the Chi-Squared Test of Association for the Stroke variable, the p-values (<0.0001) came out to be less than 0.05, as shown in Table 10. Hence, we reject the null hypothesis that the proportions of stroke values in different clusters are the same. The test confirms that the proportions of Stroke values are different for different clusters at 5% significance level and the clusters are reasonable.

Table 10 Chi-Squared Test of Association Test

Statistics for Table of stroke by CLUSTER

Statistic	DF	Value	Prob
Chi-Square	1	193.1872	<.0001
Likelihood Ratio Chi-Square	1	146.9508	<.0001
Continuity Adj. Chi-Square	1	191.4364	<.0001
Mantel-Haenszel Chi-Square	1	193.1806	<.0001
Phi Coefficient		0.0815	
Contingency Coefficient		0.0812	
Cramer's V		0.0815	

Model Interpretation

Based on the cluster analysis, we found that older people who struggle with their health status have the highest risk of having a stroke, although their health status is not as bad as people in cluster 4, who are younger but with extremely unhealthy status. In addition, younger people with high average glucose level have higher risk of having a stroke than younger people with comparatively healthier status. Therefore, we believe that age and average glucose level both have significant impacts on a person's chance of having a stroke. People should be aware of the importance of a healthy lifestyle and pursue one for their own benefit.

It is crucial for older people to be educated on the severity of stroke and how a healthier lifestyle can lower their risk of getting a stroke. Medical professionals should specifically focus their research on older people with very high glucose levels. Furthermore, the healthcare industry may provide customized treatments for different groups of people.

Section 4 - Summary and Concluding Remarks

The logistic regression model concluded that age, smoking status, average glucose level, hypertension and heart disease have significant impacts on stroke and bmi does not. The result showed that heart disease has one of the largest influences with the odds of having a stroke and average glucose level has a minimal impact, compared with other factors. Therefore, older people, people who currently smoke, people with hypertension or heart disease, and people with high average glucose level should be aware of their risk of having a stroke, especially those with heart disease.

Based on the cluster analysis, age and average glucose level are both important factors that helped conclude the four clusters, including younger people with relatively healthy status, younger people who struggle with their glucose level, older people who struggle with their healthy status, and younger people who are in extremely unhealthy status. In general, people in older age with very high glucose levels have a higher risk of getting a stroke than younger people, even though younger people might not have a better health status. Nevertheless, younger people with high average glucose level should still be aware of the importance of a healthy lifestyle and pursue one for their own benefit.

Although this study was based on a dataset with 29072 observations, which is large enough to develop complex statistical models, there are still some limitations that can be improved in order to get more accurate information in further analysis. Firstly, some of the data were collected without clear parameters which may cause our findings to be misleading. For example, there should be a testing instruction for glucose level specifying that all tests should be done at one hour after meal, because glucose can change rapidly and adding time measurement will ensure that all test results are retrieved under the same condition. Additionally, more specific information on people's smoking history should be collected as well, such as how long a person has started or quitted smoking, because people with different years of smoking history have different health conditions. Secondly, since cluster analysis is subjective, there is no correct answer when it comes to grouping and labeling observations. Different numbers of clusters may change the final result. Thirdly, in order to further understand and lower the risk of getting a stroke, new predictors like drinking habits may be introduced and analyzed.

Overall, the project was aimed to draw people's attention to their health status in order to lower their chance of getting a stroke. We learned from our analysis that older people have higher risk and therefore should be more alerted about their health status and try to live a healthier lifestyle. Younger people should start to take proactive precautions in order to prevent the disease from happening when they get older.

Section 5 - Appendix

All outputs from Appendix are generated by SAS. It provides additional information and explanations related to the logistic regression analysis and cluster analysis.

Appendix Table of Contents

A1. Data Characteristics

A1.1 Cleaned Dataset

https://drive.google.com/drive/folders/18obicV3m9GmpIBAnuzlaJb_ZIxhHB_Wn

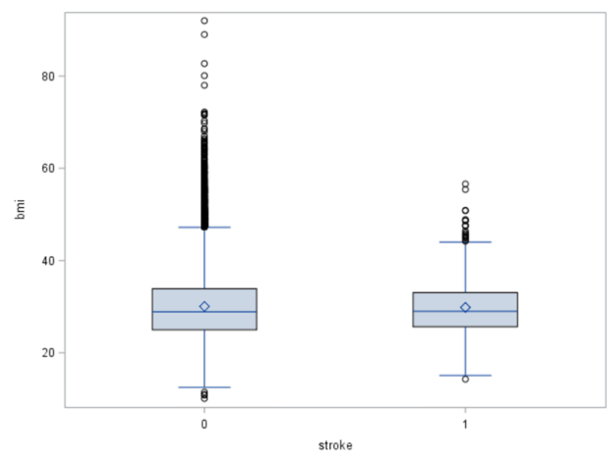

Figure A1.2 Boxplot of bmi by Stroke Experience

Table A1.3 Boxplot of Smoking Status by Stroke Experience

The FREQ Procedure

Table of smoking_status by stroke

Frequency Percent Row Pct Col Pct	stroke		
smoking_status	0	1	Total
formerly smoked	6919 23.80 97.46 24.26	180 0.62 2.54 32.85	7099 24.42
never smoked	15491 53.28 98.37 54.31	256 0.88 1.63 46.72	15747 54.17
smokes	6114 21.03 98.20 21.43	112 0.39 1.80 20.44	6226 21.42
Total	28524 98.12	548 1.88	29072 100.00

Frequency Missing = 15

A2. Logistic Regression Model

A2.1 Logistic Regression Model and Assumptions

In order to conduct the logistic regression model, we made following assumptions:

$$Y_i \sim \text{Binomial}\ (n = 1, p = f(E_i))$$

- $E_i = \beta_0 + \beta_{age} Age_i + \beta_{HBP} HBP_{i\ no} + \beta_{HD} HD_{i\ no} + \beta_{average\_glucose\_level} Average\_Glucose\_Level_i + \beta_{bmi} bmi_i + \beta_{smoking\_status} Smoking\_Status_{i\ foremerly\ smoked} + \beta_{smoking\_status} Smoking\_Status_{i\ never\ smoked}$
- f is the logistic link function, i.e. $f(E) = \frac{e^E}{1+e^E}$ß
- Y_i is whether or not patient have stroke or not
- Age is the age for ith patient
- $HBP_{i\ no}$ is dummy variable for the ith hypertension was listed in, where "yes" is the reference level, where HBP stands for hypertension
- $HD_{i\ no}$ is dummy variable for the ith heart disease was listed in, where "yes" is the reference level, where HD stands for heart disease
- Average_glucose_level is the average glucose level measured after meal for ith patient
- bmi(box mass index) is the body mass index for ith patient
- $Smoking\_Status_{i\ never\ smoked}$ and $Smoking\_Status_{i\ foremerly\ smoked}$ are two dummy variables for the ith smoking status was listed in, where "smokes" is the reference level.

Table A2.2 Full Classification Tables

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	548	0	28524	0	1.9	100.0	0.0	98.1	.
0.020	427	21149	7375	121	74.2	77.9	74.1	94.5	0.6
0.040	329	24529	3995	219	85.5	60.0	86.0	92.4	0.9
0.060	222	26294	2230	326	91.2	40.5	92.2	90.9	1.2
0.080	142	27266	1258	406	94.3	25.9	95.6	89.9	1.5
0.100	76	27819	705	472	96.0	13.9	97.5	90.3	1.7
0.120	53	28133	391	495	97.0	9.7	98.6	88.1	1.7
0.140	33	28302	222	515	97.5	6.0	99.2	87.1	1.8
0.160	16	28391	133	532	97.7	2.9	99.5	89.3	1.8
0.180	9	28448	76	539	97.9	1.6	99.7	89.4	1.9
0.200	8	28481	43	540	98.0	1.5	99.8	84.3	1.9
0.220	5	28505	19	543	98.1	0.9	99.9	79.2	1.9
0.240	2	28513	11	546	98.1	0.4	100.0	84.6	1.9
0.260	1	28518	6	547	98.1	0.2	100.0	85.7	1.9
0.280	1	28523	1	547	98.1	0.2	100.0	50.0	1.9
0.300	1	28524	0	547	98.1	0.2	100.0	0.0	1.9
0.320	0	28524	0	548	98.1	0.0	100.0	.	1.9

A3. Cluster Analysis

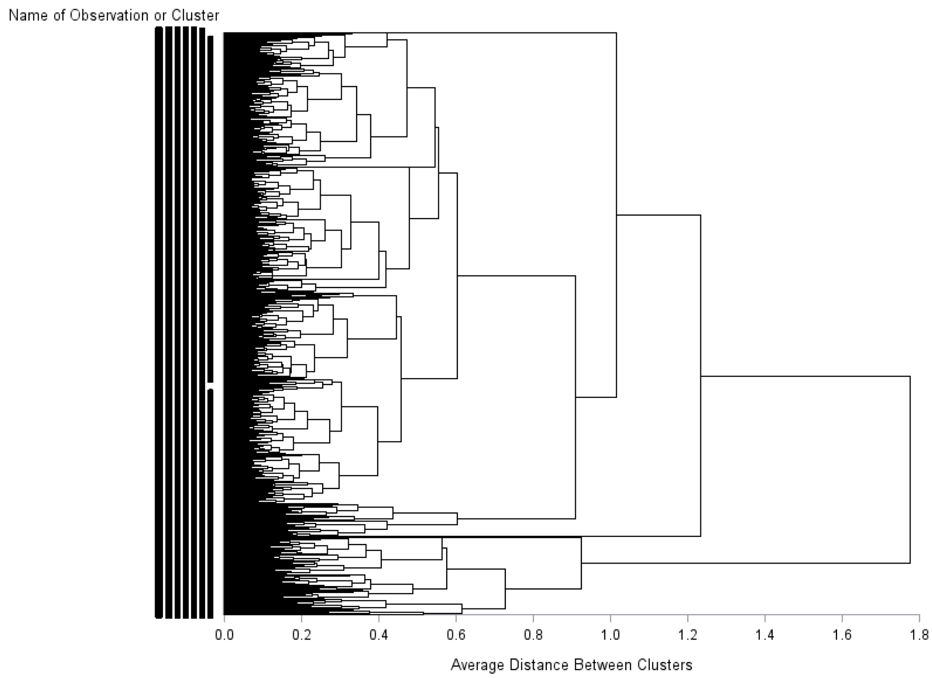

Figure A3.1 Dendrogram

Table A3.2 Partial Cluster History, as it only shows the last 20 clusters of the Cluster History.

Cluster History					
Number of Clusters	Clusters Joined		Freq	Norm RMS Distance	Tie
20	CL46	CL86	27	0.4631	
19	CL24	CL34	6695	0.4733	
18	CL119	CL28	6283	0.4793	
17	CL33	CL36	1216	0.4895	
16	CL35	CL69	276	0.5167	
15	CL19	CL41	6713	0.5469	
14	CL15	CL18	12996	0.5538	
13	CL20	CL29	1847	0.565	
12	CL13	CL17	3063	0.577	
11	CL23	CL25	1621	0.602	
10	CL14	CL21	23527	0.6044	
9	CL27	CL16	812	0.6144	
8	CL31	CL26	45	0.6522	
7	CL12	CL9	3875	0.7278	
6	CL8	OB15044	46	0.8234	
5	CL10	CL11	25148	0.9104	
4	CL6	CL7	3921	0.9248	
3	CL43	CL5	25150	1.0142	
2	CL3	OB2	25151	1.2328	
1	CL2	CL4	29072	1.7765	

Table A3.3 Calculation Difference between Each Cluster in the Cluster History

20	0.4631	Difference	
19	0.4733	0.0102	
18	0.4793	0.006	
17	0.4895	0.0102	
16	0.5167	0.0272	
15	0.5469	0.0302	
14	0.5538	0.0069	
13	0.565	0.0112	
12	0.577	0.012	
11	0.602	0.025	
10	0.6044	0.0024	
9	0.6144	0.01	
8	0.6522	0.0378	
7	0.7278	0.0756	
6	0.8234	0.0956	jump
5	0.9104	0.087	
4	0.9248	0.0144	
3	1.0142	0.0894	
2	1.2328	0.2186	jump
1	1.7765	0.5437	jump

Table A3.4 3 Clusters Clustering and 2 Clusters Clustering Means

The SAS System

The MEANS Procedure

CLUSTER	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
1	25150	age	25150	45.7165010	18.5547590	10.0000000	82.0000000
		avg_glucose_level	25150	90.6913726	21.2228056	55.0100000	175.1000000
		bmi	25150	29.4982624	6.9119126	10.1000000	92.0000000
2	3921	age	3921	60.2053048	14.5693252	10.0000000	82.0000000
		avg_glucose_level	3921	207.1743943	23.7953404	148.6400000	291.0500000
		bmi	3921	33.6047947	7.8625129	15.0000000	82.7000000
3	1	age	1	78.0000000	.	78.0000000	78.0000000
		avg_glucose_level	1	135.7300000	.	135.7300000	135.7300000
		bmi	1	89.0000000	.	89.0000000	89.0000000

The SAS System

The MEANS Procedure

CLUSTER	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
1	25151	age	25151	45.7177846	18.5555068	10.0000000	82.0000000
		avg_glucose_level	25151	90.6931633	21.2242837	55.0100000	175.1000000
		bmi	25151	29.5006282	6.9219509	10.1000000	92.0000000
2	3921	age	3921	60.2053048	14.5693252	10.0000000	82.0000000
		avg_glucose_level	3921	207.1743943	23.7953404	148.6400000	291.0500000
		bmi	3921	33.6047947	7.8625129	15.0000000	82.7000000

Section 6 - Work Cited

Day, S. M. (2013, March 18). Obesity trends in the United States. Retrieved from http://www.mortalityresearch.com/main/post/205

Defining Adult Overweight and Obesity. (2020, April 3). Retrieved from https://www.cdc.gov/obesity/adult/defining.html

Healthcare Dataset Stroke Data. (n.d.). Retrieved from https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data#train_2v.csv

Huizen, J. (2019, May 17). Blood sugar chart: Target levels throughout the day. Retrieved from https://www.medicalnewstoday.com/articles/317536#interpreting-the-results

Nall, R. (2018, September 20). Types of Strokes: Causes, Symptoms, and Treatments. Retrieved from https://www.healthline.com/health/stroke-types

The Internet Stroke Center. (n.d.). Retrieved from http://www.strokecenter.org/patients/about-stroke/stroke-statistics/

The Internet Stroke Center. (n.d.). Retrieved from http://www.strokecenter.org/patients/about-stroke/what-is-a-stroke/

Stroke Facts. (2020, January 31). Retrieved from https://www.cdc.gov/stroke/facts.htm