Predicting Strokes: What Are Your Chances?

Group 1

Shahzain Ahmed (260910106)

Emily Lawrence (260926930)

Pragna Patel (260929162)

Jonathan Steinberg (260924571)

INSY 336-002: Data Handling and Coding for Analytics

Professor Hyunji So

April 14, 2021

Problem Overview

According to the World Health Organization (WHO), strokes are the second highest cause of death and the third-highest cause of disability (Johnson et al., 2016). In the United States, a person gets a stroke every forty seconds and dies every four minutes (CDC, 2021). Strokes occur when blood flow suddenly stops flowing to the brain. Over time, due to the lack of oxygen and nutrients in the blood, brain cells die. If normal blood flow stops for an extended period of time, there is an increased likelihood of permanent brain damage and death.

The severity of a stroke can vary depending on which exact area of the brain is affected, and an individual's vision, thoughts, mobility, memory, and speech can all be delayed as a result (MedBroadcast, 2011). Due to the severe outcomes of getting a stroke, receiving emergency treatment is critical within the first three hours of symptoms occurring. These symptoms include facial droop, drifting arms, and slurred speech (CDC, 2020a).

According to the CDC, 80% of strokes are preventable (2020a). Therefore, in our project, we plan to determine which factors cause individuals to be more prone to strokes. Based on our findings, we intend on proposing practical precautionary measures one can take to reduce their risk of stroke. There are many known factors deemed by the scientific community that increase the risk of a stroke, such as old age, gender, previous stroke history, obesity, and smoking, however our dataset includes additional variables that may help to explain one's exposure to stroke.

Dataset Description

The dataset contains continuous and categorical data on 5,110 individuals, 249 of which have experienced a stroke. Divided into 12 columns, each row represents a single patient and holds data on 11 variables with a unique identifier. For the purposes of our analysis, we have

divided the variables into three subgroups; demographics, medical factors, and non-medical factors. We define demographic variables as ones that describe the characteristics of an individual, which are age, gender and body mass index (BMI) in our dataset. Variables describing the current and past health status of patients qualify as medical factors, and include average glucose level, hypertension, heart disease, and smoking status. Non-medical factors include the patient's current marital status, work type, and residence type. Finally, the last column in our dataset indicates whether or not an individual has experienced a stroke, denoted by a 1 or 0. Age, BMI, and average glucose level are continuous variables, while the remaining variables are categorical. Please refer to Appendix A for a more detailed description of each variable.

Approach Overview

The goal of our analysis is to find the variables that are most influential in causing stroke. In addition, we will compare across variable subgroups to identify which set of factors are best at predicting stroke. Before conducting descriptive analysis, we first collected our dataset from Kaggle by downloading it as a csv file and converting it into a DataFrame. We then preprocessed the data by removing missing values, creating dummy variables for categorical data, and making correlation matrices to ensure there was no multicollinearity between our independent variables. Our descriptive analysis consisted of plotting each independent variable against stroke in a categorical bar chart. Before creating bar charts for the three continuous variables (age, BMI, and average glucose level), we segmented each variable appropriately. We then created bar graphs with all possible combinations of independent variables and hues to assess how the interplay between our variables can help explain stroke. Finally, we ran three logistic regression models for each of the variable subgroups to assess which group more accurately predicts stroke. Please refer to Appendix B to view the code we used to conduct our analysis.

Data Analysis Strategy

Data-Preprocessing

We pre-processed our data by removing missing values, creating dummy variables for categorical data, and making correlation matrices. Out of our entire sample of 5,110 individuals, there were 201 null values within the BMI column, which we decided to drop in order to clean our data. In addition, we found that the smoking variable contained 1,483 unknown values. We decided not to remove all of these rows as they account for almost 25% of our data, which could potentially skew our results. Then, to ensure our data is usable for our analysis and logistic regression models, we converted our data's five categorical variables into dummy variables (gender, smoking status, marital status, work type, and residence type). Lastly, we ran three different correlation matrices for each variable subgroup to ensure there was no multicollinearity present between our independent variables. As seen in the matrices below, we found that none of the variables were highly correlated with each other.

	age	Male	Female	stroke
age	1.000000	-0.030149	0.030457	0.232331
Male	-0.030149	1.000000	-0.999579	0.006939
Female	0.030457	-0.999579	1.000000	-0.006851
stroke	0.232331	0.006939	-0.006851	1.000000

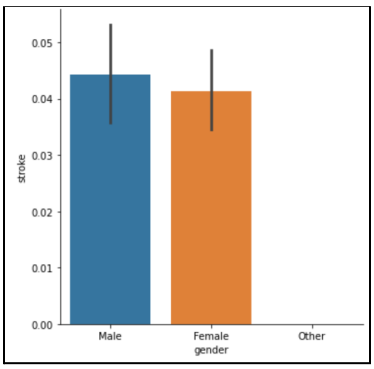
	Yes	No	Govt_job	Private	Self-employed	children	Never_worked	Urban	stroke
Yes	1.000000	-1.000000	0.137915	0.156818	0.191469	-0.545462	-0.091976	0.004989	0.105089
No	-1.000000	1.000000	-0.137915	-0.156818	-0.191469	0.545462	0.091976	-0.004989	-0.105089
Govt_job	0.137915	-0.137915	1.000000	-0.444147	-0.166136	-0.152679	-0.025745	0.010287	0.003553
Private	0.156818	-0.156818	-0.444147	1.000000	-0.501179	-0.460584	-0.077664	-0.017155	0.014934
Self-employed	0.191469	-0.191469	-0.166136	-0.501179	1.000000	-0.172285	-0.029051	0.012175	0.055356
children	-0.545462	0.545462	-0.152679	-0.460584	-0.172285	1.000000	-0.026698	-0.002790	-0.080971
Never_worked	-0.091976	0.091976	-0.025745	-0.077664	-0.029051	-0.026698	1.000000	0.023430	-0.014149
Urban	0.004989	-0.004989	0.010287	-0.017155	0.012175	-0.002790	0.023430	1.000000	0.006031
stroke	0.105089	-0.105089	0.003553	0.014934	0.055356	-0.080971	-0.014149	0.006031	1.000000

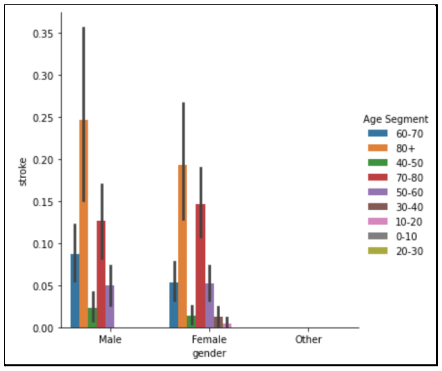
	hypertension	heart_disease	avg_glucose_level	bmi	smokes	formerly smoked	never smoked	stroke
hypertension	1.000000	0.115991	0.180543	0.167811	0.028214	0.062078	0.066717	0.142515
heart_disease	0.115991	1.000000	0.154525	0.041357	0.048686	0.071339	-0.020685	0.137938
avg_glucose_level	0.180543	0.154525	1.000000	0.175502	0.010981	0.074250	0.032085	0.138936
bmi	0.167811	0.041357	0.175502	1.000000	0.088324	0.107031	0.107964	0.042374
smokes	0.028214	0.048686	0.010981	0.088324	1.000000	-0.190555	-0.327141	0.021530
formerly smoked	0.062078	0.071339	0.074250	0.107031	-0.190555	1.000000	-0.352884	0.057320
never smoked	0.066717	-0.020685	0.032085	0.107964	-0.327141	-0.352884	1.000000	0.010723
stroke	0.142515	0.137938	0.138936	0.042374	0.021530	0.057320	0.010723	1.000000

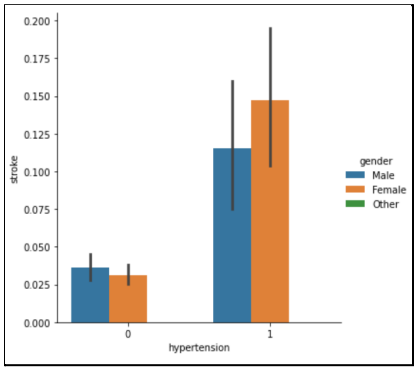
Descriptive Analysis

Gender

According to the CDC, women are disproportionately at risk of suffering from strokes. In fact, strokes are the third leading cause of death for women and kill women two times more than breast cancer (2020a). Many scientists presume that a key factor behind this staggering difference among the genders is that females tend to live many more years relative to their male counterparts (CDC, 2020a). Although our dataset consisted of 40.97% males and 59.03% females, we found conflicting results with the scientific community, where men had a slightly higher stroke rate than women. Interestingly, while stroke rate remained relatively the same at younger ages, men experienced more strokes that females in the older age categories. Although, for those who have hypertension, our findings are reversed; women experienced slightly more strokes than men.





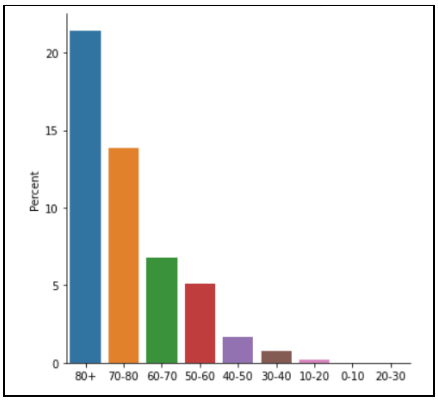


Age

According to Stanford Health Care, people over the age of 65 are significantly at a higher risk of getting a stroke than younger individuals and only 10% of stroke cases occur to those under the age of 45 (2020). For older patients, strokes are often caused by high cholesterol, diabetes, high blood pressure, and smoking patterns. However, doctors are unable to pinpoint the direct cause of the stroke for younger ages. Meanwhile, in our dataset, the average age was 43

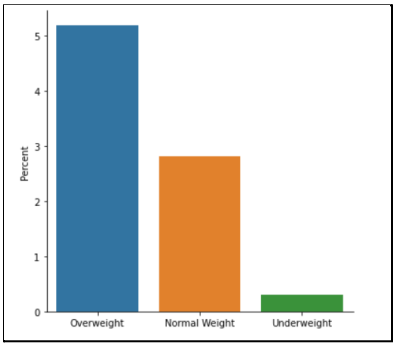
years old, where the maximum was 82 and the minimum was one-month. In comparison to the scientific literature, our findings were overwhelmingly similar. Those who were at least 70 years old had a significantly higher stroke rate than younger age groups. In particular, 21.43% of individuals over the age of 80 experienced a stroke, as well as 13.84% of individuals aged from 70-80. The group with the next highest stroke rate, 60-70, had a stroke percentage of only 6.78%.

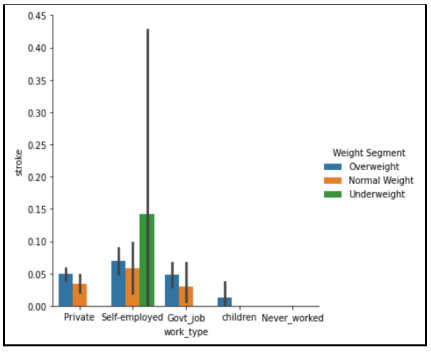
Age Segment	Stroke	Total	Percent
80+	39	182	21.43
70-80	71	513	13.84
60-70	40	590	6.78
50-60	41	805	5.09
40-50	12	705	1.70
30-40	5	631	0.79
10-20	1	485	0.21
0-10	0	459	0.00
20-30	0	539	0.00



BMI

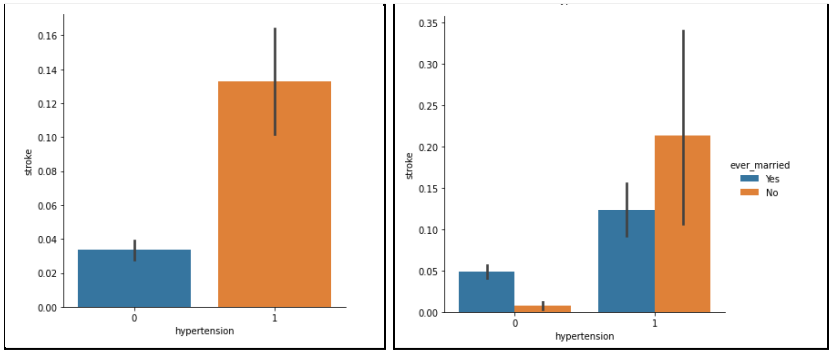
BMI is a metric that determines whether or not one's weight is proportional to their height and is calculated by dividing a person's weight by their height. Any BMI over 25 is considered overweight, under 18.4 is considered underweight, while between 18.5 to 24.9 is considered healthy (BMI Calculator, 2021). We found that, in agreement with prior scientific research, those who were considered overweight experienced almost two times the amount of strokes than people in other weight segments. Interestingly, those who were self-employed and underweight experienced the highest stroke rate among all other BMI-work type combinations. However, this finding could be skewed as we found that our dataset only contained seven people who met this criteria, one of whom had a stroke in the past.





Hypertension

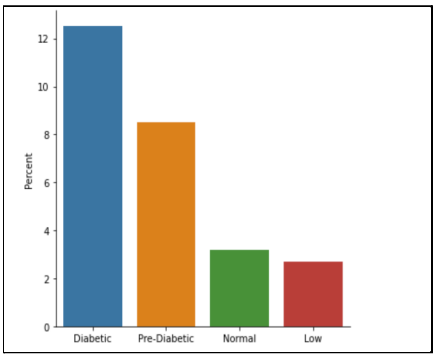
Hypertension, more commonly known as high blood pressure, occurs when an individual's blood pressure is higher than normal (CDC, 2020b). According to Harvard Health Publishing, hypertension is the leading cause of stroke; it can increase the risk of stroke by 220%. The higher the blood pressure, the greater the chance of stroke; each 10 mm rise in pressure increases the risk of stroke by 28% to 38% (2009). Our dataset supports these findings; 13% of individuals with hypertension suffered from a stroke while only 3% with normal blood pressure did. Interestingly, we found that while marriage alone is associated with higher stroke rates in our dataset, those who have hypertension have a higher stroke rate if they are not married.

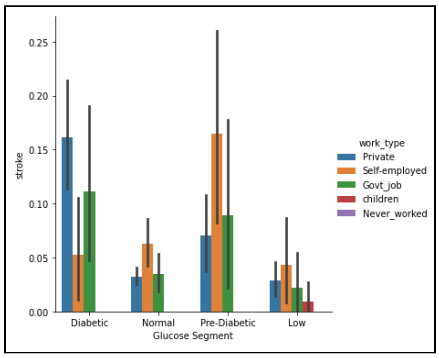


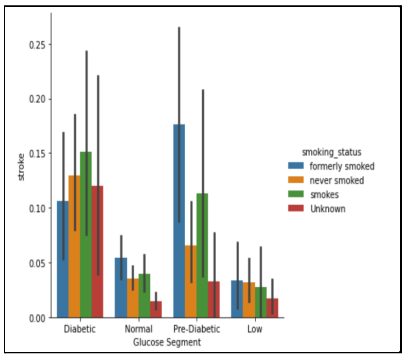
Glucose Level

High glucose levels are becoming increasingly common in patients due to sedentary lifestyles. A study done by the National Institute of Health in America suggests that people who are diabetic have a higher chance of suffering from a stroke and the severity of the stroke is much stronger (2019). In our dataset we segmented our population in four different glucose level categories: diabetic, pre-diabetic, normal and low. When we ran descriptive analysis, the effect of high glucose level on stroke was very adamant. Around 12.5% of people who were diabetic got a stroke, and 8.5% of those who were pre-diabetic also got a stroke. One interesting observation is

that diabetics were found to have more strokes if they worked a private job, while pre-diabetics had a higher stroke rate if they were self-employed. Another notable finding is those who smoke and are diabetic did not have the highest stroke rate. Rather, it was those who were pre-diabetic and had formerly smoked.

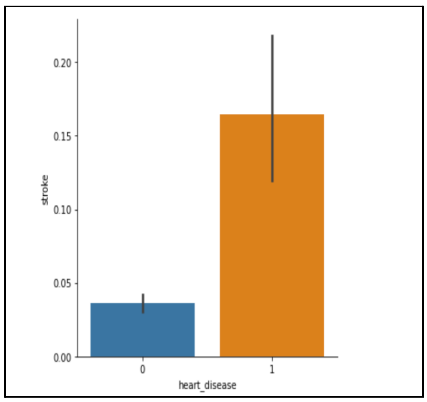


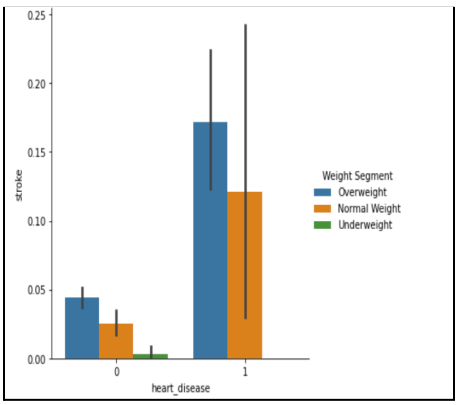


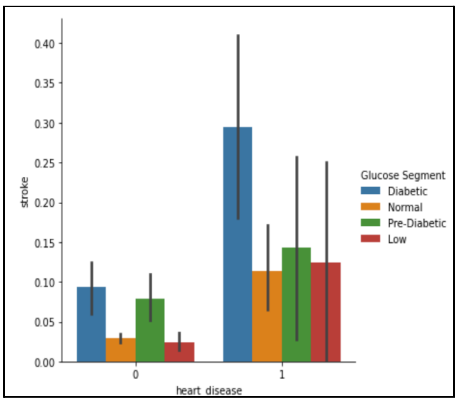


Heart Disease

Heart disease is one of the most important factors concerning stroke (CDC, 2021a). By graphing heart disease again stroke, we can see that over 15% of people who have heart disease got a stroke, whereas only 5% of people who do not have any heart complication experienced a stroke in our dataset. Research also shows that people who have diabetes can develop heart disease fifteen years earlier than those who do not have diabetes (Diabetes Canada, 2021). When incorporating glucose level as a hue, our results support this research; over 15% of people who have heart disease and diabetes experienced a stroke in our dataset, which is at least twice as large as any other heart disease-glucose level combination. In addition, we found that stroke in overweight people dramatically increases if they also have heart disease.

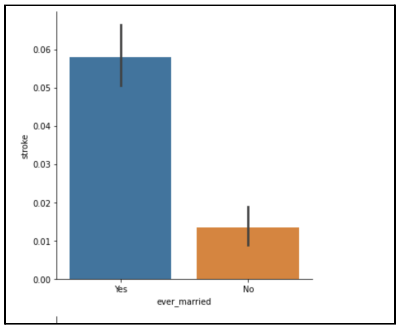


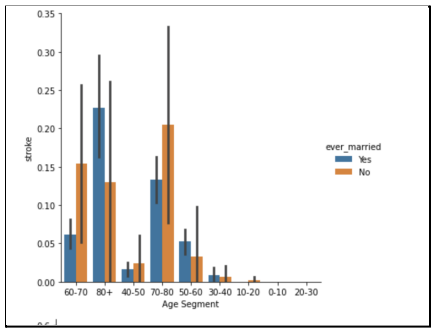




Marital Status

Stroke patients are often more vulnerable to future health complications, thereby requiring "more social or family support" (Liu et al., 2018). As a result, research has shown that marriage can significantly reduce post stroke deaths. Indeed, those who never married had a 71% greater risk of death after stroke, while those who were divorced, remarried or widowed are 23% more likely to die after stroke (Crist, 2016). On the contrary, individuals in long-term stables marriages have the lowest mortality rates as having a reliable and supporting partner had "adverse stroke outcomes" (Liu et al., 2018). Information on a person's current marital status and past history is therefore crucial to evaluating their risk of experiencing a stroke. However, as seen below, we found that 6% of patients who are currently married or were married at one point in their life suffered from a stroke, while 1.5% of patients who were never married experienced a stroke. Furthermore, our results showcased that marital status of patients over 60 years old had drastic effects on that individual's stroke rate.

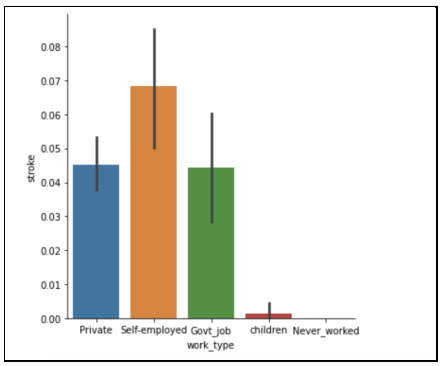


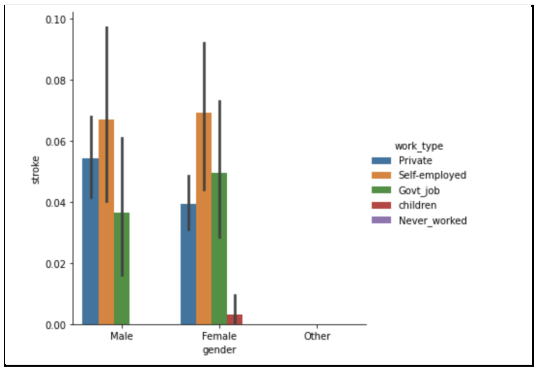


Work Type

A person's work and career can have major implications on their health, leading to overall higher stress levels and toxic lifestyle changes. Although research has shown no direct cause and effect relationship between job stress and the risk of getting a stroke, studies have indicated that high-stress levels jobs tend to foster unhealthy behaviours such as smoking, less

exercise and poor diet and are also linked with cardiovascular risk factors such as high BMI (Blaszczak, 2015). As a result, employees in high-stress level jobs are 22% more likely to experience a stroke, with female professionals garnering a 33% more likelihood of getting a stroke (Blaszczak, 2015). Oddly, there seems to be no correlation between high-stress jobs and increased risk of stroke in men, but this may be due to insufficient data on the subject. A categorical graph combining work status with stroke outcome reinforced research observations, as individuals with self-employed jobs were at a higher risk of experiencing a stroke as seen below. As opposed to observations found in the previous study, our results suggested that gender had little effect on stroke outcomes, as both men and women equally likely to get a stroke regardless of job type.

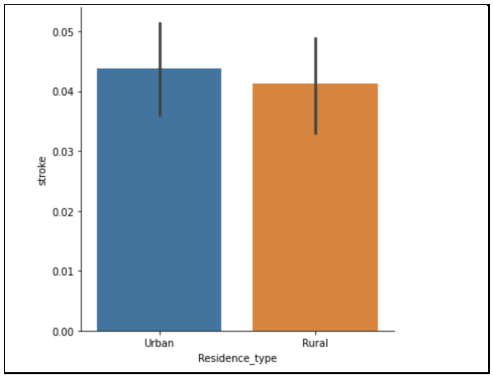




Residence Type

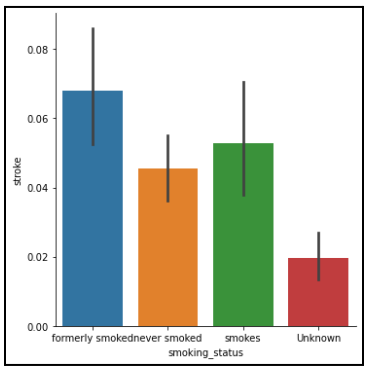
The residence type variable in our dataset specifies whether an individual lives in a rural or urban area. As an individual's residential setting can influence their lifestyle choices, this could also have an impact on a patient's overall health. Studies have shown that rural populations were more at risk of stroke. This is because rural areas often have a shortage of specialized healthcare workers for stroke patients (Leira et al., 2008). As a result, stroke patients in rural areas are not able to get the medical treatment they need, leading to increased strokes and deaths

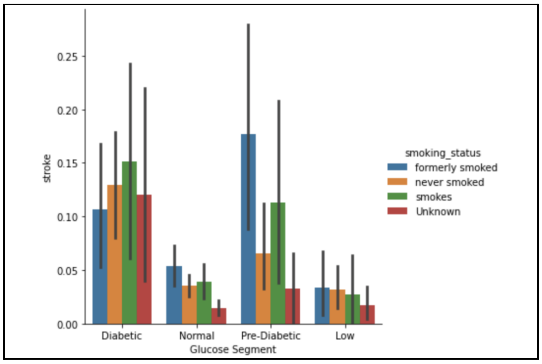
(Kapral et al., 2019). Despite research observations, our data indicated very little difference between stroke patients living in rural areas versus urban areas, a finding that was maintained throughout our analysis.



Smoking Status

Smoking habits can have serious negative effects on an individual's overall health. This is because smoking can lead to damaged blood line cells, thicken blood vessels and increased plaque buildup. Nicotine also transforms blood to be more sticky, which can cause more blood clots blocking blood to the heart and the brain (Heart disease and stroke, 2020). As a result, smoking is a major cause of heart disease and stroke resulting in 1 out of 4 deaths (Heart disease and stroke, 2020). Overall, our results were in consensus with this research, as people with current and past smoking tendencies were found to have a higher stroke rate. This further reinforces the notion that nicotine can have long term health impacts even after an individual has quit smoking. In our dataset, approximately 17.5% of pre-diabetic patients with previous smoking habits had a stroke, which was the highest among any other group. However, among current smokers, diabetic patients had the highest stroke rate.





Predictive Methodology

For our predictive methodology, we aimed to create a logistic regression model for each variable subgroup to predict whether or not an individual will suffer from a stroke. As a result, the stroke column in our dataset acts as the dependent variable, while the remaining 10 variables are the independent variables. Since experiencing a stroke or not is a binary event, we believe that logistic regression is an appropriate predictive model to use. It is important to note that the only continuous variables in our dataset are age, BMI, and average glucose level, Therefore, in order to ensure that our continuous variables are on the same scale as our dummy variables, we took the natural logarithm of each continuous variable to be used in our logistic regression.

Model 1: Demographics

The Demographics logistic regression model contains the variables that describe the characteristics of individuals in our dataset; age, gender, and BMI. According to the model's coefficients, an individual having a stroke is positively related to age and BMI. In other words, the higher the age and BMI, the more likely it is that an individual suffers from a stroke. However, given that the coefficient for age is much larger than that of BMI, we can infer that age has a much greater influence on stroke relative to BMI. In addition, the coefficient of gender was found to be 0.076949, indicating that, given the same age and BMI, males are more likely to suffer a stroke than females. In fact, the odds of a male suffering from a stroke is 1.079987 times greater than the odds for a female when age and BMI remain constant.

```
intercept: -19.332774095693665
score: 0.9574251375025463
```

	Coefs	Odds
log_age	3.822544	45.720360
log_bmi	0.282796	1.326835
Male	0.076949	1.079987

Model 2: Medical History

The Medical History logistic regression model contains the variables that describe the past and current health status of individuals in our dataset; heart disease, hypertension, glucose level, and smoking status. For smoking status, we designated those who have never smoked as the reference variable. According to the model's coefficients, stroke is positively related to heart disease, hypertension, glucose level, formerly smoked, and currently smokes respectively. For glucose level, this indicates that the higher one's blood sugar is, the greater their chance of having a stroke. By observing the odds of the model, we can see that individuals who have heart disease are 3.071916 times more likely to suffer from a stroke compared to individuals without heart complications, holding all other medical factors constant. Furthermore, individuals that have hypertension are 2.775899 times more likely to experience a stroke compared to those with normal blood pressure given the same set of medical factors. Interestingly, in comparison to people who have never smoked, people who are former smokers have a greater chance of getting a stroke than those who currently smoke according to the model. A possible explanation for this result may be that the long-term side effects of smoking have greater influence on stroke.

```
intercept: -7.811189738248201
score: 0.9574251375025463
```

	Coefs	Odds
heart_disease	1.122302	3.071916
hypertension	1.020975	2.775899
log_avg_glucose_level	0.955789	2.600723
formerly smoked	0.285501	1.330429
smokes	0.128875	1.137548
Unknown	-0.596622	0.550669

Model 3: Non-Medical Factors

The Non-Medical logistic regression model contains the variables that describe the lifestyle of individuals in our dataset; marital status, employment type, and residence type,

leaving out those who have never worked as our reference variable for work type. According to the model's coefficients, stroke is positively related to people that are married, self-employed, work private jobs, work government jobs, and live in urban areas respectively. On the other hand, stroke is negatively related with stay-at-home parents. By observing the odds of the model, we can see that individuals who are married are 2.749569 times more likely to suffer from a stroke compared to individuals who are not married, holding all other non-medical factors constant. Furthermore, individuals that live in Urban areas are 1.053797 times more likely to experience a stroke compared to those who live in rural areas with the same set of non-medical factors. Interestingly, in comparison to people who have never worked, people who are self-employed have a greater chance of getting a stroke compared to those who are privately employed, or employed by the government. On the other hand, people who are stay-at-home parents are less likely to experience a stroke compared to people who have never worked.

```
intercept: -4.35839002113096
score: 0.9574251375025463
```

	Coefs	Odds
Yes	1.011444	2.749569
Self-employed	0.786669	2.196069
Private	0.469650	1.599434
Govt_job	0.360038	1.433385
Urban	0.052400	1.053797
children	-1.328423	0.264895

Overall Findings

Based on our logistic regression models, the variables that seem to significantly increase the chances of experiencing a stroke are age, heart disease, hypertension, glucose level, and marriage. However, it is important to note that the scores of each model were found to be identical at approximately 95.74%. In theory, this score suggests that when a new data point is added to the dataset, each model will predict whether or not that individual will get a stroke with

95.74% accuracy, which is very reliable. While it is unrealistic that demographic, medical, and non-medical factors are all equally as good at predicting stroke in reality, we can infer that the predictive power of each subgroup is very similar. Therefore, while medical factors are obviously very important to consider while attempting to reduce the likelihood of getting a stroke, non-medical factors and demographics are also important to consider.

Practical Implications

To conclude our report, we would like to state key takeaways and recommendations that patients and physicians could follow to reduce risk of stroke. From five variables that have the greatest influence on stroke, we selected the three that have the most feasible solutions; hypertension, heart disease, and glucose level. Based on our findings, we believe that if individuals focus on improving these three medical factors, they can significantly reduce their chance of experiencing a stroke.

Hypertension

To lower blood pressure levels, diet changes and weight loss are two of the most effective solutions. Eating foods that are rich in whole grains, as well as fruits, vegetables and low-fat dairy products can lower your blood pressure by up to 11 mm (Mayo Clinic Staff, 2021). In addition, a reduction in sodium intake can lower your blood pressure by up to 6 mm. In general, one should limit sodium to 2,300 milligrams per day. Weight loss can also help immensely in reducing blood pressure (Mayo Clinic Staff, 2021). Therefore, it is recommended to engage in consistent physical activity throughout a given week; exercising 150 minutes a week can lower blood pressure by about 5 to 8 mm (Mayo Clinic Staff, 2021). Finally, people with hypertension should try to avoid stress-provoking situations and engaging in more relaxing activities such as yoga and meditation.

Heart Disease

In order to prevent heart disease, people should be cautious of what they eat. In particular, eating foods high in saturated fat can lead to high levels of LDL, a harmful type of cholesterol that clogs one's arteries (Diabetes Canada, 2021). In addition, it is recommended that individuals who are at risk of heart disease consult their health professional about receiving prescribed medication to protect against heart disease such as cholesterol-lowering pills (Diabetes Canada, 2021). Based on our previously conducted descriptive analysis, people that suffer from heart disease who are also overweight or diabetic should be extra-mindful, since a combination of any of these factors can increase stroke risk dramatically.

Glucose Level

One of the most effective ways to prevent high blood sugar levels is to exercise often (Diabetes Canada, 2021). Physical activity makes the body more sensitive to insulin and can lower one's blood sugar up to 24 hours after a workout (Diabetes Canada, 2021). Furthermore, people who are diabetic and pre-diabetic should be mindful of their carbohydrate intake level (Diabetes Canada, 2021). Having a high-carb diet leads to more sugar build-up in the bloodstream, directly contributing to higher glucose levels. In addition to these lifestyle changes, receiving an A1C test, meant to diagnose Type 1 or Type 2 diabetes, every twelve months is crucial to ensure that one's blood sugar is trending in normal directions (Diabetes Canada, 2021). Based on our descriptive analysis above, those who have high glucose levels should be mindful of their work type and smoking status, since certain glucose level-work type combinations have a higher stroke rates than others.

References

Blaszczak, A. (2015, October 15). *High-stress jobs may raise stroke risk*. Live Science. Retrieved April 11, 2021, from https://www.livescience.com/52482-high-stress-jobs-stroke-risk.html

Body Mass Index (BMI) Calculator. (2021). Diabetes Canada, www.diabetes.ca/managing-my-diabetes/tools---resources/body-mass-index-(bmi)-calculator. Accessed 12 Apr. 2021.

CDC. (2020a, August 5). *Women and Stroke*. Centers for Disease Control and Prevention. https://www.cdc.gov/stroke/women.htm

CDC. (2020b, May 19). *High Blood Pressure Symptoms and Causes*. Centers for Disease Control and Prevention. https://www.cdc.gov/bloodpressure/about.htm

CDC. (2021a). *Heart Disease and Stroke*. Centers for Disease Control and Prevention. https://www.cdc.gov/chronicdisease/resources/publications/factsheets/heart-disease-stroke.htm

CDC. (2021b). *Stroke Facts*. Centers for Disease Control and Prevention. https://www.cdc.gov/stroke/facts.htm

Crist, C. (2016, December 14). *Marital status, history linked to survival after stroke*. Reuters. Retrieved April 11, 2021, from https://www.reuters.com/article/us-health-stroke-marriage-idUSKBN1432T2

fedesoriano. (2021). *Stroke Prediction Dataset.* Kaggle. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Harvard Health Publishing (2009, October). *Blood pressure and your brain*. Retrieved from https://www.health.harvard.edu/heart-health/blood-pressure-and-your-brain

Diabetes Canada. (2021). *Heart disease & stroke*. Retrieved from https://www.diabetes.ca/managing-my-diabetes/preventing-complications/heart-disease---stroke

Johnson, W., Onuma, O., Owolabi, M., & Sachdev, S. (2016). Stroke: a global response is needed. *Bulletin of the World Health Organization*, *94*(9), 634–634A. https://doi.org/10.2471/blt.16.181636

Kapral, M.K. et al. (2019, February 14). *Rural-Urban differences in stroke risk factors, incidence, and mortality in people with and without Prior Stroke*. Retrieved April 11, 2021, from https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.118.004973

Leira, E.C. et al. (2008). Rural-Urban Differences in Acute Stroke Management Practices. *Archives of Neurology*, *65*(7). https://doi.org/10.1001/archneur.65.7.887

Liu, Q. et al. (2018, April). Association between marriage and outcomes in patients with acute ischemic stroke. *Journal of Neurology*, *265*(4), 942–948. https://doi.org/10.1007/s00415-018-8793-z

Mayo Clinic Staff. (2021, February 24). *10 drug-free ways to control high blood pressure*. Mayo Clinic, Retrieved from https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/high-blood-pressure/art-20046974?pg=1

MedBroadcast. (2021, March 18). *Why is stroke risk reduction so important? - Stroke Risk Reduction*. MedBroadcast. https://www.medbroadcast.com/channel/stroke-risk-reduction/i-want-to-learn-more-about-stroke-risk-reduction/why-is-stroke-risk-reduction-so-important

National Institutes of Health. (2019, July 23). *Researchers get a handle on how to control blood sugar after stroke*. U.S. Department of Health & Human Services. Retrieved https://www.nih.gov/news-events/news-releases/researchers-get-handle-how-control-blood-sugar-after-stroke

Appendix A: Variable List

Variable	Description	Type
id	Identifier of patient	Quantitative
gender	Gender	Male Female
age	Age of patient in years	Quantitative
hypertension	Whether the patient has experienced hypertension	Binary (1 = yes, 0 = no)
heart_disease	Whether the patient has a heart disease	Binary (1 = yes, 0 = no)
ever_married	Marital status	Binary (1 = yes, 0 = no)
work_type	Type of employment	Private Self-Employed Government Job
Residence_type	Area where the patient resides	Urban Rural
avg_glucose_level	Average glucose level of patient	Quantitative
bmi	Body Mass Index of patient	Quantitative
smoking_status	Past or current smoking experience	Never Smoked Formerly Smoked Smokes
stroke	Whether the patient has experienced a stroke	Binary (1 = yes, 0= no)

Source: fedesoriano (2021)

Appendix B: Technical Code

Import Modules & Data

```
# For data manipulation
import pandas as pd
import numpy as np
import math
# For graphical analysis
import seaborn as sns
import matplotlib.pyplot as plt
# For predictive analysis
from sklearn.linear_model import LogisticRegression

# Generate DataFrame
data = pd.read_csv('Stroke.csv')
data
```

Data Preprocessing

```
#Missing Values
# Original number of rows in DataFrame
len(data)

# Missing values per column
data.isnull().sum()

# Remove 201 missing observations from bmi column
data = data.dropna()
data.isnull().sum()

# Number of rows in DataFrame after removing missing values
len(data)

smoking_status_unknown = data[(data.smoking_status=='Unknown')]
len(smoking_status_unknown)

#Dummy Variabels
dummy_smoking_status = pd.get_dummies(data.smoking_status)
dummy_ever_married = pd.get_dummies(data.ever_married)
dummy_work_type = pd.get_dummies(data.work_type)
dummy_Residence_type = pd.get_dummies(data.Residence_type)
dummy_gender = pd.get_dummies(data.gender)
```

```
data = data.join(dummy_smoking_status)
data = data.join(dummy_ever_married)
data = data.join(dummy_work_type)
data = data.join(dummy_Residence_type)
data = data.join(dummy_gender)

#Correlation Matrices
def CorrMatrix(*arg):
        new_data = data[[*arg]]
        corrMatrix = new_data.corr()
        return pd.DataFrame(corrMatrix)

# Demographic
CorrMatrix('age','Male','Female','stroke')

# Medical Factors
CorrMatrix('hypertension','heart_disease','avg_glucose_level','bmi','smokes','formerly
smoked','never smoked','stroke')

# Non-Medical Factors
CorrMatrix('Yes','No','Govt_job','Private','Self-employed','children','Never_worked','Urban','strok
e')
```

Descriptive Analysis

```
#Descriptive Statistics
data.describe()

#Categorical Plots
#Categorical variables
column_headers = list(data.columns.values)
del column_headers[12:]
column_headers.remove('id')
column_headers.remove('age')
column_headers.remove('avg_glucose_level')
column_headers.remove('bmi')
column_headers.remove('stroke')
for column in column_headers:
        sns.catplot(x=column , y='stroke', kind='bar', data=data)

# Continuous variables
def Group(column):
        df = pd.DataFrame()
        df["Stroke"] = data['stroke'].groupby(data[column]).sum()
        df["Total"] = data[column].groupby(data[column]).count()
        percents = []
```

```
        for i in df.index:
        stroke = int(df.loc[i,['Stroke']])
        total = int(df.loc[i,['Total']])
      percent = np.round((stroke/total)*100,2)
        percents.append(percent)
        df['Percent'] = percents
        df = df.sort_values('Percent', ascending=False)
        barchart = sns.catplot(x=df.index.tolist() , y='Percent', kind='bar', data=df)
        return df
        return barchart

#bmi
weight_status = []
for i in data['bmi'].tolist():
        if i>=25:
        weight_status.append('Overweight')
        elif 18.5<=i<25:
        weight_status.append('Normal Weight')
        elif 18.5>i:
        weight_status.append('Underweight')
data['Weight Segment'] = weight_status
Group('Weight Segment')

#avg_glucose_level
glucose_level = []
for i in data['avg_glucose_level'].tolist():
        if 200<i:
        glucose_level.append('Diabetic')
        elif 200>=i>140:
        glucose_level.append('Pre-Diabetic')
        elif 140>=i>70:
        glucose_level.append('Normal')
        elif i<=70:
        glucose_level.append('Low')
data['Glucose Segment']=glucose_level
Group('Glucose Segment')

#age
age = []
for i in data['age'].tolist():
        if 80<=i:
        age.append('80+')
           elif 80>i>=70:
        age.append('70-80')
        elif 70>i>=60:
        age.append('60-70')
```

```
        elif 60>i>=50:
        age.append('50-60')
        elif 50>i>=40:
        age.append('40-50')
        elif 40>i>=30:
        age.append('30-40')
        elif 30>i>=20:
        age.append('20-30')
        elif 20>i>=10:
        age.append('10-20')
        else:
        age.append('0-10')
data['Age Segment']=age
Group('Age Segment')


#Variables with Hues
hues = data.columns.tolist()[1:11]
hues[1] = 'Age Segment'
hues[-3] = 'Glucose Segment'
hues[-2] = 'Weight Segment'
for i in hues:
        for j in hues:
        if i!=j:
        sns.catplot(x=i , y='stroke', hue=j, kind='bar', data=data)

# Assessing underweight and self-employed outlier
weight_work = data[['Weight Segment','work_type','stroke']]
weight_work.columns = ['weight_segment','work_type','stroke']
stroke=len(weight_work[(weight_work.weight_segment=='Underweight')&(weight_work.work_
type=='Self-employed')&(weight_work.stroke==1)])
total=len(weight_work[(weight_work.weight_segment=='Underweight')&(weight_work.work_ty
pe=='Self-employed')])
print(stroke, "out of",total,"people who are underweight and self-employed experienced a
stroke")
```

Logistics Regression

```
#Continuous Variable Transformations
def Log(column_header):
        var_list = data[column_header].tolist()
        log_var_list = []
    for i in var_list:
        log = np.log(i)
        log_var_list.append(log)
        new_log_column = "log_" + column_header
        data[new_log_column] = log_var_list
```

```
Log('age')
Log('bmi')
Log('avg_glucose_level')

# logistic regression function
def VarPredict(variables):
        y = data['stroke']
        x = data[variables]
        model = LogisticRegression(fit_intercept=True).fit(x,y)
        int = model.intercept_
        coefs = list(model.coef_)
        score = model.score(x,y)
        print("intercept:", int[0])
        print("score:", score)
        odds = []
        for coef in coefs:
        odd = math.e**(coef)
        odds.append(odd)
        df = pd.DataFrame(index = variables)
        df['Coefs'] = coefs[0]
        df['Odds'] = odds[0]
        return df.sort_values('Odds', ascending = False)

#Demograohic
X = [‘log_age’,’Male’,’log_bmi]
VarPredict(x)

#Medical Factors
x = ['hypertension','heart_disease','log_avg_glucose_level','smokes','formerly
smoked','Unknown']
VarPredict(x)

#Non-Medical Factors
x = ['Yes','Govt_job','Private','Self-employed','children','Urban']
VarPredict(x)
```