Stroke Prediction

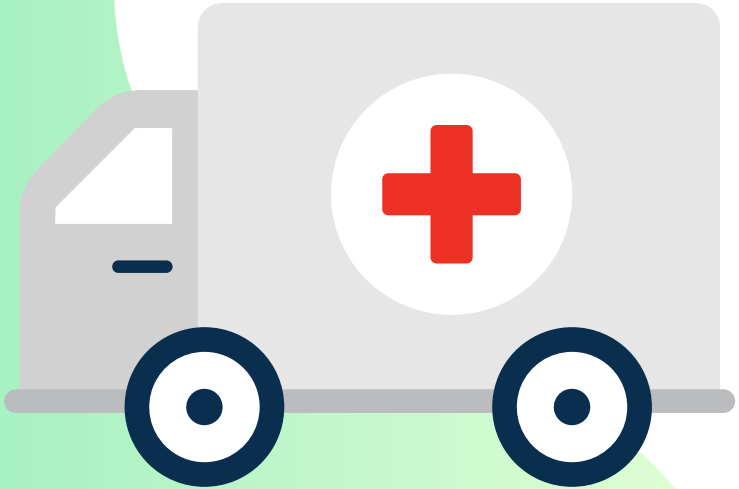
Table of Contents

00

Executive Summary

Executive Summary

The current project aims to find out if some relevant variables (i.e. gender, age, marital status, residence type, hypertension, heart disease, average glucose in blood, Body Mass Index and smoking status) are significant predictors of stroke, in order to facilitate patient screening and apply early medication or therapy to enhance treatment success rate and reduce death rate. After data cleaning and encoding, descriptive analysis, binary logistic regression model and K-means clustering (2 and 3 clusters) will be used to analyze a public Kaggle dataset about stroke prediction. Results found that age, hypertension, heart disease and diabetes are significant predictors of stroke: people who have an older age and diagnosed with these clinical features are more likely to have stroke. Hypothesis (H0 and H1) will be made based on the results and some limitations of the current study will also be discussed at the end.



01

Background & Problem Definition

≥10 person is killed by stroke

in every minute

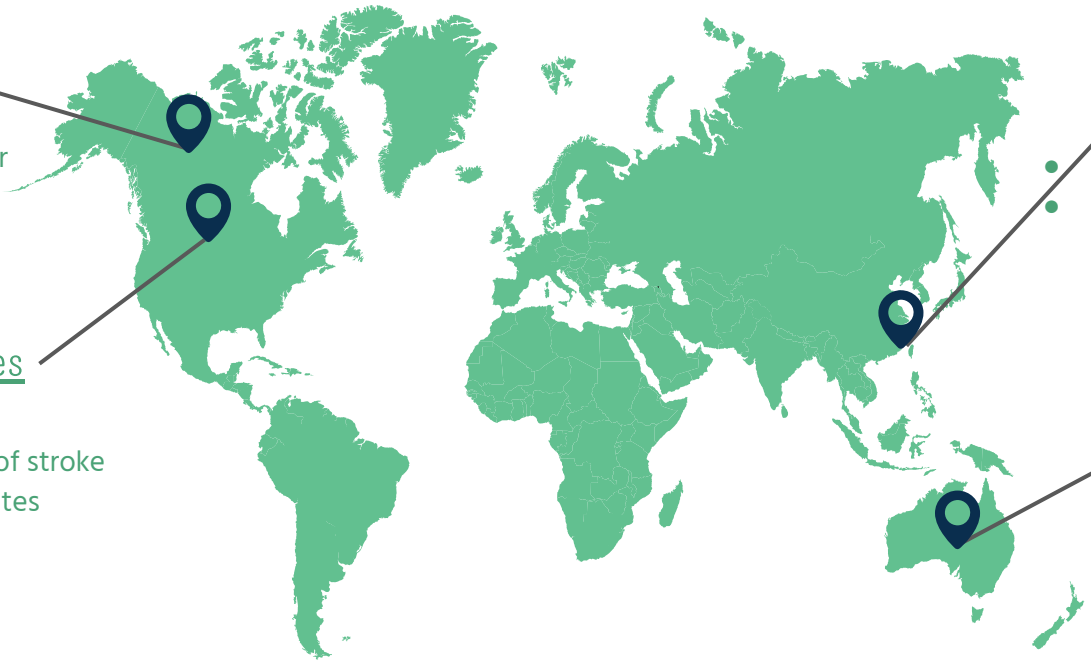
Stroke is a worldwide health issue

Canada

- No.3 killer
- 10th largest contributor to disability-adjusted life years

United States

- No.5 killer
- Someone dies of stroke in every 4 minutes



Hong Kong

- No.4 killer
- About 3,500 people are killed by stroke each year

Australia

- Underlying cause of 8,400 deaths in 2018
- Account for 5.3% of all deaths in 2018

Types of Stroke

Ischemic stroke
(缺血性中風)

- Most common type of stroke
- Happens when the brain's blood vessels become narrowed or blocked which caused by fatty deposits/blood clots/other debris
- Causing severely reduced blood flow (ischemia)

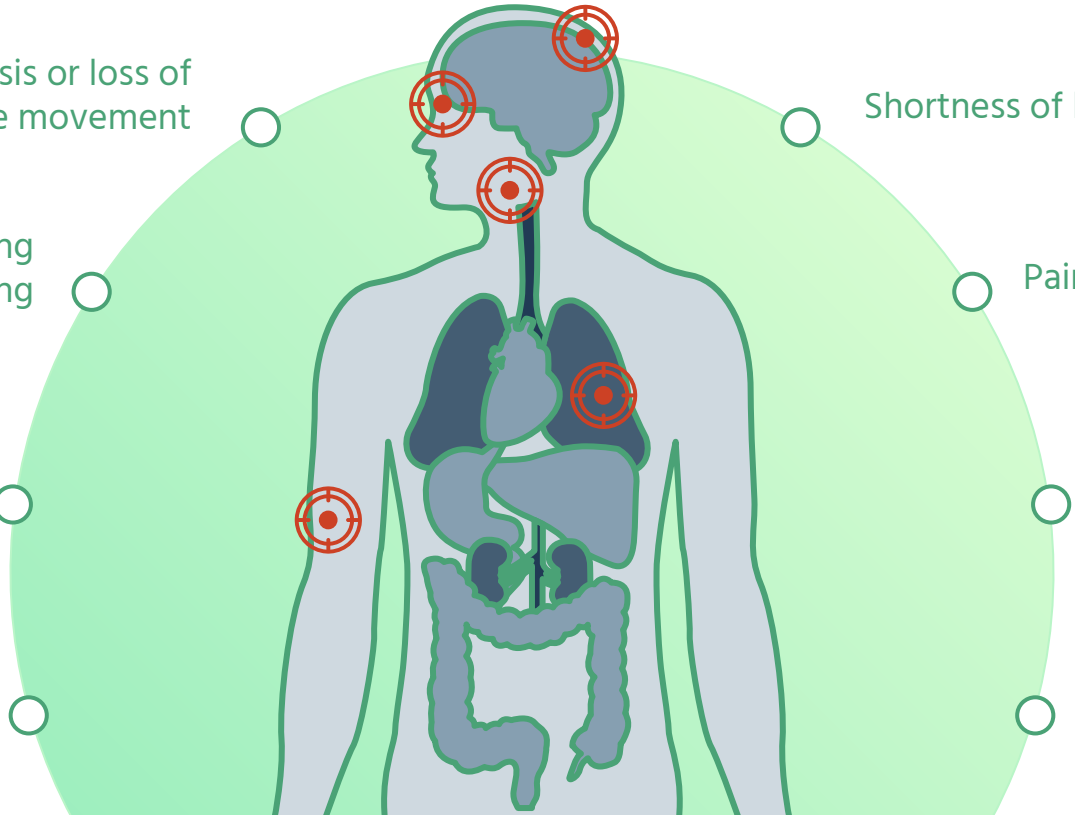
Hemorrhagic stroke
(出血性中風)

- Occurs when a blood vessel in the brain leaks or ruptures
- May due to uncontrolled high blood pressure, overtreatment with blood thinners, or bulges at weak spots in blood vessel walls

Transient ischemic attack
(短暫性腦缺血發作)

- Sometimes known as a "mini-stroke"
- Caused by a temporary decrease in blood supply to part of the brain which may last as little as five minutes
- Occurs when a clot or debris reduces or blocks blood flow to part of the nervous system

Common Stroke Symptoms & Complications (並發症)



- Paralysis or loss of muscle movement
- Difficulty talking or swallowing
- Memory loss or thinking difficulties
- Emotional problems
- Shortness of breath (氣促)
- Pain
- Vision problems
- Changes in behavior and self-care ability

Risk Factors

Lifestyle

- Being overweight/obese
- Physical inactivity
- Heavy or binge drinking
- Use of illegal drugs (e.g. cocaine, methamphetamine)

Medical

- High blood pressure
- Cigarette smoking or secondhand smoke exposure
- High cholesterol
- Diabetes
- Obstructive sleep apnea
- Cardiovascular diseases
- Personal or family history of stroke/heart attack/transient ischemic attack
- COVID-19 infection

Others

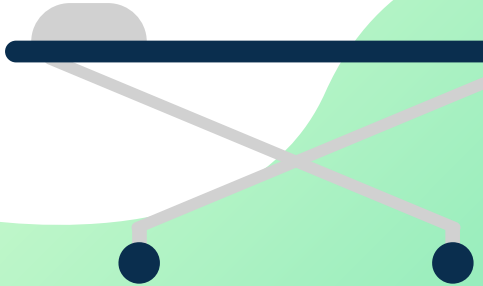
- Age — People age 55 or older have a higher risk
- Race — African Americans have a higher risk
- Sex — Men have a higher risk; but women are usually older when they have strokes and more likely to die than are men
- Hormones — Use of birth control pills/hormone therapies

Problem Definition

Considering that stroke is a prevalent health issue around the world, how can we identify significant factors which may predict stroke, so that people with high risk of stroke can self-monitor and prevent better, be screened out and diagnosed easier, and receive medication or therapy earlier to enhance treatment success rate and reduce sequelae and death rate?

02

Data Source, IV & DV



Data Acquisition: Online Public Dataset

Source: Kaggle

https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Context

The dataset is used to predict whether a patient is likely to get stroke based on several clinical features.

Sample Size: N=5,110

Independent Variables (Predictors): 10 Stroke Clinical Features

1. **Gender**
   - Male
   - Female
   - Other
2. **Age**
3. **Marital Status**
   - Have married
   - Have not married
4. **Work Type**
   - Children
   - Government job
   - Never worked
   - Private
   - Self-employed
5. **Residence Area**
   - Rural
   - Urban
6. **Hypertension**
   - Yes
   - No
7. **Heart Disease**
   - Yes
   - No
8. **Average Glucose Level in Blood**
9. **Body Mass Index (BMI)**
10. **Smoking Status**
    - Formerly smoked
    - Never smoked
    - Smokes
    - Unknown

Dependent Variable (Predicted): Stroke (yes or not)

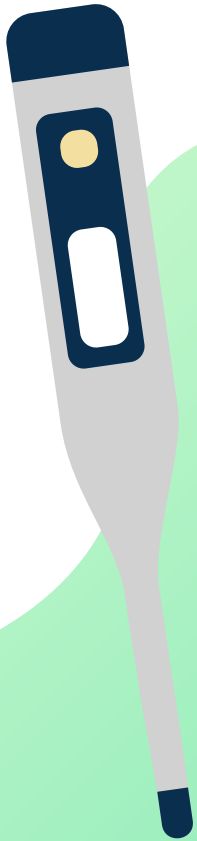
Output: Binary and categorical values (1/0)

- ➔ 1: Have stroke
- ➔ 0: Do not have stroke

Aim: Use various clinical predictors (IV/input) to predict likelihood of stroke (DV/output) and find out what are the significant predictors of stroke.

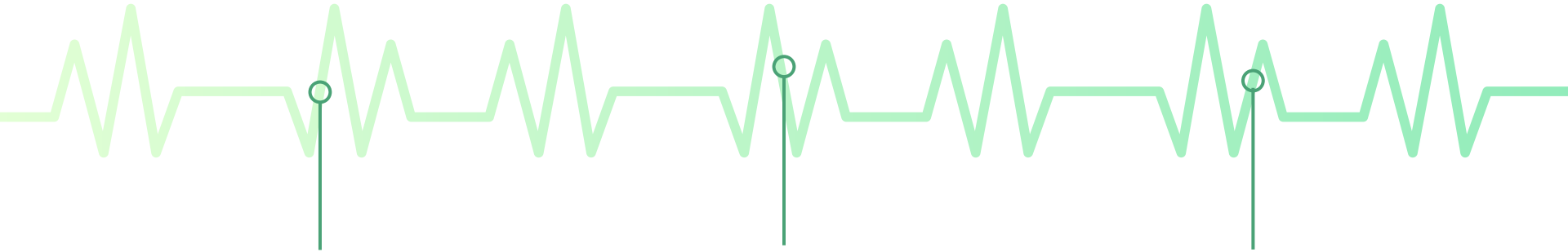
03

SPSS Data Analysis



Before Data Analysis:

Data Cleaning



Exclude work type

- Limited and overlapped information provided by this predictor (e.g. “children” and “private” as work type)
- More than 55% answer is “private”
- Not a useful/meaningful predictor
- May not help to generate key insights
- Final number of predictors (IV): 9 (original: 10)

Missing Values

- “N/A” in BMI will be replaced by mean value (i.e. 28.9)

Better Categorization

- Delete one data who stated “other” as gender (considered that there is only one and will not impact the analysis much)
- Final sample size: N=5,109 (original: N=5,110)

Before Data Analysis:

Literature Support on Data Binning/Encoding

1) Age groups
   a) Erikson's Psychosocial Developmental Stages
      i) 0-3 years old → Infancy and toddler
      ii) 3-12 years old → Childhood
      iii) 13-18 years old → Adolescence
      iv) 19-39 years old → Early adulthood
      v) 40-65 years old → Middle adulthood
      vi) 66 years old or above → Older adulthood

2) Average glucose level in blood
   a) Random blood sugar test
      i) A blood sample will be taken at a random time, regardless of when we last ate
         - ❖ Blood sugar level of 200 milligrams per deciliter (mg/dL) or higher suggests diabetes

3) Body Mass Index (BMI)
   a) BMI = Weight (kilograms) / Height (metre)*Height (metre)
      i) BMI equal or above 25 is obese
      ii) BMI between 23 and 24.9 is overweight
      iii) BMI between 18.5 and 22.9 is normal range
      iv) BMI less than 18.5 is underweight

Before Data Analysis:

Data Binning/Encoding (Demographic Factors)

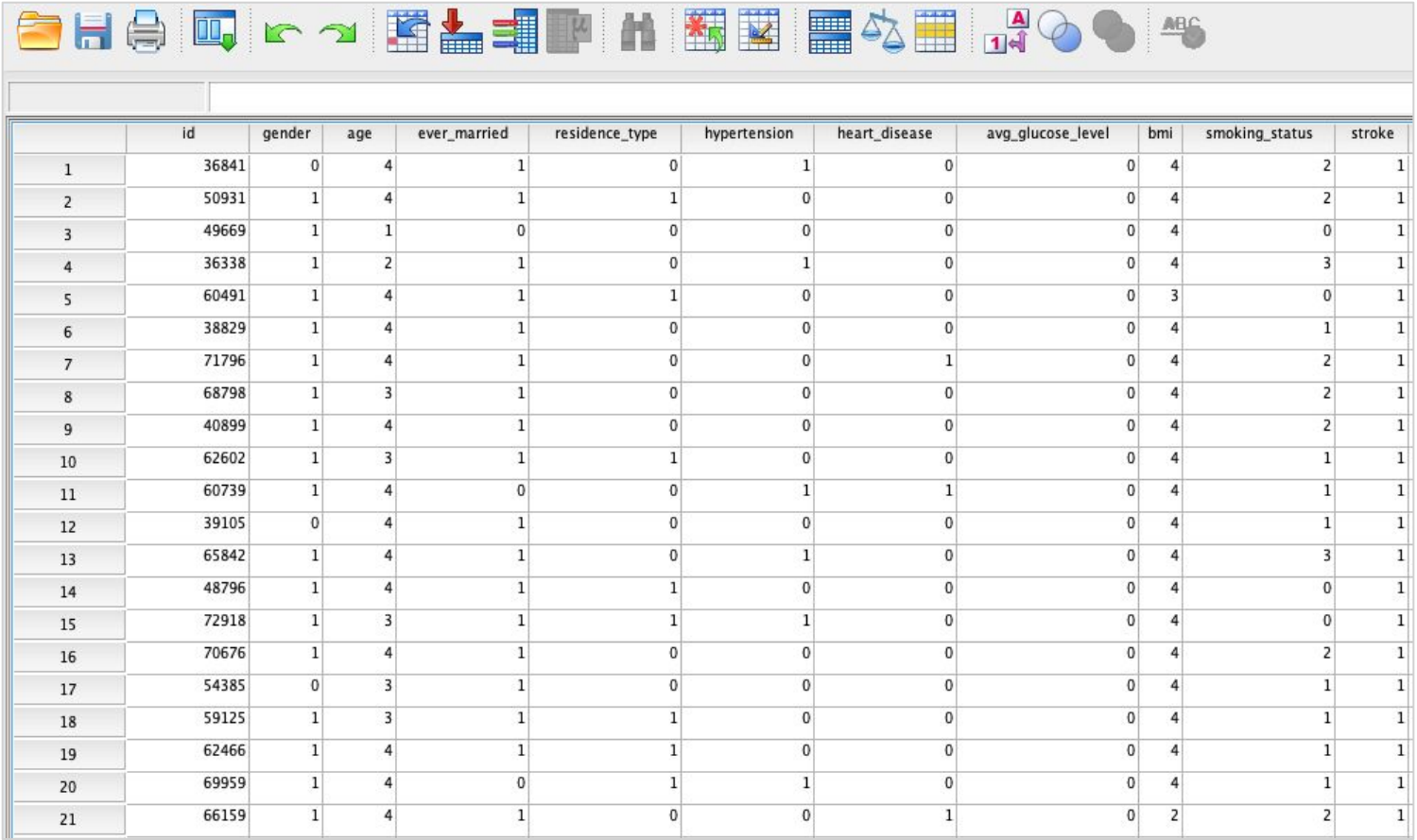
	0	1	2	3	4
Gender	Male	Female	/	/	/
Age	/	0 - 18 [infancy/childhood/adolescence]	19 - 39 [early adulthood]	40 - 65 [middle adulthood]	> 65 [older/senior adulthood]
Marital Status	Not married	Have married	/	/	/
Residence Type	Rural	Urban	/	/	/

Before Data Analysis:

Data Binning/Encoding (Health-related Factors)

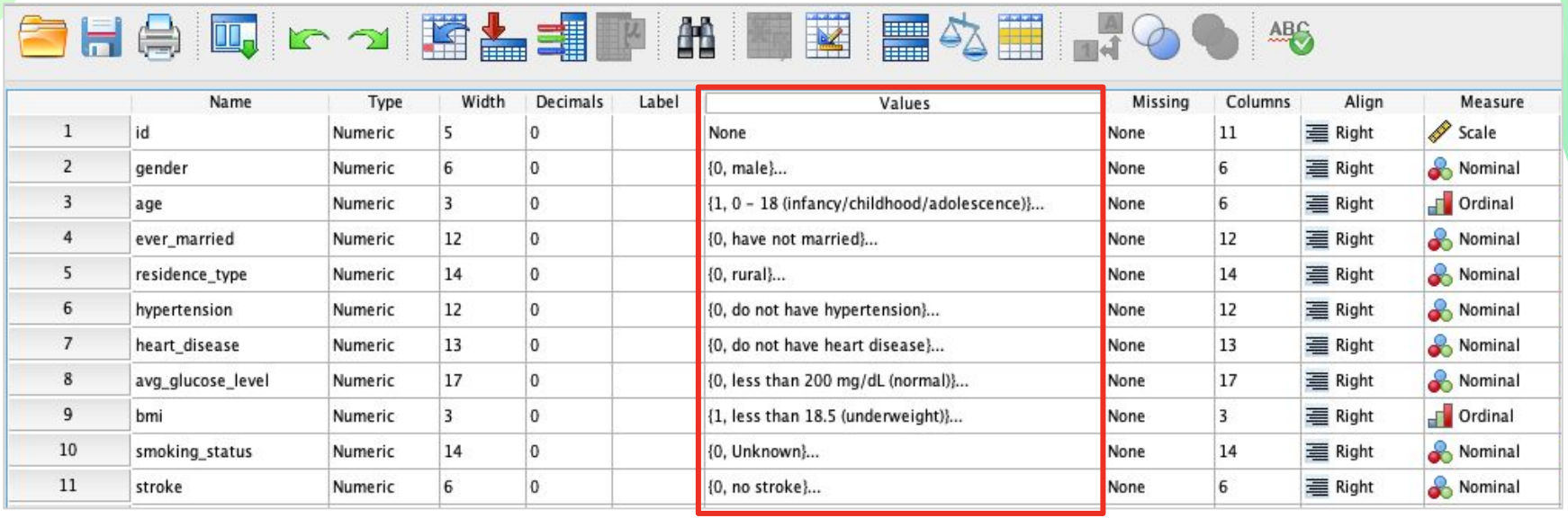
	0	1	2	3	4
Hypertension	Do not have hypertension	Have hypertension	/	/	/
Heart Disease	Do not have heart disease	Have heart disease	/	/	/
Average Glucose Level in Blood	< 200 mg/dL [normal]	≥ 200 mg/dL [diabetes]	/	/	/
Body Mass Index (BMI)	/	< 18.5 [underweight]	18.5 - 22.9 [normal range]	23 - 24.9 [overweight]	≥ 25 [obese]
Smoking Status	Unknown	Never smoked	Formerly smoked	Smokes	/

Cleaned and Coded Dataset on SPSS



	id	gender	age	ever_married	residence_type	hypertension	heart_disease	avg_glucose_level	bmi	smoking_status	stroke
1	36841	0	4	1	0	1	0	0	4	2	1
2	50931	1	4	1	1	0	0	0	4	2	1
3	49669	1	1	0	0	0	0	0	4	0	1
4	36338	1	2	1	0	1	0	0	4	3	1
5	60491	1	4	1	1	0	0	0	3	0	1
6	38829	1	4	1	0	0	0	0	4	1	1
7	71796	1	4	1	0	0	1	0	4	2	1
8	68798	1	3	1	0	0	0	0	4	2	1
9	40899	1	4	1	0	0	0	0	4	2	1
10	62602	1	3	1	1	0	0	0	4	1	1
11	60739	1	4	0	0	1	1	0	4	1	1
12	39105	0	4	1	0	0	0	0	4	1	1
13	65842	1	4	1	0	1	0	0	4	3	1
14	48796	1	4	1	1	0	0	0	4	0	1
15	72918	1	3	1	1	1	0	0	4	0	1
16	70676	1	4	1	0	0	0	0	4	2	1
17	54385	0	3	1	0	0	0	0	4	1	1
18	59125	1	3	1	1	0	0	0	4	1	1
19	62466	1	4	1	1	0	0	0	4	1	1
20	69959	1	4	0	1	1	0	0	4	1	1
21	66159	1	4	1	0	0	1	0	2	2	1

Cleaned and Coded Dataset on SPSS



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	id	Numeric	5	0		None	None	11	Right	Scale
2	gender	Numeric	6	0		{0, male}...	None	6	Right	Nominal
3	age	Numeric	3	0		{1, 0 - 18 (infancy/childhood/adolescence)}...	None	6	Right	Ordinal
4	ever_married	Numeric	12	0		{0, have not married}...	None	12	Right	Nominal
5	residence_type	Numeric	14	0		{0, rural}...	None	14	Right	Nominal
6	hypertension	Numeric	12	0		{0, do not have hypertension}...	None	12	Right	Nominal
7	heart_disease	Numeric	13	0		{0, do not have heart disease}...	None	13	Right	Nominal
8	avg_glucose_level	Numeric	17	0		{0, less than 200 mg/dL (normal)}...	None	17	Right	Nominal
9	bmi	Numeric	3	0		{1, less than 18.5 (underweight)}...	None	3	Right	Ordinal
10	smoking_status	Numeric	14	0		{0, Unknown}...	None	14	Right	Nominal
11	stroke	Numeric	6	0		{0, no stroke}...	None	6	Right	Nominal

Data Analysis

Using SPSS

01 Descriptive Statistics

02 Binary Logistic Regression

03 K-Means Clustering

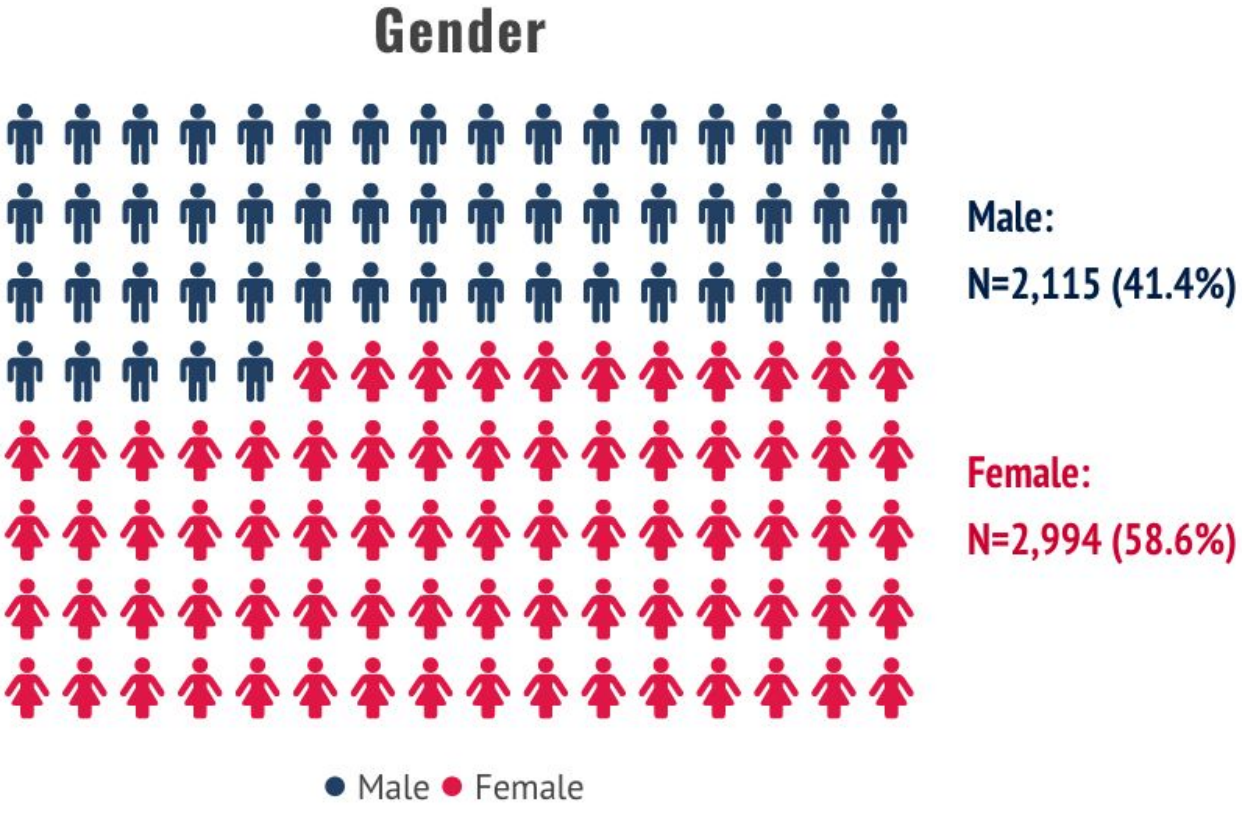
(*significance level set at 0.05)

1. Descriptive Statistics

- ➔ Show data distributions
- ➔ Central tendency & variability

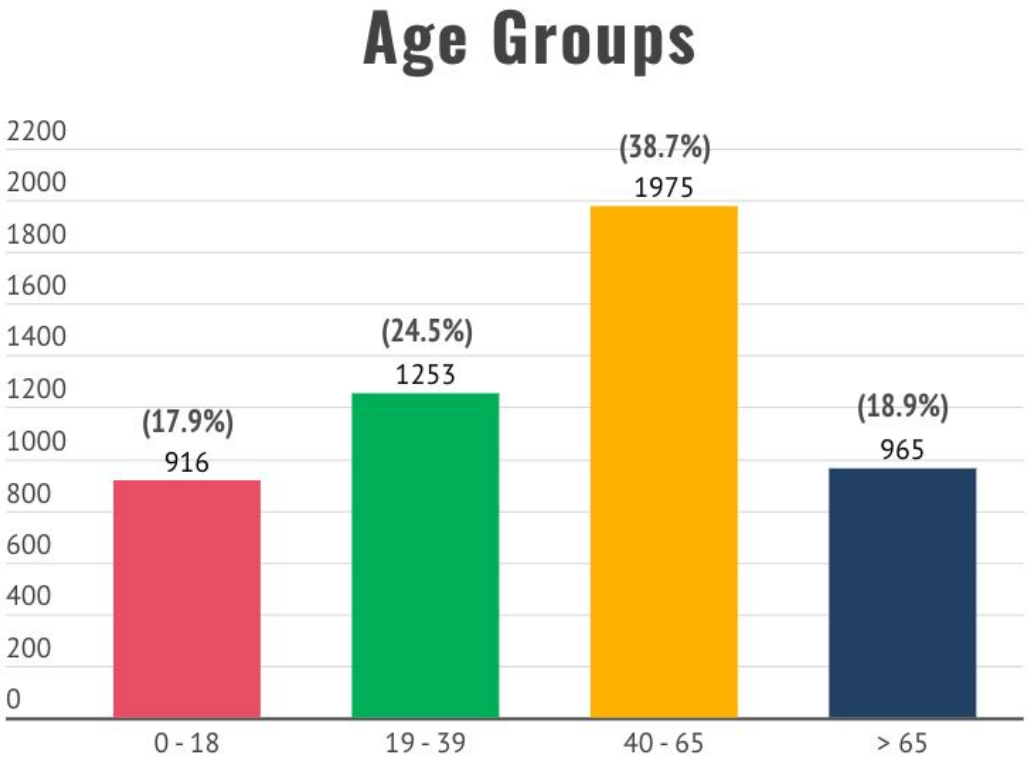
1. Descriptive Analysis

(IV) Demographic Variables



1. Descriptive Analysis

(IV) Demographic Variables



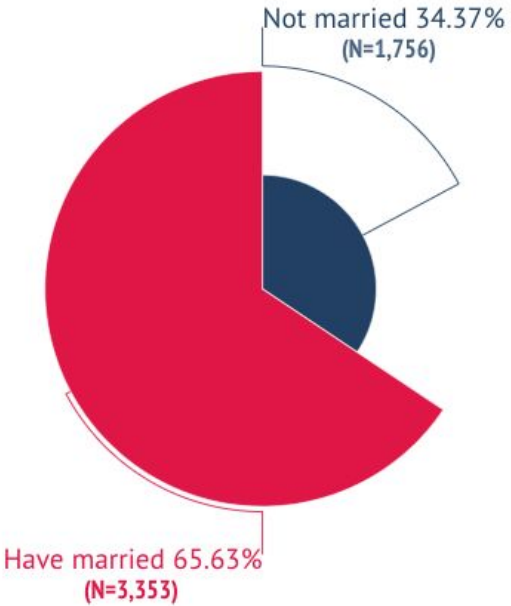
Age:

- Mean: 43.23
- Median: 45
- Range (Min - Max): 0.08 - 82
- Mode: 40 - 65 (middle adulthood)

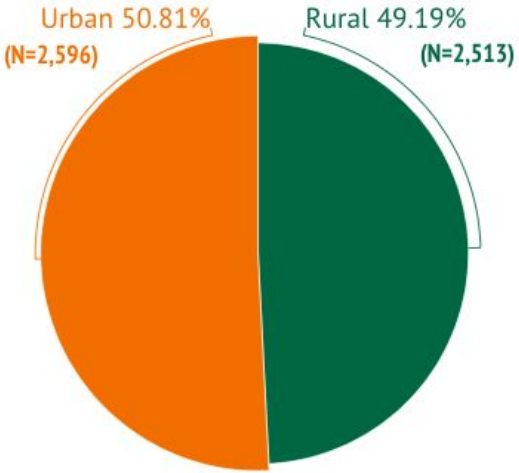
1. Descriptive Analysis

(IV) Demographic Variables

Marital Status

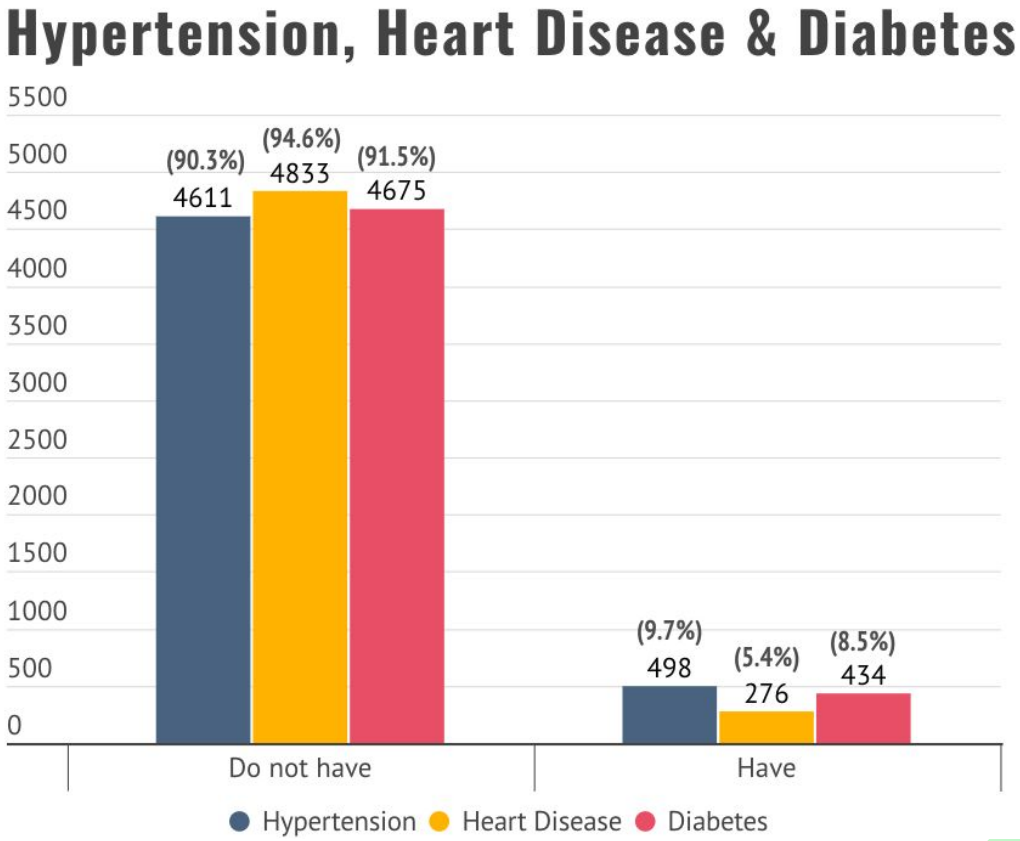


Residence Type



1. Descriptive Analysis

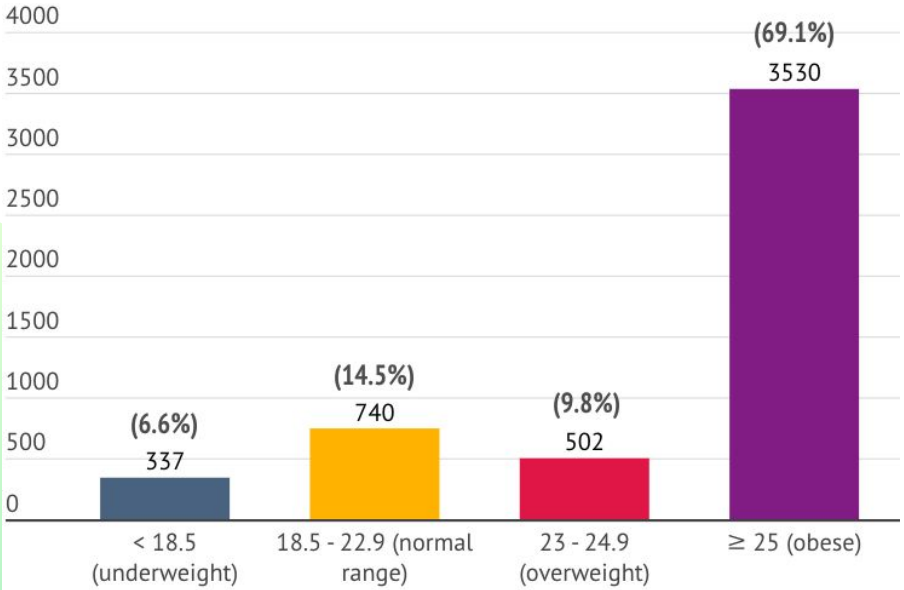
(IV) Health-related Variables



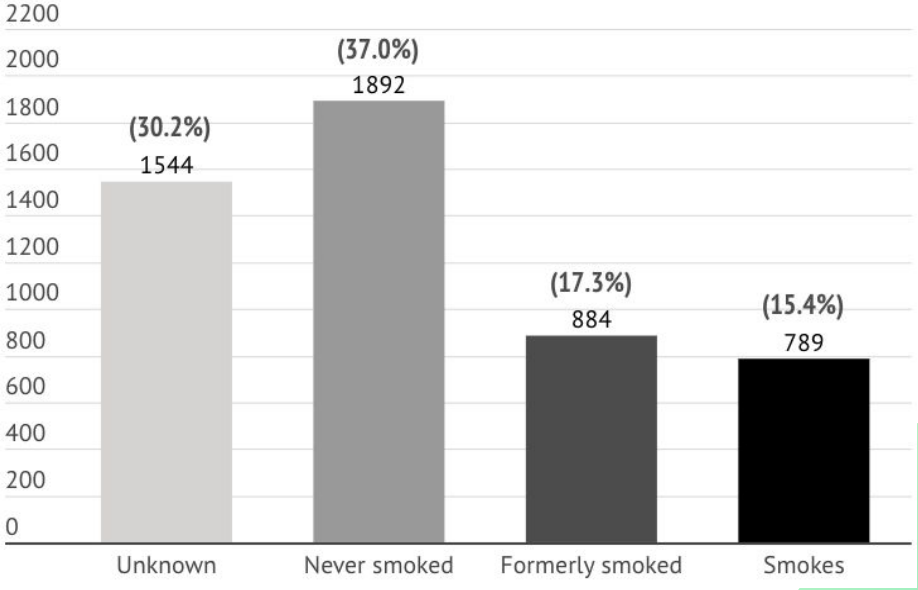
1. Descriptive Analysis

(IV) Health-related Variables

Body Mass Index (BMI)

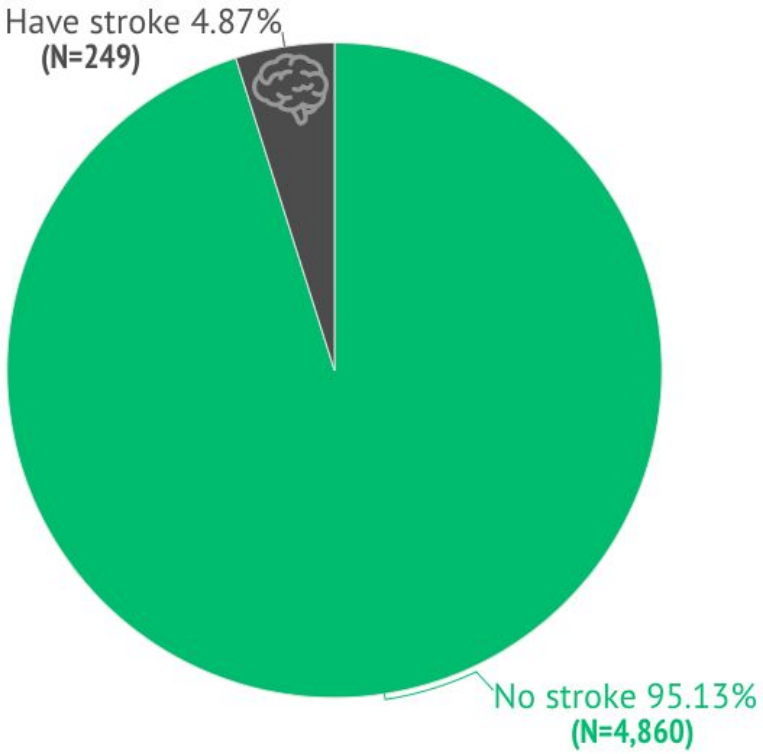


Smoking Status



1. Descriptive Analysis

Dependent Variable (DV)



1. Descriptive Analysis

(Supplement) SPSS Raw Statistics Table

Statistics

		gender	age	ever_married	residence_type	hypertension	heart_disease	avg_glucose_level	bmi	smoking_status	stroke
N	Valid	5109	5109	5109	5109	5109	5109	5109	5109	5109	5109
	Missing	0	0	0	0	0	0	0	0	0	0
Mean		.59	43.23	.66	.51	.10	.05	106.14	28.89	1.18	.05
Median		1.00	45.00	1.00	1.00	.00	.00	91.88	28.40	1.00	.00
Mode		1	78	1	1	0	0	94	29	1	0
Std. Deviation		.493	22.614	.475	.500	.297	.226	45.285	7.698	1.030	.215
Skewness		-.349	-.137	-.658	-.033	2.715	3.947	1.573	1.076	.484	4.193
Std. Error of Skewness		.034	.034	.034	.034	.034	.034	.034	.034	.034	.034
Kurtosis		-1.879	-.991	-1.567	-2.000	5.373	13.582	1.682	3.623	-.902	15.586
Std. Error of Kurtosis		.069	.069	.069	.069	.069	.069	.069	.069	.069	.069
Range		1	82	1	1	1	1	217	87	3	1
Minimum		0	0	0	0	0	0	55	10	0	0
Maximum		1	82	1	1	1	1	272	98	3	1
Percentiles	25	.00	25.00	.00	.00	.00	.00	77.24	23.80	.00	.00
	50	1.00	45.00	1.00	1.00	.00	.00	91.88	28.40	1.00	.00
	75	1.00	61.00	1.00	1.00	.00	.00	114.09	32.80	2.00	.00

Descriptive Analysis

Key Findings

- People who have hypertension/ heart disease/diabetes only account for a small proportion.
- Obesity and smoking behaviors (no matter currently or previously) are common issues.
- Most people (>95%) in the dataset had no stroke.

2. Binary Logistic Regression

- ➔ Given that the DV/output is categorical data (0/1)
- ➔ Find significant predictors of stroke

2. Binary Logistic Regression

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	353.971	9	.000
	Block	353.971	9	.000
	Model	353.971	9	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1636.301[a]	.067	.207

a. Estimation terminated at iteration number 8 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8.651	8	.373

Results Interpretation:

1. Fittingness of the model
   a. The full model is fitting significantly better than the null model
   b. Rejects the null hypothesis that the models are fitting equally well
2. Compare variations of the same model
   a. Useful to compare variations of the same model with a good R-Square value between 0.2 and 0.4
3. Fittingness of the model to the data
   a. The p-value non-significance is indicating a good fitting model to the data

(*significance level set at 0.05)

2. Binary Logistic Regression

Classification Table[a]

Observed			Predicted		
			stroke		Percentage Correct
			no stroke	stroke	
Step 1	stroke	no stroke	4860	0	100.0
		stroke	249	0	.0
	Overall Percentage				95.1

a. The cut value is .500

Results Interpretation:

1. Show correspondence on not having a stroke
   a. 4860 samples are correctly predicted by the model to express a likelihood of having no stroke
   b. Accuracy rate is 100% when predicting those who do not have a stroke
2. Missed classifications on having a stroke
   a. None was correctly predicted by the model to express a likelihood of having stroke
   b. Accuracy rate is 0% when predicting those who have a stroke
3. Overall accuracy rate is high
   a. > 95% of the sample were correctly predicted to fall into their respective groups

(*significance level set at 0.05)

2. Binary Logistic Regression

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	gender	-.030	.140	.046	1	.830	.970	.737	1.277
	age	1.438	.117	150.816	1	.000	4.213	3.349	5.299
	ever_married	-.106	.218	.239	1	.625	.899	.587	1.377
	residence_type	.129	.137	.894	1	.344	1.138	.870	1.488
	hypertension	.490	.161	9.222	1	.002	1.632	1.190	2.239
	heart_disease	.441	.189	5.431	1	.020	1.554	1.073	2.252
	avg_glucose_level	.357	.173	4.266	1	.039	1.430	1.018	2.007
	bmi	.148	.115	1.660	1	.198	1.159	.926	1.451
	smoking_status	.054	.071	.585	1	.444	1.056	.919	1.214
	Constant	-8.256	.597	191.191	1	.000	.000		

a. Variable(s) entered on step 1: gender, age, ever_married, residence_type, hypertension, heart_disease, avg_glucose_level, bmi, smoking_status.

Remarks:
We cannot directly look at relationship between predictors and DV since it is a non-linear relationship. But, we can modeling the relationship between predictors and DV through the concept of log-odds, which makes the relationship linear through converting the probabilities to log-odds.

Results Interpretation:

1. Significant positive relationship between:
   a. Age and stroke → getting older may increase likelihood of having a stroke
   b. Hypertension and stroke → having hypertension may increase likelihood of having a stroke
   c. Heart disease and stroke → having heart disease may increase likelihood of having a stroke
   d. Diabetes (i.e. exceed normal average glucose level in blood) and stroke → having diabetes may increase likelihood of having a stroke

(*significance level set at 0.05)

Logistic Regression

Key Findings

- The current model is a good fitting model to the data, which also have a significant improvement and fit over the null model, and able to compare variations of the same model.
- The current logistic regression model is doing a good job of predicting those who do not have a stroke, but doing a lousy job of predicting those who have a stroke.
- Age, hypertension, heart disease and diabetes are significant predictors of stroke (DV).

3. K-Means Clustering

➔ Additional analysis to have stroke patients segmentations with 2 and 3 clusters

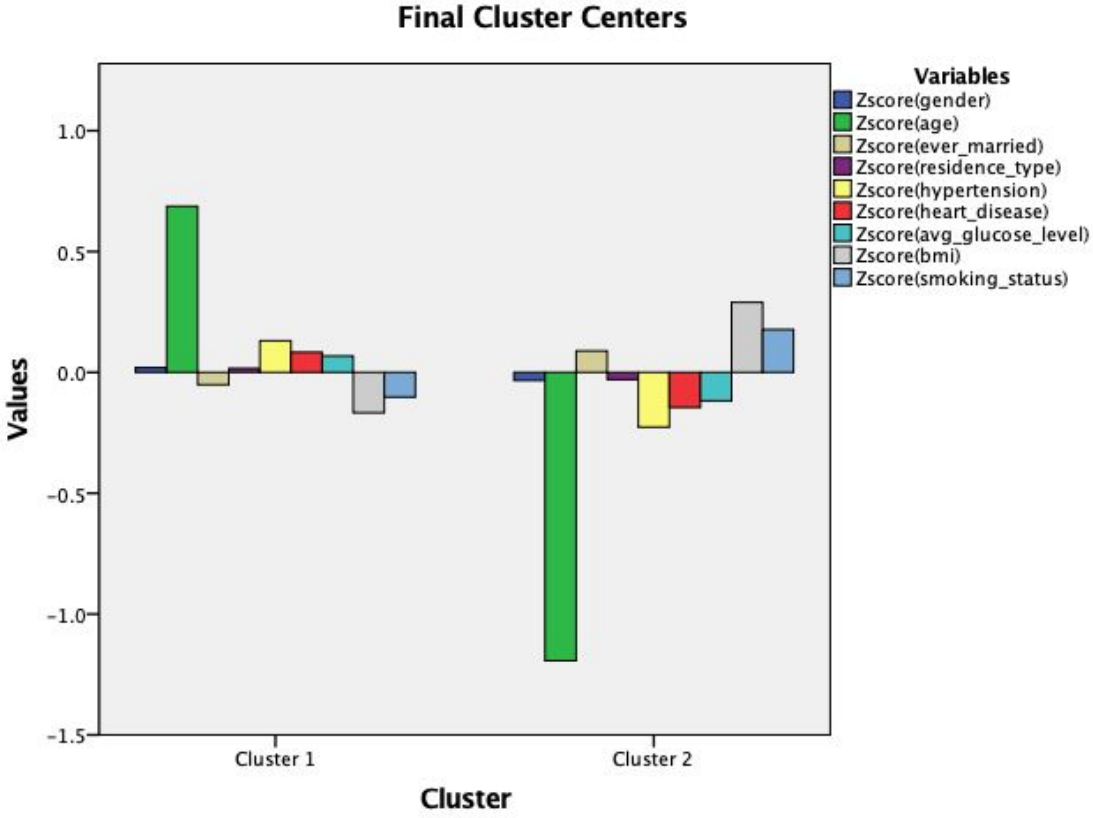
3. K-Means Clustering
(2 Clusters)

Final Cluster Centers

	Cluster	
	1	2
Zscore(gender)	.01950	-.03386
Zscore(age)	.68740	-1.19351
Zscore (ever_married)	-.05116	.08883
Zscore (residence_type)	.01695	-.02944
Zscore (hypertension)	.13052	-.22662
Zscore (heart_disease)	.08356	-.14508
Zscore (avg_glucose_level)	.06756	-.11731
Zscore(bmi)	-.16680	.28961
Zscore (smoking_status)	-.10248	.17793

Number of Cases in each Cluster

Cluster	1	158.000
	2	91.000
Valid		249.000
Missing		.000



3. K-Means Clustering

(2 Clusters)

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zscore(gender)	.164	1	1.003	247	.164	.686
Zscore(age)	204.286	1	.177	247	1154.285	.000
Zscore (ever_married)	1.132	1	.999	247	1.132	.288
Zscore (residence_type)	.124	1	1.004	247	.124	.725
Zscore (hypertension)	7.365	1	.974	247	7.560	.006
Zscore (heart_disease)	3.019	1	.992	247	3.044	.082
Zscore (avg_glucose_level)	1.973	1	.996	247	1.981	.161
Zscore(bmi)	12.029	1	.955	247	12.591	.000
Zscore (smoking_status)	4.540	1	.986	247	4.606	.033

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Results Interpretation:

1. Among all stroke people in the dataset (N=249), cluster 1 tend to be older people who have a slight likelihood of hypertension, but a lower BMI score and low or unknown smoking behaviors.
2. Among all stroke people in the dataset (N=249), cluster 2 tend to be younger people who have a slight likelihood of not having hypertension, but a higher BMI score and likelihood of smoking behaviors.

(*significance level set at 0.05)

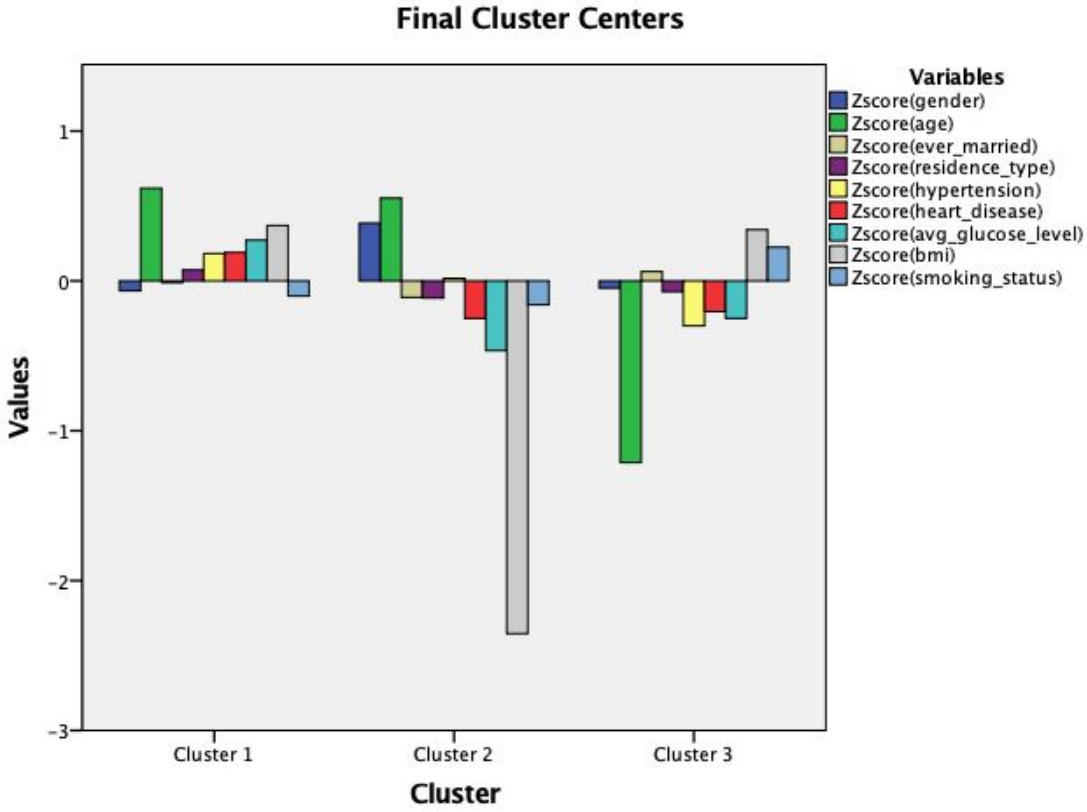
3. K-Means Clustering
(3 Clusters)

Final Cluster Centers

	Cluster		
	1	2	3
Zscore(gender)	-.06531	.38525	-.04852
Zscore(age)	.61888	.55326	-1.21167
Zscore(ever_married)	-.01193	-.10904	.06247
Zscore(residence_type)	.07367	-.11482	-.07240
Zscore(hypertension)	.18271	.01734	-.29967
Zscore(heart_disease)	.18975	-.24954	-.20485
Zscore(avg_glucose_level)	.27117	-.46514	-.24959
Zscore(bmi)	.37043	-2.35357	.34218
Zscore(smoking_status)	-.10073	-.15991	.22499

Number of Cases in each Cluster

Cluster	1	133.000
	2	33.000
	3	83.000
Valid		249.000
Missing		.000



3. K-Means Clustering
(3 Clusters)

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zscore(gender)	2.830	2	.985	246	2.873	.058
Zscore(age)	91.449	2	.265	246	345.556	.000
Zscore (ever_married)	.368	2	1.005	246	.366	.694
Zscore (residence_type)	.796	2	1.002	246	.795	.453
Zscore (hypertension)	5.952	2	.960	246	6.201	.002
Zscore (heart_disease)	5.163	2	.966	246	5.344	.005
Zscore (avg_glucose_level)	11.045	2	.918	246	12.027	.000
Zscore(bmi)	105.382	2	.151	246	696.218	.000
Zscore (smoking_status)	3.198	2	.982	246	3.256	.040

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

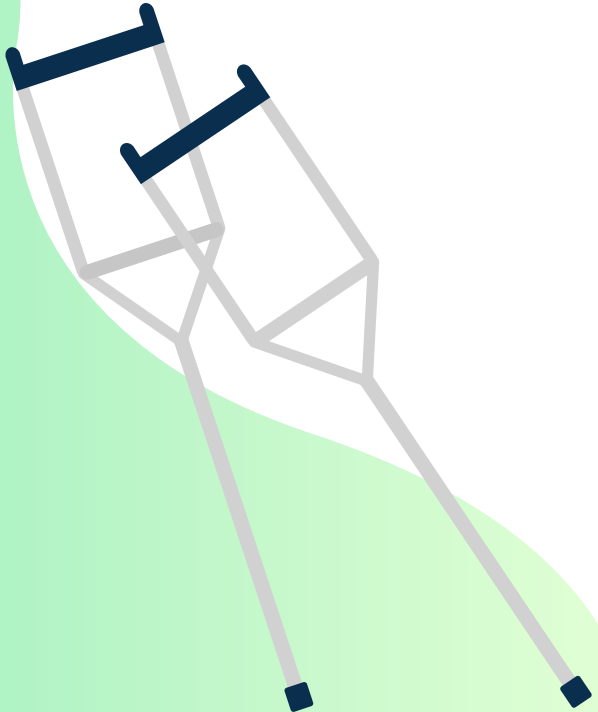
Results Interpretation:

1. Among all stroke people in the dataset (N=249), cluster 1 tend to be older people who have a slight likelihood of hypertension, heart disease, higher average glucose level in blood and BMI score, but low or unknown smoking behaviors.

2. Among all stroke people in the dataset (N=249), cluster 2 tend to be older people who have a slight likelihood of not having heart disease, lower average glucose level in blood and BMI score, and low or unknown smoking behaviors.

3. Among all stroke people in the dataset (N=249), cluster 3 tend to be younger people who have a slight likelihood of not having hypertension and heart disease, and a lower average glucose level in blood, but a higher BMI score and likelihood of smoking behaviors.

(*significance level set at 0.05)



04
Hypothesis & Discussion

Hypothesis

Null Hypothesis (H0)

based on variables included on the Kaggle dataset

Gender, age, marital status, residence type, hypertension, heart disease, diabetes, BMI, smoking status are predictors of stroke.

Alternative Hypothesis (H1)

based on logistic regression results

Age, hypertension, heart disease and diabetes can significantly predict stroke, whereas gender, marital status, residence type, BMI and smoking status cannot predict stroke.

Problem Solving & Recommendations

Problems we are trying to solve:

How can we identify significant factors which may predict stroke?

Solutions we are providing:

Using SPSS logistic regression model, age, hypertension, heart disease and diabetes are found to be significant predictors of stroke.

Recommendations

1. Always keep a healthy lifestyle.
2. To evaluate risk of having a stroke with the gradual increase of ages, we should pay more attention in monitoring health status and regularly check if we have clinical features of hypertension/heart disease/diabetes.
3. Have regular body check and consult a professional doctor if needed.

Discussions

- Why smoking status is not a significant predictor of stroke?
  - Potential reason: more than 30% responses in the dataset are “unknown” on the smoking status → may affect results of the model
- Why accuracy rate of the logistic regression model on predicting stroke is poor (0%)?
  - Potential reason: stroke samples (N=249) only account for a very small proportion in the dataset; most people (>95%) in the dataset had no stroke → limited stroke data available may affect the results

Discussions

Limitations

Output/Dependent variable of logistic regression model is limited to categorical data (i.e. only answer 0/1 or yes/no questions)

- ★ Results can only predict whether a person have a stroke or not, but fail to provide an estimated probability (e.g. number) to predict the likelihood of a person to have stroke

Glucose level in blood can be varied in time and meal status, unstandardized data collection time among the participants (e.g. some may scored much higher after meal and some may scored much lower before breakfast) may lead to bias

- ★ Predictive and explanatory power of the model may be affected/limited

The current study only analyzed relationship between 9 independent variables and stroke (DV), but stroke can be caused and influenced by many other comprehensive risk factors (e.g. use of illegal drugs, family stroke history, COVID-19 infection)

- ★ Stroke may not be predicted/prevented easily just based on the analyzed variables and results of the current study

05

Personal Learnings

Knowledge Applications on This Project

How to deal with missing value

Identify problem/goal and choose data analysis method

Data cleaning and binning/encoding

Data visualization and storytelling

Identify IV and DV

Generate hypothesis (H0 and H1)

Identify data bias and research limitations

Always try multiple models

References

American Stroke Association. (2021). *About Stroke*. www.stroke.org. https://www.stroke.org/en/about-stroke.

Australian Institute of Health and Welfare. (2020, July 23). *Stroke*. Australian Institute of Health and Welfare. https://www.aihw.gov.au/reports/australias-health/stroke.

Centre for Health Protection. (2019). *Centre for Health Protection, Department of Health - Body Mass Index Chart*. Centre for Health Protection - Body Mass Index Chart. https://www.chp.gov.hk/en/resources/e_health_topics/pdfwav_11012.html.

Fedesoriano. (2021, January 26). *Stroke Prediction Dataset*. Kaggle. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.

G-W Online. (n.d.). *Lifespan Development, 2nd Edition page 24*. G-W Online. https://www.g-wonlinetextbooks.com/lifespan-development-2018/44.

Holland, K. (2019, October 16). *Stroke: Symptoms, Causes, Treatment, Types & More*. Healthline. https://www.healthline.com/health/stroke#symptoms-in-women.

Hong Kong Stroke Fund. (n.d.). *Hong Kong Stroke Fund - About us*. Hong Kong Stroke Fund. http://www.strokefund.org/eng/aboutus_part1.php.

Mayo Clinic. (2020, October 30). *Diabetes*. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451.

Mayo Clinic. (2021, February 9). *Stroke*. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113.

Public Health Agency of Canada. (2019, December 9). *Stroke in Canada: Highlights from the Canadian Chronic Disease Surveillance System*. Government of Canada. https://www.canada.ca/en/public-health/services/publications/diseases-conditions/stroke-canada-fact-sheet.html.

YouTube. (2015). *K-Means Cluster Analysis in Spss*. *YouTube*. https://www.youtube.com/watch?v=YRue69W-dYU.

YouTube. (2018). *Binary logistic regression using Spss (2018)*. *YouTube*. https://www.youtube.com/watch?v=H_48AcV0qlY.

Thanks