Final Project: Stroke Data Manipulation with R

Huynh, Lucas
Lee, Joseph
Song, Daniel

STAT 1601

Dataset Summary: Background and Description

Summary of Stroke Dataset:

The source of the dataset comes the Electronic Health Record (EHR) provided by McKinsey and Company, a healthcare global management consulting firm. The dataset was cleaned, refined and obtained through Kaggle. Relevant parameters/variables are given in the dataset for the means of predicting the likelihood of a stroke during the rest of the patient's lifetime.

Motivation for Choosing the Stroke Dataset:

The World Health Organization cites stroke as accounting for 11% of all global deaths, establishing it as the 2nd highest leading cause of death. Thus, stroke is a relevant and pertinent health issue that is critical to investigate. By doing statistical analysis on the dataset, we hope to better educate ourselves and our peers on what risk factors are most relevant in predicting the likelihood of stroke.

Relevant Parameters/Variables

Categorical Variables:

- Gender: male or female
- Hypertension (cat. binary)
  - 0: No hypertension
  - 1: Has Hypertension
- Heart Disease (cat. binary)
  - 0: No Heart Disease
  - 1: Has Heart Disease
- Smoking Status
  - Never Smoked
  - Smokes
  - Formerly Smoked
- Marital status (omitted)
- Work type (omitted)
- Residence type (omitted)

Numerical Variables:

- Age: age of patient
- Average Glucose Level: blood sugar level
- Body Mass Index (BMI): ratio of weight (kg) to height (m)
- Id: unique for each person (omitted)

Data Preparation

- ❖ ID data was removed as it was irrelevant to analysis.
- ❖ Marriage status, Residence, and Type of Work were removed to focus on the health related variables in predicting stroke.
- ❖ Heart disease, hypertension, and stroke data were altered for easier legibility and understanding.
  - ➢ i.e for hypertension, 0 was changed to no hypertension and 1 was changed to hypertension.
- ❖ Rows with unknown or missing values were removed
- ❖ BMI values were converted from char to double for analysis purposes.
- ❖ 70% of strokes occur in those ages 65+, thus data was filtered to focus on the specified age range.

Numeric Data Summary

	Age	Avg Glucose Level mg/dL	BMI
Min	65.00	55.32	14.10
1st Quartile	69.00	79.97	26.00
Median	74.00	99.27	28.80
Mean	73.84	125.84	29.56
3rd Quartile	79.00	190.60	32.80
Max	82.00	271.74	54.60

Frequency Tables

Disease Status	Smoking Status	Frequency
Heart Disease	Formerly Smoked	50
Heart Disease	Never Smoked	53
Heart Disease	Smokes	27
No Heart Disease	Formerly Smoked	231
No Heart Disease	Never Smoked	336
No Heart Disease	Smokes	89

Stroke History	Blood Pressure	Frequency
No Stroke	Hypertension	142
No Stroke	No Hypertension	527
Stroke	Hypertension	42
Stroke	No Hypertension	75

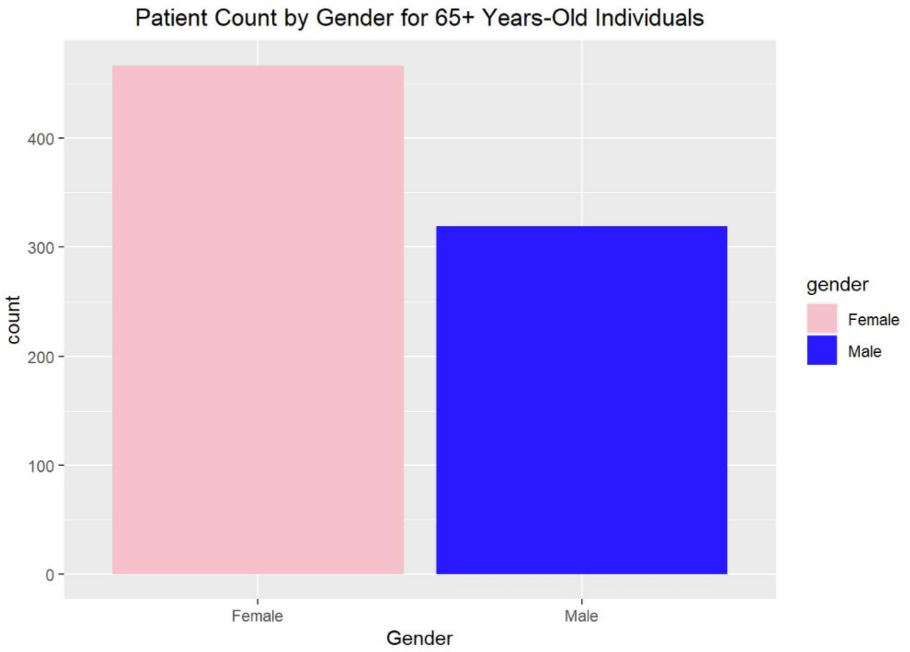
Two-Way Tables

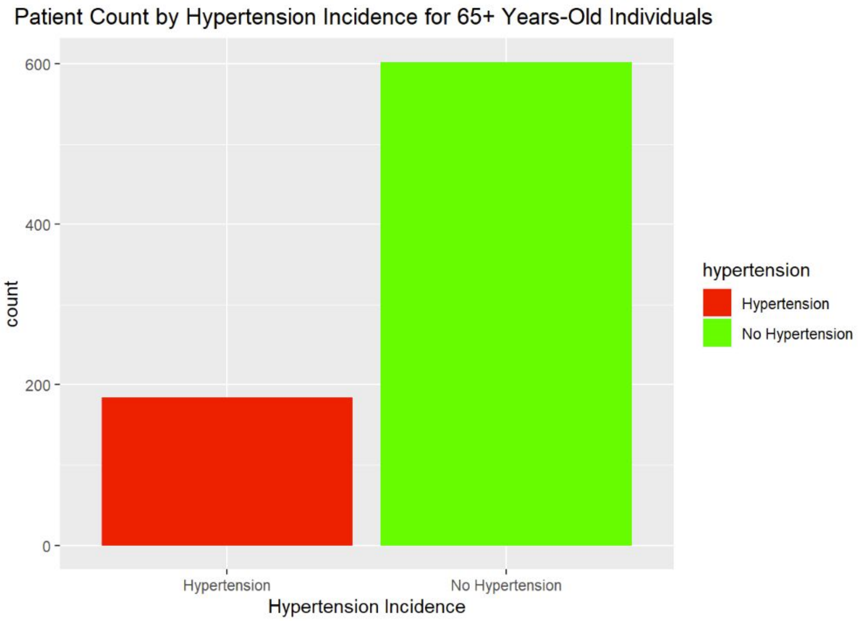
	Heart Disease	No Heart Disease
Female	57	410
Male	73	246

	Hypertension	No Hypertension
Female	108	359
Male	76	243

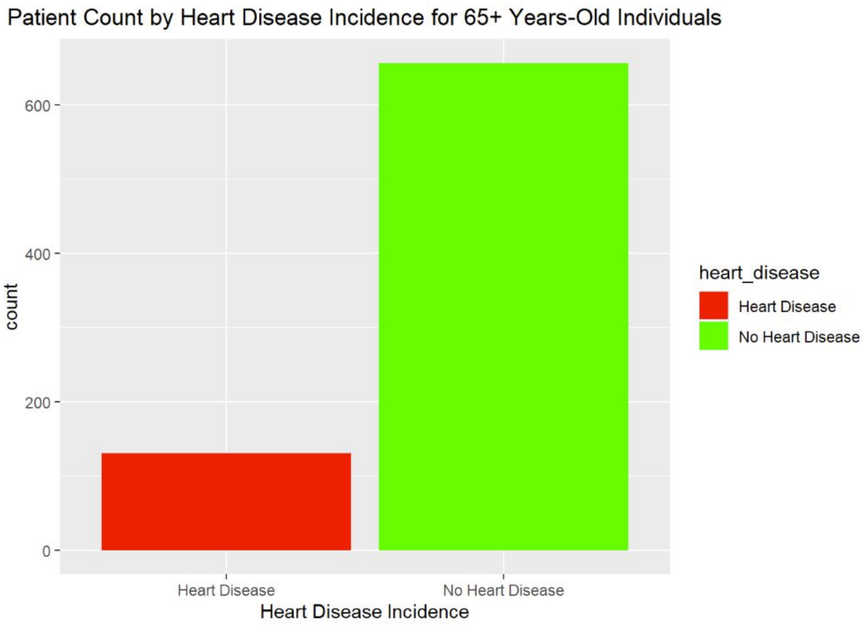
	Stroke	No Stroke
Female	70	397
Male	47	272

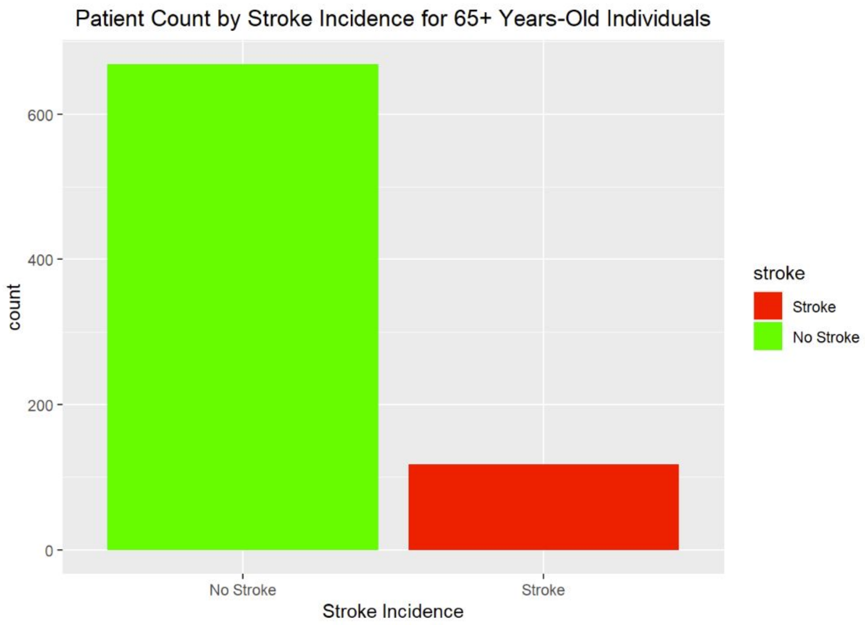
Data Visualization: Categorical



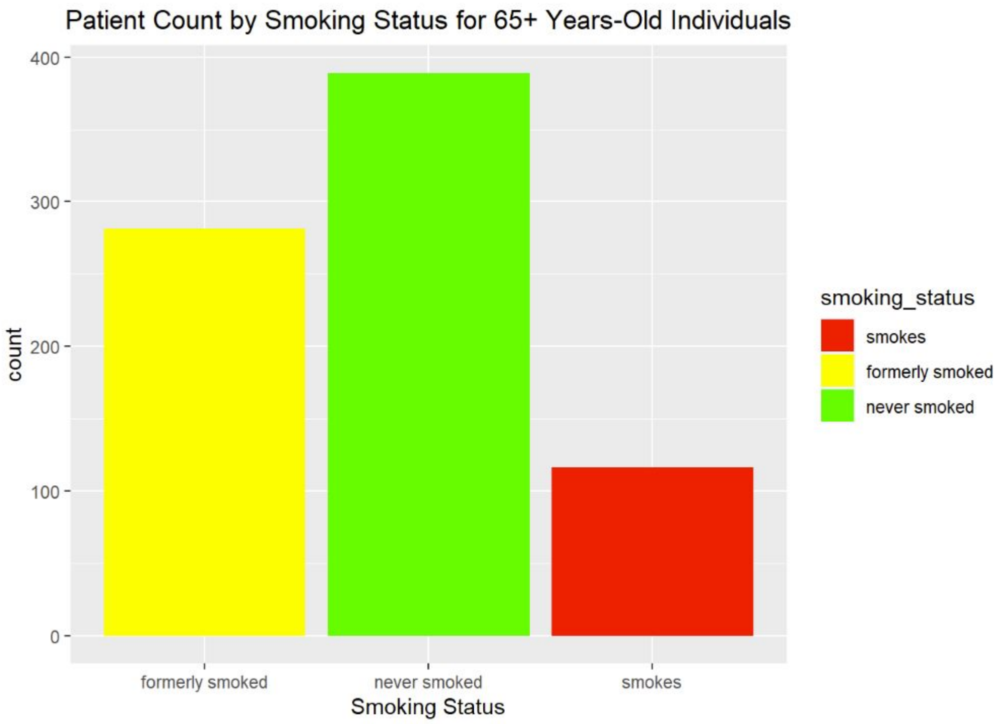


Data Visualization: Categorical

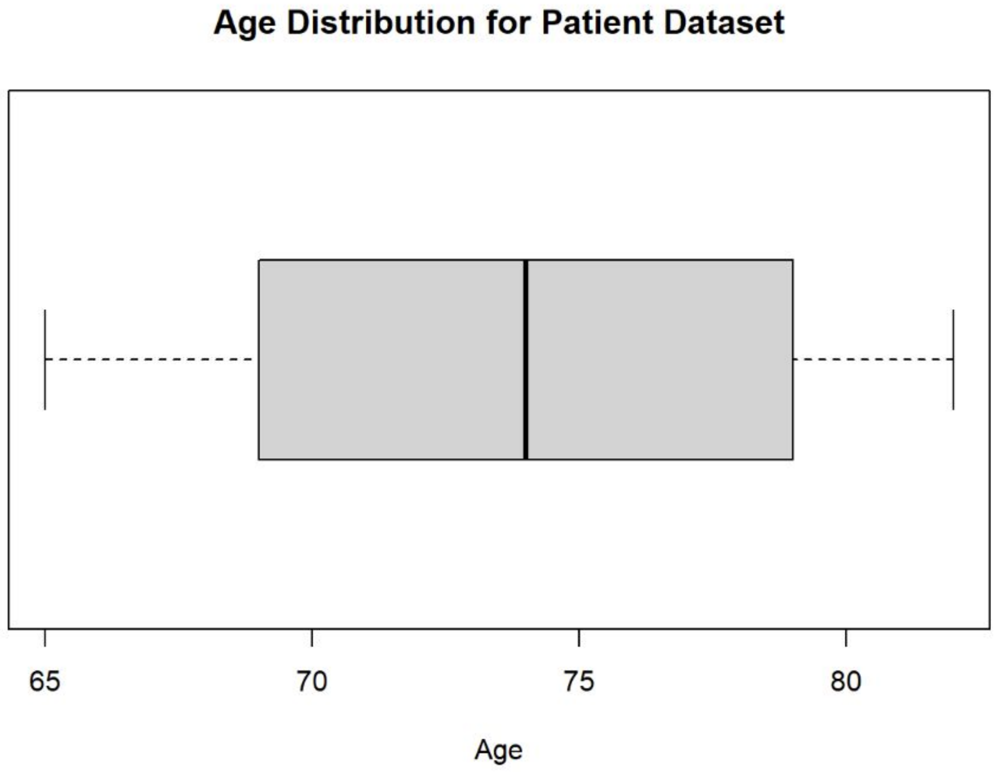




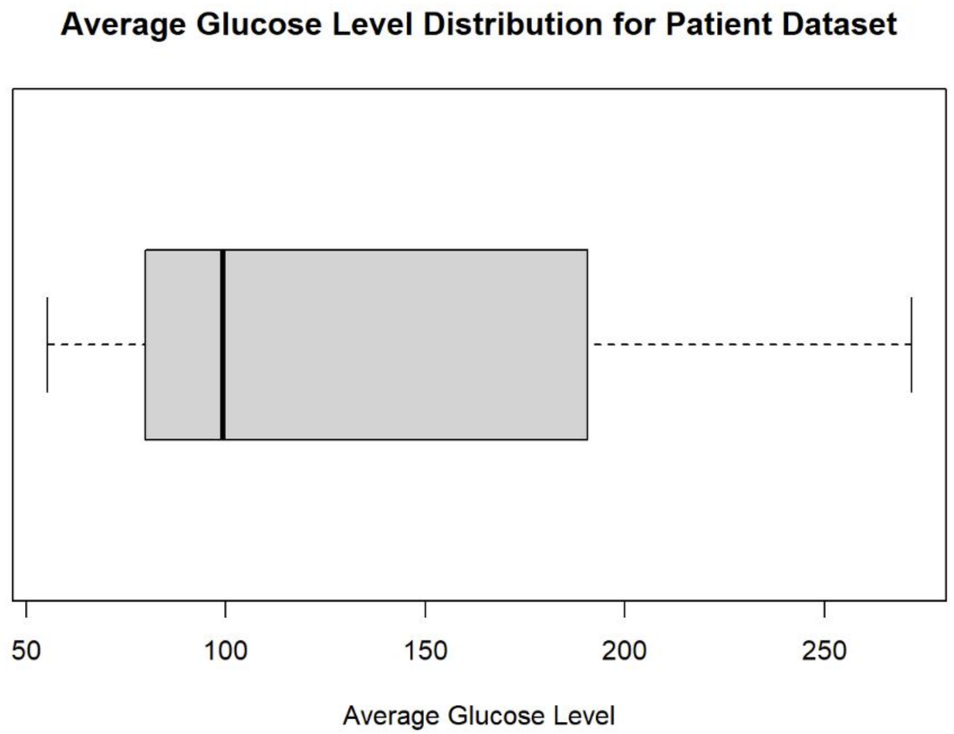
Data Visualization: Categorical



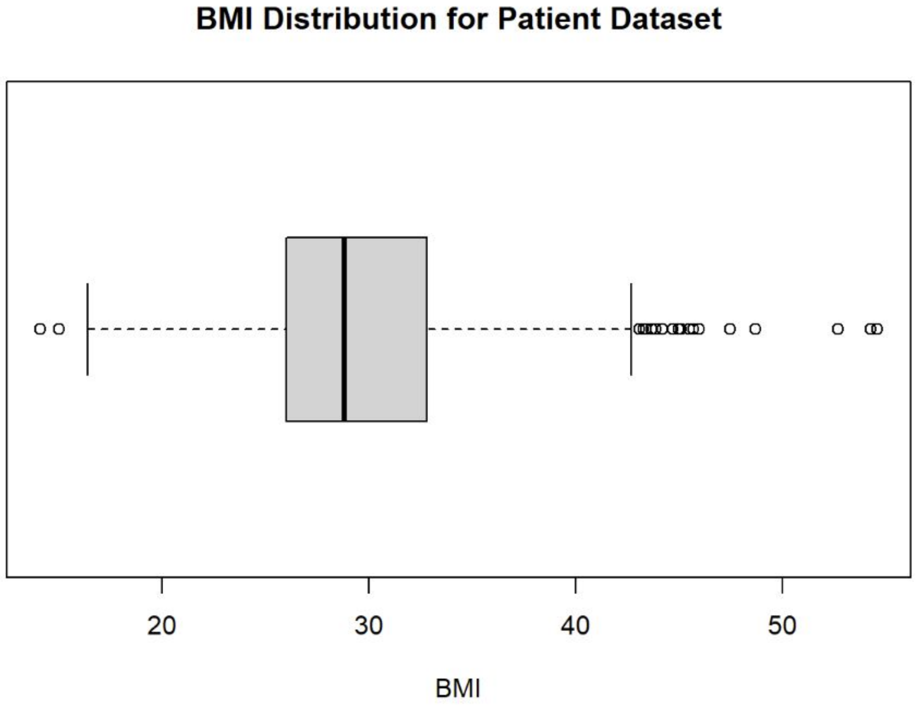
Data Visualization: Numeric



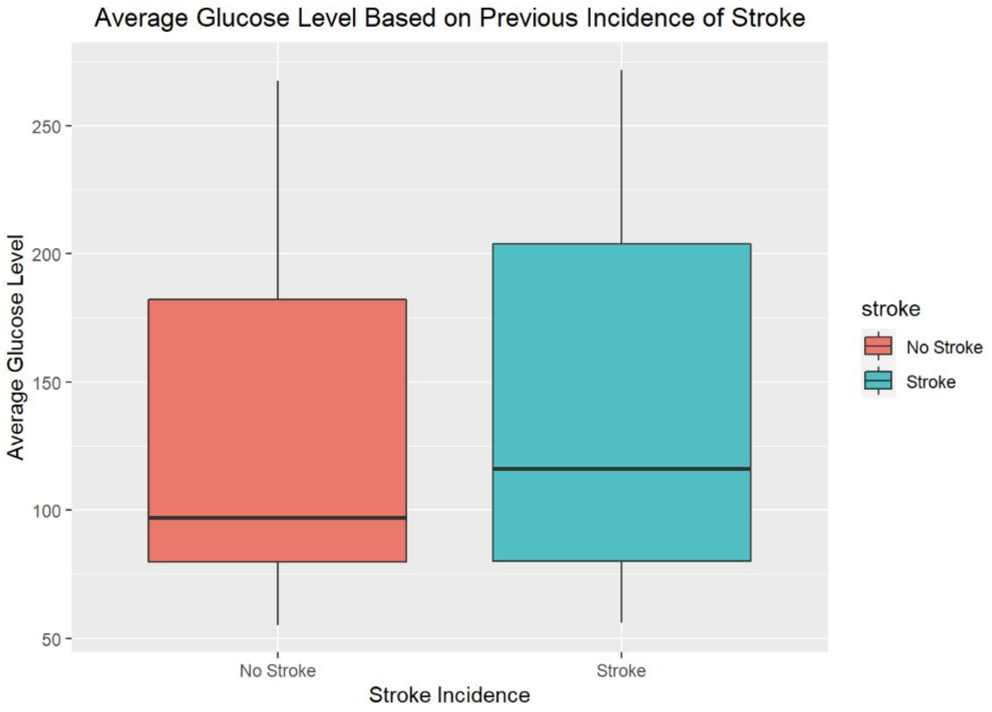
Data Visualization: Numeric



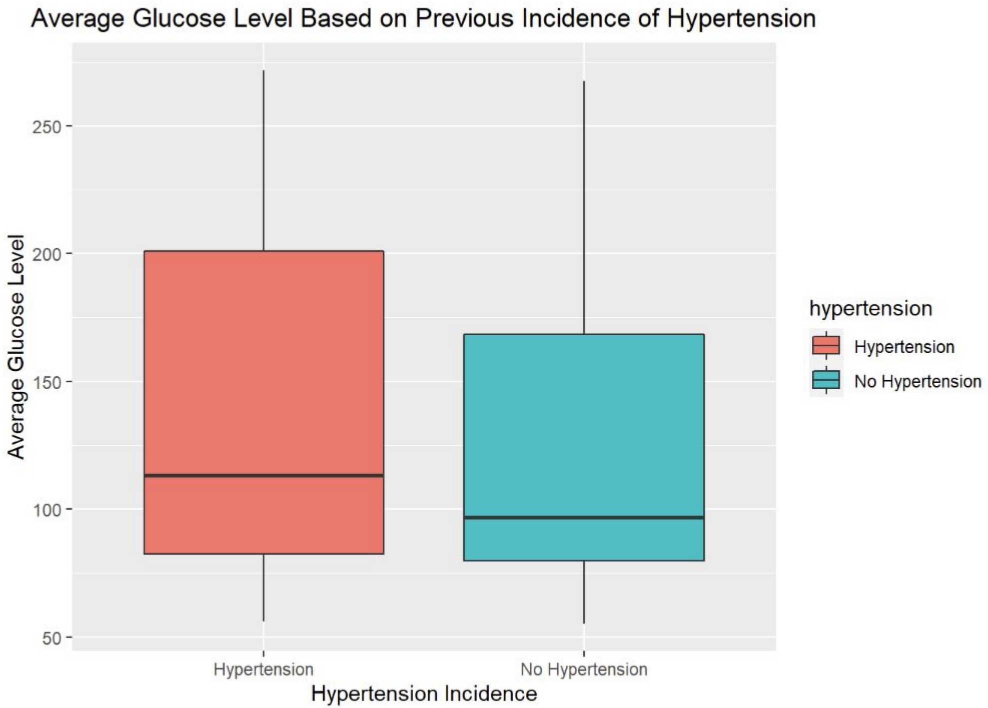
Data Visualization: Numeric



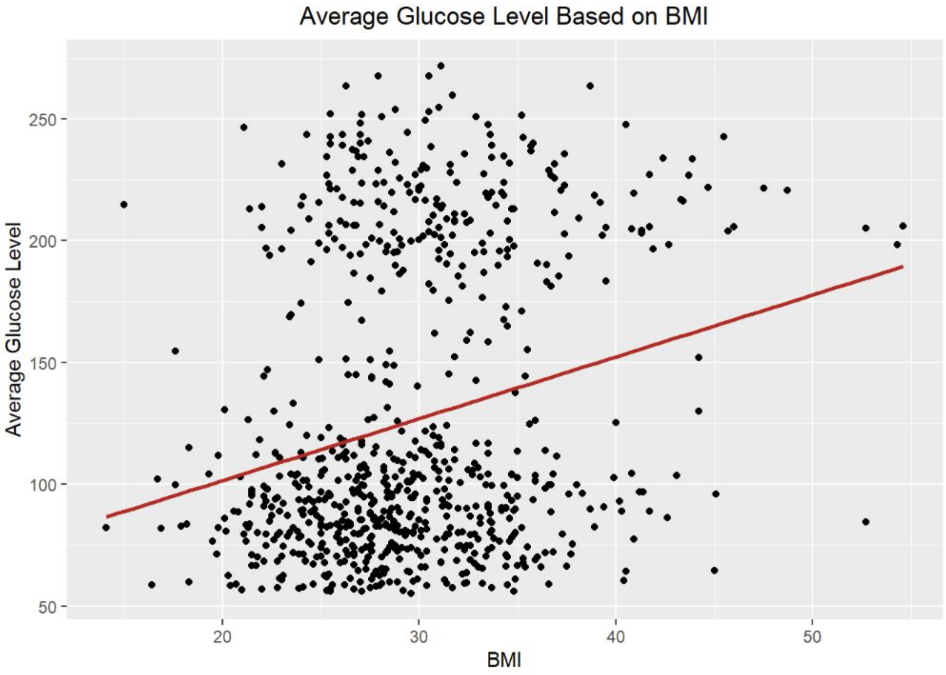
Data Visualization: Side-By-Side Boxplot



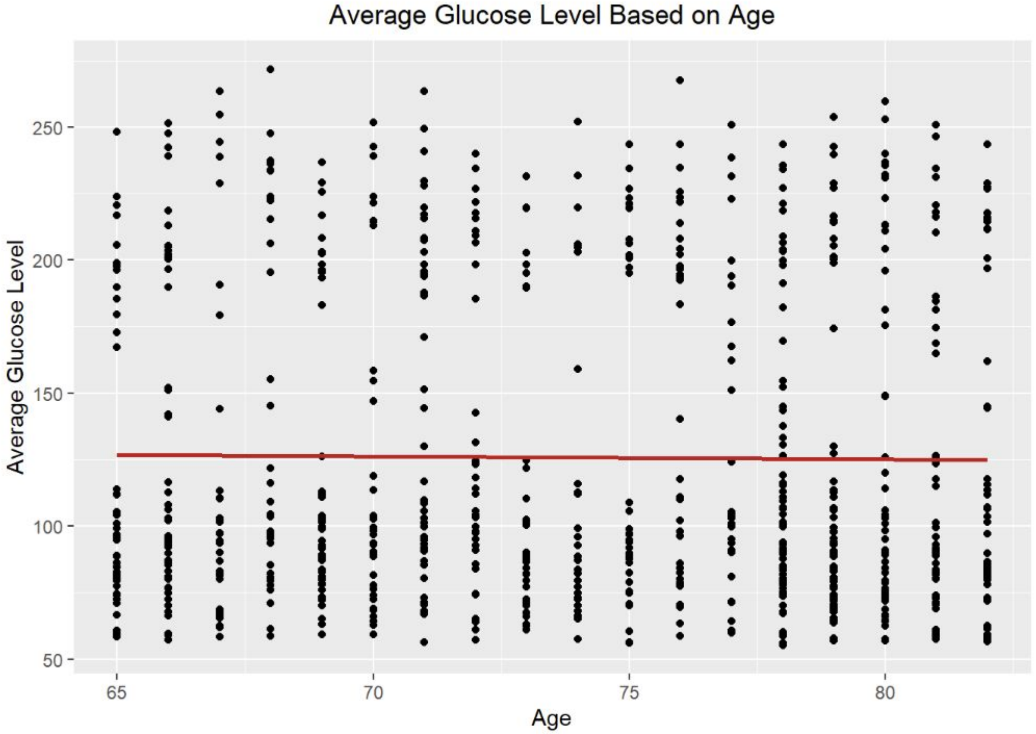
Data Visualization: Side-By-Side Boxplot



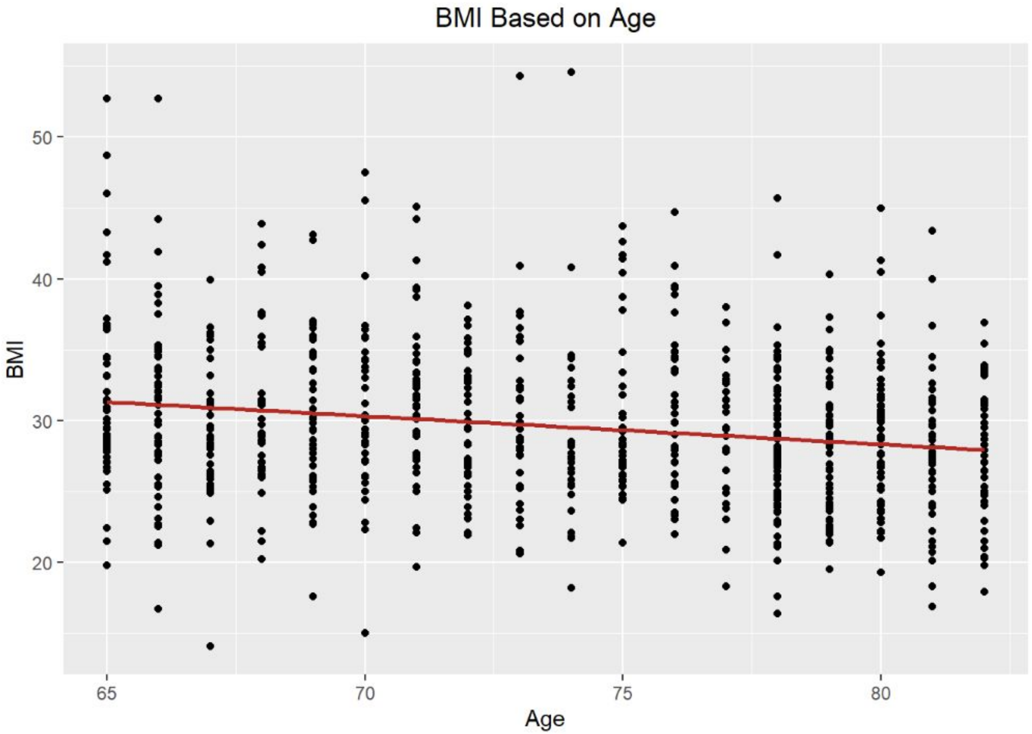
Data Visualization: Scatterplot



Data Visualization: Scatterplot

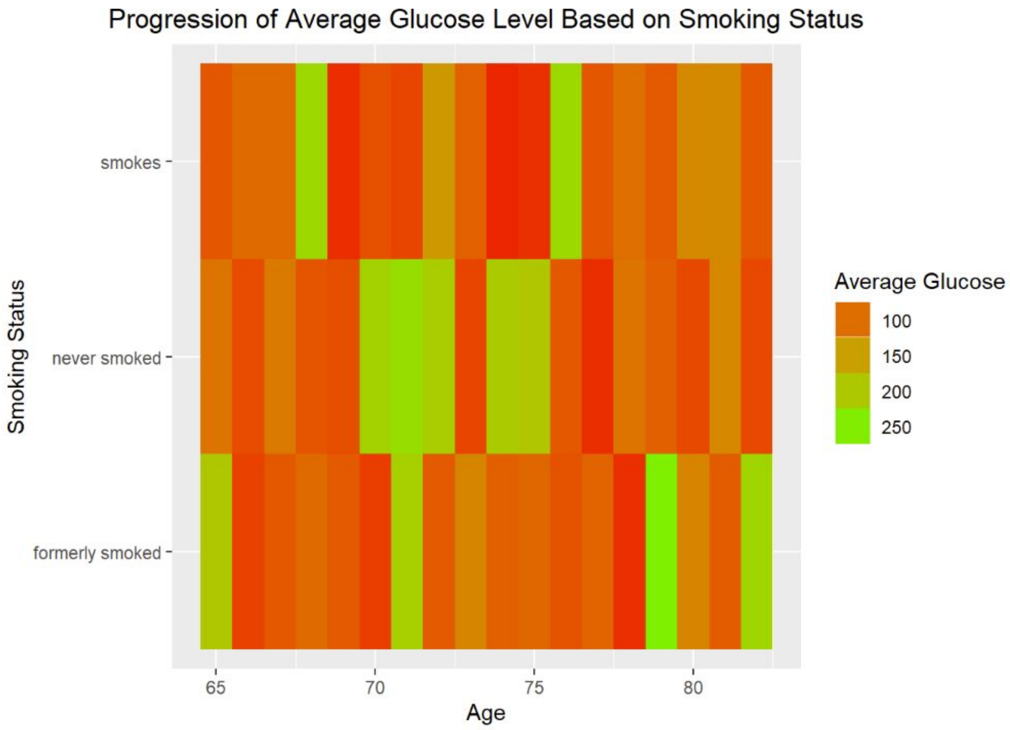


Data Visualization: Scatterplot



Data Visualization: 3-Variable Facet Wrap

Data Visualization: Heatmap (Special Graph)



Simple Linear Regression

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -100.58  -44.52  -21.67   52.49  146.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  50.8111    11.2145   4.531 6.79e-06 ***
## bmi           2.5381     0.3727   6.810 1.95e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.62 on 784 degrees of freedom
## Multiple R-squared:  0.05584,    Adjusted R-squared:  0.05464
## F-statistic: 46.37 on 1 and 784 DF,  p-value: 1.948e-11
```

Simple Linear Regression Model: Average Glucose Level = 2.5381(BMI) + 50.8111

Interpretation(s):

Slope –> For every additional unit of BMI, the model predicts that there is a 2.5381 (in mg/dL) increase in average glucose level.

Intercept –> If BMI was set to be 0, the model predicts that there would be a baseline average glucose level of 50.8111 mg/dL.

R-Squared –> Since the Multiple R-Squared value is equal to 0.05584, 5.584% of the variability in Average Glucose Level can be explained by our model. In other words, 5.584% of the variation in Average Glucose Level can be predicted from BMI.

Simple Linear Regression

```
cor(new_dat$avg_glucose_level,new_dat$bmi)
```

```
## [1] 0.2363122
```

There is a weak, positive, scattered correlation between average glucose level and bmi.

Predicting Values (Interpolation)

Recommended BMI for older individuals according to the NIH: 25-27

Predicting Average Glucose Level for an individual with a BMI of 25

```
new=data.frame(bmi=25)
predict(model1,new)
```

```
##        1
## 114.2626
```

For an individual with a BMI of 25, the model predicts an average glucose level of 114.2626 mg/dL

Predicting Average Glucose Level for an individual with a BMI of 27

```
new1=data.frame(bmi=27)
predict(model1,new1)
```

```
##        1
## 119.3387
```

For an individual with a BMI of 27, the model predicts an average glucose level of 119.3387 mg/dL .

Logistic Regression

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9413  -0.6101  -0.5027  -0.3915   2.3384
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -8.011308   1.742216  -4.598 4.26e-06 ***
## age                0.076691   0.020374   3.764 0.000167 ***
## bmi               -0.004352   0.019237  -0.226 0.821031
## avg_glucose_level  0.005107   0.001640   3.113 0.001850 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 661.37  on 785  degrees of freedom
## Residual deviance: 636.10  on 782  degrees of freedom
## AIC: 644.1
##
## Number of Fisher Scoring iterations: 5
```

Logistic Regression

```
exp(coef(logit$finalModel))
```

```
##      (Intercept)              age              bmi avg_glucose_level
##     0.0003316906     1.0797081943     0.9956577271      1.0051197235
```

Interpretation(s):

Holding BMI and average glucose level constant, the model predicts that for every one year increase in age, the odds of a stroke incidence increase by 7.97%.

Holding age and average glucose level constant, the model predicts that for every unit increase in BMI, the odds of a stroke incidence decreases by 0.44%.

Holding age and BMI constant, the model predicts that for every 1 mg/dL . increase in average glucose level, the odds of a stroke incidence increases by 0.51%.

Predicting Values (Interpolation)

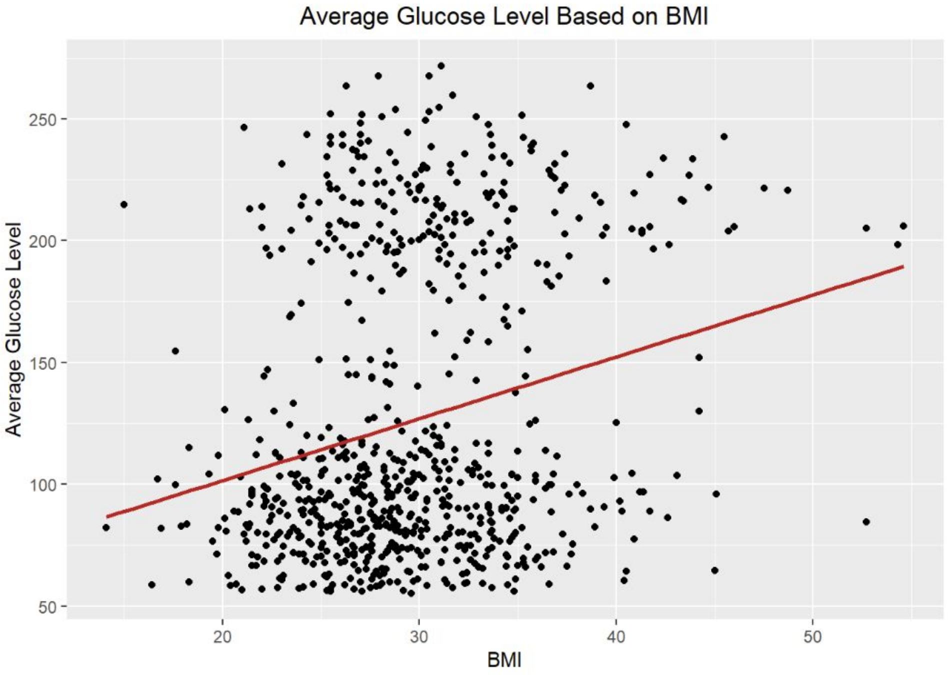
Age = 65 since it is the baseline age of the dataset. Average Glucose Level = 100 mg/dL since it is the recommended level for senior individuals. BMI = 26 since it is the average between 25 and 27 (recommended level as listed above).

```
logitdata<-data.frame(age=65,avg_glucose_level=100,bmi=26)
predict(logit,logitdata)
```

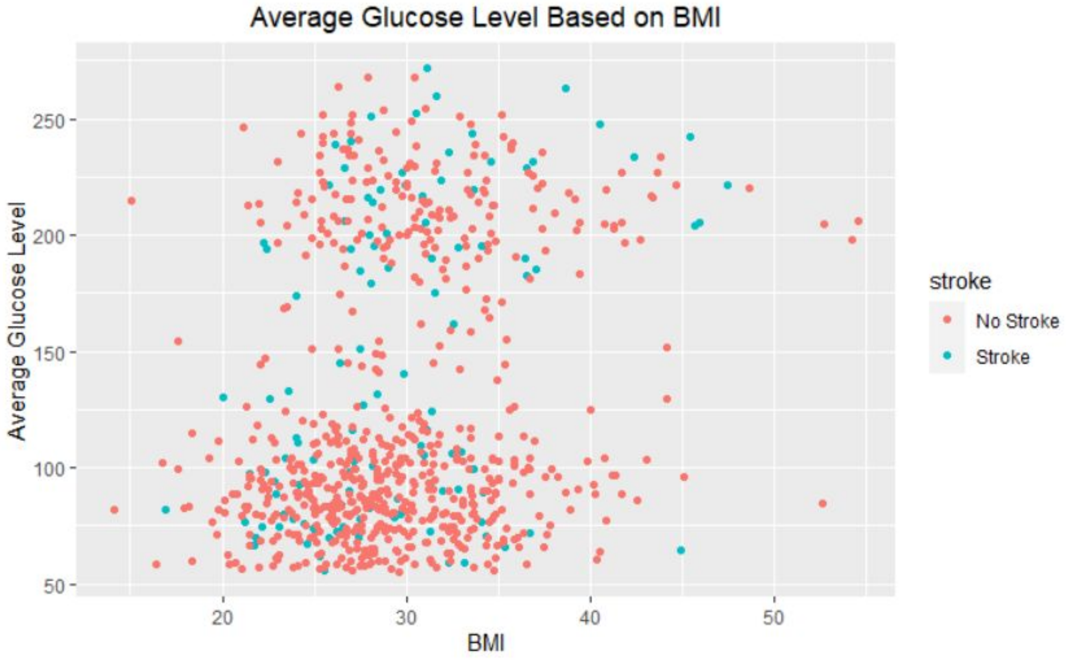
```
## [1] No Stroke
## Levels: No Stroke Stroke
```

The logistic model predicts No Stroke for an individual with average levels of the relevant variables.

Classification: Reusing an Old Graph



Classification: Modified Graph



Classification: New Patient

Gender: Male

Age: 80

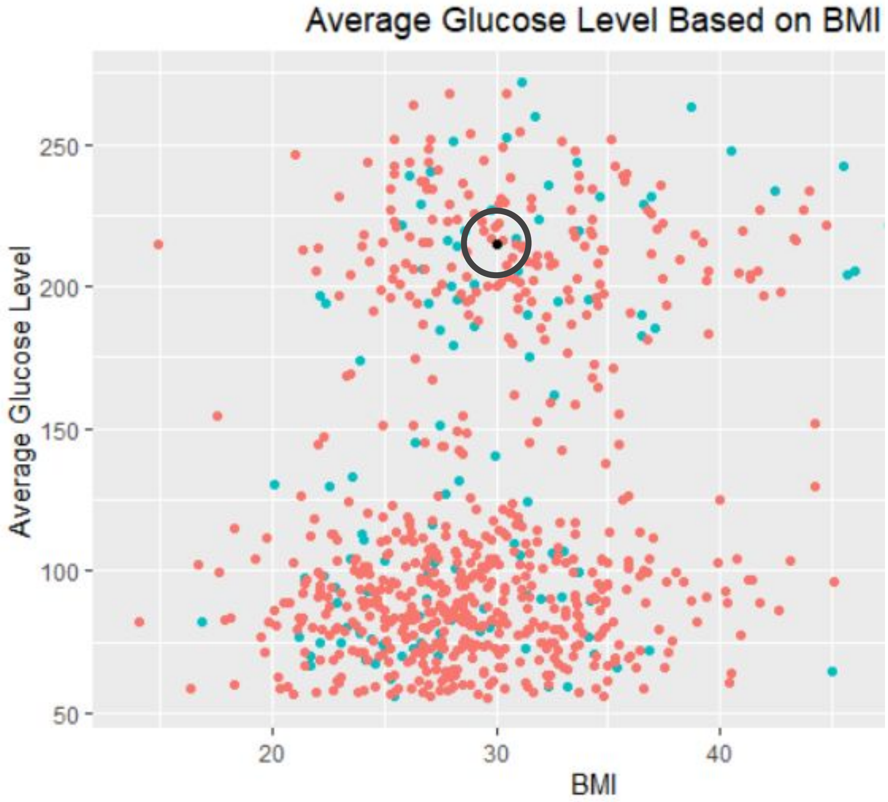
Hypertension: Yes

Heart Disease: Yes

Average Glucose Level: 215

BMI: 30

Classification: Where He Lies



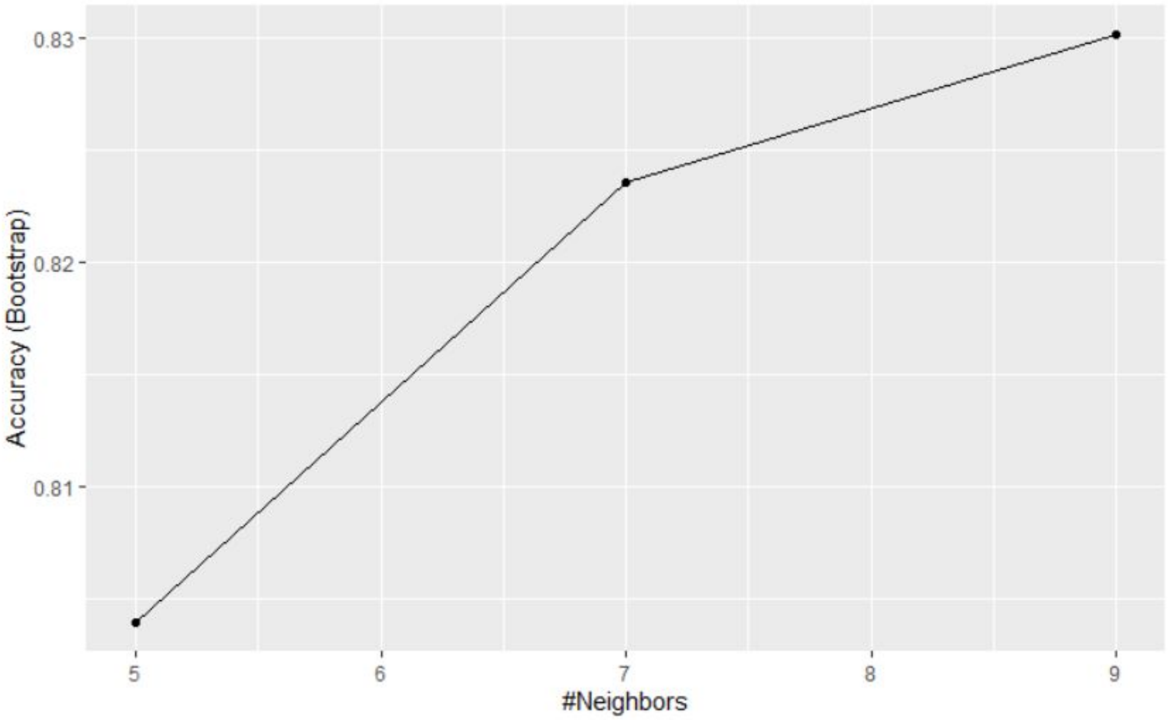
K-Nearest Neighbor (KNN) Training/Result

```{r}
knn_model = train(stroke~bmi+avg_glucose_level, new_dat, method="knn")
knn_model$finalModel
ggplot(knn_model)
```

```
9-nearest neighbor model
Training set outcome distribution:

No Stroke    Stroke
      669       117
```

K-Nearest Neighbor Graphical Result



Prediction with KNN

```{r}
predict(knn_model, new_pat)
```

```
[1] No Stroke
Levels: No Stroke Stroke
```

References

Fedesoriano. (2021, January 26). *Stroke prediction dataset*. Kaggle. Retrieved May 4, 2022, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Kelly-Hayes, M. (2010, October). *Influence of age and health behaviors on stroke risk: Lessons from longitudinal studies*. Journal of the American Geriatrics Society. Retrieved May 4, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3006180/#:~:text=The%20risk%20increases%20with%20age,will%20result%20in%20death1.