Midterm Exam - Individual Part

## #BANL6900

#Midterm Exam - Individual Part

#Complete this part on your own. Sharing your work or receiving someone’s work will be penalized. When done, save this document and then Render it to produce a Word document.

#Submit both this file (.qmd) and Word file on Canvas.

## #NAME: Sumukh. Ramagiri

## #Questions

#Run the following code to read a dataset. This dataset is about healthcare management. The dataset provides information about 5000 patients.

1) Inspect the data and write two questions (one normal and one unusual) that can be answered using this dataset. Make sure your questions are non-trivial. For example, what is the mean BMI is trivial, too simple. Consider relationships between at least two variables, or influential factors (more than two variables).

HINT:

**Normal questions -> questions related to diseases**

What is the impact of age and heart disease on the likelihood of having a stroke, considering different BMI categories as a moderating factor?

This question seeks to understand the interplay between age, pre-existing heart conditions, and body mass index (BMI) on stroke incidence. It hypothesizes that older individuals with heart disease in certain BMI categories (either underweight, normal, overweight, or obese) may exhibit different risks of experiencing a stroke. This question can help us reveal critical insights into how age and heart conditions, when combined with nutritional status as indicated by BMI, affect stroke risk.

**Unusual questions -> questions related to anything other than diseases**

Does the average glucose level vary significantly between different work types within the subset of individuals who have never smoked, and how does this variation compare between those with and without hypertension?

This question aims to explore the relationship between occupational stress or activity levels (inferred from work type) and metabolic health (as indicated by average glucose levels) in a population that has eliminated the confounding factor of smoking. Furthermore, it examines whether hypertension status alters this relationship, suggesting that blood pressure control may interact with occupational factors and metabolic health in a nuanced manner. This approach takes a less direct path to understanding health outcomes, focusing on metabolic indicators across lifestyle and health conditions.

2) Explain how you can answer your questions identified in Question-1. Write some R code below (if possible) to answer them. If the R code is to be too long or difficult, just explain your method in your own words.

Call:  
glm(formula = stroke ~ age + heart\_disease + bmi\_category + age:heart\_disease,   
 family = binomial(link = "logit"), data = d)  
  
Deviance Residuals:   
 Min 1Q Median 3Q Max   
-0.8028 -0.3329 -0.1720 -0.0782 3.7446   
  
Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) -8.505975 1.083394 -7.851 4.12e-15 \*\*\*  
age 0.074537 0.005407 13.785 < 2e-16 \*\*\*  
heart\_disease 2.933978 1.275346 2.301 0.0214 \*   
bmi\_categoryNormal 1.002764 1.039999 0.964 0.3349   
bmi\_categoryOverweight 1.397652 1.029607 1.357 0.1746   
bmi\_categoryObese 1.168874 1.030032 1.135 0.2565   
age:heart\_disease -0.035424 0.017840 -1.986 0.0471 \*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
 Null deviance: 1990.4 on 5109 degrees of freedom  
Residual deviance: 1600.9 on 5103 degrees of freedom  
AIC: 1614.9  
  
Number of Fisher Scoring iterations: 8

Call:  
lm(formula = avg\_glucose\_level ~ work\_type \* hypertension, data = non\_smokers)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-79.24 -30.08 -13.69 11.07 160.37   
  
Coefficients: (2 not defined because of singularities)  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 103.8563 6.2544 16.605 < 2e-16 \*\*\*  
work\_typeGovt\_job 2.3683 6.9018 0.343 0.732   
work\_typeNever\_worked -5.9849 13.7841 -0.434 0.664   
work\_typePrivate -0.2059 6.4087 -0.032 0.974   
work\_typeSelf-employed 2.3789 6.8665 0.346 0.729   
hypertension 28.9637 6.2527 4.632 3.87e-06 \*\*\*  
work\_typeGovt\_job:hypertension -3.4412 10.4758 -0.328 0.743   
work\_typeNever\_worked:hypertension NA NA NA NA   
work\_typePrivate:hypertension -5.5456 7.5696 -0.733 0.464   
work\_typeSelf-employed:hypertension NA NA NA NA   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 45.96 on 1884 degrees of freedom  
Multiple R-squared: 0.03417, Adjusted R-squared: 0.03058   
F-statistic: 9.522 on 7 and 1884 DF, p-value: 1.146e-11

3) Are there any missing data in the dataset? Which columns are affected? Suggest a method to handle the missing data issue. Write the R code to handle the missing value problem below. HINT: summary()

id gender age hypertension   
 0 0 0 0   
 heart\_disease ever\_married work\_type Residence\_type   
 0 0 0 0   
avg\_glucose\_level bmi smoking\_status stroke   
 0 0 0 0   
 health\_bills bmi\_category   
 201 0

[1] 0

4) Run the following code to generate a visualization. Now comment on what you see in this scatter plot.

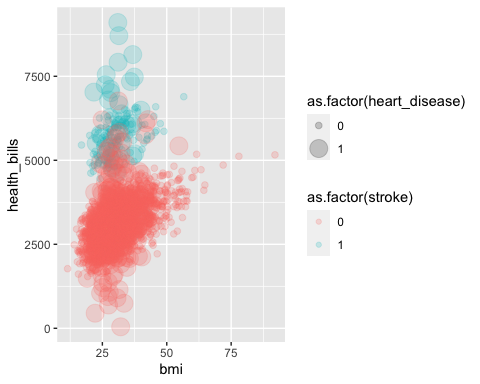
Warning: package 'dplyr' was built under R version 4.2.3

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

Warning: Using size for a discrete variable is not advised.



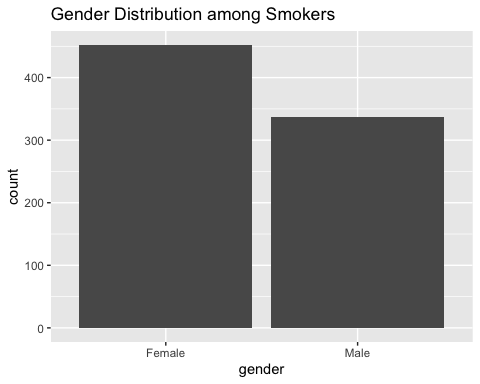
5) Assume that your company sells a product that helps people quit smoking. You will target a specific group of smoking people in your advertisements. Explain how you could identify characteristics of this people (age, gender, etc.). For example, which gender is the majority of the smoking people? Mostly married, or single? What is the majority age group of smoking people? Do calculations, tables or visualizations to support your explanation.

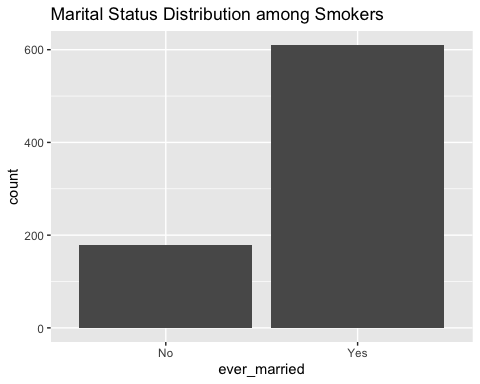
never smoked formerly smoked smokes Unknown   
 1892 885 789 1544

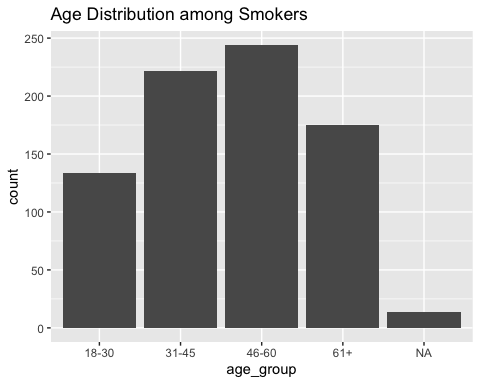
never smoked formerly smoked smokes Unknown  
 Female 1229 477 452 836  
 Male 663 407 337 708  
 Other 0 1 0 0

never smoked formerly smoked smokes Unknown  
 No 530 147 179 901  
 Yes 1362 738 610 643

never smoked formerly smoked smokes Unknown  
 18-30 301 69 134 150  
 31-45 448 158 222 220  
 46-60 468 254 244 222  
 61+ 510 376 175 243







\_\_\_\_\_\_ End of the Exam \_\_\_\_\_

NOW render this file to get a Word document. Edit the Word document to look like a business report. Then, submit both this .qmd file and the Word document on Canvas.