Topic ideas

STA 210 - Project

Bayes' Harem - Christina Wang, Ethan Song, Kat Cottrell, David Goh

```
library(tidyverse)
library(tidymodels)
library(knitr)
library(ggfortify)
```

Project idea 1

Introduction and data

- State the source of the data set.
- Describe when and how it was originally collected (by the original data curator, not necessarily how you found the data)
- Describe the observations and the general characteristics being measured in the data

Research question

- Describe a research question you're interested in answering using this data.

Glimpse of data

- Use the `glimpse` function to provide an overview of the dataset

```
# add code to load and glimpse data here
```

Project idea 2

Introduction and data

- State the source of the data set.
- Describe when and how it was originally collected (by the original data curator, not necessarily how you found the data)
- Describe the observations and the general characteristics being measured in the data

Research question

- Describe a research question you're interested in answering using this data.

Glimpse of data

- Use the `glimpse` function to provide an overview of the dataset

```
# add code to load and glimpse data here
```

Project idea 3

> **!** Important
>
> Project idea 3 is optional. If you decide to submit only 2 ideas, please delete the section headings below and leave a note below stating so. If you decide to submit the 3rd idea, please delete this callout.

Introduction and data

- State the source of the data set.

This dataset was published in this form by Kamil Pytlak, and it is taken from the Centers of Disease Control and Prevention (CDC) in the USA. It was retrieved from Pytlak's post on Kaggle.com. Citation: Kamil Pytlak. "Personal Key Indicators of Heart Disease". Kaggle. February 15 2022. Retrieved on May 25 2022 from https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/versions/2?resource=download.

- Describe when and how it was originally collected (by the original data curator, not necessarily how you found the data)

The data was collected by the Behavioral Risk Factor Surveillance System (BRFSS) which conducts **annual telephone surveys about the health status of U.S. residents in all 50 states, the District of Columbia and in the 3 US territories.** This data set includes data from **2020**, consisting of 401958 rows and 279 columns. Of note is that the classes are not balanced, and the author advises fixing the weights or undersampling. (Source: Pytlak, 2022)

- Describe the observations and the general characteristics being measured in the data

The key indicators for the risk of heart disease include smoking, diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol. In addition to these, the data records other information such as history of stroke, mental health, age category, race, general health, sleep time, kidney disease and skin cancer.

Research question

- Describe a research question you're interested in answering using this data.

What can we conclude about a person's likelihood of heart disease from their (1) BMI, (2) sex, (3) age category, (4) sleeping hours, (5) race, and (6) whether they are physically active?

Note that whether someone has heart disease is a binary variable. I have another research question prepared with the outcome variable set to a person's BMI, which is a continuous variable:

What can we conclude about a person's BMI from their (1) alcohol consumption, (2) sex, (3) age category, (4) sleeping hours, (5) race, and (6) whether they are physically active?

Glimpse of data

- Use the `glimpse` function to provide an overview of the dataset

```
heart <- read_csv("data-3/heart_2020_cleaned.csv")
```

```
Rows: 319795 Columns: 18
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (14): HeartDisease, Smoking, AlcoholDrinking, Stroke, DiffWalking, Sex, ...
dbl  (4): BMI, PhysicalHealth, MentalHealth, SleepTime

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(heart)
```

```
Rows: 319,795
Columns: 18
$ HeartDisease     <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No"~
$ BMI              <dbl> 16.60, 20.34, 26.58, 24.21, 23.71, 28.87, 21.63, 31.6~
$ Smoking          <chr> "Yes", "No", "Yes", "No", "No", "Yes", "No", "Yes", "~
$ AlcoholDrinking  <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No",~
$ Stroke           <chr> "No", "Yes", "No", "No", "No", "No", "No", "No", "No"~
$ PhysicalHealth   <dbl> 3, 0, 20, 0, 28, 6, 15, 5, 0, 0, 30, 0, 0, 7, 0, 1, 5~
$ MentalHealth     <dbl> 30, 0, 30, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 30, 0, 2,~
$ DiffWalking      <chr> "No", "No", "No", "No", "Yes", "Yes", "No", "Yes", "N~
$ Sex              <chr> "Female", "Female", "Male", "Female", "Female", "Fema~
$ AgeCategory      <chr> "55-59", "80 or older", "65-69", "75-79", "40-44", "7~
$ Race             <chr> "White", "White", "White", "White", "White", "Black",~
$ Diabetic         <chr> "Yes", "No", "Yes", "No", "No", "No", "No", "Yes", "N~
```

```
$ PhysicalActivity <chr> "Yes", "Yes", "Yes", "No", "Yes", "No", "Yes", "No", ~
$ GenHealth        <chr> "Very good", "Very good", "Fair", "Good", "Very good"~
$ SleepTime        <dbl> 5, 7, 8, 6, 8, 12, 4, 9, 5, 10, 15, 5, 8, 7, 5, 6, 10~
$ Asthma           <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", "Yes", "~
$ KidneyDisease    <chr> "No", "No", "No", "No", "No", "No", "No", "No", "Yes"~
$ SkinCancer       <chr> "Yes", "No", "No", "Yes", "No", "No", "Yes", "No", "N~
```