Report: Exploratory Data Analysis

Applied Economics & Data Analysis in R (Econ 680/880, Dr. Venoo Kakar)

AUTHOR
Kevin Liu

About the data

This dataset was retrieved from Kaggle, which is a web-based platform where data scientists collaborate with others to work and learn from datasets and/or data analysis related challenges. The data contains public health information on U.S. residents containing various factors that could influence heart disease. There are 319,795 observations in this dataset, and contain 18 variables that affect an individual's risk for heart disease. This data is from the year 2020.

Questions

1. What are the lifestyle choices that affect an individual's risk for heart disease?
2. How does heart disease prevalence vary across different age groups?
3. Who is at the greatest risk for heart disease?

Exploratory Data Analysis

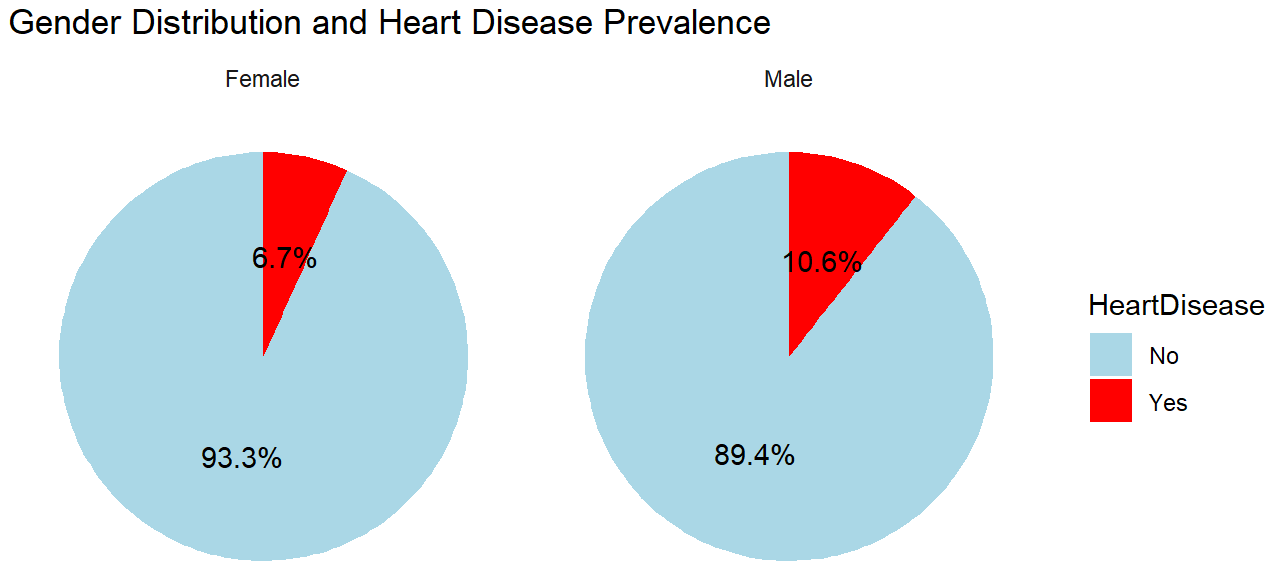
▶ Code

```
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
```

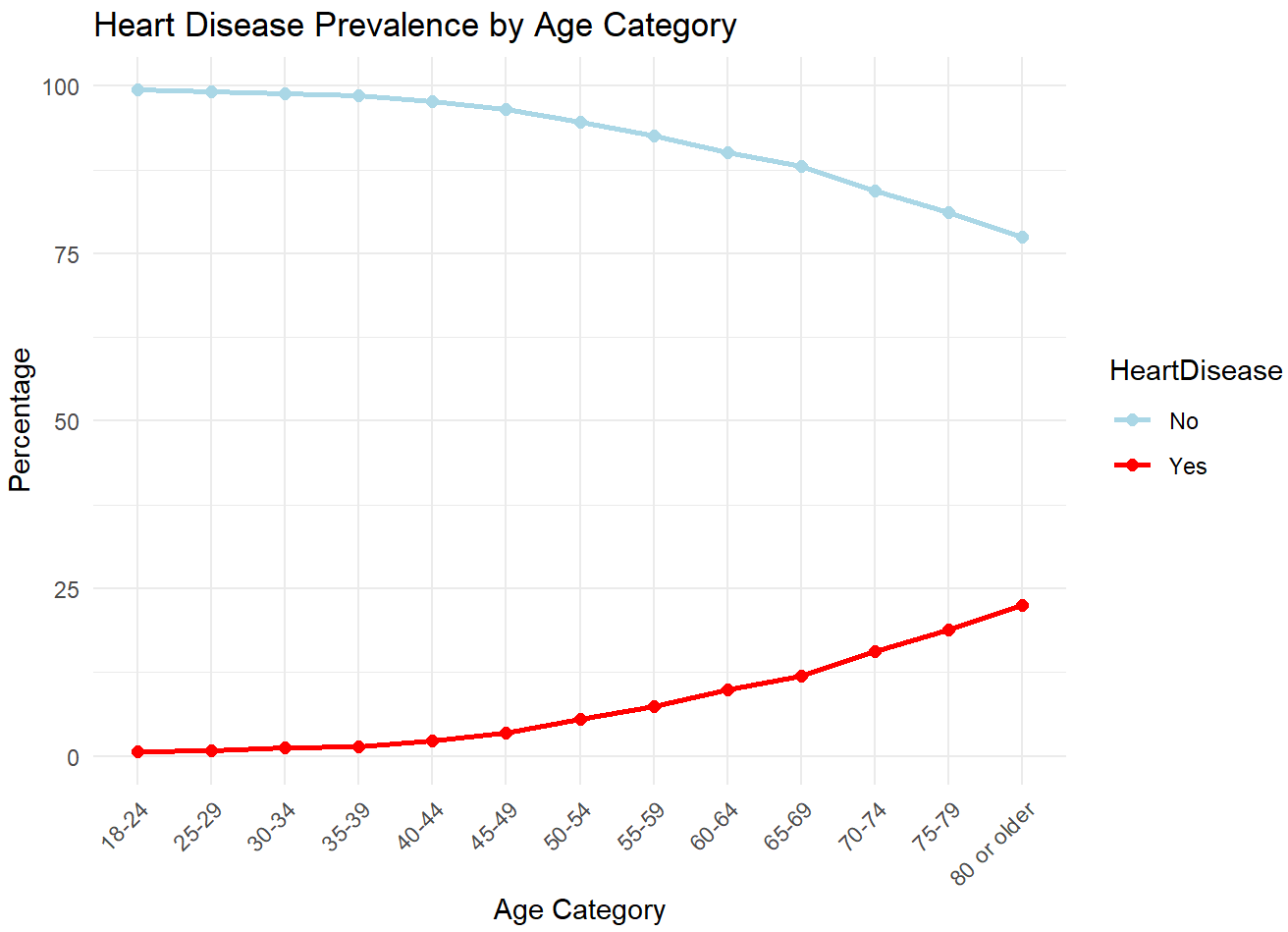
▶ Code

```
     Sex      n
1 Female 167805
2   Male 151990
```

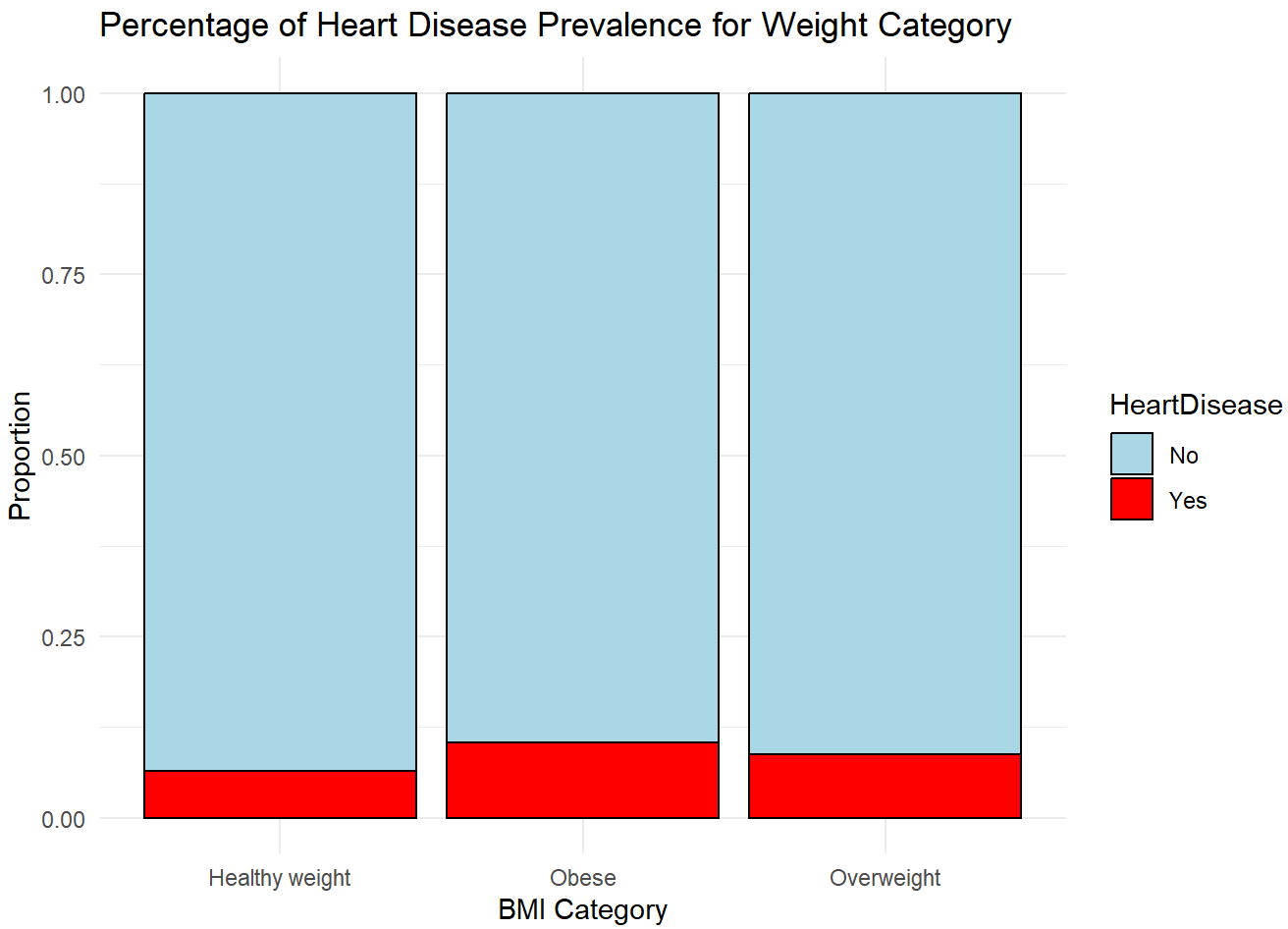
▶ Code


Gender Distribution and Heart Disease Prevalence
Female
Male
6.7%
93.3%
10.6%
89.4%
HeartDisease
No
Yes

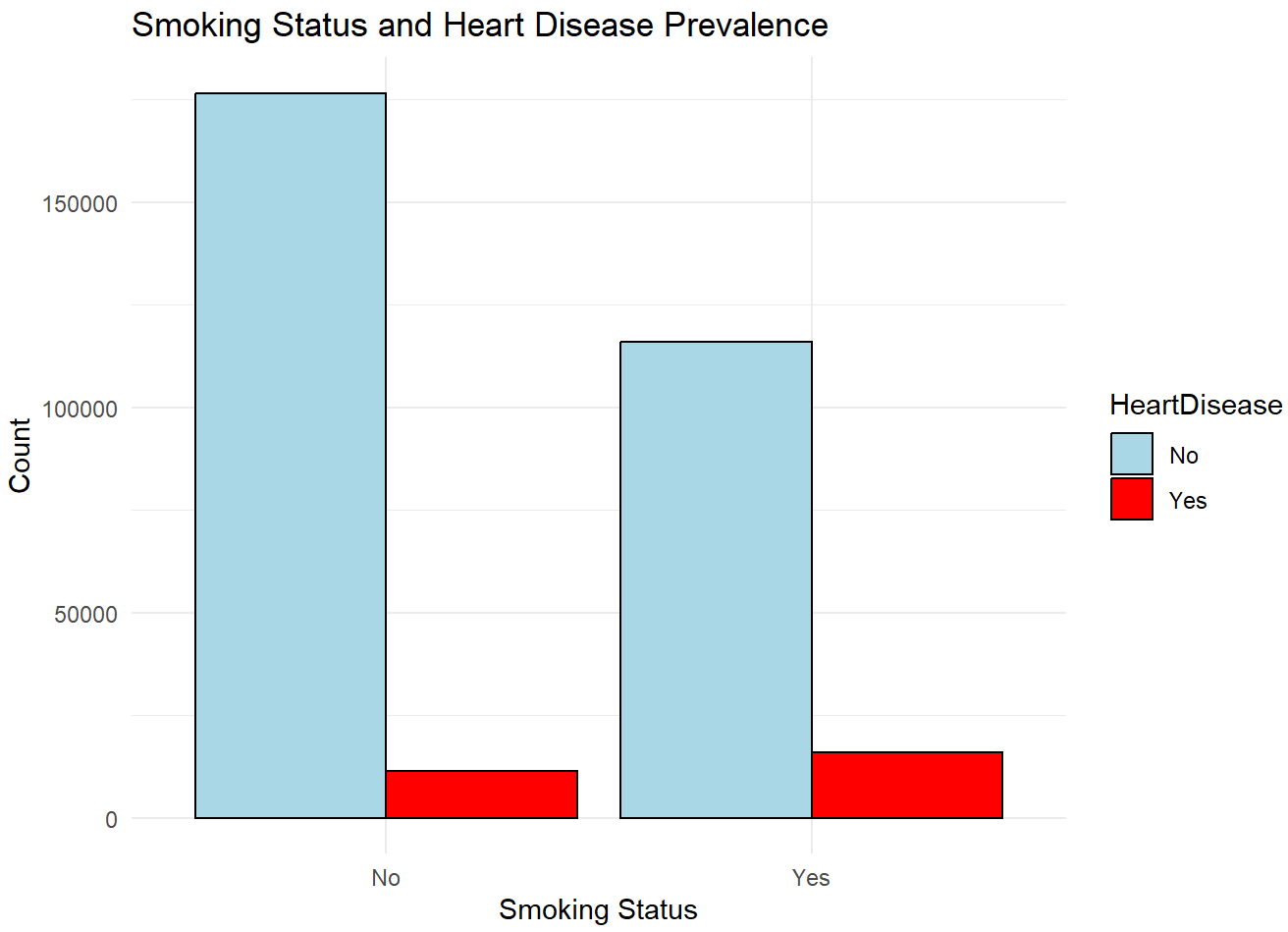
▶ Code


Heart Disease Prevalence by Age Category
100
75
50
25
0
Percentage
18-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80 or older
Age Category
HeartDisease
No
Yes

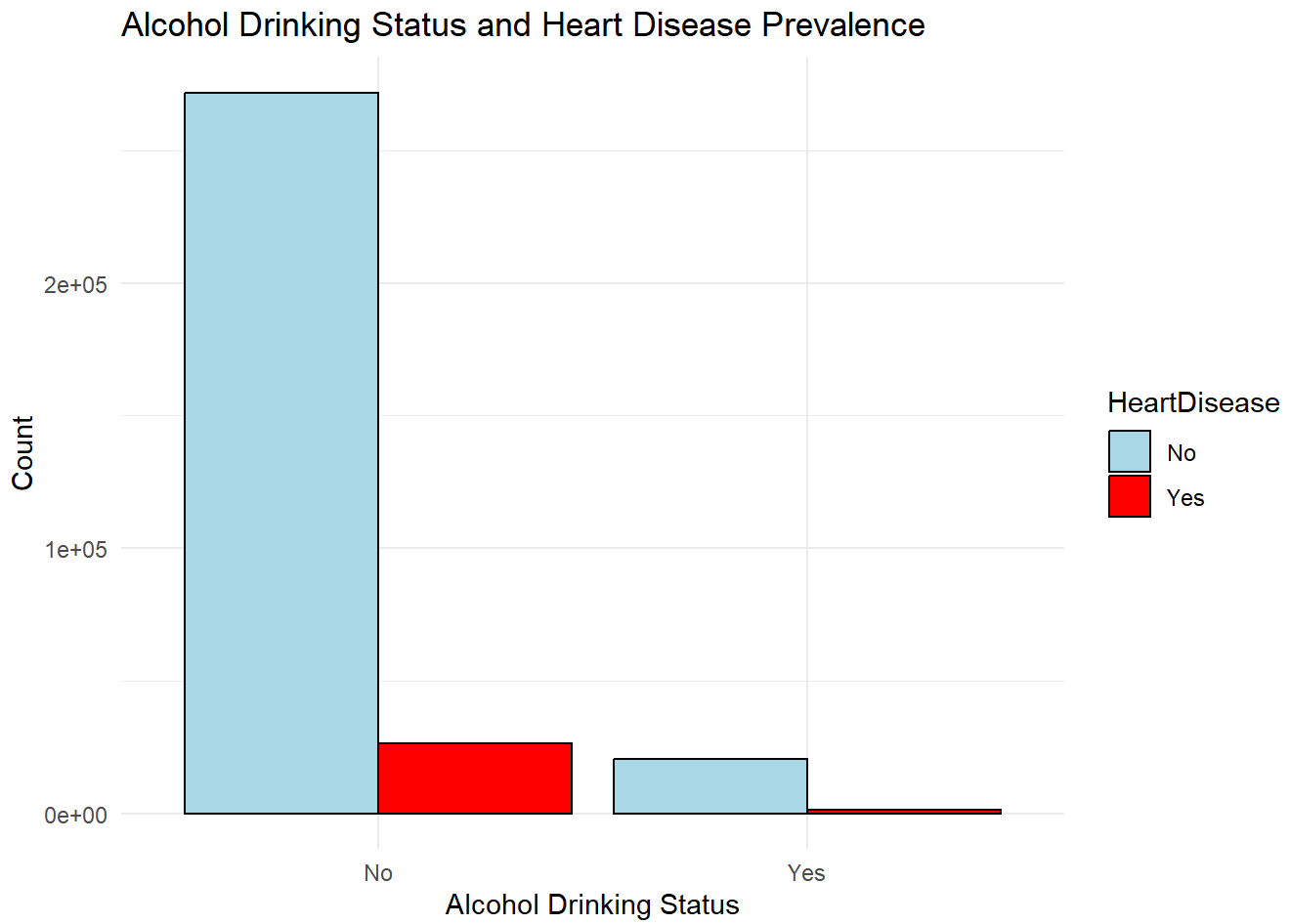
▶ Code


Percentage of Heart Disease Prevalence for Weight Category
1.00
0.75
0.50
0.25
0.00
Proportion
Healthy weight
Obese
Overweight
BMI Category
HeartDisease
No
Yes

▶ Code


Smoking Status and Heart Disease Prevalence
150000
100000
50000
0
Count
No
Yes
Smoking Status
HeartDisease
No
Yes

▶ Code



Summary

Age, Weight, and Smoking are the 3 most consistent indicators of an individual's risk for heart disease. Males also have a slightly higher prevalence for heart disease. An individual that would be at greatest risk would be a male who regularly smokes, is of the "Obese", or "Overweight" BMI categories, is not physically active, and is of an older age category. The greater the age category, the greater the risk becomes.