Midterm Project

Jacob Burke

13/10/2019

Hawainn Health Survey Analysis

For this project, I am going to clean and organize a raw survey data set, to prepare for my 'modeling team' who are going to run logistic regression modeling to observe the risks of diabetes diagnosis, given a number of predictor variables also from this data set pertaining to the individuals.

First things first, we need to read in the raw data set.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(asciiSetupReader)
library(dplyr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
hawaii <- read_ascii_setup("samadult.dat", "SAMADULT.sps")

hawaii <- as_tibble(hawaii)
```

Next what I want to look at is a frequency table value corresponding to the number of NA values for each survey question(variable) and the number of unique answers also shown. For analysis, I want to have a section of variables with strong response rates and significant enough variance

```
## getting frequencie of NA and unique answers for each survey question
hawaiiNA <- NA

for(i in 1:488)
{
 hawaiiNA <- rbind(hawaiiNA, sum(is.na(hawaii[ , i])))
}

hawaii_unique <- NA

for(i in 1:488){
```

```
  hawaii_unique <- rbind(hawaii_unique, length(t(unique(hawaii[,i]))))
}


hawaii_answers <- cbind(hawaiiNA, hawaii_unique)
colnames(hawaii_answers) <- c("NACount", "UniqueAnswers")

## frequency table
hawaii_answers <- hawaii_answers[2:489, ]

hawaii_answers <- t(hawaii_answers)
```

This ends with a summary matrix looking like this (header).

```
print(head(t(hawaii_answers)))
```

```
##  NACount UniqueAnswers
##     1500             5
##        0             4
##     2063             4
##     2063             3
##     2063             3
##        0             4
```

Now, I want to look specifically at people who have been diagnosed with diabetes. To narrow down my data, I need to isolate the variable pertaining on diabetes diagnosis (column 47). I also want to check the NA values to see if there is sufficient data to work with. ** It is important to note that for the purpose of this investigation, we are making the assumption that column 47 "Have you ever been told you have diabetes. . . " is a pointer towards a definitive diagnosis for individuals that answered 'yes'.

From there, I'll select a number of variables that will be interesting to investigate into whether or not they have effects on the risks of developing diabetes.

```
diabetes <- hawaii[, 47]

print(hawaii_answers[, 47])
```

```
##       NACount UniqueAnswers
##             0             5
```

We can see that there is no missing data, now to check for inconsistent data entries.

```
colnames(diabetes) <- "Diabetes"

table(diabetes)
```

```
## diabetes
##    1    2    3    7    9
##  381 2111   95    1    2
```

From here we can see that these questions were answered fairly consistently. For the purpose of data consistency and our modeling, we are going to set "2" to "0", as our baseline, which is individuals whom have never been told they have diabetes. In addition, for the purpose of this analysis, we will recode "3" of being borderline diagnosed, to the "1" of being diagnosed with some sort of diabetes. Patients who answered "7" (refused to answer) will be recoded to an "NA" and values "9" of patients who were not certain, will be recoded to the "0" baseline value.

```r
diabetes[diabetes == 2] <- 0
diabetes[diabetes == 3] <- 1
diabetes[diabetes == 7] <- NA
diabetes[diabetes == 9] <- 0
```

Next we want to pick out a number of variables from the survey data that should be controlled for when modelling diabetes diagnosis, as well as certain variables that would be interesting to observe whether or not they hold any affects on diabetes diagnosis.

Firstly, physical attributes. Here I'm going to extract the columns from the raw data, and clean any inconsistencies.

```r
## Physical attributes

physical <- hawaii[, 315:317]

physical <- as_tibble(physical)

## Renaming columns

colnames(physical) <- c("Height_(in)", "Weight_(lbs)", "BMI")

## Setting NA values to 96, 97, 98, 99 values entered in height (these were either refused answers, or

physical$`Height_(in)`[physical$`Height_(in)` == 96 | physical$`Height_(in)` ==97
                       | physical$`Height_(in)`== 98
                       | physical$`Height_(in)` ==99] <- NA

## Setting unanswered Weight values to NA

physical$`Weight_(lbs)`[physical$`Weight_(lbs)` == 996 | physical$`Weight_(lbs)` ==997
                       | physical$`Weight_(lbs)` == 998
                       | physical$`Weight_(lbs)` ==999] <- NA

## Setting unanswered BMI Values to NA

physical$BMI[physical$BMI == 5] <- NA
```

Now that I have a clean physical attributes dataset, I want to extract a few other variables. Survey questions that involve quantifying alcohol consumption (column 313), general health questions (columns 224 - 226), smoking(column 291 - 300), sleep indicators (columns 472 - 476), and joint pain (columns 58-78).

Now to extract the alcohol consumption, general health, sleep, and joint data, and clean missing values.

```r
## 1) Alcohol

alcohol <- hawaii[,313]

## setting unknown drinking category to NA

alcohol[alcohol == 10] <- NA
alcohol <- as_tibble(alcohol)
colnames(alcohol) <- "Alcohol_Consum"

## 2) General Health
```

```r
health <- hawaii[,224:226]

## Looking at the health data, I've decided to use the "number of work days lost in the past 12 months"

health <- health[,1]
colnames(health) <- ("Sick_Days_General")
health[health == 999] <- NA

## 3) Sleep

sleep <- hawaii[,472:476]
## Looking at the sleep data, "Hours of sleep" will be the survey question I'm choosing as my sleeping


sleep <- sleep[,1]
colnames(sleep) <- ("Hours_Sleep_Night")
sleep[sleep == 97 | sleep == 98 | sleep == 99] <- NA

## 4) Joint pain

joint <- hawaii[, 58:78]

## Looking at this section, "joint pain last 30 days" will be the variable subset I'll use as my joint

joint <- joint[,1]

## cleaning values that were not answered, or did not know

joint[joint == 7 | joint == 9] <- NA
colnames(joint) <- "Joint_Pain"
```

Now, I'll combine my individual data columns of interest, into a data frame that I now can use to run EDA and modelling on (after removing NA values).

```r
data <- cbind(diabetes, physical, alcohol, health, sleep, joint)

## To remove final NA values

data <- na.omit(data)

## Tidy data (preview)

print(head(data))
```

```
##   Diabetes Height_(in) Weight_(lbs) BMI Alcohol_Consum Sick_Days_General
## 1        0          71          240   4              6                 0
## 2        1          64          160   3              1                 1
## 3        0          63          115   2              6                 0
## 5        0          69          170   3              5                 0
## 7        1          63          190   4              2                 0
## 9        0          65          189   4              2                10
##   Hours_Sleep_Night Joint_Pain
## 1                 4          1
## 2                 8          1
```

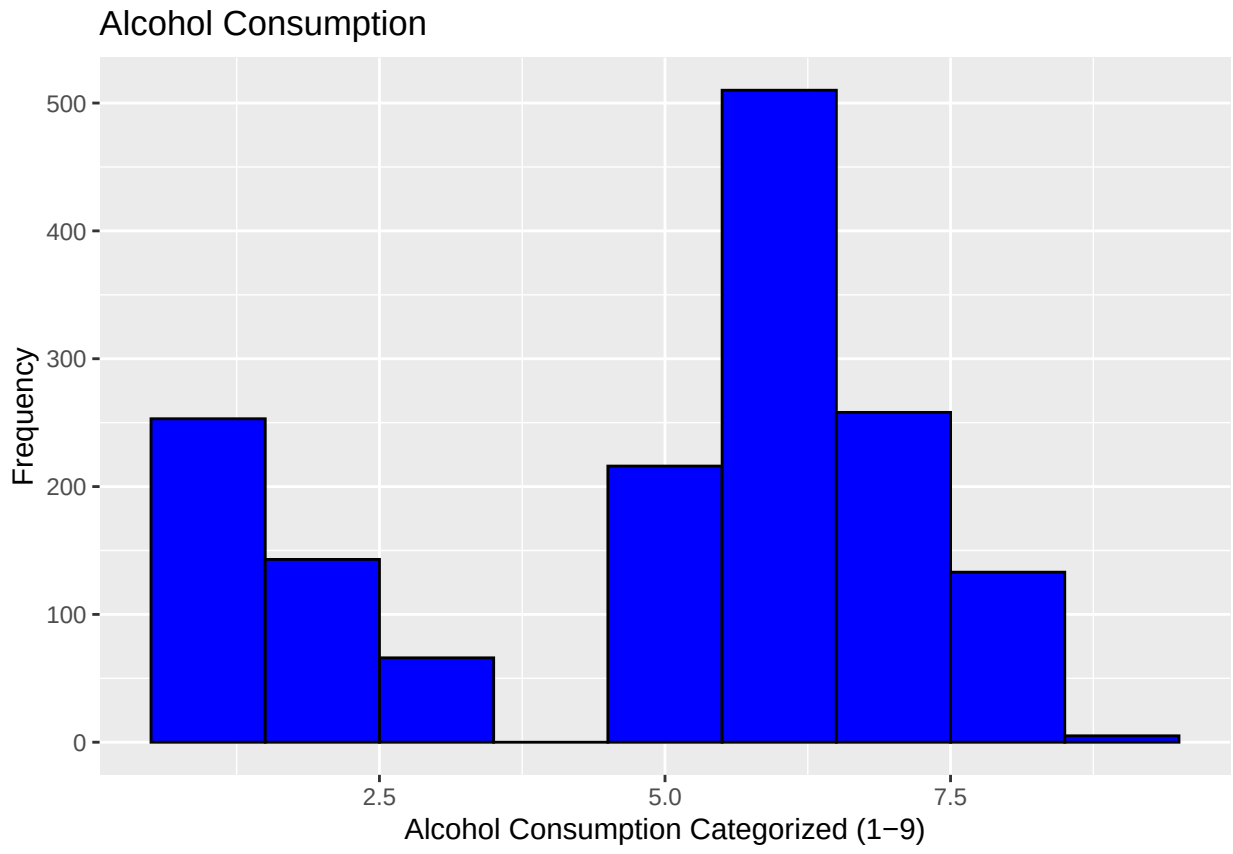
```
## 3                6          2
## 5                6          2
## 7                6          1
## 9                6          1
```

EDA

To look further into the data, I'm going to begin to run some intial EDA, just to see what patterns or certain observations come up.

Here we're first looking at multiple histograms, to see the different distributions of variables. Given that each variable of interest here is coded as a categorical variable, I also have an output table for each variable's catagorization respectively.

```
## Alcohol

Number <- (1:9)

Category <- c("Lifetime abstainer", "Former infrequent",
      "Former regular", 'Former, unknown frequency', 'Current infrequent', 'Current light',  'Current m

alcohol_cat <- cbind(Category,Number)

ggplot() + geom_histogram(aes(x = data$Alcohol_Consum), binwidth = 1, color = 'black', fill = 'blue') +
```




```
kable(alcohol_cat, digits = 2, format = "latex", booktabs=TRUE, caption = "Alcohol Categorization") %>%
```

Table 1: Alcohol Categorization

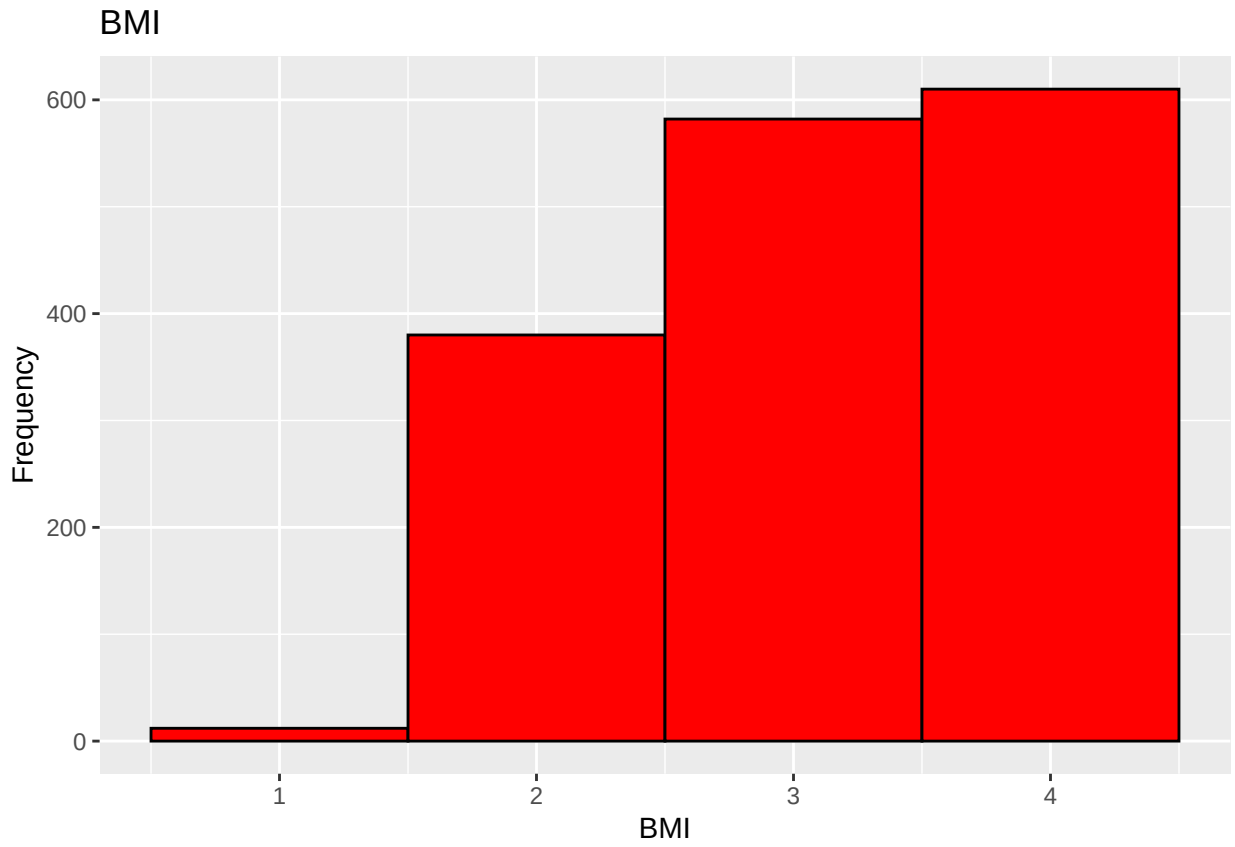
Category	Number
Lifetime abstainer	1
Former infrequent	2
Former regular	3
Former, unknown frequency	4
Current infrequent	5
Current light	6
Current moderate	7
Current heavier	8
Current drinker, frequency/level unknown	9

In regards to alcohol consumption, we have essentially a split between people who were either abstainers for most of their lives or former drinkers, and and then a larger subset who are current drinkers, with most identifying on the 'light' side, while still a significant amount on the moderate-heavy side.

```
## BMI

Number <- (1:4)

Category <- c('Underwieght', 'Healthy Weight', 'Overweight', 'Obese')

BMI_cat <- cbind(Category,Number)


ggplot() + geom_histogram(aes(x = data$BMI), binwidth = 1, color = 'black', fill = 'red') + labs(x = "BI
```

Table 2: BMI Categorization

Category	Number
Underwieght	1
Healthy Weight	2
Overweight	3
Obese	4




```
kable(BMI_cat, digits = 2, format = "latex", booktabs=TRUE, caption = "BMI Categorization") %>% kable_s
```

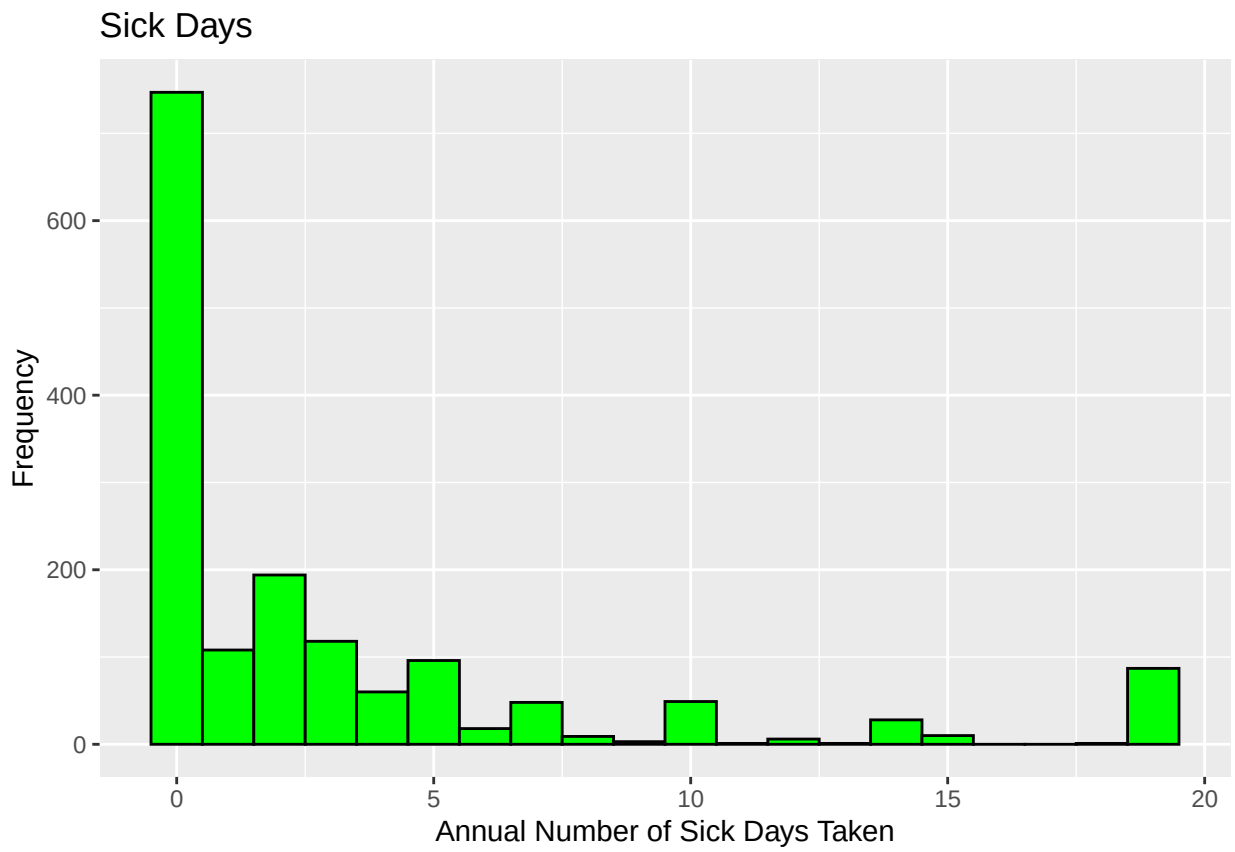
From here we can see that there is a significant skew of these individuals either overweight, or obese. This could suggest we're looking at quite a few individuals in this data set with less than optimal health.

```
## Health

Number <- (1:2)
Number[1] <- "0-18"
Number[2] <- "19"

Category <- c('0-18 days', '19 or more days')

Health_cat <- cbind(Category,Number)
```

Table 3: Sick Days Categorization

Category	Number
0-18 days	0-18
19 or more days	19

```
ggplot() + geom_histogram(aes(x = data$Sick_Days_General), binwidth = 1, color = 'black', fill = 'green
```




```
kable(Health_cat, digits = 2, format = "latex", booktabs=TRUE, caption = "Sick Days Categorization") %>%
```

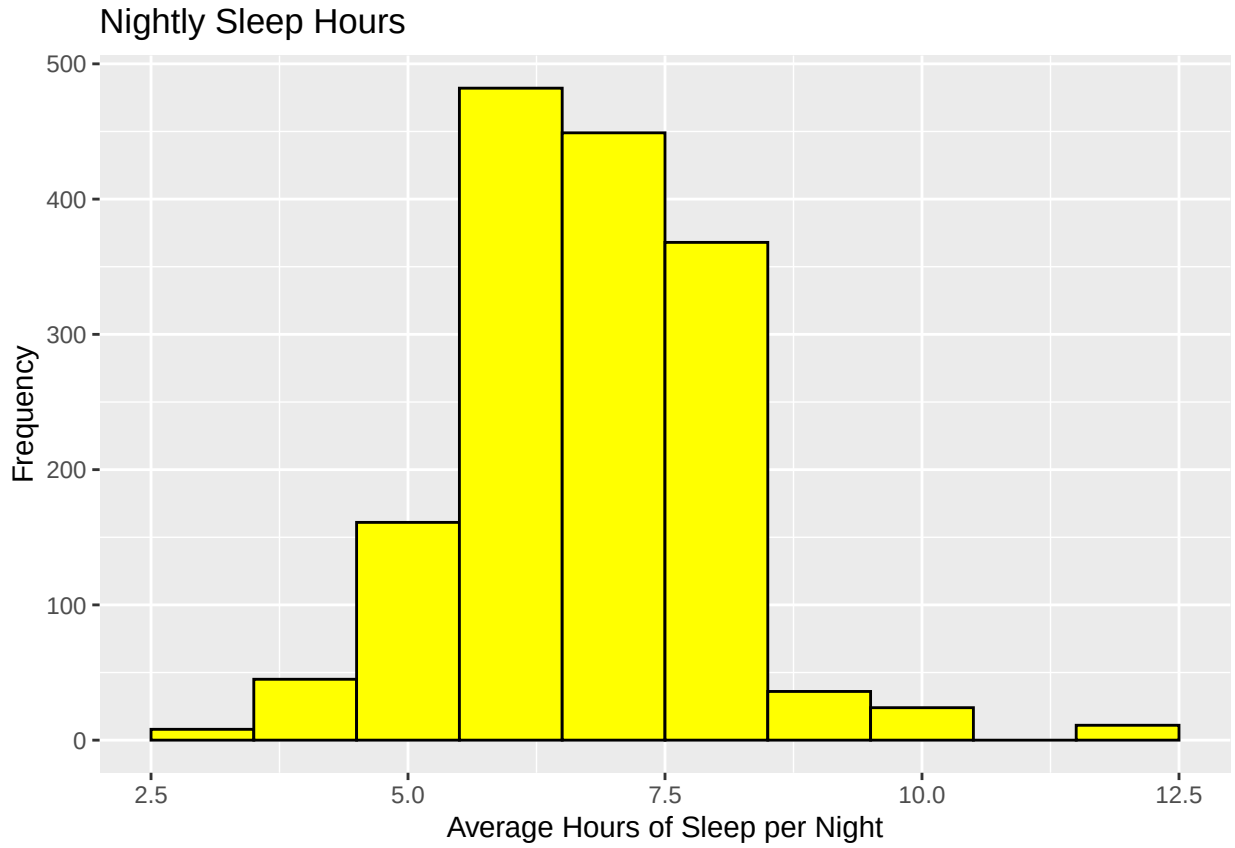
From here we can see that most individuals missed little to know time from work due to illnesses. There are a significant few between 2-5 sick days taken, and then a significant outlier subset of a few people who claim they missed more than 15 days.

```
## Sleep

Number <- (1:3)

Number[1] <- "3"
Number[2] <- "4-11"
Number[3] <- "12"
```

Table 4: Sleep Amounts Categorization

Category	Number
1-3 hours	3
4-11 hours	4-11
12 or more hours	12

```
Category <- c("1-3 hours", "4-11 hours",
      "12 or more hours")

Sleep_cat <- cbind(Category,Number)

ggplot() + geom_histogram(aes(x = data$Hours_Sleep_Night), binwidth = 1, color = 'black', fill = 'yellow
```




```
kable(Sleep_cat, digits = 2, format = "latex", booktabs=TRUE, caption = "Sleep Amounts Categorization")
```

From here, we can see that most average at just about 6 hours of sleep a night, with the majority of individuals hovering between 5-8 hours.

Modeling Analysis (Next Steps for Modeling Team)

From here, my modeling team can take this cleaned and explored data set to start model analysis, with the main goal of analysis to create a logistic regression model using diabetes diagnosis as the outcome variable, with the other variables in this data set as predictors, to see what intepretations can be made. **It is

important to note that for the purposes of this analysis, we have made the assumption that answering "yes" to the question of whether or not one has ever been told they have diabetes, points towards a definitive diabetes diagnosis.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.