**Section 2: Final Code**

\*The data set that I am working with is from the CDC, and is called the 2019 Behavioral Risk Factor Surveillance System.

It is derived from survey data from annual telephone surveys, and considers the prevalence of factors that pose a risk

of future health problems. It was pre-made to be suitable for direct SAS import, along with cleaning and formatting.;

\*After visually inspecting the data set, and referencing the data structure here:

https://www.cdc.gov/brfss/annual\_data/2019/pdf/codebook19\_llcp-v2-508.HTML.

The data set is huge -- over 400,000 observations, across 342 different variables. Before I can start exploring questions,

I'll need to trim it down to relevant variables.

Also, the variable code names are very clumsy and counterintuitive, so I will rename them to categorize.

For analytic purposes, it will be easier if I dichotomize them into 1/0 arrangements, based on risk factors.

Thus: MAXDRNKS (maximum # of drinks in a single session) gt 4, then "heavy" =1. DRNK3GE5 (binge drinking occasions) gt 1,

then "binge"=1. AVEDRNK3 (average # of drinks per day) gt 2, then "high"=1. MENTHLTH (#days with bad mental health in a

month) gt 1 then "sad"=1. ADDEPEV3 (clinical diagnosis of depression) = 1 (yes), then "depressed"=1.

Since I can't perform analytics on missing data, I'm deleting those who didn't respond, those who refused to answer,

those who responded "don't know.";

**Data** Project;

set llcp2019 (keep= MENTHLTH ADDEPEV3 AVEDRNK3 DRNK3GE5 MAXDRNKS BIRTHSEX);

if MAXDRNKS eq **.** or MAXDRNKS eq **77** or MAXDRNKS eq **99** then delete;

else if MAXDRNKS LT **5** then Heavy=**0**;

else if MAXDRNKS ge **5** then Heavy=**1**;

if AVEDRNK3 eq **.** or AVEDRNK3 eq **77** or AVEDRNK3 eq **99** then delete;

else if AVEDRNK3 lt **2** then High=**0**;

else if AVEDRNK3 ge **2** then High=**1**;

if DRNK3GE5 eq **.** or DRNK3GE5 eq **77** or DRNK3GE5 eq **99** then delete;

else if DRNK3GE5 eq **88** then Binge=**0**;

else if DRNK3GE5 ge **1** then Binge=**1**;

if MENTHLTH eq **.** or MENTHLTH eq **77** or MENTHLTH eq **99** then delete;

else if MENTHLTH eq **88** then Sad=**0**;

else if MENTHLTH ge **1** then Sad=**1**;

if ADDEPEV3 eq **.** or ADDEPEV3 eq **7** or ADDEPEV3 eq **9** then delete;

else if ADDEPEV3 eq **2** then Depressed=**0**;

else if ADDEPEV3 eq **1** then Depressed=**1**;

if BIRTHSEX eq **.** or BIRTHSEX eq **7** or BIRTHSEX eq **9** then delete;

else if BIRTHSEX eq **1** then Gender= "Male";

else if BIRTHSEX eq **2** then Gender= "Female";

**run**;

\*Now, the data is stripped down to only the variables relating to drinking and depression.

The resulting data has 184,000 observations, all of which are numeric. Let's see what we are left with;

**proc** **freq** data=project;

tables gender heavy high binge sad depressed;

**run**;

\*The data summary tables indicate that around 48% or subjects were female, 21% of subjects were heavy daily drinkers, 54.5% had

higher than recommended consumption when they drank, 25% had binged at least once in the last month,

36% reported having mental health issues in the past month, and 17% had been diagnosed with depression.

\*Now, I want to see if there is variation between men and women on these topics;

**proc** **freq** data=project;

tables gender \* (heavy high binge sad depressed);

**run**;

\*Interesting -- men comprised 60% of the high daily drinkers, 72% of the heavy drinking days, and 60% of the binges. However,

women were hovering around 60% for both sad days and depression diagnoses.

This can be seen in the next two charts;

title "Heavy Drinking by Gender";

**proc** **gchart** data=project;

pattern1 value=solid color=pink;

pattern2 value=solid color=vlib;

where heavy =**1**;

pie gender;

**run**;

title "Depression by Gender";

**proc** **gchart** data=project;

pattern1 value=solid color=pink;

pattern2 value=solid color=vlib;

where depressed =**1**;

pie gender;

**run**;

\*It occurs to me that there is likely overlap between these two categories (sad/depressed and heavy/high/binge), and

that if there isn't, there might be some data entry errors. For example, how could someone be diagnosed as depressed, but not

be experiencing more than 1 day a month of sadness? I'm going to subset to look at this more closely;

**data** combination;

set project (keep= gender heavy high binge depressed sad);

if depressed = **1** and sad =**1** then Verysad=**1**;

else Verysad=**0**;

if heavy = **1** and high =**1** and binge=**1** then Alcoholic=**1**;

else Alcoholic=**0**;

**run**;

\*Now we have only 30,000 observations. I'm going to look at the same graph as above, to see if alcoholism

and heavy drinking are distributed similarly;

title "Heavy Drinking by Gender";

**proc** **gchart** data=combination;

pattern1 value=solid color=pink;

pattern2 value=solid color=vlib;

where alcoholic =**1**;

pie gender;

**run**;

title "Depression by Gender";

**proc** **gchart** data=combination;

pattern1 value=solid color=pink;

pattern2 value=solid color=vlib;

where Verysad =**1**;

pie gender;

**run**;

\*Turns out, they are almost identical.;

\*I'm curious, given the parameters I defined to use as categories, what the average number of depressed days, number of

drinks consumed, and binge days are;

**data** averages;

set llcp2019 (keep=DRNK3GE5 MAXDRNKS AVEDRNK3 MENTHLTH);

binges= DRNK3GE5;

saddays= MENTHLTH;

if DRNK3GE5 eq **.** or DRNK3GE5 eq **77** or DRNK3GE5 eq **99** then delete;

else if DRNK3GE5 eq **88** then binges=**0**;

if MAXDRNKS eq **.** or MAXDRNKS eq **77** or MAXDRNKS eq **99** then delete;

if AVEDRNK3 eq **.** or AVEDRNK3 eq **77** or AVEDRNK3 eq **99** then delete;

if MENTHLTH eq **.** or MENTHLTH eq **77** or MENTHLTH eq **99** then delete;

else if MENTHLTH eq **88** then saddays =**0**;

**run**;

**proc** **means** data=averages;

var saddays AVEDRNK3 binges MAXDRNKS;

**run**;

\*Note that above, gender and ADDEPEV3 are categorical.

Looks like the parameters being used are reasonable-- the average number of Sad Days is 3.5, the average number of daily drinks

is 2.2, the average number of binges is 1.1, and the average most drinks on a single occasion is 3.3.

\*Now I need to address the question originally posed: Does depression correlate with heavy drinking. I'll use some of the

new variables I've defined and create combinations;

**data** combination2;

set averages (keep= binges saddays MAXDRNKS AVEDRNK3);

alcoholic= binges + MAXDRNKS + AVEDRNK3;

**run**;

\*Now, I can simply see if the combined total of average drinks plus binges plus MAXDRNKS correlates

with an increased number of Sad Days. First, a scatter plot;

title "Correlation Between Alcohol Consumption and Depression";

**proc** **sgplot** data=combination2;

reg x=saddays y=alcoholic;

**run**;

\*This produced a very weird plot, with a regression line indicating virtually no correlation.;

**proc** **reg** data=combination2;

model alcoholic=saddays;

**run**;

\*This confirmed that there was virtually no correlation, with an r-square value of .015. I'm going to print a subset of my data

to make sure that it looks how it should in my head;

**proc** **print** data=combination2 (obs=**50**);

**run**;

\*The data is fine as it is. Let's try a different approach;

**proc** **sort** data=combination out=project2;

by descending alcoholic descending Verysad;

**run**;

**proc** **freq** data=project2 order=data;

tables Verysad\*alcoholic/relrisk CMH1;

**run**;

\*This is interesting, when using the composite dummy variables from the "combination" dataset, the odds ratio is 1.3, or that

people who are "VerySad" are 1.3x more likely to be "Alcoholic". I wonder why there was no correlation when it wasn't

dichotomized.

**proc** **freq** data=project2 order=data;

tables Gender\*Verysad\*alcoholic/relrisk CMH1;

**run**;

\*When stratified by gender, we see that among men, being Very Sad carries an odds ratio of 1.5 for alcoholism, whereas for

women, being very sad carries an odds ratio of 1.9.;

**proc** **logistic** data=combination descending;

class depressed/param=ref ref=first;

model alcoholic=depressed;

**run**;

\*Okay, perhaps I was correct above in speculating the MENTHLTH variable wasn't appropriate for this analysis. When I applied

logistic regression to just depression (clinical diagnosis), there was an odds ratio of 1.23.