**Name: Juan Villegas**

**Lab 5A: Monitoring the Future Study**

**Part I: Preparing the Data**

For this lab, you will begin working with the Monitoring the Future data for 2019, the most current year for which data are available. You will continue working with these data next week in Lab 5B: Monitoring the Future Study.

1. a. In your Data folder, create a new folder named Lab 5. Go to Blackboard. In the Data folder you will find two files: MTFData\_2019.csv and MTFCodebook\_2019. Save them into your Lab 5 folder.

b. Open a new Excel spreadsheet, save it as follows:

* File Name: Lab5Data
* Save as Type: CSV (comma delimited)
* Click **OK** and then **Yes** to questions. Check your Lab 5 folder and you should see Lab5Data listed.

c. Open the file MTFData\_2019.csv.

2. Open the Codebook. Find the code V# corresponding to the variables in the bulleted list below. **You want to get the variables from the Core data set – there should be over 13,000 observations, not 2,000 – 4,000 observations.** Copy the column corresponding to each V# into your newly created file Lab5Data.csv. Here’s how:

* Go to the top of the column, Click to highlight the entire column.
* Press Ctrl c to copy, then go to your new Lab5Data spreadsheet.
* Click in the top cell in Column A, press Ctrl v to copy the data into the first column.
* Replace the corresponding V# with a variable name. Choose a name **without any spaces** (you can use underscores or periods for spaces). Keep the variable names short but informative. When this is completed, **save** the **.csv file**. (You will have to answer yes to a question, then x out of the file and don’t save a second time.)
* Repeat the process for the other 6 variables.

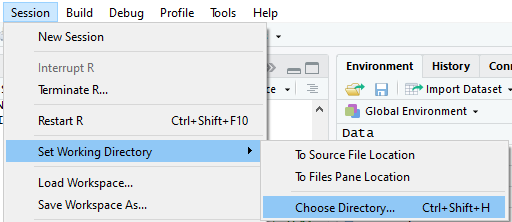
**Variables:**

* Sex (R’S SEX)
* Father’s Educational Level (FATHR EDUC LEVEL)
* Mother’s Educational Level (MOTHR EDUC LEVEL)
* Political Preference (R’S POLTL PRFNC)
* Religious Importance (RLGN IMP R’S LF)
* Ticket (moving violation) (#X/12MO R TCKTD)
* Drive (how far) (DRIVE>200 MI/WK

3. a. Open RStudio. (Save any old work that you want to return to later.)

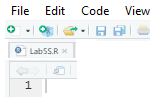
b. Set up the following:

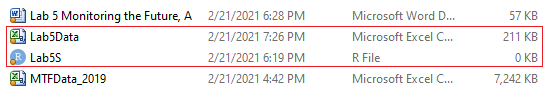
* Set your working directory to the Lab 5 folder. Here’s how:



Use the browser to find your Lab 5 folder, select it and click Open.

* Open a new script: File>New File>R Script. Save it as Lab5S to the Lab 5 folder: File>Save As>Lab5S. (Notice: you will be automatically saving into your Lab 5 folder because you have set Lab 5 as your working directory.) Your script tab should now be named Lab5S.R instead of “untitled” as shown below.



* Look in your Lab 5 folder. It should contain files named something like what is shown below. 

c. Read your newly created csv file into RStudio as a data frame:

DataLab5 <- read.csv(file.choose())

Your browser will now go directly to the Lab 5 folder and you can import

your data from Lab5Data.csv.

Alternatively, go to the Environment panel, click Import Data, make sure Heading Yes is selected.

Check that you have entered your data correctly by running the command head(DataLab5. (**Warning:** Do type DataLab5 and then Run or R will try to print 13,000+ rows of data!)

d. For now, the variable values are all numeric. Use the Codebook and give the numeric values labels. (How did you handle the missing values?)

For example, if you named the first variable Sex, then here’s how to label the values:

DataLab5$Sex <-factor(DataLab5$Sex, levels = c(1, 2),

labels = c(“Male”, “Female”))

**Note:** the missing values (-9) will not be labeled and hence, will not end up

as rows in your tables. That is exactly how you want to treat the missing

values.

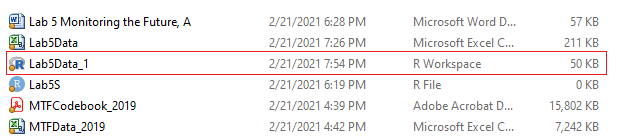
**Complete the labeling for all the variables.** Then you can save your work as an .RData file in part (e). That file will save all of your labeling.

e. Save a copy of this data frame as an .RData file.

save(DataLab5, file = "DataLab5\_1.RData")

Check your Lab 5 folder to see how this file appears in your folder. Here’s

how it looks in my Lab 5 folder.



**Note:** I named the saved file DataLab5\_1RData. The reason I named this file DataLab5\_1 is that after I have done more work on my data, I may decide to save another version of the data frame, and this time I will name the file DataLab5\_2.RData. This way as I work on complicated data, I can save various versions and if I make a mistake in coding, I can go back to an earlier saved file instead of starting over. I can also email the .RData file to team members and they can open my data file with the changes that I have made.

**Here’s how to load an .RData file into RStudio:**

Set the working directory to the folder that contains the data frame.

load(file.choose())

Select DataLab5\_1

Press Ctrl Enter

You should see DataLab5 listed in your Global Environment

**Part II: Analysis**

4. What percentage of the respondents were Male? Female? (You will need to make a table from your Lab5Data$Sex (or whatever you named this variable).

Male Female

6123 6540.

Male Female

48.35% 51.65%

5. Were males or females more likely to get tickets? Support your answer with appropriate percentages. (For this you will need to make a two-way table and calculate either row or column percentages, which ever makes sense to answer this question.

Text

Description automatically generated

Text

Description automatically generated

Text, letter

Description automatically generated

Comparing the percentages from SexTicket\_Row, Males were most likely to get a ticket in every category other than the ‘None’ category.

Comparing the percentages from SexTicket\_Column, Males still beat Females in every column other than None.

I think that the This proves that Males were most likely to get Tickets from this data.

6. a. Make a two-way table for Drive and Ticket.

A picture containing text, receipt

Description automatically generated

b. Is there an association between how far students drive and their getting tickets? (**You need to run a chi-square test in order to answer this question.**)

Text

Description automatically generated

Since my P-Value, 2e-16, is a lot less than .05, I have to reject the Null hypothesis. There is an association b/w how far a student drives and receiving a ticket.

c. Make table of the conditional distributions of tickets for each level of driving distance. (Here we are treating driving distance as the explanatory variable and ticket as the response variable.)

Text

Description automatically generated

d. Make a graphic display for part (c).

Chart, histogram

Description automatically generated

e. Write a short paragraph summarizing what can be learned from the output used to answer parts (c) and (d).

From the Graph and Table, the viewer can tell that there is an associate between driving distance and tickets. From the visual display (The Boxplot), If the viewer looked at the “None” category, they are able to tell that People that drove longer distances had a smaller Percentage of not getting a ticket compared to the people that drove less.

The viewer gets the same message from looking at the “1 ticket” “2 Tickets” “3 Tickets” & “4Tickets” categories because the longer the distance driven, the higher the percentage was for tickets.

**Part III: About the MTF Study**

7. Spend some time conducting some research via the Codebook and the Internet on the MTF study. Describe some of your findings in one or two **well-written** paragraphs. If you put some effort into these paragraphs, you may be able to use them in your final technical report later in the semester.

After researching Monitoring the Future, I learned that this project is focused on surveying High School Seniors. This project tends to focus on surveys regarding drug use amongst American High Schools & young adults. The research is done by the University of Michigan.

I learned that this research group first started in 1975 and has sent out questionnaires yearly by mail.

The survey that is sent out annually is anonymous for 8th-10th graders but, it is confidential for High school seniors.

Something interest from this project is that while they are known for their drug surveys, they have also expanded on their research by adding things like college attendance, military service, civilian employment and unemployment.

Giving a graph a quick look, I learned that the regular use of cigarettes was at an all-time high for High School seniors in 1997. The percentage of use was 24.6… almost 25%. The percentage has drastically dropped down to 4.2% in 2017 but, the use of marijuana has linearly started to rise.