**Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Chapter 2 Lab Cover Sheet**

**\*\* First load the data in post\_open\_refine\_clean.dta**

## load necessary packages

install.packages("foreign")

library(foreign)

## set working directory (use the path where you saved the file post\_open\_refine\_clean.dta)

setwd("C:/Users/…PPOL560/PPOL560\_data")

## load data

data <- read\_dta("post\_open\_refine\_clean.dta")

**\*\* The codebook for this data is Codebook for EXIT POLL survey.docx**

1. Use the following to create dummy variables for Arlington and Prince William Counties. How many observations are from each county?

Arlington <- dataset[which(dataset$precinct == "AR49" | dataset$precinct == "AR22" | dataset$precinct == "AR2"

| dataset$precinct == "AR18" | dataset$precinct == "41" | dataset$precinct == "16"

| dataset$precinct == "4" | dataset$precinct == "17" | dataset$precinct == "31"

| dataset$precinct == "48" | (dataset$precinct == "2" & dataset$state == 4)),]

table(Arlington$precinct)

16 17 2 31 4 41 48 AR18 AR2 AR22 AR49

32 35 76 22 47 88 84 26 19 32 14

PrinceWilliam <- dataset[which(dataset$precinct == "PW 101" | dataset$precinct == "PW 104" | dataset$precinct == "PW 401"

| dataset$precinct == "PW101" | dataset$precinct == "PW104" | dataset$precinct == "PW402"

| dataset$precinct == "PW406" | dataset$precinct == "401" | dataset$precinct == "402" |

(dataset$precinct == "104" & dataset$state == 4)),]

table(PrinceWilliam$precinct)

104 401 402 PW 101 PW 104 PW 401 PW101 PW104 PW402 PW406

28 23 14 15 2 9 25 10 37 25

1. Create dummy variables for each state/DC. How many observations are in DC, Maryland, Ohio and Virginia?

DC <- length(which(dataset$state == 1))

DC

768

MD <- length(which(dataset$state == 2))

MD

369

OH <- length(which(dataset$state == 3))

OH

547

VA <- length(which(dataset$state == 4))

VA

664

1. Convert the year\_born variable into age. Be sure to check for and correct for data errors. What is the average age of all observations in the data set? The minimum and maximum?

##minimum age is 17

##maximum age is 95

colnames(dataset)[80] <- "age"

dataset$age <- sapply(dataset$age, function(x) 2016-x)

##note that age that was greater than 120 was removed

dataset <- subset(dataset, dataset$age < 120)

mean(dataset$age)

##mean is 43.069

1. What is the distribution of the gender variable? Create a male dummy variable and indicate the distribution of this variable. Compare distribution of your male variable to the distribution of the gender variable.

table(dataset$gender)

1 2 3

845 1001 4

dataset[which(dataset$gender == 1),'male'] <- 1

dataset[which(dataset$gender == 2),'male'] <- 0

dataset[which(dataset$gender == 3),'male'] <- 0

table(dataset$male)

0 1

1005 845

For the male variable, all males were assigned a value of 1,and everything else was assigned a value of 0

1. Provide descriptive stats for Trump and Clinton feeling thermometer. Is there anything you need to adjust?

##clinton's therm max is 200 --> remove data error

dataset$therm\_clinton[which(dataset$therm\_clinton > 100)] <- NA

summary(dataset$therm\_clinton)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0.00 20.00 70.00 56.48 85.00 100.00 116

summary(dataset$therm\_trump)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0.00 0.00 0.00 18.28 25.00 100.00 146

1. What is the distribution of the education variable? Is there any adjustment you would need to make if you will use this as a continuous variable in a regression model?

table(dataset$education)

1 2 3 4 5 6 7

15 117 234 9 122 639 706

##you would have to somehow turn the discrete bins into the number of years person received education

##the problem is that the categories are discrete and not continuous

library(haven)

##dont use foreign library

##loading data

dataset <- read\_dta("post\_open\_refine\_clean.dta")

##Number 1

Arlington <- dataset[which(dataset$precinct == "AR49" | dataset$precinct == "AR22" | dataset$precinct == "AR2"

| dataset$precinct == "AR18" | dataset$precinct == "41" | dataset$precinct == "16"

| dataset$precinct == "4" | dataset$precinct == "17" | dataset$precinct == "31"

| dataset$precinct == "48" | (dataset$precinct == "2" & dataset$state == 4)),]

table(Arlington$precinct)

PrinceWilliam <- dataset[which(dataset$precinct == "PW 101" | dataset$precinct == "PW 104" | dataset$precinct == "PW 401"

| dataset$precinct == "PW101" | dataset$precinct == "PW104" | dataset$precinct == "PW402"

| dataset$precinct == "PW406" | dataset$precinct == "401" | dataset$precinct == "402" |

(dataset$precinct == "104" & dataset$state == 4)),]

table(PrinceWilliam$precinct)

##Number 2

DC <- length(which(dataset$state == 1))

DC

MD <- length(which(dataset$state == 2))

MD

OH <- length(which(dataset$state == 3))

OH

VA <- length(which(dataset$state == 4))

VA

##Number 3

##minimum age is 17

##maximum age is 95

colnames(dataset)[80] <- "age"

dataset$age <- sapply(dataset$age, function(x) 2016-x)

##note that age that was greater than 120 was removed

dataset <- subset(dataset, dataset$age < 120)

mean(dataset$age)

##mean is 43.069

##number 4

table(dataset$gender)

dataset[which(dataset$gender == 1),'male'] <- 1

dataset[which(dataset$gender == 2),'male'] <- 0

dataset[which(dataset$gender == 3),'male'] <- 0

table(dataset$male)

##number 5

##clinton's therm max is 200 --> remove data error

dataset$therm\_clinton[which(dataset$therm\_clinton > 100)] <- NA

summary(dataset$therm\_clinton)

summary(dataset$therm\_trump)

##number 6

table(dataset$education)

##you would have to somehow turn the discrete bins into the number of years person received education

##the problem is that the categories are discrete and not continuous