PS - Oct. Solutions

JZ

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

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

```
library(haven)
exit <- read_dta("https://github.com/zilinskyjan/R-stata-tutorials/blob/master/homework/31116399_Florida
```

R Answers

Part A: Descriptives

1. How many respondents were 65 years old or older? What was their proportion in the sample?

One option:

```
exit %>% count(AGE65)
```

```
## # A tibble: 7 x 2
##        AGE65     n
##    <dbl+lbl> <int>
## 1  1 [18-24]   261
## 2  2 [25-29]   232
## 3  3 [30-39]   431
## 4  4 [40-49]   462
## 5  5 [50-64]   927
## 6  6 [65+]     816
## 7 NA            18
```

```
# Another option:
# exit %>% count(AGE8)
```

Add a column with proportions:

```
exit %>%
  count(AGE65) %>%
  mutate(proportion = n / sum(n))
```

```
## # A tibble: 7 x 3
##        AGE65     n proportion
```

```
##    <dbl+lbl> <int>      <dbl>
## 1  1 [18-24]   261    0.0829
## 2  2 [25-29]   232    0.0737
## 3  3 [30-39]   431    0.137
## 4  4 [40-49]   462    0.147
## 5  5 [50-64]   927    0.295
## 6  6 [65+]     816    0.259
## 7 NA            18    0.00572
```

Another possibility:

```
table(exit$AGE65)
```

```
##
##   1   2   3   4   5   6
## 261 232 431 462 927 816
```

```
table(exit$AGE65) / sum(table(exit$AGE65))
```

```
##
##          1          2          3          4          5          6
## 0.08341323 0.07414509 0.13774369 0.14765101 0.29626079 0.26078619
```

Even better to run:

```
exit %>%
  count(AGE65) %>%
  filter(!is.na(AGE65)) %>%
  mutate(proportion = n / sum(n))
```

```
## # A tibble: 6 x 3
##       AGE65     n proportion
##   <dbl+lbl> <int>      <dbl>
## 1 1 [18-24]   261     0.0834
## 2 2 [25-29]   232     0.0741
## 3 3 [30-39]   431     0.138
## 4 4 [40-49]   462     0.148
## 5 5 [50-64]   927     0.296
## 6 6 [65+]     816     0.261
```

2. What was the proportion of Hispanic respondents who:

- Lived in cities with pop. over 50,000?
- Lived in suburbs?
- Lived in small cities or rural areas?

(Hint: Look at the variation in the variable labeled `SIZEPLC3`.)

```
exit %>% filter(latino==1) %>%
  group_by(SIZEPLC3) %>%
  summarize(n = n()) %>%
  mutate(proportion = n / sum(n))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 3
##                 SIZEPLC3     n proportion
##                <dbl+lbl> <int>      <dbl>
## 1 1 [Cities over 50,000]   256    0.494
## 2 2 [Suburbs]              257    0.496
```

```
## 3 3 [Small Cities/Rural]     5    0.00965
```

Alternative code:

```
exit %>% filter(latino==1) %>%
  group_by(SIZEPLC3) %>%
  tally() %>%
  mutate(proportion = n / sum(n))
```

```
## # A tibble: 3 x 3
##                 SIZEPLC3     n proportion
##                <dbl+lbl> <int>      <dbl>
## 1 1 [Cities over 50,000]   256    0.494
## 2 2 [Suburbs]              257    0.496
## 3 3 [Small Cities/Rural]     5    0.00965
```

3. Prepare a table displaying the proportion of voters who said they were first-time (midterm election)

```r
exit %>% group_by(FTVOTER1) %>% tally()
```

```
## # A tibble: 3 x 2
##    FTVOTER1     n
##   <dbl+lbl> <int>
## 1   1 [Yes]   209
## 2   2 [No]    781
## 3  NA        2157
```

```
exit %>% group_by(FTVOTER1,sex) %>%
    filter(!is.na(FTVOTER1),!is.na(sex)) %>%
    tally() %>%
    mutate(prop = n / sum(n))
```

```
## # A tibble: 4 x 4
## # Groups:   FTVOTER1 [2]
##    FTVOTER1        sex     n  prop
##   <dbl+lbl>  <dbl+lbl> <int> <dbl>
## 1   1 [Yes] 1 [Male]      78 0.373
## 2   1 [Yes] 2 [Female]   131 0.627
## 3   2 [No]  1 [Male]     349 0.448
## 4   2 [No]  2 [Female]   430 0.552
```

Why is it important to remove the missing observations from the denominator?

- Most student correctly say that 78 of male voters were first-time voters.
- But we cannot divide 78 by the total number of male respondents (1398).
- Rather, we must divide 78 by 427 (i.e. the number of male respondents for who we know their FT-voting status).

```
exit %>% count(sex)
```

```
## # A tibble: 3 x 2
##           sex     n
##     <dbl+lbl> <int>
## 1  1 [Male]    1398
## 2  2 [Female]  1741
## 3 NA              8
```

```
exit %>% filter(!is.na(FTVOTER1)) %>% count(sex)
```

```
## # A tibble: 3 x 2
##           sex     n
##     <dbl+lbl> <int>
## 1  1 [Male]     427
## 2  2 [Female]   561
## 3 NA              2
```

4. What was the proportion of Democrats who said in 2018 that Donald Trump should not be impeached and removed from office?

```
exit %>% count(party,IMPEACH1)
```

```
## # A tibble: 16 x 3
##                              party  IMPEACH1     n
##                          <dbl+lbl> <dbl+lbl> <int>
##  1  1 [Democrat]                    1 [Yes]    256
##  2  1 [Democrat]                    2 [No]      56
##  3  1 [Democrat]                    9 [Omit]    22
##  4  1 [Democrat]                   NA          712
##  5  2 [Republican]                  1 [Yes]     20
##  6  2 [Republican]                  2 [No]     300
##  7  2 [Republican]                  9 [Omit]     2
##  8  2 [Republican]                 NA          705
##  9  3 [Independent/Something else]  1 [Yes]    114
## 10  3 [Independent/Something else]  2 [No]     169
## 11  3 [Independent/Something else]  9 [Omit]    31
## 12  3 [Independent/Something else] NA          637
## 13 NA                              1 [Yes]      6
## 14 NA                              2 [No]       5
## 15 NA                              9 [Omit]     1
## 16 NA                             NA          111
```

Let's limit our attention to Democrats:

```
exit %>% count(party,IMPEACH1) %>% filter(party==1)
```

```
## # A tibble: 4 x 3
##          party  IMPEACH1     n
##      <dbl+lbl> <dbl+lbl> <int>
## 1 1 [Democrat]  1 [Yes]    256
## 2 1 [Democrat]  2 [No]      56
## 3 1 [Democrat]  9 [Omit]    22
## 4 1 [Democrat] NA          712
```

Part B

1. Is there an association between 2016 vote choice and the type of area where a voter lives?

```
exit %>% count(SIZEPLC3,VOTE2016)
```

```
## # A tibble: 17 x 3
##                  SIZEPLC3             VOTE2016     n
##                <dbl+lbl>            <dbl+lbl> <int>
##  1 1 [Cities over 50,000]  1 [Hillary Clinton]   170
##  2 1 [Cities over 50,000]  2 [Donald Trump]      115
```

```
##  3 1 [Cities over 50,000]  3 [Other]              20
##  4 1 [Cities over 50,000]  4 [Didn't vote]       32
##  5 1 [Cities over 50,000]  9 [Omit]               3
##  6 1 [Cities over 50,000] NA                    794
##  7 2 [Suburbs]             1 [Hillary Clinton]  222
##  8 2 [Suburbs]             2 [Donald Trump]     271
##  9 2 [Suburbs]             3 [Other]             34
## 10 2 [Suburbs]             4 [Didn't vote]       40
## 11 2 [Suburbs]             9 [Omit]              12
## 12 2 [Suburbs]            NA                   1266
## 13 3 [Small Cities/Rural]  1 [Hillary Clinton]   19
## 14 3 [Small Cities/Rural]  2 [Donald Trump]      30
## 15 3 [Small Cities/Rural]  3 [Other]              4
## 16 3 [Small Cities/Rural]  4 [Didn't vote]        4
## 17 3 [Small Cities/Rural] NA                    111
```

Run a regression:

```
exit$votedForCliton2016 <- ifelse(exit$VOTE2016==1,1,0)

model_vote <- lm(votedForCliton2016 ~ factor(SIZEPLC3), data= exit)
```

2. What percentage of voters said that when choosing their candidate for the US Senate, "Donald Trump was not a factor"?

```
exit %>% count(fortrump)
```

```
## # A tibble: 5 x 2
##                                     fortrump     n
##                                    <dbl+lbl> <int>
## 1  1 [To express support for Donald Trump]     189
## 2  2 [To express opposition to Donald Trump]   249
## 3  3 [Donald Trump was not a factor]           334
## 4  9 [Omit]                                     28
## 5 NA                                          2347
```

3. What percentage voters said that, in their vote for the US Senate, Bill Nelson's vote against Brett Kavanaugh's confirmation "not a factor at all?"

```
exit %>% count(KAVFL18)
```

```
## # A tibble: 6 x 2
##                          KAVFL18     n
##                         <dbl+lbl> <int>
## 1  1 [The most important factor]    64
## 2  2 [An important factor]         226
## 3  3 [A minor factor]              120
## 4  4 [Not a factor at all]         265
## 5  9 [Omit]                         72
## 6 NA                              2400
```

Part C:

```
exit$motivated <- ifelse(exit$RUSSIA18==2,1,0)
```

```
exit <- exit %>% mutate(politcally_motivated = ifelse(RUSSIA18==2,1,0))
```

Note: Republicans are `party==2`:

```
mod <- lm(motivated ~ factor(party) + factor(AGE8),
              data = exit %>% filter(!is.na(RUSSIA18)))
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = motivated ~ factor(party) + factor(AGE8), data = exit %>%
##     filter(!is.na(RUSSIA18)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8587 -0.3345  0.1773  0.4264  0.7663
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.33454    0.05549   6.029 2.37e-09 ***
## factor(party)2  0.50565    0.03583  14.113  < 2e-16 ***
## factor(party)3  0.22062    0.03676   6.002 2.79e-09 ***
## factor(AGE8)2  -0.03962    0.07376  -0.537    0.591
## factor(AGE8)3  -0.10083    0.06600  -1.528    0.127
## factor(AGE8)4  -0.08811    0.07835  -1.125    0.261
## factor(AGE8)5   0.01846    0.07391   0.250    0.803
## factor(AGE8)6  -0.05155    0.06050  -0.852    0.394
## factor(AGE8)7  -0.01745    0.06984  -0.250    0.803
## factor(AGE8)8  -0.03441    0.05969  -0.576    0.564
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4514 on 935 degrees of freedom
##   (23 observations deleted due to missingness)
## Multiple R-squared:  0.1879, Adjusted R-squared:  0.1801
## F-statistic: 24.04 on 9 and 935 DF,  p-value: < 2.2e-16
```

What are the key things to notice here?

- Higher values of `party` do not mean that a respondent is "more Republican".
- The values are not ordered in a meaningful way (1 = Democrat; 2 = Republican; 3 = "Independent).
- Even if the values were ordered in an ideological manner, it still doesn't mean that treating the variable as continuous is necessarily a defensible choice.
- The values are unordered, so you *must* include a series of indicator/dummy variables, if you wish to control for partisanship.
- One way to achieve that is to add `factor(party)` to your regression forumula.

STATA Answers

```
tab AGE8 or tab AGE65
```

```
gen hispanic = (latino==1) if !mi(latino)
```

```
tabstat hispanic, by(SIZEPLC3)
```

```
tab FTVOTER1 sex
```

```
tab party IMPEACH1, row nof
```

Part B:

B1: `tab SIZEPLC3 VOTE2016 [aw=weight], row nof`

B2: `tab fortrump`

B3: `tab KAVFL18`

Part C:

```
gen motivated = (RUSSIA18==2) if !mi(RUSSIA18)
```

```
reg motivated i.AGE8 ib3.party
```

It is important to note that:

- Higher values of party do not mean that a respondent is "more Republican".
- The values are of the variable not ordered in a meaningful way (1 = Democrat; 2 = Republican; 3 = "Independent). And even if the values were ordered ideologically, it still doesn't mean that treating the variable as continuous is necessarily a defensible choice.
- Given that the values are unordered, then you *must* include a series of indicator/dummy variables.
- One way to achieve that is to use the "i." operator.