Assignment 1

YOUR NAME GOES HERE

Due 2/10

Overview

You will load, clean, manipulate, and explore data from Koch and Nicholson's article (2016), "Death and Turnout". This is part 1 of 2 toward a complete replication of Table 4 from the article.

There are four relevant datasets. All datasets are at Deming's GitHub page: HERE:

- `bes05_short.dta`
- `bes10_short.dta`
- `ukregion_cas.tab`
- `0501districtdata.tab`

Throughout, you should use `dplyr` functions and syntax whenever possible.

Get started

1. Load the following packages in the `setup` chunk. (You may need to install some of them first):

- `tidyverse` (contains `dplyr` and `ggplot2`)
- `readr` (for importing `.tab` formatted data)
- `haven` (for importing `.dta` formatted data)
- `here` (recommended but not required. You might read about how `here()` works.)

2. Import the four datasets above in the `setup` chunk.

3. Take some time to get to know the four datasets before moving on. Examine their dimensions, variable names, etc. You may also wish to examine the `koch-nicholson_codebook` at GitHub, which contains variables descriptions for `bes05` and `bes10`.

Clean and Manipulate

2005 BES data

4. Examine the variables in `bes05`. They are not very informative. The `koch-nicholson_codebook` at GitHub contains more informative names. Rename the variables according to codebook.

```
# Rename variables
bes05 <- bes05 %>%
  rename(region = pre_q1,
         labor_iraq = pre_q13,
         conserve_iraq = pre_q23,
         partyid = pre_q29,
         party_strength = pre_q33,
         likelyvote = pre_q34,
         blair_competent = pre_q50,
         executive_approval = pre_q68,
         gov_party_approve = pre_q84,
         perception_economy = pre_q92,
         attention = pre_q141,
         birthyr = pre_q148,
         education = pre_q156,
         income = pre_q163,
         race = pre_q174,
         gender = pre_q180,
         marital_status = pre_q158,
         british_iraq = pre_q128,
         weights = pre_w8)
```

5. Some of the variables' current values actually denote missing data: "no response", "don't know", etc. Examine the codebook for the variables below. Recode them so that missing values are denoted as NA:

- `party_strength`
- `labor_iraq`
- `perception_economy`
- `likelyvote`

```
# Recode missing values
bes05 <- bes05 %>%
  mutate(party_strength = if_else(party_strength == 4, NA_real_, party_strength),
         labor_iraq = if_else(labor_iraq == 6, NA_real_, labor_iraq),
         perception_economy = ifelse(perception_economy == 6, NA_real_, perception_economy),
         likelyvote = ifelse(likelyvote == 12, NA_real_, likelyvote))
```

6. Rename `executive_approval` to `pmtherm`. Also rename `labor_iraq` to `pmwar`.

```
# Rename EXECUTIVE_APPROVAL and LABOR_IRAQ
bes05 <- bes05 %>%
  rename(pmtherm = executive_approval,
         pmwar = labor_iraq)
```

7. Create two new variables. Add them to `bes05`. Here are the variable definitions:

- `year`: equals 2005
- `age`: equals 2005 minus individuals' birth year

```
# Add AGE and YEAR variables
bes05 <- bes05 %>%
  mutate(year = 2005,
         age = 2005 - birthyr)
```

2010 BES Data

8. Examine the variables in `bes10`. They are not very informative. The `koch-nicholson_codebook` at GitHub contains more informative names. Rename the variables according to codebook.

```r
# Rename variables
bes10 <- bes10 %>%
  rename(region = aaq1,
         labor_afghan = aaq13,
         conserve_afghan = aaq22,
         partyid = aaq28,
         party_strength = aaq32,
         likelyvote = aaq33,
         brown_competent = aaq81,
         executive_approval = aaq52,
         gov_party_approve = aaq63,
         perception_economy = aaq87,
         attention = aaq131,
         birthyr = aaq151,
         education = aaq159,
         income = aaq166,
         race = aaq177,
         gender = aaq186,
         marital_status = aaq161,
         british_afghan = aaq116,
         weights = w8_f)
```

9. Some of the variables' current values actually denote missing data: "no response", "don't know", etc. Examine the codebook for the variables below. Recode them so that missing values are denoted as NA:

- `party_strength`
- `labor_afghan`
- `perception_economy`
- `likelyvote`
- `income`

```r
# Recode missing values
bes10 <- bes10 %>%
  mutate(party_strength = if_else(party_strength == 4, NA_real_, party_strength),
         labor_afghan = ifelse(labor_afghan == 6, NA_real_, labor_afghan),
         perception_economy = ifelse(perception_economy == 6, NA_real_, perception_economy),
         likelyvote = ifelse(likelyvote == 12, NA_real_, likelyvote),
         income = ifelse(income == 17, NA_real_, income))
```

10. Rename `executive_approval` to `pmtherm`. Also rename `labor_iraq` to `pmwar`.

```r
# Rename EXECUTIVE_APPROVAL and LABOR_AFGHAN
bes10 <- bes10 %>%
  rename(pmtherm = executive_approval,
         pmwar = labor_afghan)
```

11. Create two new variables. Add them to `bes10`. Here are the variable definitions:

- year: equals 2010
- age: equals individuals' birth year

```
# Add AGE and YEAR variables
bes10 <- bes10 %>%
  mutate(year = 2010,
         age = birthyr)
```

Append and Merge

12. Append bes10 to bes05. On doing so, you may wish to save the appended dataframe to your computer as a means of backing up your work.

```
# Append
bes0510 <- bind_rows(bes05, bes10)
```

```
## Warning: `..1$partyid` and `..2$partyid` have conflicting value labels.
## i Labels for these values will be taken from `..1$partyid`.
## x Values: 10

## Warning: `..1$pmtherm` and `..2$pmtherm` have conflicting value labels.
## i Labels for these values will be taken from `..1$pmtherm`.
## x Values: 0, 10, and 12

## Warning: `..1$gov_party_approve` and `..2$gov_party_approve` have conflicting value
## labels.
## i Labels for these values will be taken from `..1$gov_party_approve`.
## x Values: 0, 10, and 12

## Warning: `..1$attention` and `..2$attention` have conflicting value labels.
## i Labels for these values will be taken from `..1$attention`.
## x Values: 0, 10, and 12

## Warning: `..1$birthyr` and `..2$birthyr` have conflicting value labels.
## i Labels for these values will be taken from `..1$birthyr`.
## x Values: 89

## Warning: `..1$education` and `..2$education` have conflicting value labels.
## i Labels for these values will be taken from `..1$education`.
## x Values: 10, 12, and 99
```

13. Merge the appended dataframe above and the data on casualties by UK region (ukregion_cas). The aim is to produce a dataframe that matches each individual in the BES data to the number of casualties in their UK region for 2005 and 2010. (Hint: You should merge using two "key" variables.)

```
# Merge
bes0510casmerge <- bes0510 %>%
  left_join(casualties, by = c("region","year"))
```

14. Merge the merged dataframe above and the data on UK district democgraphics (districtdata). The aim is to produce a dataframe that matches each individual to their district demographics for 2005 and 2010. (Hint: You should merge using two "key" variables.)

```r
# Merge
bes_final_data <- bes0510casmerge %>%
  left_join(districts, by = c("region","year"))
```

## Manipulate (Again)

15. Create the following five variables. Add them to the dataframe that you created above. Here are the variable definitions:

- `white`: dummy variable that equals 1 if an individual is white and 0 otherwise.
- `female`: dummy variable that equals 1 if an individual is female and otherwise.
- `low_attention`: dummy variable that equals 1 if an individuals' political attention is less than 4 and 0 otherwise.
- `married`: dummy variable that equals 1 if an individual is married and 0 otherwise.
- `partstrength`: dummy variable that equals 1 if and individuals "very strongly" identifies with a political party and 0 otherwise.

```r
# Create WHITE, FEMALE, LOW_ATTENTION, MARRIAGE, and PARTSTRENGTH variables.
bes_final_data <- bes_final_data %>%
  mutate(white = ifelse(race == 1, 1, 0),
         female = ifelse(gender == 2, 1, 0),
         low_attention = ifelse(attention < 4, 1, 0),
         married = ifelse(marital_status == 1, 1, 0),
         partstrength = ifelse(party_strength == 1, 1, 0))
```

Explore

16. Generate summary statistics for the following variables. For each, generate the minimum, maximum, median, mean, standard deviation, and number of observations. See if you can use `dplyr`'s summary functionality to create new dataframe of summary statistics.

- `likelyvote`
- `region_cas`
- `low_attention`
- `female`
- `married`
- `income`
- `education`
- `age`
- `white`
- `partstrength`
- `perception_economy`
- `pmtherm`
- `pmwar`
- `unemploy_rate`
- `income_pc`
- `pct_white`

```r
my_vars <- c("likelyvote", "region_cas", "low_attention", "female", "married", "income",
             "education", "age", "white", "partstrength", "perception_economy", "pmtherm",
             "pmwar", "unemploy_rate", "income_pc", "pct_white")
```

```r
summary_df <- bes_final_data %>%
  mutate(across(everything(), as.numeric)) %>%
  summarise(across(everything(), list(
    min = ~min(., na.rm = TRUE),
    max = ~max(., na.rm = TRUE),
    mean = ~mean(., na.rm = TRUE),
    median = ~median(., na.rm = TRUE),
    sd = ~sd(., na.rm = TRUE),
    n = ~sum(!is.na(.))
  ), .names = "{.col}__{.fn}"))
```

```
## Warning: There was 1 warning in ‘mutate()‘.
## i In argument: ‘across(everything(), as.numeric)‘.
## Caused by warning:
## ! NAs introduced by coercion
```

```
## Warning: There were 2 warnings in ‘summarise()‘.
## The first warning was:
## i In argument: ‘across(...)‘.
## Caused by warning in ‘min()‘:
## ! no non-missing arguments to min; returning Inf
## i Run ‘dplyr::last_dplyr_warnings()‘ to see the 1 remaining warning.
```

```r
summary_df <- summary_df %>%
  pivot_longer(everything(), names_to = c("Variable", "Statistic"), names_sep = "__") %>%
  pivot_wider(names_from = "Statistic", values_from = "value")


summary_df %>%
  knitr::kable(
    format = "latex",
    align = "l",
    booktabs = TRUE,
    longtable = TRUE,
    linesep = "",
    digits = 2,
    ) %>%
  kableExtra::kable_styling(
      position = "left",
      latex_options = c("striped", "repeat_header"),
      stripe_color = "gray!15"
    )
```

Variable	min	max	mean	median	sd	n
besid	1.00	7813.00	3907.91	3908.00	2255.98	15586
region	1.00	11.00	6.19	7.00	3.00	11195
pmwar	1.00	5.00	3.81	4.00	1.21	10965
conserve_iraq	1.00	6.00	3.80	4.00	1.40	7793
partyid	1.00	11.00	4.14	2.00	3.71	7291
party_strength	1.00	3.00	2.23	2.00	0.72	6424
likelyvote	0.00	10.00	8.43	10.00	2.87	11046

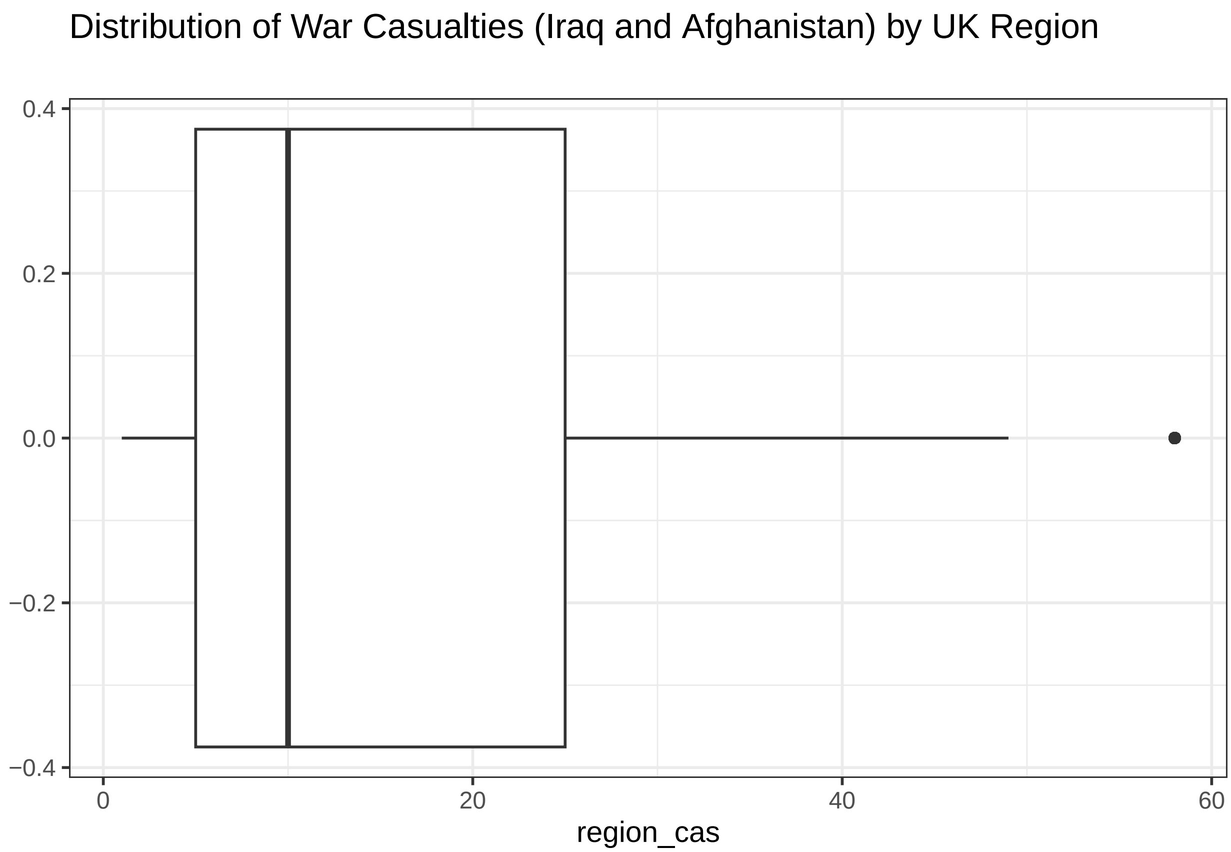
(continued)

Variable	min	max	mean	median	sd	n
blair_competent	0.00	999.00	495.62	12.00	496.81	7793
pmtherm	0.00	10.00	3.81	4.00	3.21	10871
gov_party_approve	0.00	10.00	3.52	3.00	2.95	10842
perception_economy	1.00	5.00	2.58	3.00	1.03	10590
attention	0.00	10.00	6.20	7.00	2.52	11070
birthyr	1.00	89.00	59.38	59.00	14.58	11187
education	1.00	18.00	7.65	7.00	5.42	9348
income	1.00	16.00	6.80	6.00	3.80	10682
race	1.00	5.00	1.10	1.00	0.53	11195
gender	1.00	2.00	1.49	1.00	0.50	11195
marital_status	1.00	6.00	2.47	1.00	1.99	11195
british_iraq	1.00	5.00	3.14	3.00	1.04	7793
weights	0.06	4.40	0.99	0.93	0.41	11195
year	2005.00	2010.00	2007.50	2007.50	2.50	15586
age	1.00	2004.00	1371.64	1933.00	868.04	11187
conserve_afghan	1.00	5.00	3.18	3.00	1.10	2889
brown_competent	0.00	10.00	3.73	3.00	3.17	3301
british_afghan	1.00	4.00	2.95	3.00	0.92	3059
region_cas	1.00	58.00	17.70	10.00	17.15	11195
area	Inf	-Inf	NaN	NA	NA	0
population	2515479.00	8634750.00	5791904.80	5295000.00	1642459.51	11195
income_pc	11332.00	19465.85	13871.20	13376.00	1828.89	11195
unemploy_rate	3.20	9.80	5.47	5.00	1.80	11195
pctwhite	0.45	0.98	0.90	0.93	0.10	11195
white	0.00	1.00	0.96	1.00	0.19	11195
female	0.00	1.00	0.49	0.00	0.50	11195
low_attention	0.00	1.00	0.17	0.00	0.37	11070
married	0.00	1.00	0.56	1.00	0.50	11195
partstrength	0.00	1.00	0.17	0.00	0.37	6424

17. Generate seperate visulizations of the distributions of `region_cas`, `low_attention`, and `likelyvote`. Be sure to select visualizations that are appropriate for variables' type. Use `ggplot2` syntax.

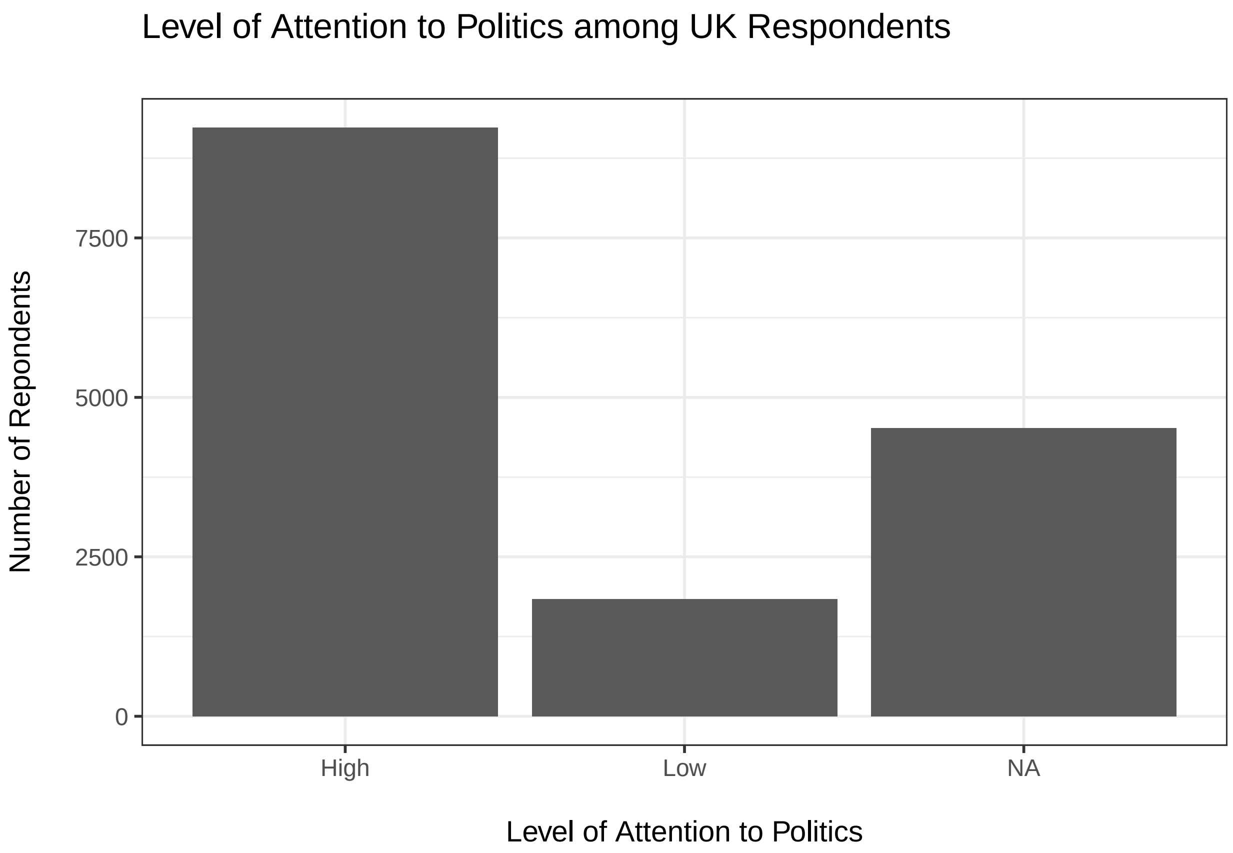
```r
p1 <- ggplot(bes_final_data, aes(region_cas)) +
  geom_boxplot() +
  theme_bw() +
  labs(title = "Distribution of War Casualties (Iraq and Afghanistan) by UK Region\n")
p1
```

```
## Warning: Removed 4391 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```


Distribution of War Casualties (Iraq and Afghanistan) by UK Region

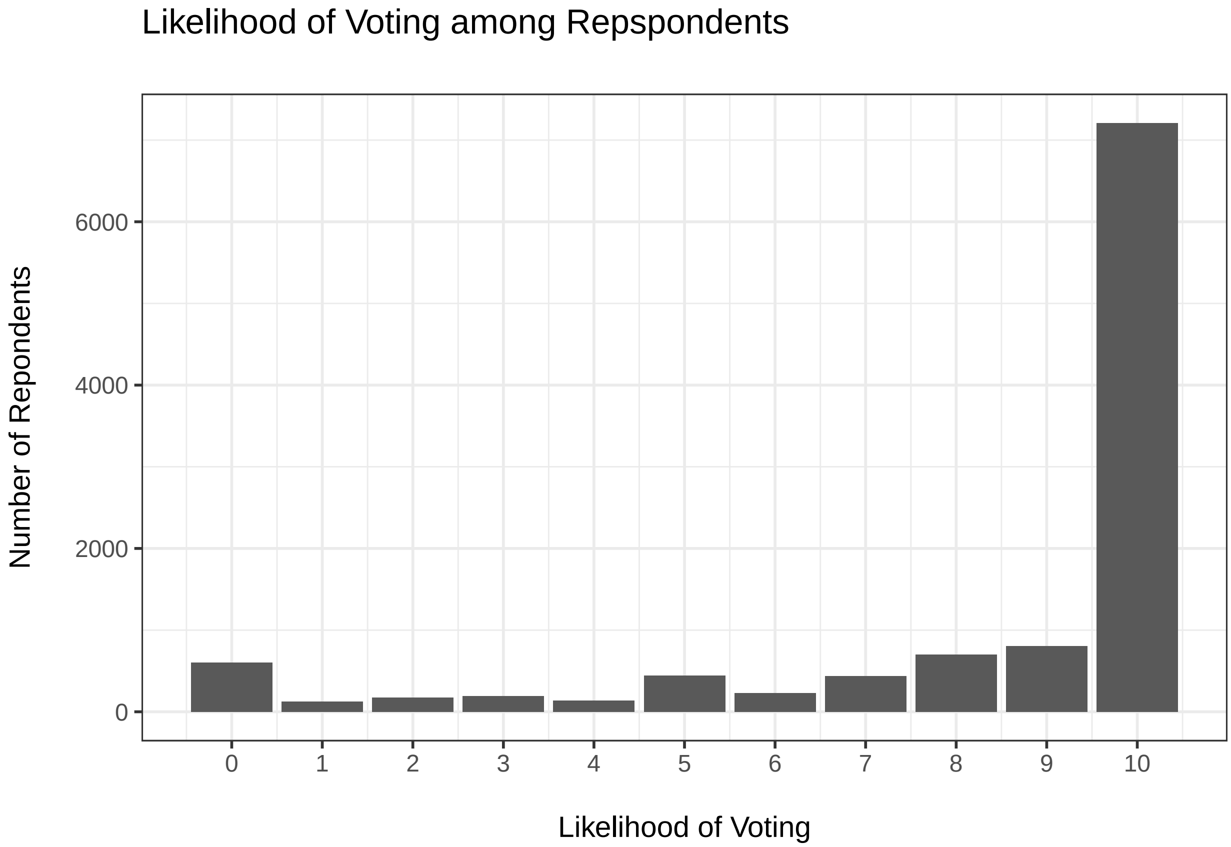
```r
bes_final_data <- bes_final_data %>%
  mutate(low_attention = as.factor(low_attention))

p2 <- ggplot(bes_final_data, aes(as.factor(low_attention))) +
  geom_bar() +
  scale_x_discrete(labels = c("High", "Low")) +
  theme_bw() +
  labs(title = "Level of Attention to Politics among UK Respondents\n",
       x = "\nLevel of Attention to Politics",
       y = "Number of Repondents\n")
p2
```



```
p3 <- ggplot(bes_final_data, aes(as.numeric(likelyvote))) +
  geom_bar() +
  scale_x_continuous(breaks = seq(0,10,1)) +
  theme_bw() +
  labs(title = "Likelihood of Voting among Repspondents\n",
       x = "\nLikelihood of Voting",
       y = "Number of Repondents\n")
p3
```

```
## Warning: Removed 4540 rows containing non-finite outside the scale range
## (‘stat_count()‘).
```


Likelihood of Voting among Repspondents
Number of Repondents
6000
4000
2000
0
0
1
2
3
4
5
6
7
8
9
10
Likelihood of Voting