Homework 1

Analysis of Behavioral Risk Factor Surveillance System Data

Your Name Here

Background

The Behavioral Risk Factor Surveillance System (BRFSS) is a long-term effort administered by the CDC to collect data on behaviors affecting physical and mental health, past and present health conditions, and access to healthcare among U.S. residents. The BRFSS comprises telephone surveys of U.S. residents conducted annually since 1984; in the last decade, over half a million interviews have been conducted each year. This is the largest such data collection effort in the world, and many countries have developed similar programs. The objective of the program is to support monitoring and analysis of factors influencing public health in the United States.

Each year, a standard survey questionnaire is developed that includes a core component comprising questions about: demographic and household information; health-related perceptions, conditions, and behaviors; substance use; and diet. Trained interviewers in each state call randomly selected telephone (landline and cell) numbers and administer the questionnaire; the phone numbers are chosen so as to obtain a representative sample of all households with telephone numbers. Take a moment to read about the 2019 survey here.

In this assignment you'll import and subsample the BRFSS 2019 data and perform a simple descriptive analysis exploring associations between adverse childhood experiences, health perceptions, tobacco use, and depressive disorders. This is an opportunity to practice:

- review of data documentation
- data assessment and critical thinking about data collection
- dataframe transformations in R
- communicating and interpreting grouped summaries

IMPORTANT: we've set `eval=FALSE` in each of the R code chunks below so that we can render the homework without filling in valid R solutions. As you fill in answers and progress through the assignment, make sure to set `eval=TRUE` to the code chunks you have completed.

Data import and assessment

The cell below imports select columns from the 2019 dataset as an R `tibble`. The file Frame. The file is big, so this may take an the cell and then have a quick look at the first few rows and columns.

```
# store variable names of interest
selected_vars <- c("_SEX", "_AGEG5YR",
                   "GENHLTH", "ACEPRISN",
                   "ACEDRUGS", "ACEDRINK",
                   "ACEDEPRS", "ADDEPEV3",
                   "_SMOKER3", "_LLCPWT")

# import full 2019 BRFSS dataset
brfss <- read_csv("data/brfss2019.zip",
                  col_select = selected_vars)

# invert sampling weights
brfss$`_LLCPWT` = 1/brfss$`_LLCPWT`

# print first few rows
head(brfss)
```

Question 1: Row and column information

Now that you've imported the data, you should verify that the dimensions conform to the format you expect based on data documentation and ensure you understand what each row and each column represents.

Check the number of records (interviews conducted) reported and variables measured for 2019 by reviewing the surveillance summaries by year, and then answer the following questions in a few sentences:

- Does the number of rows match the number of reported records?
- How many columns were imported, and how many columns are reported in the full dataset?
- What does each row in the `brfss` dataframe represent?
- What does each column in the `brfss` dataframe represent

Type your answer here, replacing this text.

Question 2: Sampling design and data collection

Skim the overview documentation for the 2019 BRFSS data. Focus specifically the 'Background' and 'Data Collection' sections, read selectively for relevant details, and answer the following questions in a few sentences:

i. Who conducts the interviews and how long does a typical interview last?

ii. Who does an interviewer speak to in each household?

iii. What criteria must a person meet to be interviewed?

iv. Who *can't* appear in the survey? Give two examples.

v. What is the study population (*i.e.*, all individuals who could possibly be sampled)?

vi. Does the data contain any identifying information?

Type your answer here, replacing this text.

Question 3: Variable descriptions

You'll work with the small subset variables imported above: sex, age, general health self-assessment, smoking status, depressive disorder, and adverse childhood experiences (ACEs). The names of these variables as they appear in the raw dataset are defined in the cell in which you imported the data as `selected_vars`. It is often useful, and therefore good practice, to include a brief description of each variable at the outset of any reported analyses, both for your own clarity and for that of any potential readers. Open the 2019 BRFSS codebook in your browser and use text searching to locate each of the variable names of interest. Read the codebook entries and fill in the second column in the table below with a one-sentence description of each variable identified in `selected_vars`. Rephrase the descriptions in your own words – do not copy the codebook descriptions verbatim.

Variable name	Description
GENHLTH	
_SEX	
_AGEG5YR	
ACEPRISN	
ACEDRUGS	
ACEDRINK	
ACEDEPRS	
ADDEPEV3	
_SMOKER3	

Question 4: Fixing variable names

Notice that some variable names start with underscores and others are hard to identify given the acronyms. When variables start with underscores, numbers or include spaces, you must wrap the variable in backticks when referring to that variable in selecting function, e.g. `brfss |> select(`_AGEG5YR`)`. This can get a bit annoying and can lead to unexpected errors. Use the `rename` function to rename any variables containing an underscore to the same variable name but with the underscore removed. We've gone ahead and renamed `_LLCPWT` to `sampling_weights` which is a much more descriptive name for the variable.

```
## rename other variables below or modify code above
brfss <- brfss |>
  rename(sampling_weights = `_LLCPWT`)
```

Subsampling

To simplify life a little, we'll draw a large random sample of the rows and work with that in place of the full dataset. This is known as **subsampling**.

The cell below draws a random subsample of 10k records. Because the subsample is randomly drawn, we should not expect it to vary in any systematic way from the overall dataset, and distinct subsamples should have similar properties – therefore, results downstream should be similar to an analysis of the full dataset, and should also be possible to replicate using distinct subsamples.

```
# for reproducibility
set.seed(32221)

# randomly sample 10k records
# Sample 10k records with weights
samp <- brfss %>%
  slice_sample(n = 10000, weight_by = sampling_weights)
```

Asides:

- Notice that the random number generator seed is set before carrying out this task – this ensures that every time the cell is run, the same subsample is drawn. As a result, the computations in this notebook are *reproducible*: when I run the notebook on my computer, I get the same results as you get when you run the notebook on your computer.

- Notice also that *sampling weights* provided with the dataset are used to draw a weighted sample. Some respondents are more likely to be selected than others from the general population of U.S. adults with phone numbers, so the BRFSS calculates derived weights that are inversely proportional to estimates of the probability that the respondent is included in the survey. This is a somewhat sophisticated calculation, however if you're interested, you can read about how these weights are calculated and why in the overview documentation you used to answer the questions above. We use the sampling weights in drawing the subsample so that we get a representative sample of U.S. adults with phone numbers.
- Notice the missing values. How many entries are missing in each column? The cell below computes the proportion of missing values for each of the selected variables. We'll return to this issue later on.

```
# proportions of missingness
colMeans(is.na(samp))
```

Tidying

In the following series of questions you'll tidy up the subsample by performing these steps:

- selecting columns of interest;
- replacing coded values of question responses with responses;
- defining new variables based on existing ones;

The goal of this is to produce a clean version of the dataset that is well-organized, intuitive to navigate, and ready for analysis.

The variable entries are coded numerically to represent certain responses but the variables are actually categorical. These should be replaced by more informative entries. We can use the codebook to determine which number means what, and replace the values with the appropriate factor accordingly.

The cell below replaces the numeric values for `age_group` by a *factor variable*, the datatype in R for categorical variables. Below we've replaced the numeric values with their meanings, illustrating how to create factors.

```
# dictionary representing variable coding
age_codes <- c(
  "1" = "18-24", "2" = "25-29", "3" = "30-34",
  "4" = "35-39", "5" = "40-44", "6" = "45-49",
  "7" = "50-54", "8" = "55-59", "9" = "60-64",
```

```
  "10" = "65-69", "11" = "70-74", "12" = "75-79",
  "13" = "80+", "14" = "Unsure/refused/missing"
)


# recode age categories
samp_mod1 = samp |>
  mutate(age_group = factor(age_group,
                            levels=names(age_codes),
                            labels=age_codes))

# check result
head(samp_mod1)
```

Question 5: Recoding variables

Following the example immediately above and referring to the 2019 BRFSS codebook, replace the numeric codings with response categories for each of the following variables:

- `_SEX`
- `GENHLTH`
- `_SMOKER3` Note above that your variable names will have changed based on your choice in the previous question. Above, the first modification (slicing) was stored as `samp_mod1`, and was a function of `samp`. You'll follow this pattern, creating `samp_mod2`, `samp_mod3`, and so on so that each step (modification) of your data manipulations is stored separately, for easy troubleshooting.

i. Recode the sex variable: define a new dataframe `samp_mod2` that is the same as `samp_mod1` but with the sex variable recoded as `M` and `F`.

ii. Recode `GENHLTH` variable: define a new dataframe `samp_mod3` that is the same as `samp_mod2` but with the `GENHLTH` variable recoded as `Excellent`, `Very good`, `Good`, `Fair`, `Poor`, `Unsure`, and `Refused`.

iii. Recode the `SMOKER3` variable: define a new dataframe `samp_mod4` that is the same as `samp_mod3` but with `SMOKER3` recoded as `Daily`, `Some days`, `Former`, `Never`, and `Unsure/refused/missing`.

iv. Print the first few rows of `samp_mod4`.

```
# define dictionary for sex
sex_codes = ...

# recode sex
samp_mod2 = ...

# define dictionary for health
health_codes = ...

# recode health
samp_mod3 = ...

# define dictionary for smoking
smoke_codes = ...

# recode smoking
samp_mod4 = ...

# print a few rows
...
```

Question 6: Value replacement

Now all the variables *except* the adverse childhood experience and depressive disorder question responses are represented in an interpretable way. In the codebook, note that the answer key is identical for all the remaining ACE variables. We can leverage this fact to create concise code to change the variable names.

i. First, we want to change variables, so we'll use the `mutate` function. Since we want to mutate multiple variables at once, we will combine this with the `across` function which specifies which variables we want to mutate across and how to mutate them. For example, in the `penguins` dataset, if we wanted to multiple all variables measuring lengths by 10 we could do

```
penguins |> mutate(across(contains("length"), \(x) x*10))
```

the lecture slides provide additional examples. What selecting function can you use to get the ACE variables?

ii. In this case, we of course don't want to multiple each variable by 10, but rather change the values the ACE variables. We'll do this using the `case_match` function. Read the

documentation for the `case_match` function and look at the examples they provide. Define a function called `recode_ace` which takes an ACE variable and recodes the variables according to the answer key `Yes`, `No`, `Unsure`, `Refused`.

Define a new dataframe `samp_mod5` that is the same as `samp_mod4` but with the ACE variables recoded, by using `mutate`, `across` and your `recode_ace` function.

```
...
```

Question 7: Define ACE indicator variable

Downstream analysis of ACEs will be facilitated by having an indicator variable that is a `1` if the respondent answered 'Yes' to any ACE question, and a `0` otherwise – that way, you can easily count the number of respondents reporting ACEs by summing up the indicator or compute the proportion by taking an average.

To this end, define a new logical variable, `adverse_conditions`: did the respondent answer yes to any of the adverse childhood condition questions?

Store the result as `samp_mod7`, and print the first few rows using `head()`.

```
...
```

Question 8: Define missingness indicator variable

As you saw earlier, there are some missing values for the ACE questions. These arise whenever a respondent is not asked these questions. In fact, answers are missing for nearly 80% of the respondents in our subsample. We should keep track of this information. Define a missing indicator, `adverse_missing`: is a response missing for at least one of the ACE questions? Store the result as `samp_mod8`

```
...
# check using head()
```

Question 9: Filter respondents who did not answer ACE questions

Since values are missing for the ACE question if a respondent was not asked, we can remove these observations and do any analysis *conditional on respondents having been asked the ACE questions*. Use your indicator variable `adverse_missing` to filter out respondents who were not asked the ACE questions.

Note that this dramatically limits the scope of inference for subsequent analyses to only those locations where the ACE module was included in the survey.

```
samp_mod9 <- ...
```

Question 10: Define depression indicator variable

It will prove similarly helpful to define an indicator for reported depression:

- `depression`: did the respondent report having been diagnosed with a depressive disorder?

Follow the same strategy as above for the ACE variables, and store the result as `samp_mod10`. See if you can perform the calculation of the new variable in a single line of code. Print the first few rows.

```
samp_mod10 <- ...

# define new variable
```

Question 11: Final dataset

For the final dataset, drop the respondent answers to individual questions, the missingness indicator, and select just the derived indicator variables along with general health, sex, age, and smoking status.

See if you can perform both operations (slicing and renaming) in a single chain. Store the result as `final_brfss_data`.

```
# slice and rename
final_brfss_data <- ...

# check using head()
```

Descriptive analysis

Now that you have a clean dataset, you'll use grouping and aggregation to compute several summary statistics that will help you explore whether there is an apparent association between experiencing adverse childhood conditions and self-reported health, smoking status, and depressive disorders in areas where the ACE module was administered.

The basic strategy will be to calculate the proportions of respondents who answered yes to one of the adverse experience questions when respondents are grouped by the other variables.

Question 12: Proportion of respondents reporting ACEs

Calculate the overall proportion of respondents in the subsample that reported experiencing at least one adverse condition (given that they answered the ACE questions). Store the result as `mean_ace` and print.

```
# proportion of respondents reporting at least one adverse condition
mean_ace <- ...

# print
print(mean_ace)
```

Question 13: Proportion of respondents reporting ACEs by health

Does the proportion of respondents who reported experiencing adverse childhood conditions vary by general health?

Compute the porportion of respondents reporiting adverse childhood conditions separately by general health self-rating. Notice that the depression variable is dropped so that the result doesn't also report the proportion of respondents reporting having been diagnosed with a depressive disorder. Notice also that the proportion of missing values for respondents indicating each general health rating is shown.

```
# proportions grouped by general health
# Group by general health
ace_health <- final_brfss_data %>%
  group_by(general_health) %>%
  summarise(adverse_conditions = mean(adverse_conditions))
print(ace_health, n=10)
```

Question 14: Association between smoking status and ACEs

Does the proportion of respondents who reported experiencing adverse childhood conditions vary by smoking status?

Now calculate the proportion of respondents reporting ACEs by smoking status (be sure the rows are arranged in appropriate order of smoking status) and store as `ace_smoking`.

```
# proportions grouped by smoking status
ace_smoking <- ...

# print
print(ace_smoking)
```

Question 15: Association between depression and ACEs

Does the proportion of respondents who reported experiencing adverse childhood conditions vary according to depressive disorder?

Calculate the proportion of respondents reporting ACEs by whether respondents had been diagnosed with a depressive disorder and store as `ace_depression`.

```
# proportions grouped by having experienced depression

ace_depression <- ...

# print
print(ace_depression, n=10)
```

Question 16: Exploring subgroupings

Does the apparent association between general health and ACEs persist after accounting for sex?

Repeat the calculation of the proportion of respondents reporting ACEs by general health rating, but also group by sex. Store the result as `ace_health_sex`.

```
# group by general health and sex
ace_health_sex <- ...
```

Use `pivot_wider` to put the sex as 2 different columns at the top of your dataframe and print it for easier readability.

```
...
```

Question 17

Even after rearrangement, the table in the last question is a little tricky to read (it can be difficult to visually scan tables). Instead, we'll use the original longer data frame, `ace_health_sex`, to create a visual representation. Create a barplot with `ggplot` using the `geom_col` geometry. For the aesthetics (`aes`) the y-value will be proportion of respondents reporting ACEs, the x-value will be the general health rating, and the `fill` color will be according to `sex`. In `geom_col`, be sure to specify `position="dodge"` to place male and female counts next to one another for each health category (rather than stacked).

```
# coerce indices to columns for plotting
...
```

Question 18: Scatter plot

Use the example above to plot the proportion of respondents reporting ACEs against smoking status for men and women.

```
...
```

Communicating results

Here you'll be asked to reflect briefly on your findings.

Question 19: Summary

Is there an observed association between reporting ACEs and general health, smoking status, and depression among survey respondents who answered the ACE questions?

Write a two to three sentence answer to the above question summarizing your findings. State an answer to the question in your first sentence, and then in your second/third sentences describe exactly what you observed in the foregoing descriptive analysis of the BRFSS data. Be precise, but also concise. There is no need to describe any of the data manipulations, survey design, or the like.

Type your answer here, replacing this text.

Question 20: Scope of inference

Recall from the overview documentation all the care that the BRFSS dedicates to collecting a representative sample of the U.S. adult population with phone numbers. Do you think that your findings provide evidence of an association among the general public (not just the individuals survey)? Why or why not? Answer in two sentences.

Type your answer here, replacing this text.

Question 20: Ethical Considerations

The BRFSS data includes information about traumatic and sensitive experiences.

i. How might your findings impact different stakeholders (e.g., individuals with ACEs, healthcare providers, policymakers)? Is there any potential for harm or misinterpretation? Answer in two or three sentences.

Type your answer here, replacing this text.

ii. The survey asks respondents to recall potentially traumatic childhood events. What are some possible consequences of asking this of respondents? Do the potential benefits outweigh potential harms? What kinds of response bias might arise as a result of the questions? Answer in two or three sentences.

Type your answer here, replacing this text.

Question 21

Notice that the language 'association' is non-causual: we don't say that ACEs cause (or don't cause) poorer health outcomes. This is intentional, because the BRFSS data are what are known as 'observational' data, *i.e.* not originating from a controlled experiment. There could be unobserved factors that explain the association. In a few sentences, propose a few potential unobserved factors that could explain this association between adverse childhood experiences and health.

Answer in two or three sentences.

Type your answer here, replacing this text.