CANCER SURVIVORSHIP OUTCOMES IN INDIVIDUALS WITH LUNG CANCER: AN ANALYSIS OF THE BEHAVIORAL RISK FACTOR SURVEILLANCE SYSTEM

Uyen Nguyen (gmd8sq)
Andy Ortiz (eao7r)
Lee Ann Johnson (lj6gd)
JD Pinto (jp5ph)

Github: https://github.com/yoowhyeeen/DS-5100.git

Introduction

After breast cancer and prostate cancer, lung cancer is the second most diagnosed cancer in the United States (U.S.) for both men and women. In 2021, approximately 235,760 new cases will be diagnosed (SEIGEL, 2021). Unfortunately, lung cancer remains the leading cause of cancer-related deaths with approximately 131,880 new deaths estimated for 2021 (SEIGEL, 2021). Regardless of survival, a significant number of individuals with lung cancer receive cancer care each year.

Although lung cancer claims the lives of more Americans than any other type of cancer, relatively little is understood about lung cancer survivors when compared to other types of cancer where individuals live longer after diagnosis. Relatively short survival times after a diagnosis have limited studies in this area. Individuals are typically considered to be survivors of cancer beginning at the time of a cancer diagnosis. Due to the high number of new lung cancer cases, and thus high number of new lung cancer survivors each year, it is vital to understand lung cancer survivorship outcomes.

The overall purpose of this project was to create predictive models to identify patient characteristics associated with guideline-concordant survivorship care. More specifically, in a nationally representative sample of individuals with lung cancer, we will 1) describe the demographics, physical, and mental health characteristics, 2) examine associations between demographics, physical, and mental health characteristics and the survivorship outcomes (cancer care summaries, written cancer care summaries, and health insurance coverage), 3) model which demographic physical, and mental health characteristics predict survivorship outcomes.

Data Set

To explore cancer survivorship outcomes in a nationally represented sample of individuals with lung cancer, the Behavioral Risk Factor Surveillance System (BRFSS) data set was selected. This dataset is created through an annual telephone-based survey collected in each U.S. state and territory. The primary questions are standardized between states, and include information pertaining to use of preventative services, health-related risk behaviors, and chronic health conditions. In addition to the standard questions, states can opt to administer other approved and standardized health-related modules.

BRFSS data is published annually on the Centers for Disease Control (CDC) website (https://www.cdc.gov/brfss/index.html). Published materials for each year include the

questionnaires, a data dictionary that identifies questions and answer codes, and the data. All data are deidentified and freely available for download. Each year of data is available as a separate file. For this project, to ensure an adequate sample size of individuals diagnosed with lung cancer, data from the following years were selected: 2016, 2017, 2018, 2019, and 2020.

Data Structures

The data structure used for this project is a 2-D array with 18 columns and 945 rows. For the purpose of this experiment, we believe that the dataset is large enough.

Data Pre-processing

Initially, data from only the year 2020 was selected. After initial data exploration, it was determined that the sample size needed to be larger, thus data for 2016, 2017, 2018, and 2019 were then included into the dataframe. Data for the project is contained in .zip files maintained by the CDC. Each .zip file contains the data formatted for the statistical software package SAS. The first step in data-preprocessing was to convert the SAS files into a .cvs file. In order to save memory, we then deleted the SAS files, only keeping the .csv files for use in analysis.

The dataframes contained more data than was needed for the purpose of this project, thus the next step was to reduce the number of columns and rows to include only what was needed for analysis. We chose to retain 14 predictor variables and 4 response variables, leaving 18 rows in each dataframe. The predictor variables were demographic, physical, and mental health characteristics and included the following columns: state of residence, sex, age, education level, marital status, employment status, income level, race, mental health in the last 30 days, physical health in the last 30 days, activities of daily living (ADLs) in the last 30 days, difficulty with mobility, difficulty with cognition, and difficulty performing errands. The following response variables were selected.

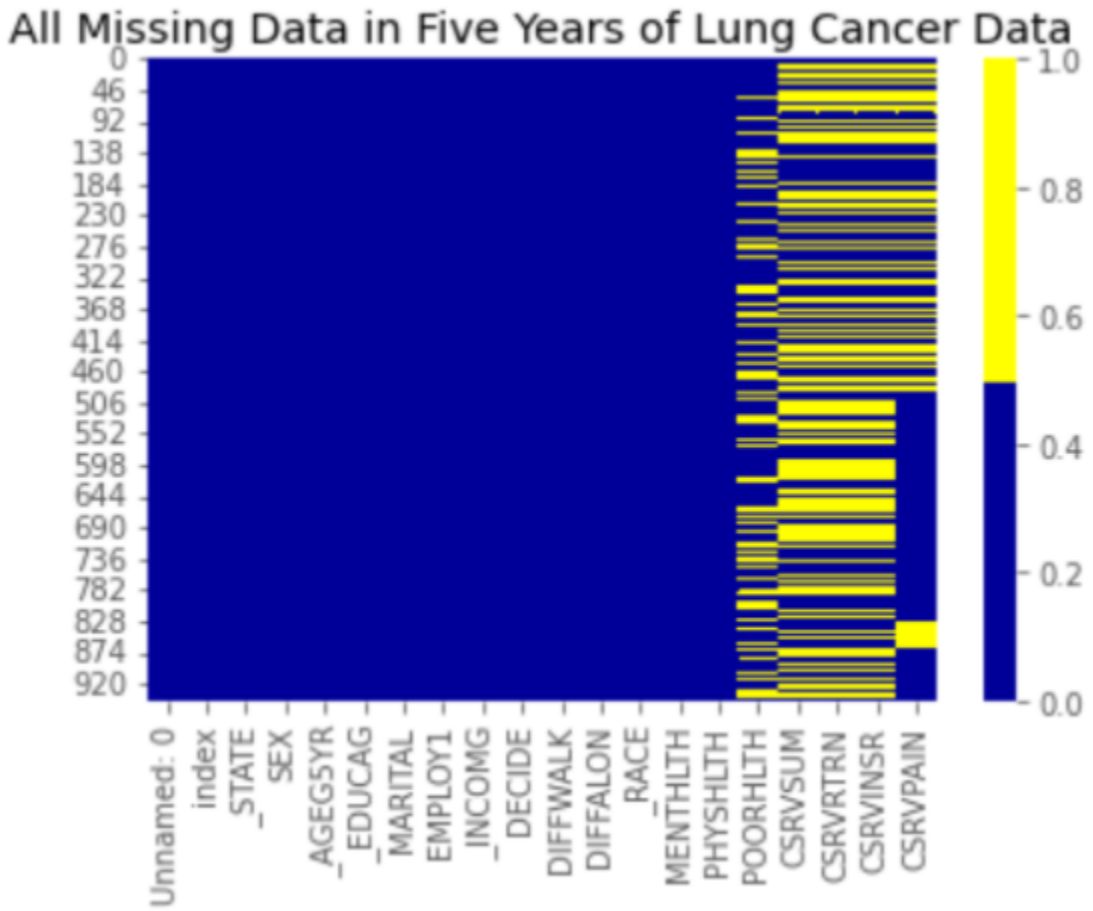
Variable Type	Variable Name	Survey Question
Written Summary	CSRVSUM	Did any doctor, nurse, or other health professional ever give you a written summary of all the cancer treatments that you received?
Routine Cancer Check-Up	CSRVRTRN	Have you ever received instructions from a doctor, nurse, or other health professional about where you should return or who you should see for routine cancer check-ups after completing treatment for cancer?
Adequate Health Insurance Coverage	CSRVINSR	With your most recent diagnosis of cancer, did you have health insurance that paid for all or part of your cancer treatment? (“Health insurance” also includes Medicare, Medicaid, or other types of state health programs.)
Cancer Pain	SCRVPAIN	Do you currently have physical pain caused by your cancer or cancer treatment?

Several additional steps were taken prior to merging the five dataframes. Column names were reviewed for consistency between survey years. Though most column names were consistent across years, some were not. See the following image for an example of variation in column name. To address this issue, columns in each individual year of data were renamed for consistency.

_SEX	X or SEXVAR	X or SEXVAR	SEX1	SEX	SEX

Additionally, we used the variable for cancer type to select only those individuals who self-identified as having lung cancer. To do this, we selected for the column "CNCRTYP1" and then selected only those coded 24 (lung cancer). After these changes, all five dataframes were merged into one.

To continue data pre-processing, we explored missing data within the merged dataframe. The following image depicts identified missing data in the 14 predictor variables and 4 response variables. This dataframe included 945 rows as no missing data was removed.

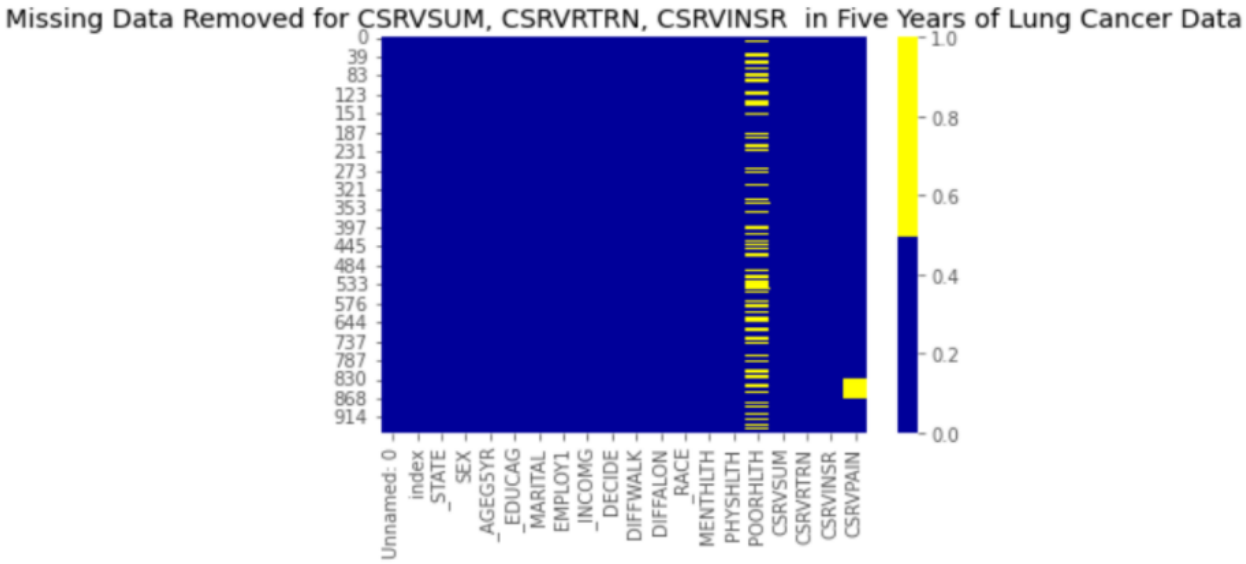


For nearly all 14 predictor variables, data was complete. This is because all states administer the modules for basic demographic information. There was some missing data in one of the predictor variables (POORHLTH). More missing data was identified in the selected response variables, Not

all states administered the module that contain cancer survivorship outcomes. Three of the response variables, written summary, routine check-up, and adequate health insurance coverage, had a different pattern of missing data when compared to the response variable of cancer pain. Thus, further exploration of missing data was conducted according to this pattern.

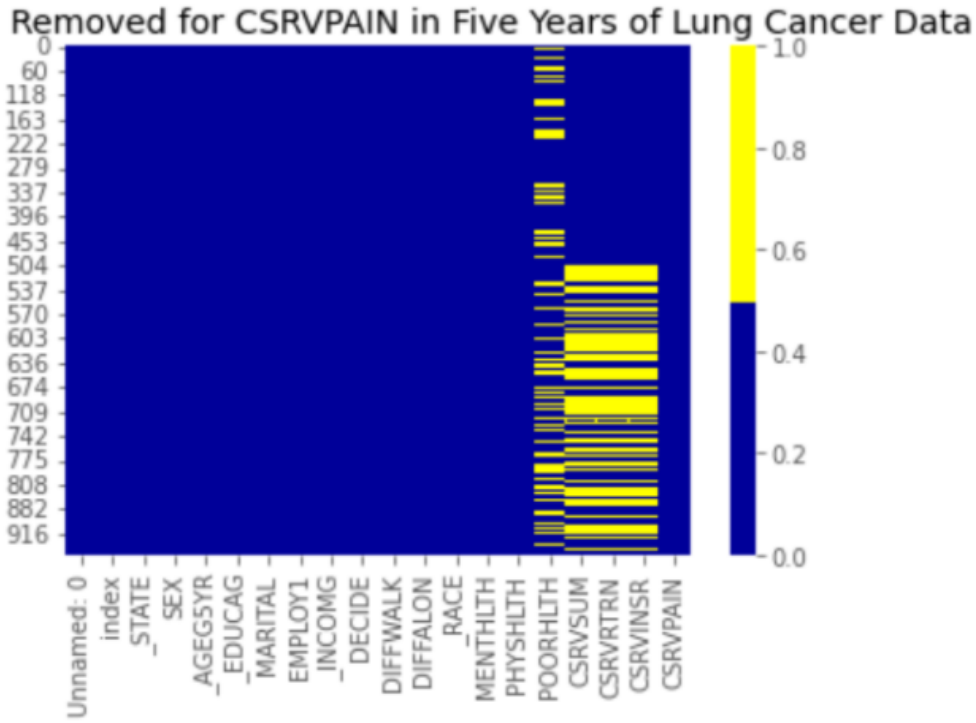
Next, missing data was analyzed for each of the response variables. When rows with missing data for the response variables written summary, routine check-up, and health insurance were identified and removed, the dataframe was reduced to 458 rows.

```
(458, 20)
```



Missing data for the response variable of cancer pain was examined. When missing data was removed for this variable, 689 rows remained.

(689, 20)



After patterns of missing data were visualized, it was decided that rows with missing data would not be deleted from the dataframe. The rationale for this decision was that removing rows with any missing data would unnecessarily reduce the overall size of the dataframe.. Because a logistic regression would be conducted separately for each response variable, the rows with missing data for that response variable would not be included in the analysis (due to no response).

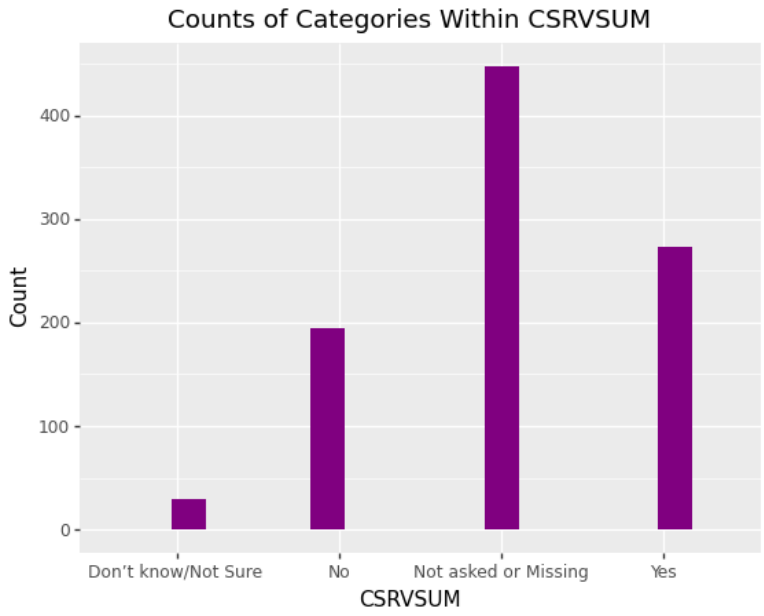
Data Analysis and Processing

Four different methods were used for data analysis and processing. These methods included 1) the creation and examination of visualizations of the predictor variables, 2) the creation and examination of visualizations for the response variables, 3) assessing the assumptions and tests for the logistic regression, and finally, 4) conducting a logistic regression for each of the selected response variables.

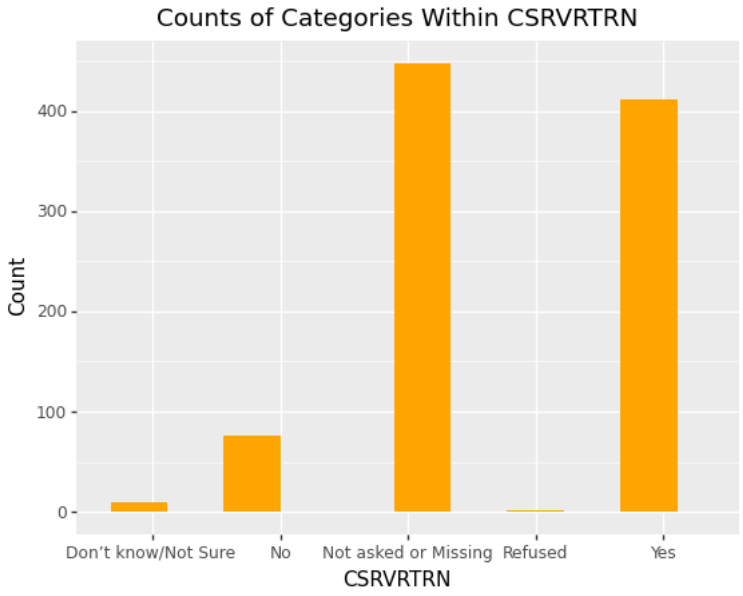
Data Analysis and Processing: Visualization of Response Variables

In this section, we show the visualizations of the response variables.

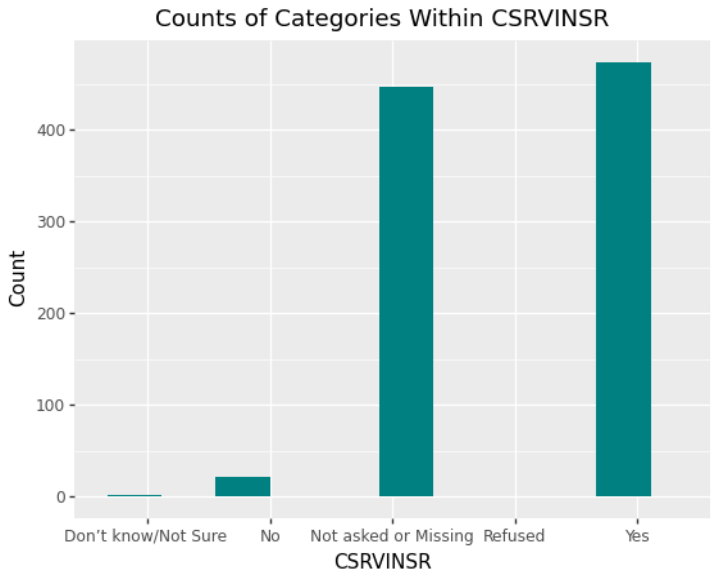
The visual below is of the response variable CSRVSUM. This response variable is the answer to the question, “Did any doctor, nurse, or other health professional ever give you a written summary of all the cancer treatments that you received?”



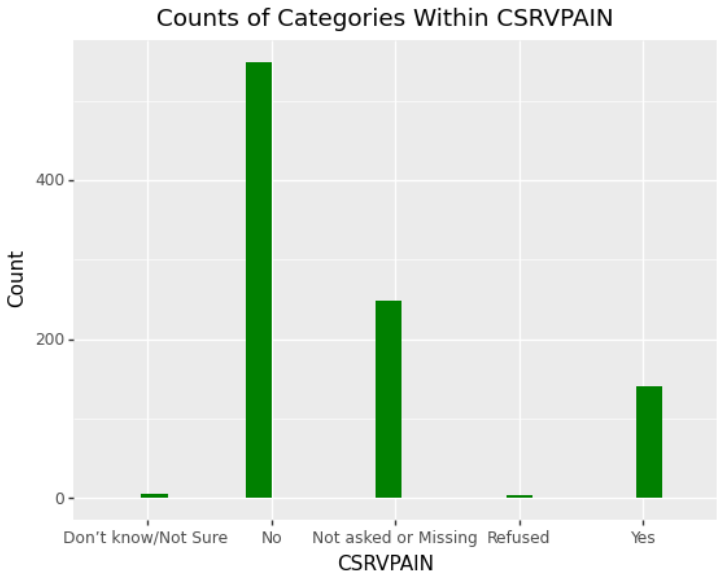
The visual below is of the response variable CSRVRTRN. This response variable is the answer to the question, "Have you ever received instructions from a doctor, nurse, or other health professional about where you should return or who you should see for routine cancer check-ups after completing treatment for cancer?" Of the actual responses, most answered yes.



The visual below is of the response variable CSRVINSR. This response variable is the answer to the question, "Have you ever received instructions from a doctor, nurse, or other health professional about where you should return or who you should see for routine cancer check-ups after completing treatment for cancer?" In terms of yes or no responses, the majority of individuals answered yes.



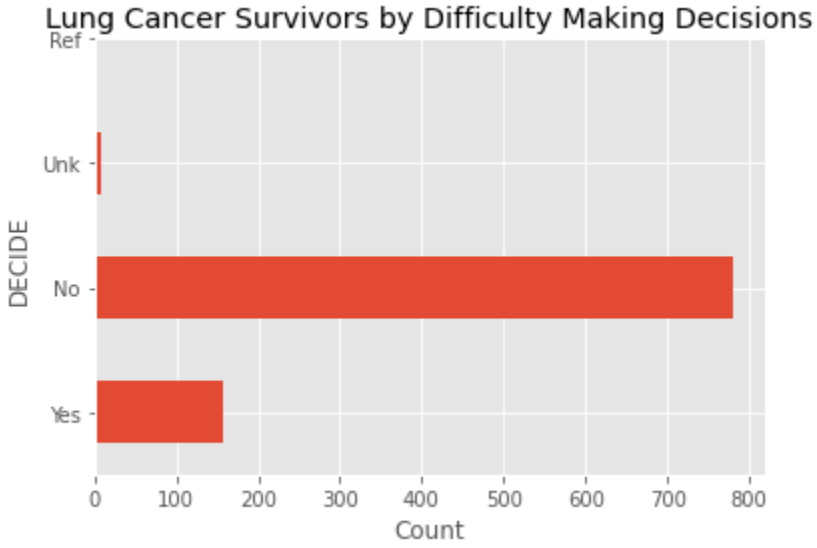
The visual below is of the response variable CSRVPAIN. This response variable is the answer to the question, "Do you currently have physical pain caused by your cancer or cancer treatment?" As you can see, there are a number of missing responses but the overwhelming majority of cancer survivors do not have physical pain caused by cancer treatment, which is surprising.



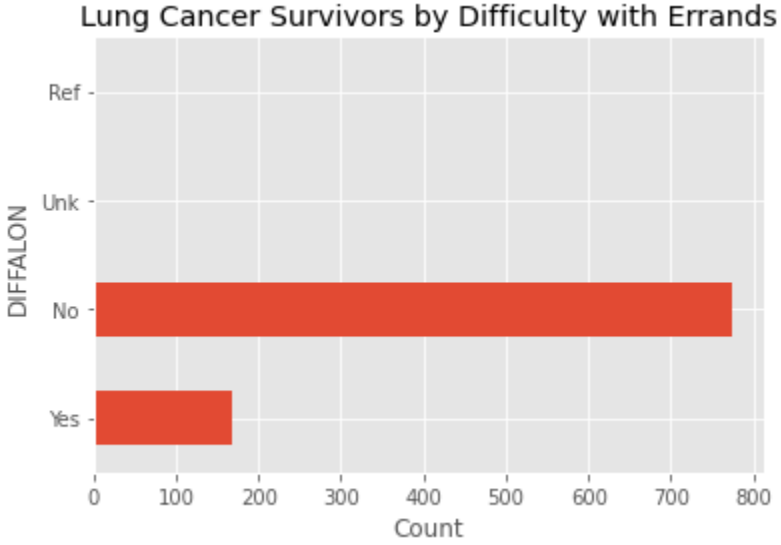
Data Analysis and Processing: Visualization of Predictor Variables

The graphs below are visualizations of the predictor variables. These visualizations helped us to better understand the dataset and search for important possible relationships.

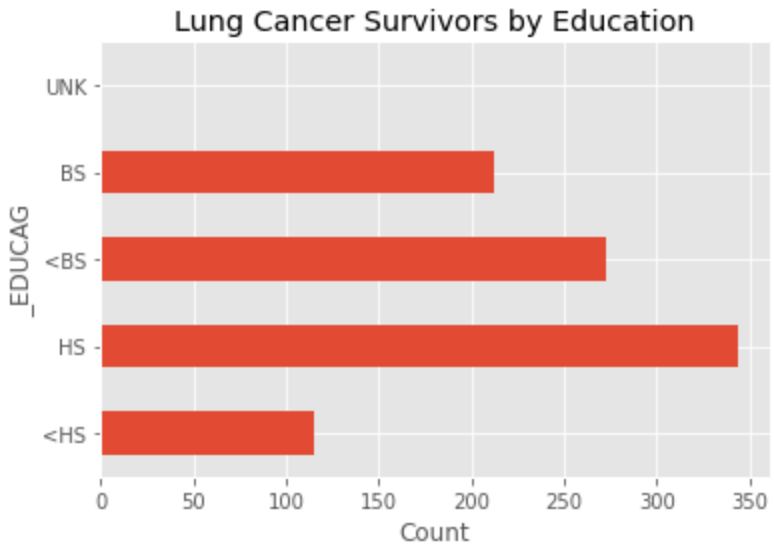
The chart below shows lung cancer survivors with difficulty making decisions. The majority of the responses were no. This means that the majority of lung cancer survivors in our dataset did not have difficulty making decisions.



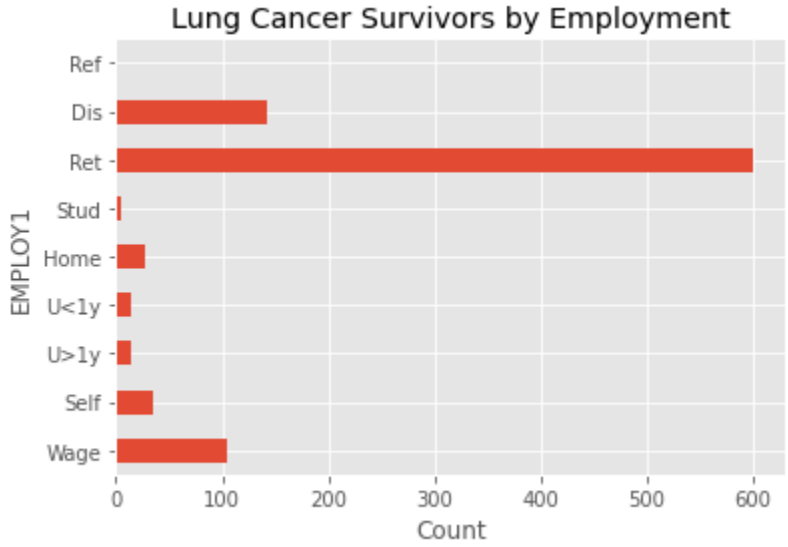
The chart below shows lung cancer survivors with difficulty with errands. The majority of the responses were no. This means that the majority of lung cancer survivors in our dataset did not have difficulty running errands.



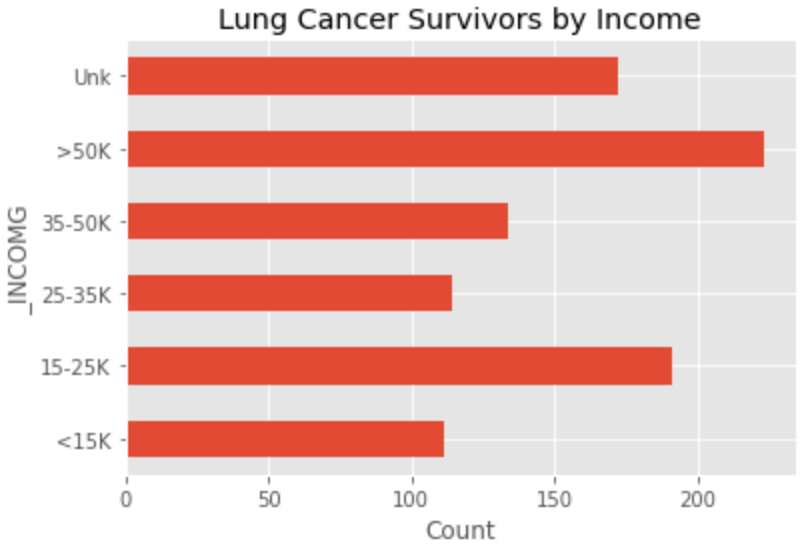
The chart below shows lung cancer survivors by level of education. The majority of respondents had a maximum High School level of education. This means that the majority of lung cancer survivors in our dataset only had a high school education.



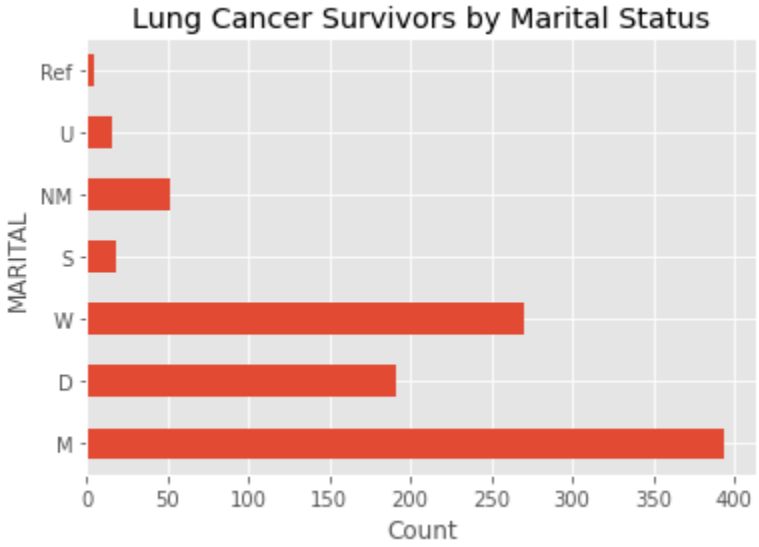
The chart below shows lung cancer survivors by employment status. The majority of respondents were retired. This means that the majority of lung cancer survivors in our dataset were retired.



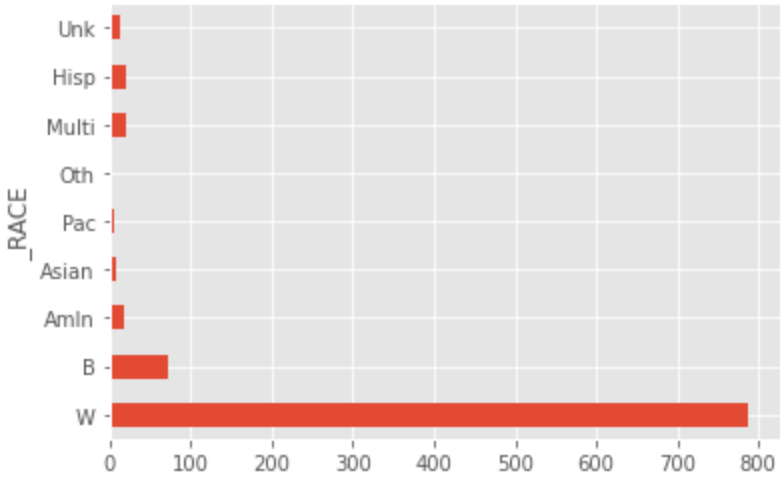
The chart below shows lung cancer survivors by income level. The majority of respondents had income of over 50k annually.



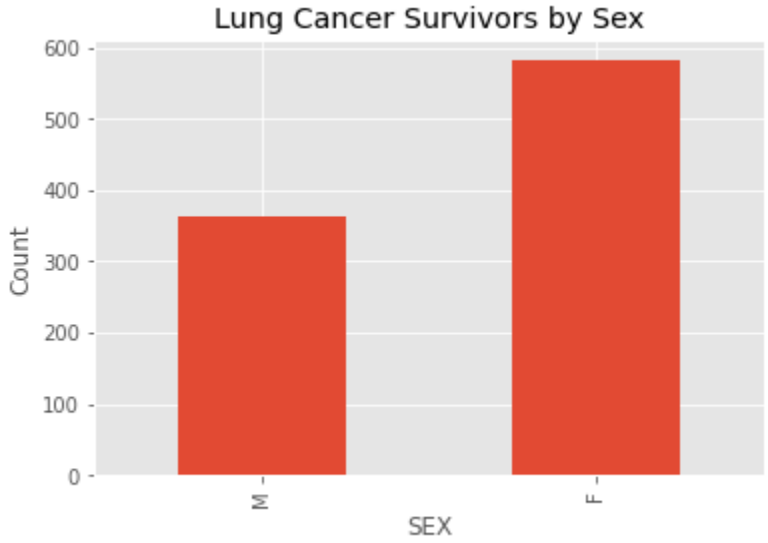
The chart below shows lung cancer survivors by marital status. The majority of respondents were married.



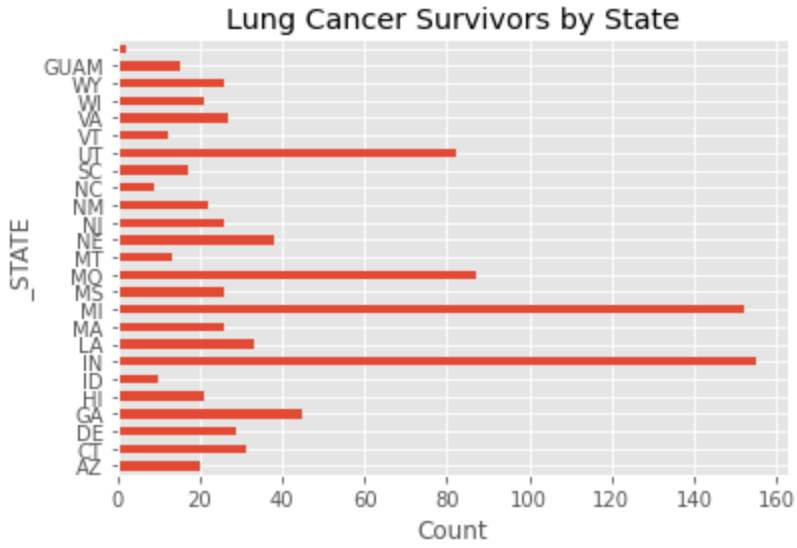
The chart below shows lung cancer survivors by race. The majority of respondents were white.



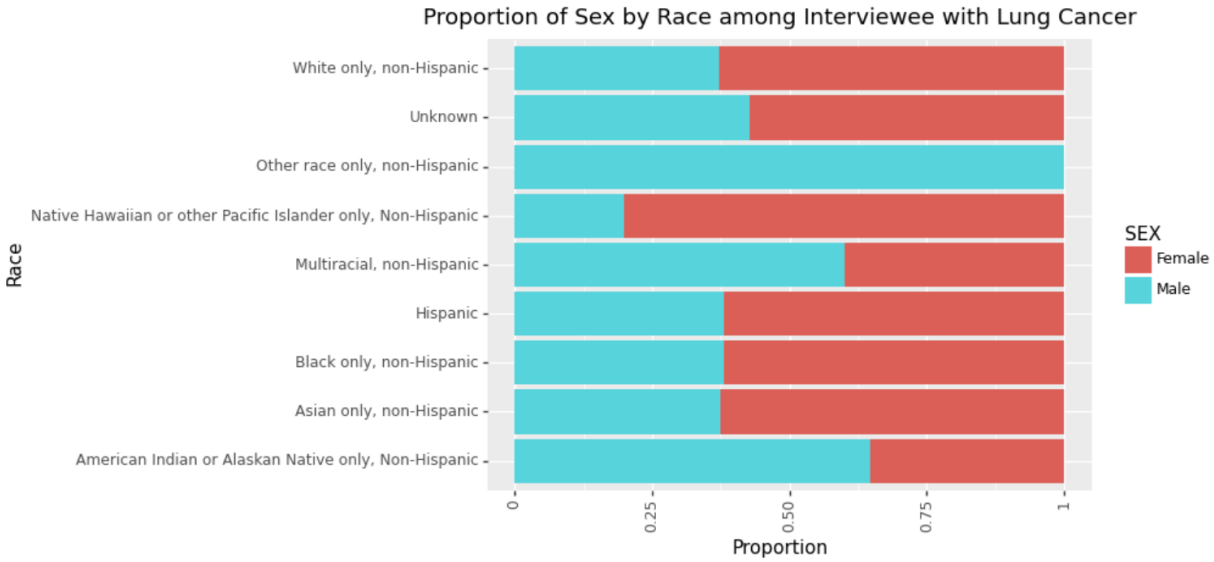
The chart below shows lung cancer survivors by sex. The majority of respondents were female.



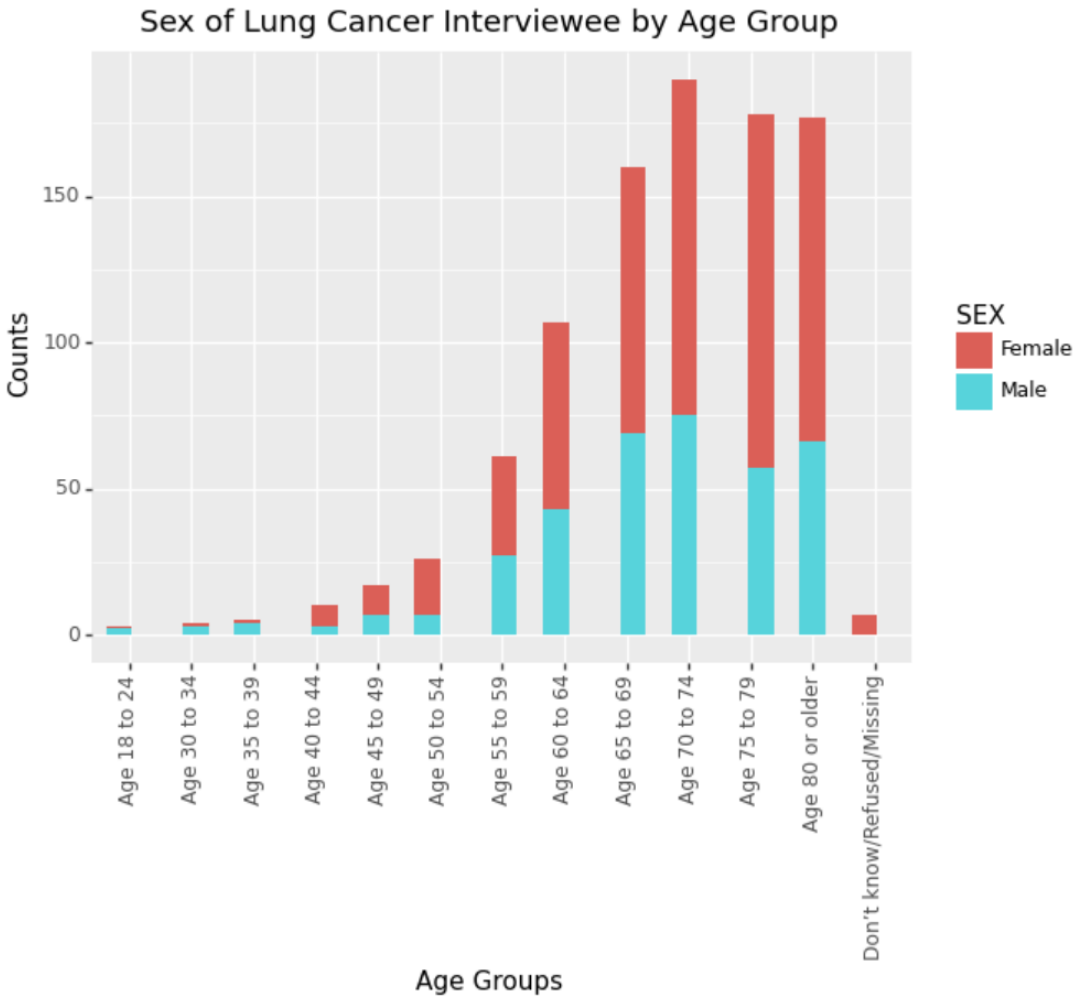
The chart below shows lung cancer survivors by state. There is a spike in respondents from Indiana and Michigan.



This table shows the proportion of men and women by race. In most racial categories, there were more women than men. Of note, more men identified as multi-racial and American Indian when compared to women.



This table demonstrates the proportion of men and women by age category. Women are more likely to have lung cancer than men (source: https://healthcare.utah.edu/the-scope/shows.php?shows=0_ziut6aix) You can see from this table that the majority of participants were over age 60, which is typical of lung cancer. Most participants were between the ages of 70 and 74.



We had several assumptions while working with the dataset. Our first assumption was that the dataset did not contain the same individuals each year and we got a new sample of different participants in each interview. Another assumption would be the participants interviewed did not forget or lie about their information on the interview. Since participants can hang up at any point during the voluntary survey and might not pick up when called back, information might not be completed or lost during the interview and not all information will be accurate. However, since our survey data was collected nationally, we hope that having a nationally representative sample will help us, overall, in data analysis and we don't have too much to worry about data being misrepresented.

Data Analysis and Processing: Logistic Regression

The logistic below is of the response variable CSRVSUM. This response variable is the answer to the question, "Did any doctor, nurse, or other health professional ever give you a written summary of all the cancer treatments that you received?" In this regression model we found that those with a

college education were less likely to receive a written summary.

```
Call:
glm(formula = CSRVSUM ~ X_EDUCAG, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2907  -0.9609  -0.8628   1.1909   1.5288

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.2624     0.2974   0.882  0.37771
X_EDUCAG2    -0.2941     0.3467  -0.848  0.39629
X_EDUCAG3    -0.7957     0.3529  -2.255  0.02415 *
X_EDUCAG4    -1.0587     0.3893  -2.720  0.00654 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 497.68  on 364  degrees of freedom
Residual deviance: 486.15  on 361  degrees of freedom
AIC: 494.15

Number of Fisher Scoring iterations: 4
```

```
"p-value =  0.0212117864944179"
```

```
"Accuracy 0.58695652173913"
```

```
  FALSE TRUE
1    33   24
2    14   21
```

Although the Education predictor was significant, and the model had a significant p-value, the model's accuracy (0.29), false positive rate (0.42), and false negative rate (0.4) were weak.

In the next regression model below, the predictor was binned levels of income (INCOMG).

```
Call:
glm(formula = CSRVSUM ~ X_INCOMG, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2033  -1.0365  -0.8607   1.3252   1.5315

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.06062    0.34832   0.174   0.8618
X_INCOMG2   -0.40155    0.41809  -0.960   0.3368
X_INCOMG3   -0.62702    0.44268  -1.416   0.1567
X_INCOMG4   -0.24295    0.44120  -0.551   0.5819
X_INCOMG5   -0.86297    0.42075  -2.051   0.0403 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 411.73  on 306  degrees of freedom
Residual deviance: 405.93  on 302  degrees of freedom
AIC: 415.93

Number of Fisher Scoring iterations: 4
```

```
"p-value =  0.121789798652368"
```

Although the Income predictor was significant, the model did not have a significant p-value.

Our model selection was not exhaustive, but of the many we tried, the above model with education as a predictor of a cancer survivor receiving a written summary of his or her treatment was the only significant one.

Testing

For our project we created two unit tests. The first tested if the SAS files were correctly imported and converted to a dataframe that could be used with pandas. The second tested if unneeded large files existed. We ran both tests and they returned OK without error.

```
# function that checks for files to delete
def cleanup():
  import os
  try:
    os.remove("LLCP2016.XPT")
    os.remove("LLCP2017.XPT")
    os.remove("LLCP2018.XPT")
    os.remove("LLCP2019.XPT")
    os.remove("LLCP2020.XPT")
    return 'Files were removed.'
  except:
    return 'These files have already been removed.'
```

```
# function that converts XPT/SAS formatted files to pandas dataframes
def converter():
  try:
    brfss2016 = pd.read_sas("LLCP2016.XPT", format='xport')

    brfss2017 = pd.read_sas("LLCP2017.XPT", format='xport')

    brfss2018 = pd.read_sas("LLCP2018.XPT", format='xport')

    brfss2019 = pd.read_sas("LLCP2019.XPT", format='xport')

    brfss2020 = pd.read_sas("LLCP2020.XPT", format='xport')
    return 'All files were converted.'
  except:
    return 'File conversion was interrupted.'
```

```
import unittest
# from Group3ProjectCode.ipynb import cleanup, converter

class ProjectFunctionTestCase(unittest.TestCase):  # inherit from unittest.TestCase

  def test_cleanup(self):
    self.assertEqual(cleanup(),'These files have already been removed.')

  def test_converter(self):
    self.assertEqual(converter(),'File conversion was interrupted.')


if __name__ == "__main__":
  # unittest.main()
  unittest.main(argv=[''], exit = False)
```

Results

Though we tested several models with our selected predictor and outcome variables, only one model was significant. Although this might initially seem disappointing, these findings illustrate that within this particular sample of individuals with lung cancer, there were no identified health disparities for the outcomes of routine cancer check-up, adequate health insurance coverage, or cancer pain. We did find, however, that level of education influences the receipt of a written summary of cancer care. Those with a college level education or higher were less likely to receive a written summary of cancer care.

These results should be interpreted with caution. The data for the BRFSS survey is cross-sectional, thus causality cannot be assumed. Due to the high mortality rates of lung cancer, the cross-sectional nature of data collection might not capture those individuals with late stage disease who are too sick to complete a long survey. Stage of disease is not captured by BRFSS so we do not know if the stage of disease is skewed to earlier stages of lung cancer. To improve our understanding of the outcome variables, longitudinal studies that follow lung cancer survivors over time may be

able to identify which predictor variables are related to our selected response variables at different points in the cancer trajectory. Collecting stage of disease would also provide information about how severity of disease might influence our selected outcome variables. Additionally, the sample lacks racial diversity. Racial disparities in lung cancer outcomes have been well documented, but were not found in our logistic regression models. One solution would be to oversample for underrepresented racial groups.

How the Results Can be Used by Others

A cancer diagnosis is an overwhelming time for both the individual diagnosed with cancer as well as their support system. Often, during the initial appointment to discuss the care plan, individuals and families are too distressed and overwhelmed to process the information. Because of this, written summaries of the cancer treatments should be provided so that the plan can be reviewed at a later time, when stress is not as high.

In this project, we found that those with more education were not as likely to receive these cancer summaries. There are several possibilities that may explain this phenomenon. Those individuals may waive a written summary or the healthcare provider may assume these individuals understood the care plan during appointments and not provide a written summary. Retrospective studies at the primary care level are needed to determine the cause of our findings.

In the meantime, providers and health care clinics should be more cognizant of providing written summaries to all individuals diagnosed with lung cancer. A quality improvement project that examines the process of delivering written summaries is one way to address this gap in care. These projects could focus on ensuring patients have online access to written summaries of care or creating new sections in electronic health records to document patient receipt of written care plans.

Ways to Improve, Expand, or Add Functionality

Several steps can be taken to improve, expand, and add functionality to our project. More specifically, interactivity could be added with respect to each year of the survey. This could be accomplished by allowing users to input certain years of data to quickly obtain visuals of descriptive statistics for each variable by year in addition to the merged data. This would be an especially useful function for the map feature so that users could quickly see changes in state of residence across years.

Another way to improve functionality would be to create an opportunity for user input to generate information about each of the response variables. Users could choose to look at the data in aggregate, or input specific single years or a combination of years to assess outcome variables across multiple years of data.

Because the BRFSS is administered every year, these additional features in functionality could be extended to include each additional year of data as it is published on the website. Although we only chose to include data from 2016 to 2020 in our analysis, this additional functionality could be added for the years prior to 2016 as well.

****

```
The original CDC BRFSS data download links are
https://www.cdc.gov/brfss/annual_data/2016/files/LLCP2016XPT.zip
https://www.cdc.gov/brfss/annual_data/2017/files/LLCP2017XPT.zip
https://www.cdc.gov/brfss/annual_data/2018/files/LLCP2018XPT.zip
https://www.cdc.gov/brfss/annual_data/2019/files/LLCP2019XPT.zip
https://www.cdc.gov/brfss/annual_data/2020/files/LLCP2020XPT.zip
```

****

Presentation slides:
https://docs.google.com/presentation/d/1K6PcbH1uIusE1T3RgqQTD_BuzIQVer5O/edit?usp=sharing&ouid=109805413678459814568&rtpof=true&sd=true