**DSC 520 Final Project Template**

Name: Christine Hathaway

Date: January 13, 2019

**Section 1**

* Explain what your interests are in the data sets identified: I work at a health insurance company, and the Affordable Care Act (ACA), or Obamacare as it is also known, is of great interest to the healthcare industry as a whole, and of course to the health insurance side of the industry. For my final project, I found six data sets that have information regarding ACA metrics. Five of the data sets are from DataWorld (U.S. Dept of Health & Human Services, 2016) and can be found at <https://data.world/gswider/aca-state-data>. A sixth data set is from Kaggle (U.S. Dept of Health & Human Services, 2017)and can be found at https://www.k aggle.com/hhs/health-insurance/version/1. The data sets contain information on coverage for both individuals and employer groups, Medicaid, Medicare, information at the state level as well as nationwide, and comparisons for coverage before and after the ACA was implemented. I want to use the data to answer questions regarding how coverage has changed since the ACA was put into place, if there was improvement, if more or fewer people received coverage, if rates increased or decreased, and if types of coverage changed.
* Identify the Packages that are needed for your project: In order to look at the data, I know I will need to import packages for “pastec” and “psych” to look at the descriptive statistics using the stat.desc() and describe() functions. Using these functions, as well as the str() function, should help me see the various categories and types of data contained in each one. I will also need to import the ggplot2 package to create various plots and graphs of the data.
* Original source where the data was obtained is cited and, if possible, hyperlinked: The data sets were published by the U.S. Department of Health and Human Services, and contain metrics on the newly enacted ACA after its implementation. The datasets from data.world were collected from the years 2010 and 2015 to compare before and after metrics for the ACA implementation.
* Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.): It was collected from various sources, including the Census, academic studies, and the department’s own estimates. The dataset from Kaggle is from several years. Rates and coverage variables are from the years of 2010 and 2015, while the Medicaid and Medicare expansion variables are from the years of 2013 and 2016. The data shows the changes in coverage due to the ACA and the expansion of the Medicare programs across all of the states. In all of the data sets, missing values are indicated by a blank cell.

# Works Cited

* U.S. Dept of Health & Human Services. (2016, December 27). *ACA State Data*. Retrieved from data.world: https://data.world/gswider/aca-state-data
* U.S. Dept of Health & Human Services. (2017, March 2). *Health Insurance Coverage: Coverage rates before and after the Affordable Care Act*. Retrieved from Kaggle: https://www.kaggle.com/hhs/health-insurance/version/1

**Section 2**

* Provide an introduction that explains the problem statement you are addressing: Why would someone be interested in this? The Affordable Care Act (ACA) changed healthcare, but it is unclear if the general population has actually benefitted from these changes. Health insurance has become a heated political topic, and both political parties make claims for and against the ACA. A look at real data can help support or refute whether the ACA laws have made any difference in who has insurance coverage. It can show the impact of the Medicaid expansion in states where expansion occurred, and the utilization of the new Marketplace that was established.
* Provide a concise explanation of how you plan to address this problem statement: I intend to address this problem by looking at data captured before and after the ACA was enacted. The data sets include information regarding employer group coverage changes, Medicaid expansion changes, and changes in the individual marketplace.
* Discuss how your proposed approach will address (fully or partially) this problem: By looking at the changes in the data, I can see if more or less people are covered by insurance, the number of individuals using the new marketplace, and the changes that have occurred in Medicaid and Medicare. This should offer some insight into how the ACA laws impacted insurance coverage for Americans.
* List at least 6 research questions you aim to answer.

1. How has the uninsured rate changed since the ACA was enacted?

2. How did Medicaid enrollment change for states that expanded vs those that didn’t?

3. What percentage of individuals with prior existing conditions selected a marketplace plan?

4. Individuals with marketplace plans receiving tax credits and cost sharing reductions, vs individuals eligible for tax credit but purchasing off market insurance?

5. What is the change in premiums, comparing 2000-2010 rate changes to 2010-2015 rate changes?

6. What is the trend or change in Medical Loss Ratio (MLR) rebates for insurance providers?

7. How has Medicare been impacted, as in free services utilized, the “donut hole gap” in prescription coverage, or changes in hospital readmission rates?

* Explain how your analysis may help the consumer of your research findings: My analysis could help consumers understand what changes have been brought about as a result of the ACA. It will also show the impact of some of those changes, how coverage has changed for patients and the impact to providers.
* What types of plots and tables will help you to illustrate the ﬁndings to your research questions? I plan to use histograms, scatterplot, and boxplots to start looking at the data. I also plan to use summary and descriptive statistical functions to examine the data.
* What do you not know how to do right now that you need to learn to answer your research questions? I think there are more tables that I could use that we have not been introduced to yet. I need to become more proficient with regression to make some comparisons with the variables.

**Section 3**

* Data importing and cleaning steps are explained in the text and in the DataCamp exercises (tell me why you are doing the data cleaning activities that you perform) and follow a logical process.
* With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.
* What do you not know how to do right now that you need to learn to import and cleanup your dataset?

**Section 4**

Discuss how you plan to uncover new information in the data that is not self-evident. I plan to use some regression models, including single, multiple, and logistic regression models. These will help me test whether my results are significant, and also help reveal any correlation among the variables and observations in the data. I am also using the histogram and plot functions on the model variables to graphically look at the model results.

What are different ways you could look at this data to answer the questions you want to answer? I am using descriptive analysis to look at the data. This includes functions such as summary, stat.desc, glimpse, and str. I am also using the order function for some of the variables to see what the highest and lowest observations are and where they occur.

Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information, learn and use an R package not covered so far in the course? I have created new variables and used the melt function to rearrange the way the data is displayed. I have also created subsets of the data to focus the descriptive statistics to just those variables I am interested in.

How could you summarize your data to answer key questions? Using the techniques described above, I am isolating the data down to the variables that are most closely aligned with the questions that I am trying to answer. Using statistics, regression models, graphs and plots will help me analyze the data to answer my questions.

What types of plots and tables will help you to illustrate the findings to your questions? I am using histograms, lines, scatterplots, and regression models.

What do you not know how to do right now that you need to learn to answer your questions? I need to learn more about logistic regression in order to analyze its results for some of my models.

Do you plan on incorporating any machine learning techniques? I don’t know enough about machine learning to know if it will help with my research at this point.

**Section 5 Summary**

The Affordable Care Act (ACA) changed healthcare, but it is unclear if the general population has actually benefitted from these changes. Looking at real data, I addressed some of the changes and impact of the ACA. I examined if there were any differences in who has insurance coverage before and after the ACA was enacted, and some of the impact of the Medicaid changes.

I looked at data captured before and after the ACA was enacted. The data sets include information regarding employer group coverage changes, Medicaid expansion changes, and changes in the individual marketplace. I used descriptive statistics, graphs, plots, and some regression models to compare and contrast the data.

Massachusetts had the smallest percentage change in the uninsured rate, while Nevada had the highest. The median change in the uninsured rate was a decrease of 5.0%. States with expanded Medicaid programs saw much larger enrollment percentages than states that did not expand, some even doubling enrollment. Hawaii saw the smallest percentage of individuals plan with preexisting conditions selecting a marketplace, while Florida had the largest. New York had the highest growth in premiums for employee coverage, while Florida had the lowest.

My analysis showed consumers the changes that have been brought about as a result of the ACA. It showed the states with the biggest impact of those changes, the change in premiums for employer provided coverage, and changes in Medicaid usage.

My analysis is limited by my lack of experience, and really knowing how to dig into the information. Someone with more insight could have better ideas of how to analyze the data that is available, and what some of the numbers really mean. For example, what have changes been over the last 20 years, are the changes we are seeing entirely related to the ACA, or does inflation and cost of living have more impact? Without that data, my analysis is limited. My interpretation of the regression models and some of the graphs is also not as robust as it should be.