**DSC 520 Final Project**

Name: Michael Gonzalez

Date: 04/18/2019

**Section 1**

* Explain what your interests are in the data sets identified.

The first data set is Adults with diabetes in California from the healthdata.gov website. My interest is personal, since my family history has diabetes and the data is from my home state. The second data set is Accidents, Fatalities, and Rates, 1995 through 2014 .U.S. General Aviation from the data.gov website. My interest is due to recent news report of plane crashes and accidents. The third data set is Oregon’s worker’s compensation claims data 1968 to 2017 from the open data network website. My interest is from my work, where at times, we deal with transferring calls of fellow employees to our worker’s compensation department.

* What is the target audience for this research?

The target audience for the three data sets are the public. The first data set is the public who has diabetes and local health groups to use for promoting healthy living. The second data set is the public who travel on planes and watch dog groups. The third data set is public to show the citizens of Oregon, how much money was spent on worker’s compensation claims.

* Identify the Packages that are needed for your project.

The packages that I think, I will use are ggplot2 for visualization, pastecs for descriptive statistics and RMarkdown for creating final project report.

* Original source where the data was obtained is cited and, if possible, hyperlinked.

The first data set: <https://healthdata.gov/dataset/adults-diabetes-100-lghc-indicator>

The second data set: <https://catalog.data.gov/dataset/accidents-fatalities-and-rates-1995-through-2014-u-s-general-aviation>

The third data set: <https://www.opendatanetwork.com/dataset/data.oregon.gov/p8ud-dzhp>

* 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.).

The source data for the first data set was taken from a telephone survey and it was collected for the state’s health department campaigns. The source data from the second data set was taken from the government’s own data about plane accidents. The source data of the second data set has many areas with missing data. The source data from the third data set was taken from the state’s worker’s compensation claims. The source data of the third data set has a few areas of missing data and it used to figure out how much money was spent on this claims.

**Section 2**

* Provide an introduction that explains the problem statement you are addressing. Why would someone be interested in this?

There are many people with diabetes, living in California, which is putting a strain on the health care system. Diabetes is becoming a major problem and is accounting to the rising prices of health insurance plans. This is interesting for people who care about the well-being of all Californian and lowering the cost of health insurance plans.

* Provide a concise explanation of how you plan to address this problem statement.

After looking through the results, I plan to make a list of what areas of the data that will help in creating a campaign. The idea of the plan is to inform and help people with diabetes to live healthier lives in California. It will also inform people about affordable and healthy meals to avoid getting diabetes.

* Discuss how your proposed approach will address (fully or partially) this problem.

The proposed campaign will promote healthy lifestyles to the focused groups that in turn will lower the number of people with diabetes. By informing people about free outdoor actives and affordable healthy meals that can be made easily. These approaches will hopefully lower the effects of diabetes and the cost of health insurance plans.

* List at least 6 research questions you aim to answer.

1. Can informing people with similar backgrounds to change their lifestyle that can lead to having diabetes in the future?
2. By creating campaigns to promote healthy habits and lifestyles lower the number of people with diabetes?
3. Do people who make less than $15,000 a year, more likely to have diabetes than others?
4. Why young people with college degrees, are less likely to have diabetes?
5. Is it common for both men and women to have diabetes?
6. Are there more people who have diabetes now than before in California?

* Explain how your analysis may help the consumer of your research findings (recall you target audience from Section 1).

The main target audience is the public, but this also help local city government who want to lower their health budgets. Another consumer of the research findings will be big corporations who want to lower their company’s health insurance costs. These research findings will appeal to anyone who wants to live a better life with diabetes and have a life without diabetes.

* What types of plots and tables will help you to illustrate the ﬁndings to your research questions?

I want to use plots and tables that give the data in straightforward way. I am thinking of using histograms and bar graphs to give visual represent that will be easy to understand. I will use simple tables will common words that is easy to understand the research findings.

* What do you not know how to do right now that you need to learn to answer your research questions?

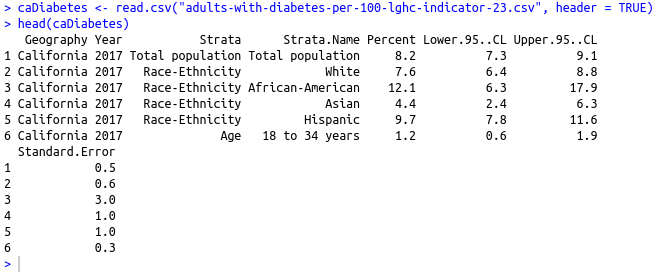
I still need to get a better understanding on interpreting the data’s p-values and null hypothesis test. I also need to grasp the concept of correctly interpreting a correlation, if there is one that is connected to the results.

**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.

The reason why I am cleaning my data, is to make sure that I will be able to get good results. If you do not remove the missing data, then you will get errors as your result. For example, when doing a correlation test, you must have all the missing data taken out, so you can get any results. Also, if you want to find the mean of a variable and a missing data is there, you will just get the missing data NA error message.

* 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?

I do not know what tools, I can to use for cleaning large data sets. In week 5’s assignments, I used the Hmisc package to clean my data set. I wonder if there is a function or a tool to do the cleaning of large data sets. I will do some more research on techniques that deal with cleaning large data sets.

**Section 4**

* Discuss how you plan to uncover new information in the data that is not self-evident.
* What are different ways you could look at this data to answer the questions you want to answer?
* 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? Explain.
* How could you summarize your data to answer key questions?
* What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).
* What do you not know how to do right now that you need to learn to answer your questions?
* Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Suggestion from the course professor: Some additional questions you may want to consider asking yourself as you work through this section of the project:

1. What features could you ﬁlter on?
2. How could arranging your data in different ways help?
3. Can you reduce your data by selecting only certain variables?
4. Could creating new variables add new insights?
5. Could summary statistics at different categorical levels tell you more?
6. How can you incorporate the pipe (%>%) operator to make your code more efﬁcient?

**Section 5 Summary**

* Overall, write a coherent narrative that tells a story with the data as you complete this section.
* Summarize the problem statement you addressed.
* Summarize how you addressed this problem statement (the data used and the methodology employed).
* Summarize the interesting insights that your analysis provided.
* Summarize the implications to the consumer (target audience) of your analysis.
* Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.
* In addition, submit your completed Project using R Markdown or provide a link to where it can also be downloaded from and/or viewed.