**DSC 520 Final Project Template**

This handout is meant to provide you with some structure for your final project. Note that there are two parts to this handout. First is the description of what you will be doing in each section and their general due dates. Second is the template itself with the various headings.

You are free to adjust this as needed. However, given what you’ll be doing, it is advantageous to you to just follow the template as it will help keep you focused. Should you have any questions, please ask!

**Part 1 – Week 10 – Getting Started**

* Provide an introduction that explains the problem statement you are addressing. Why would someone be interested in this?
  + ***Summer Olympics:* We are going to look at this data to determine if individual country or person has a larger impact on Olympic victories. It will go into detail to look at winners of the Summer Olympics who have won multiple Olympic medals, from their various countries. It will also look at the detailed information about where the events took place. It will also go into detail about the specific sports. People will be interested in this because it allows people to understand when and where medals were won, and if country has an impact.**
* Draft 5-10 Research questions that focus on the problem statement.
  + ***Olympics:***
    - How has the Olympic games changed over the years?
    - Are there countries that have a better chance at winning games?
    - Do we see specific sports that are dominated by countries or areas?
    - Are we more likely to see repeat winners if they participate multiple times?
    - Does winning one medal type lead to winning a different one?
* Provide a concise explanation of how you plan to address this problem statement.
  + **For each data set I will break down the information and see if there is a correlation between various data points. It will allow me to gain a better understanding of the data as whole. I will then go further to look into the information and the variables. Gain my understanding on how I would address the issue in the first place, and the relationship with the variables. I will then decide what points will be more valuable when looking at the relationships. This detailed data information will provide support for the question and address how to advance it further.**
* Discuss how your proposed approach will address (fully or partially) this problem.
  + **My approach will look at the detailed data points for each area. It will allow me to address the problem statement and answer the questions. Through the process of pulling specific datapoints, I will be able to build various correlation plots, and see if there are relationships between different data points. I will also readjust problem questions through this process to better hone in on the information needed to see correlation.**
* Do some digging on a dataset that you can use to address the issue.
  + Original source where the data was obtained is cited and, if possible, hyperlinked.
    - **The various datasets were pulled from Kaggle. There were 3 different datasets I pulled: Summer Olympics, 2012 vs 2016 Election Results, and Baseball Hitting Statistics.**
  + 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.).
    - **Baseball Hitting: 22 variables with 101332 observations; N/A values: IBB, HBP, SH, SF, GIDP**
    - **Summer Olympics: 9 variables with 31165 observations; N/A values: NONE**
    - **Election Results: 21 variables with 3141 observations; N/A values: N/A values: total votes 2012, votes dem 2012, votes gop 2012, county fips, state fips, per dem 2012, per gop 2012, diff2012, per point diff 2012**
* Identify the packages that are needed for your project.
  + **Summary(), head(), cor(), plot(), summarize(), table()**
* What types of plots and tables will help you to illustrate the ﬁndings to your research questions?
  + **Scatterplot, Matrix Table, Line Graph, Bar Graph**
* What do you not know how to do right now that you need to learn to answer your research questions?
  + **I feel very confident when looking at the datasets that I have grabbed. I may have to do more research in determining linear models, and further grasping correlation based off charts.**

**Part 2 – Week 11 – Cleaning Your Data and Exploratory Data Analysis**

* 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 first thing I did was upload the data, where I took a summary() and head(). From there I created vectors based off what data points I wanted to look at from the set. From there I took a summary() of the new and cleaned dataset. I was going to remove the “discipline” data point, however I decided against it because it was something that could still be referenced for further information. But, I will not really use it in my analysis.**
* 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.

Table

Description automatically generated

* What do you not know how to do right now that you need to learn to import and cleanup your dataset?
  + **I am pretty set on what I have right now. Everything looks nice and clean. The only thing I may look into is I am still considering removing the discipline section because I am unsure of the value it provides. But, I am still considering my options here and will make a decision prior to final project.**
* Discuss how you plan to uncover new information in the data that is not self-evident.
  + **My biggest takeaway would to be to look into the data further. Through various things like comparisons or even diving deeper into specific data points. It is important to view the entire picture when working with data, and in my experience it key and essential to think about the data as you are working. I plan on utilizing various tools to look further to see what some peers are doing or other individuals. Even though we are all using different data, it is important to remember there are still similarities between datasets, or questions being asked.**
* What are different ways you could look at this data to answer the questions you want to answer?
  + **To answer the questions, I proposed for this dataset I would look at the various different functions I can utilize, as well as graph features. By utilizing these functions and graphs I will be able to get a better outlook at my data. Through these various features I can actually dive in and see relationships that have been created and get a better understanding of what I am looking at. My questions are very straightforward and can be better understood if I create correlations with the data.**
* 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.
  + **After looking at my data multiple times, I decided to create various different variables in relationship to the data frame. This was because I was working mainly with categorical data, and need to create values. So thus, I developed values with wins, medals, and sports.**
* How could you summarize your data to answer key questions?
  + **For my dataset I plan on first looking at correlation through graphing, I’ll then look into other data points from looking at this information. I will also run several graphs to see if there is a linear relationship between the data points, which would answer several of my questions. These graphing points and visual relationships will show if there is a correlation.**
* 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.).
  + **Bar Graph to show value of wins**
  + **Line Graph to show changes over time**
  + **Tables to show relationships to values**
* What do you not know how to do right now that you need to learn to answer your questions?
  + **My only question is how much data needs to be shown in the final presentation. Based off some of the assignments we have had we could have 500pages worth of data information. Knowing how much data has been made up of this Summer Olympics dataset my thoughts are that this could be very overwhelming for someone looking at the information.**
* Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.
  + **Machine learning is not necessary for this analytical project. I am looking at the relationships between various variables and how they impact each other. When looking at mainly statistical information it is not necessary to utilize machine learning techniques to answer questions. As I go further, I may consider it, but as of right now I don’t think it is appropriate. But, if someone has a suggestion, I am more than willing to listen.**

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?

**Part 3 – Week 12**

* Overall, write a coherent narrative that tells a story with the data as you complete this section.
  + **The data provides several insights on the various different aspects of the Olympic games from 1896 to 2012. The data provides valuable information that can create correlation and relationships between the data points. Due to the fact that the data is many categorical, the data set was looked at in mainly graph formats. The data tells a story about how over the years the Olympic games have grown significantly. It also presents clear information that ties with historical events why there are dips in various data points. The Olympics have grown in participation, sports, and countries over the past 100 years.**
* Summarize the problem statement you addressed.
  + **The problem statement was to see visually if there is a relationship between athletes, sports and countries over the past 100 years of the Olympics. The goal was to see if there is a better chance for a specific country or athlete to win the games. It was also to look at which countries have dominated the games, and if the athletes have something to do with that. Rather than looking at gold medal wins, we looked at all medals wins, because these are considered to be winners as well. Looking at the various medal totals and athlete wins, we were able to properly conclude various relationships between the data points.**
* Summarize how you addressed this problem statement (the data used and the methodology employed).
  + **The problem statement was addressed through the process of combining various data points in tables and graphs, through this process it was more easily analyzed. The first step in the process was to group the Olympic games by year. From there I analyzed the changes in athletes over that time period. Which showed a clear increase over the past 100 years in athlete participation, it also showed the various dips that can be attributed to historical events i.e. WWI and WWII. Still utilizing the year group I then analyzed the increase in sport discipline over time. Which showed the increase in sports as time has gone on. Then year was utilized to see the various changes in gender over time as well. Showing the clear increase in women since the beginning.**

**The next point that was analyzed was the winners over time via both countries and athletes. Through looking at the data points it showed the top winning countries per athlete wins. As well as showing just overall the various winning countries. When analyzing the breakdown of top winners it shows clearly that the USA has historically been the top winner. The top winning countries were then broken down by sports to see if there have been countries that have historically dominated sports, which showed that the USA has dominated several sports. Proving that the USA is the top performing country in the Olympics.**

**The final subject that was analyzed was the top performing athletes. This was looked at through the particular athlete and sporting events they participated with. The important thing to note about this data outlook is that it provides you with an idea of who are repeat winners in various different events. The data shows the connection of top winning athletes in particular sports, and the table gives an outlook of their relationship with various countries.**

* Summarize the interesting insights that your analysis provided.
  + **Through the vast history of the Olympics, of over 100 years. The United States one of the still relatively young countries around the world has dominated.**
  + **Michael Phelps is the top performing Olympic athlete of all-time**
  + **Large name 1st world countries are typically the top performers**
  + **Gender has increased significantly over the years. There is still a gap with women, but they have caught up.**
  + **Through the plots you can see the significance of historical events impact on Olympic events. This can be seen with years like: Great Depression, WWI, WWII. This seemed to send women on an increased trajectory.**
* Summarize the implications to the consumer (target audience) of your analysis.
  + **This is shows various sports data points for Olympic fans. It allows fans to gain a better understanding of how their country has performed, as well as changes over time in the Olympic games. It also provides a great understanding of the various athletes and their performances in specific sports. Overall it shows fans that the top performing country has been the USA, and that their swimmer Michael Phelps has been the top athlete. This information is based off medal wins.**
* Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.
  + **The data utilized mainly categorical information. Which made analysis much more difficult. If someone was more experienced, they would’ve had an easier time analyzing the information. Going off of that information having more data would’ve made the analysis much easier as well. Not specifically number of athletes but, other information like weight, height, or athletic numbers/ performance. These data points would’ve made it easier to predict potential winners and outcomes. I would’ve been able to go further into depth about the specifics of various athletes and teams, and why they were successful.**
* In addition, submit your completed Project using R Markdown or provide a link to where it can also be downloaded from and/or viewed.

Name:   
Date:   
Title:   
  
**Section 1**

* Introduction
* Research questions
* Approach
* How your approach addresses (fully or partially) the problem.
* Data
* Required Packages
* Plots and Table Needs
* Questions for future steps.

**Section 2**

* How to import and clean my data
* What does the final data set look like?
* Questions for future steps.
* What information is not self-evident?
* What are different ways you could look at this data?
* How do you plan to slice and dice the data?
* 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?
* Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.
* Questions for future steps.

**Section 3**

* A story / narrative that emerged from your data. Follow this structure.
  + Introduction.
  + The problem statement you addressed.
  + How you addressed this problem statement
  + Analysis.
  + Implications.
  + Limitations.
  + Concluding Remarks