***DAP - Capstone Checklist***

* **Introduction**
* **Question - what your analysis is striving to answer**

**User Story:**

* Create a definition statement / question for your DAP Capstone example:
  + I want to build a data analysis that answers the following question about (insert some or other subject here): My question is this ...(insert your question here ...).
* Write a Definition of the Audience / type of personas that might be interested in this analysis
* What datasets do I need to help me answer the above question for my particular audience?
* Choose any publicly Internet-available dataset(s)
* What research do I need to do to find data sources for my Capstone?
* List resources researched: (links to websites / materials)
* Note what you like about each resource and why you chose it
* Name other reference sources you want to consult

**Your capstone: (Requirements)**

You will take your capstone project through the data Analytics Lifecycle, to show-case your achievements. You are required to use (1) Python (2) Excel or SQL, (3), and Tableau in this process. In your presentation explain how you used the above 3 tools in your processes for your capstone. (Where applicable explain how / when you used Pandas, Numpy, Matplotlib, Web Scraping, Jupyter Notebooks, Git/GitHub)

* 1000 cell minimum
  + Can be one dataset or multiple datasets that equal 1000 cells in total.
  + Example: 30 columns x 34 rows = 1020 cells
  + Collect and wrangle that data in a programmatic fashion (see the process below)
* Generate charts, graphs or reports to visualize some insightful aspects from the data
* What other materials could enhance my presentation around this subject? Images, tables, graphics, descriptive information ?
* And finally present your Capstone project to tell the Data Story
  + In your presentation illustrate how you performed a particular step in the data wrangling / analytics process.
  + Make sure that you’re answering how you questioned your analysis and what your analysis is striving to answer.
  + Include the biggest challenge you overcame in creating the capstone.

#### Step #1 – Discovery

* Find data that addresses your question.
* Become familiar with your data so that you know how you will end up using it.
* Identify trends, patterns and some data cells / sections that might cause issues in analysis.

#### Step #2 – Structuring

* Take your raw data and transform it to what you can work with.

Unstructured data is often text-heavy and contains things such as Dates, Numbers, ID codes, etc.

Example: - When using info scrapped from a website, you might parse HTML code, pull out what you need, and discard the rest.

#### Step #3 – Cleaning

* Removes outliers that can potentially skew your results when analyzing the data.
* Changes any null values and standardizes the data format to improve quality and consistency.
* Identifies duplicate values, standardizes systems of measurements, fixes structural errors and typos, and validates the data to make it easier to handle.

#### 

#### Step #4 – Enriching

Deciding if you need to add to the data by combining raw data with additional data from other sources.

Example: Combining two or more databases of customer information to fill in gaps in the data.

Enriching the data is an optional step that you only need to take if your current data doesn’t meet your requirements.

#### Step #5 – Validating

Making sure that the data that you have is of the quality necessary to complete your project.

* The rules of data validation require repetitive programming processes that help to verify the – Quality, Consistency, Accuracy, Security, and Authenticity of data.

#### Step #6 – Publishing

* Creating your analysis and presenting it to the public.
* You can deposit the data into a new architecture or database.
* We will display our data story using Tableau.
* Record yourself presenting your data. Presentation length should typically be at least 8 minutes, but no more than 20 minutes, in duration.

Watch these videos to see how to incorporate these items into your presentation and get great ideas:

<https://www.youtube.com/watch?v=os6DAkXuJU0>   
  
<https://youtu.be/ZEYEmlNYOq0>

**Suggested Links to find Data Sets:**

1. <https://www.geonames.org/>
2. <https://www.kaggle.com/>
3. <https://www.dataworld.com/>
4. <https://www2.ed.gov/about/inits/ed/edfacts/data-files/school-status-data.html>
5. <https://dataverse.harvard.edu/>
6. <https://data.census.gov/>
7. <https://www.dol.gov/agencies/whd/data>
8. <https://www.dol.gov/agencies/vets/latest-numbers>
9. <https://www2.ccwdata.org/web/guest/data-dictionaries>
10. <https://resdac.org/cms-data/files/hedis-rif-measures>
11. <https://data.gov/>

**Suggested Topics:**

* Immigrating to or from the U.S.
* Minimum Wage throughout the U.S. and/or decades
* Average age of individuals getting married
* Most common names (first / last names)
* Most watched Netflix shows
* Comparing the different streaming networks
* Salaries
* Wine costs
* Farming per region (different agriculture, soil required, etc.)
* Most effective diets
* Single family incomes
* Daylight savings time vs. Standard time (time change, health effects, opinions, etc.)
* Best rated vehicles
* Weather variations
* Sport with the most injuries / types of injuries
* Favorite kids toys throughout the decades
* Centenarian - People who live to be 100 years old (their typical diet, lifestyles, where most live, etc.)
* Holiday costs
* Fitbit data
* Is Tornado Alley changing