Task: Divide my code into these 5 sections: Step 1, Acquiring Data Step 2A, Exploring Data Step 2B, Pre-Processing Data Step 3, Analyze Data Step 4, Reporting Insights

***FINAL PROJECT GUIDELINES***

Several steps are outlined below, but you mustn't view this as an entirely linear process. Remember that the science component in data science is the creation of a hypothesis based on the exploration and testing of that hypothesis through analysis. You may need to go through many of these steps multiple times before you arrive at meaningful hypotheses or conclusions.

Step 1: Explore the dataset(s)

In this step, you should explore what is present in the data and how the data is organized. If using multiple datasets, you'll need to determine what common features allow you to merge the datasets. You'll want to answer the following questions:

1. Are there quality issues in the dataset (noisy, missing data, etc.)?

2. How will you clean and/or transform the raw data for analysis?

Step 2: Identify three research questions

Now that you better understand the data, you will want to form a research question that interests you. The research question should be broad enough to interest a reader but narrow enough to answer the question with the data. Some examples:

1. Too Narrow: What is the GDP of the U.S. for 2011? This is just asking for a fact or a single data point.

2. Too Broad: What is the primary reason for global poverty? This could be a Ph.D. thesis and would still be way too broad. What data will you use to answer this question? Given the variety of datasets, would it be defendable even if a single dataset offered an answer?

3. Perfect (just right): Can you use sentiment analysis on comments about movies on Twitter to predict its box office earnings? If you have or can obtain, tweets that refer to a variety of movies and you have their box office earnings, this is a question you can potentially answer well.

Based on the dataset, can you provide a brief description of the dataset and post three research questions, one of which can be answered using machine learning techniques like classification, clustering, or regression in Scikit-learn? One uses the Python package NLTK to do natural language processing on the description, and the final one must be answered by getting data from the web using Twitter API.

Step 3: Describe your dataset(s) and research question(s) in detail at the end. Revise them if necessary.

CHECKPOINT.

Step 4: Identify your research methods

Based on your research question, you can now describe how you will answer that question. Your methods will depend on the question being asked. For example, if you are looking for a relationship between two items (say, CO2 emissions and GDP), you may wish to use scatter plots and statistical correlation. To predict an outcome based on input data, you must identify the appropriate methods from the week on Machine Learning. Be sure to document, in your notebook, your research methods.

Step 5: Present your findings

In this step, you can begin to report your findings. What did you learn from the data, and how do your findings help answer your research question? Use visualizations to present these findings.

Step 6: Identify limitations to your conclusions

Rarely will a single data analysis conclusively answer a research question. Here, you need to identify possible limitations. For example, are your results limited to a certain area, city, or country? Are you making assumptions about the data which may or may not be valid (e.g., that students in one term are equally qualified as students in another)? Document these limitations.

Step 7: Identify future work

You may have clear next steps at the end of your analysis. For example, you may wish to perform the same analysis in another setting or think your findings lead to new questions. Feel free to document these next steps.

Step 8: Present your work!

Fill in the slide template below to present your project. Feel free to add more slides, but aim to keep your presentation below 20 total slides. We also recommend you attach a PDF of your Jupyter Notebook to this presentation in one PDF file to be uploaded. As you are writing, please follow the general guidelines for project reports and presentations below:

1. This is academic writing. Keep it formal and coherent as a self-contained entity. Anyone reading your presentation should understand the question, approach, and results fully.

2. Be professional. You should be comfortable giving this presentation to the general public, your boss, or your academic advisor.

3. Write for a diverse audience, including:

a. General Public: Reads only the title and abstract, looking for high-level points mainly for conversational purposes.

b. A company CEO: Reads introduction, research question, findings, and conclusions looking for business value and related details.

c. An academic advisor (or company CTO): Reads the full presentation AND your Jupyter Notebook paying particular attention to technical coherence, academic value, and technical data science strengths.

The final presentation should also be for the above-mentioned three audiences. Think of the diversity of the audience. The whole point is to tell a story - so you should motivate a reader to care based on the question you are exploring, answer that question clearly and concisely, provide an honest appraisal of your results, and give the reader valuable insights. Use charts whenever possible. Avoid slides with a lot of text and bullet points - break the slide into multiple slides when this happens. Be concise!

***FINAL PPT TEMPLATE***

1. Your Project Title

2. Abstract: Summarize your questions and findings in 1 brief paragraph (4-6 sentences max). Your abstract needs to include: what dataset, what question, what method was used, and the findings.

3. Motivation: Describe the problem you want to solve with the data. It may relate closely to your research question, but your goal is to make your audience care about the project/problem you are trying to solve. You need to articulate the problem you are exploring and why (and for whom) insight would be valuable.

4. Dataset(s): Describe your dataset(s) here. You should say what data is in the dataset, how much data is, and where you found the dataset (if applicable).

5. Data Preparation and Cleaning: At a high level, what did you need to do to prepare the data for analysis? Describe what problems, if any, did you encounter with the dataset?

6. Research Question(s): What research question do you aim to answer using the dataset? Be sure the research question is well defined (see project description for details).

7. Methods: What methods did you use to analyze the data, and why are they appropriate? Be sure to adequately but briefly describe your methods.

8. Findings: <Feel free to replicate this slide to show multiple findings>

Present your findings. Include at least one visualization in your presentation (feel free to include more). The visualization should be honest, accessible, and elegant for a general audience.

You need not come to a definitive conclusion but say how your findings relate to your research question.

9. Limitations: If applicable, describe limitations to your findings. For example, you might note that these results were true for British Premier League players but may not apply to other leagues because of differences in league structures.

Or you may note that your data has inherent limitations. For example, you may not have access to the number of Twitter followers per user, so you assume all users are equally influential. If you had the number of followers, you could weigh the impact of their tweet’s sentiment by their influence (# of followers).

10. Conclusions: Report your overall conclusions, preferably a conclusion per research question

11. Acknowledgements: Where did you get your data? Did you use other informal analyses to inform your work? Did you get feedback on your work from friends or colleagues? Etc. If you had no one give you feedback and you collected the data yourself, say so.

12. References: If applicable, report any references you used in your work. For example, you may have used a research paper from X to help guide your analysis. You should cite that work here. If you did all the work on your own, please state this.

1. What are the latest tweets mentioning the most popular Netflix shows? This could provide insights into public sentiment towards these shows.

3. What is the public sentiment toward the top-rated Netflix shows? This requires us to fetch data from the Twitter API for tweets related to the top-rated Netflix shows and perform sentiment analysis.

5. What is the public opinion about these shows or movies on Twitter? This would involve collecting tweets that mention the shows or movies in the dataset and performing sentiment analysis on those tweets to gauge public opinion.

6. Can we find the general sentiment (positive, negative, neutral) on Twitter about the most recent Netflix Original released? This would require getting data from the Twitter API, preprocessing the tweets, and performing sentiment analysis.

7. Based on recent tweets, can we analyze the audience's sentiment toward these Netflix titles? For this, we would need to gather data from Twitter using the Twitter API, then perform sentiment analysis on the tweets that mention the Netflix titles.

2. How do Twitter sentiments towards specific Netflix titles correlate with their release\_year, rating, or listed\_in categories? This would require collecting additional data from Twitter based on the titles in the dataset.

4. Can we perform sentiment analysis on the show or movie descriptions to determine whether the description conveys a positive, negative, or neutral sentiment? This can be achieved using text processing techniques and sentiment analysis tools in the NLTK package.

8. How do the most recent Netflix shows (based on date\_added) fare in terms of popularity on Twitter? This would involve getting data from Twitter API.

Based on the dataset, can you provide a brief description of the dataset and post three research questions, one of which can be answered using machine learning techniques like classification, clustering, or regression in Scikit-learn? One uses the Python package NLTK to do natural language processing on the description, and the final one must be answered by getting data from the web using Twitter API.

Based on the dataset, can you post three questions?

1. The first question is about using machine learning techniques like classification, clustering, or regression in Scikit-learn.

2. The second question is about using the Python package NLTK to do natural language processing on the description of the games.

3. The final question is about getting data from the web using Twitter API.

What are the top 5 countries producing the most content on Netflix?

What is the distribution of movie durations? Is there a specific duration range that is more common?

Can we predict the rating of a movie/show based on its description using natural language processing and machine learning techniques?

1. Machine Learning with Scikit-Learn: Can we predict whether a show is a movie or a TV show based on its description, using classification techniques in Scikit-learn?

Approach: We'll first pre-process the text data in the description column by cleaning it (removing special characters, converting it to lowercase, etc.), tokenizing it, and converting it into a numerical format that a machine learning model can use. We'll then train a classification model (for example, a logistic regression or support vector machine) using Scikit-learn to predict the type column based on the description. We can assess the model's performance using metrics such as accuracy, precision, recall, or F1-score.

Among these eight questions of Web Data using Twitter API, are there some of them that are overlapped, and you can merge them?

Handle these null values as follows:

1. To populate null values in the 'director' column based on the 'cast' column, we'll need to create a mapping of directors to casts and use this to fill in the missing values. This might not fill all the null values if some casts don't have a corresponding director in the data. For the remaining null values, we'll fill with 'Unknown.'

2. The null values are in the 'country' column. Similar to the 'director' column, we'll populate the null values based on the relationship with the 'director' column. If impossible, we'll fill the remaining nulls with 'Unknown.'

3. use backfill (bfill) to replace missing 'date\_added' values with the next value in the column

4. Drop the rows where 'duration' is missing.

5. use Scikit-Learn to create a predictive model that estimates 'rating' of a Netflix movie or TV show based on its attributes 'type,' 'director,' 'cast,' 'country,' 'date\_added,' 'release\_year,' 'duration,' and 'listed\_in.' And use this model to fill in the missing four values in the 'rating' column in our original dataset. (Do we need to exclude these four rows with missing 'rating' from both our training set and testing set?)