Capstone Project Overview

Apply what you’ve learned in data science by exploring and analyzing a dataset, either from the listed provided or a different one of your choosing. Your goal is to answer specific questions about the data, create visualizations, and share insights. You will submit one Jupyter notebook containing the requirements below.

Project Steps and Requirements

1. **Dataset Selection and Proposal**
    - Choose a dataset from the list provided or another approved dataset.
    - Define one main question to guide your analysis.
2. **Data Cleaning and Preparation**
    - Handle missing data, and identify any outliers.
    - Include at least two methods to clean or preprocess your data.

Project Requirements (cont'd)

3. **Exploratory Data Analysis (EDA)**
   - Use two or more visualizations to explore relationships or patterns.
   - Answer at least one question based on your EDA.
4. **Analysis and Insights**
   - Apply one analysis technique (e.g., regression or classification).
   - Summarize key findings in one or two paragraphs.
5. **Conclusion**
   - Describe your main insights and any surprising findings.

Suggested Guiding Questions for Datasets Below

1. **Iris Dataset**: What features help most in distinguishing iris species?
2. **Titanic Dataset**: Which factors most influence survival?
3. **Penguins Dataset**: How do penguin species differ in size and location?
4. **Wine Quality Dataset**: Which attributes correlate most with wine quality?
5. **Boston Housing Dataset**: What influences housing prices the most?
6. **Netflix Dataset**: How have genres changed over time?
7. **Students Performance**: What factors most affect test scores?
8. **Supermarket Sales**: What trends are seen in sales across days or times?
9. **NYC Taxi Rides**: What influences trip durations most?
10. **Chocolate Ratings**: What attributes correlate with higher chocolate ratings?

Rubric

- **Proposal** (10 points): Dataset and main question defined.
- **Data Cleaning** (15 points): Missing data handled, data cleaned effectively.
- **EDA** (20 points): Visualizations provided with clear analysis.
- **Analysis** (20 points): Analysis technique applied and explained.
- **Conclusion** (15 points): Key insights summarized effectively.
- **Presentation** (20 points): Well-organized and clear notebook.

Tips for Success

- Start by understanding your dataset thoroughly.
- Break down your tasks into manageable segments.
- Regularly document your findings and challenges.

Next Steps

Start by drafting a project proposal that includes:

- A brief description of the dataset.
- The main objectives of your project.
- Any initial hypotheses or questions you aim to explore.

Datasets

Here are some beginner-friendly datasets that are great for Python students learning data science:

1. Iris Dataset

- **Description**: Classic dataset for classification tasks involving different types of iris flowers.
- **Size**: Small (150 samples, 4 features).
- **Use Cases**: Classification, basic machine learning.
- **Source**: UCI Machine Learning Repository

2. Titanic Dataset

- **Description**: Contains information about passengers on the Titanic, used to predict survival.
- **Size**: Medium (891 samples, 12 features).
- **Use Cases**: Classification, exploratory data analysis (EDA), handling missing data.
- **Source**: Kaggle

3. Penguins Dataset

- **Description**: Data on penguin species, size measurements, and island locations.
- **Size**: Small.
- **Use Cases**: Classification, data visualization.
- **Source**: palmerpenguins

4. Wine Quality Dataset

- **Description**: Data on red and white wine quality based on physicochemical tests.
- **Size**: Medium (6,497 samples).
- **Use Cases**: Regression, classification.
- **Source**: UCI Machine Learning Repository

5. Boston Housing Dataset

- **Description**: Contains data on housing prices in Boston.
- **Size**: Medium (506 samples, 13 features).
- **Use Cases**: Regression, data visualization.
- **Source**: Kaggle

6. Netflix Movies and TV Shows Dataset

- **Description**: Information about Netflix titles, including genre, release year, and rating.
- **Size**: Medium.
- **Use Cases**: Data cleaning, visualization, exploratory analysis.
- **Source**: Kaggle

7. Students Performance Dataset

- **Description**: Data on students’ grades, gender, parental education, and test scores.
- **Size**: Small.
- **Use Cases**: Classification, correlation analysis.
- **Source**: Kaggle

8. Supermarket Sales Dataset

- **Description**: Transaction data from a supermarket.
- **Size**: Small.
- **Use Cases**: Time series analysis, data visualization.
- **Source**: Kaggle

9. NYC Taxi Rides Dataset

- **Description**: Data on taxi rides in New York City.
- **Size**: Large.
- **Use Cases**: Time series, spatial data analysis.
- **Source**: Kaggle

10. Chocolate Ratings Dataset

- **Description**: Data on different types of chocolate bars and their ratings.
- **Size**: Small.
- **Use Cases**: Data visualization, correlation analysis.
- **Source**: Kaggle