Course Two

Get Started with Python

Instructions

Use this PACE strategy document to record decisions and reflections as you work through this end-of-course project. You can use this document as a guide to consider your responses and reflections at different stages of the data analytical process. Additionally, the PACE strategy documents can be used as a resource when working on future projects.

Course Project Recap

Regardless of which track you have chosen to complete, your goals for this project are:

- Complete the questions in the Course 2 PACE strategy document
- Answer the questions in the Jupyter notebook project file
- Complete coding prep work on project's Jupyter notebook
- Summarize the column Dtypes
- Communicate important findings in the form of an executive summary

Relevant Interview Questions

Completing the end-of-course project will help you respond these types of questions that are often asked during the interview process:

- Describe the steps you would take to clean and transform an unstructured data set.

  To clean and transform an unstructured data set I would follow these steps:

  - **Data Understanding:**
    - **Assess Data Sources:** Identify the sources of the unstructured data (e.g., text files, social media feeds, emails).
    - **Explore Data:** Get a sense of the data by examining its structure, content, and quality.

  - **Data Cleaning:**

- **Remove Noise:** Eliminate irrelevant data such as stop words, punctuation, and special characters.
- **Standardize Data:** Convert data to a consistent format (e.g., lowercasing text, removing duplicates).
- **Handle Missing Values:** Address missing data by filling in, removing, or estimating values.
- **Correct Errors:** Identify and correct spelling mistakes, formatting issues, and other inaccuracies.

- **Data Transformation:**
  - **Tokenization:** Break down text data into smaller units, such as words or sentences.
  - **Normalization:** Convert data to a standard format, such as stemming or lemmatization for text data.
  - **Feature Extraction:** Derive meaningful features from the data (e.g., keywords, sentiment scores, named entities).
  - **Vectorization:** Convert text data into numerical format using techniques like TF-IDF, word embeddings, or one-hot encoding.
- **Data Integration:**
  - **Merge Data Sources:** Combine multiple data sources to create a unified dataset.
  - **Align Data:** Ensure that data from different sources is aligned and consistent.
- **Data Validation:**
  - **Validate Transformations:** Check the transformed data for accuracy and consistency.
  - **Perform Quality Checks:** Conduct quality checks to ensure data integrity and reliability.
- **Documentation:**
  - **Document Steps:** Keep detailed documentation of the cleaning and transformation process.
  - **Create Metadata:** Generate metadata to describe the cleaned and transformed data.

- What specific things might you look for as part of your cleaning process?
    - Locate the relevant information in the context of the analysis
    - Convert it into a consistent format:
        - identify missing data, duplicated entries, inconsistent data
        - identify noise, irrelevant data, errors and typos
    - Define the Data Types, normalize data, and check for integrity
    - Handle Categorical Data by encoding it into numerical values if required
    - Observe and understand the outlier values
- What are some of the outliers, anomalies, or unusual things you might look for in the data cleaning process that might impact analyses or ability to create insights?
    - Extreme Values:
        - Data points that are significantly higher or lower than the rest of the data. These can skew statistical analyses and models.
    - Inconsistent Data Entries:
        - Records that don't align with the expected format or values. For example, a numerical value where a categorical value is expected.
    - Unexpected Null Values:
        - Missing data in fields where you would normally expect complete information. This can indicate a problem with data collection or entry.
    - Temporal Anomalies:
        - Dates and times that don't make sense, such as future dates in historical data or timestamps that don't follow a logical sequence.
    - Duplicate Records:
        - Multiple entries for the same entity that can lead to overrepresentation in the data set.
    - Category Imbalance:
        - Categories within a variable that have disproportionately few or many records, which can affect the performance of certain models.
    - Irregular Patterns:
        - Unusual patterns or sequences that don't fit the expected trend. For example, a sudden spike in sales data without a clear reason.
    - Anomalous Relationships:

- Data points that don't follow the established relationships between variables. For example, a negative age value or a gender value that doesn't align with typical demographic data.

❖ Outlier Ratios:

- Ratios or calculated values that are significantly different from the majority. For example, an extremely high or low debt-to-income ratio in financial data.

❖ Uncommon Values:

- Rare values within a categorical variable that can indicate data entry errors or special cases that need to be handled separately.

❖ Geospatial Anomalies:

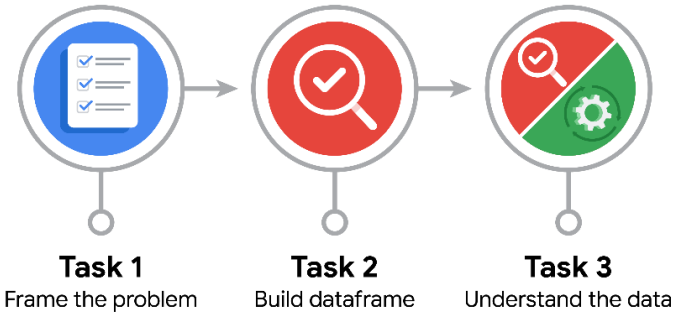
- Location data that doesn't make sense, such as coordinates that fall outside the expected geographical area.

❖ Sensor or Instrument Errors:

In sensor data, look for readings that indicate malfunctioning or calibration issues.

Reference Guide

This project has three tasks; the visual below identifies how the stages of PACE are incorporated across those tasks.



Data Project Questions & Considerations

PACE: Plan Stage

- How can you best prepare to understand and organize the provided information?

First is important to get familiar with the TikTok application, create an account and understand the upload and the search process, Review or create a Data dictionary of the fields and ask for the business rules on handling the claims

- What follow-along and self-review codebooks will help you perform this work?

Panda fundamentals
DataFrames
Masking and Grouping

- What are some additional activities a resourceful learner would perform before starting to code?

Get examples on the manual processing of different claim types to understand the process flow

Gather the function resources, conduct research on Python functions and classes that will be useful or necessary for the project

Use Gen AI tools to prompt for recommendations on how to tackle the problem, compare against my original idea, leverage from this information

PACE: Analyze Stage

- Will the available information be sufficient to achieve the goal based on your intuition and the analysis of the variables?

Yes, the dataset contains the detailed transcription of the video text, plus the necessary variables to process the claims:

```
#   Column                  Non-Null Count  Dtype
---  ------                 --------------  -----
 0   #                       19382 non-null  int64
 1   claim_status            19084 non-null  object
 2   video_id                19382 non-null  int64
 3   video_duration_sec      19382 non-null  int64
 4   video_transcription_text 19084 non-null  object
 5   verified_status         19382 non-null  object
 6   author_ban_status       19382 non-null  object
 7   video_view_count        19084 non-null  float64
 8   video_like_count        19084 non-null  float64
 9   video_share_count       19084 non-null  float64
 10  video_download_count    19084 non-null  float64
 11  video_comment_count     19084 non-null  float64
```

- How would you build summary dataframe statistics and assess the min and max range of the data?

I would use the pandas function describe() which acts of dataframes and provides:

Count: The number of non-null values.
Mean: The average of the values.
Std: The standard deviation of the values.
Min: The minimum value.
25%: The 25th percentile.
50% (Median): The 50th percentile or median.
75%: The 75th percentile.
Max: The maximum value.

- Do the averages of any of the data variables look unusual? Can you describe the interval data?

	#	video_id	video_duration_sec	video_view_count	video_like_count	video_share_count	video_download_count	video_comment_count
count	19382.000000	1.938200e+04	19382.000000	19084.000000	19084.000000	19084.000000	19084.000000	19084.000000
mean	9691.500000	5.627454e+09	32.421732	254708.558688	84304.636030	16735.248323	1049.429627	349.312146
std	5595.245794	2.536440e+09	16.229967	322893.280814	133420.546814	32036.174350	2004.299894	799.638865
min	1.000000	1.234959e+09	5.000000	20.000000	0.000000	0.000000	0.000000	0.000000
25%	4846.250000	3.430417e+09	18.000000	4942.500000	810.750000	115.000000	7.000000	1.000000
50%	9691.500000	5.618664e+09	32.000000	9954.500000	3403.500000	717.000000	46.000000	9.000000
75%	14536.750000	7.843960e+09	47.000000	504327.000000	125020.000000	18222.000000	1156.250000	292.000000
max	19382.000000	9.999873e+09	60.000000	999817.000000	657830.000000	256130.000000	14994.000000	9599.000000

c

PACE: Construct Stage

Note: The Construct stage does not apply to this workflow. The PACE framework can be adapted to fit the specific requirements of any project.

PACE: Execute Stage

- Given your current knowledge of the data, what would you initially recommend to your manager to investigate further prior to performing exploratory data analysis?

Understanding the outlier values in **video_view_count** and **video_like_count**

- What data initially presents as containing anomalies?

There are two cases where the mean is significantly higher than the median: video_view_count and video_like_count there are outlier values in these two fields that need attention.

- What additional types of data could strengthen this dataset?

Timestamp

detailed analysis of video_transcription_text:

Data Transformation:

- Tokenization: Break down text data into smaller units, such as words or sentences.
- Normalization: Convert data to a standard format, such as stemming or lemmatization for text data.
- Feature Extraction: Derive meaningful features from the data (e.g., keywords, sentiment scores, named entities).
- Vectorization: Convert text data into numerical format using techniques like TF-IDF, word embeddings, or one-hot encoding.