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?

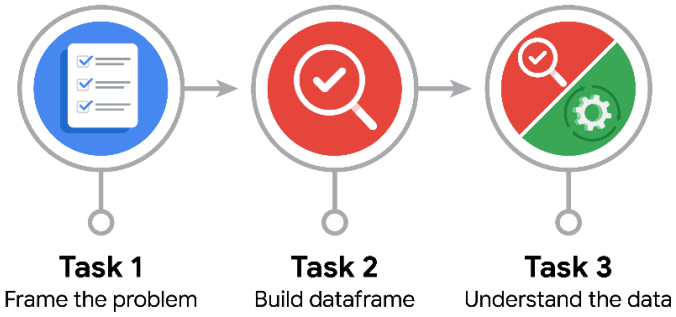
    - Extreme Values:
        - Data points that are significantly higher or lower than the rest of the data. These can skew statistical analyses and models.
    - Inconsistent Data Entries:
        - Records that don't align with the expected format or values. For example, a numerical value where a categorical 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?

Get familiar with the WAZE application, use it for a couple rides, review the data dictionary, read the project relevant documentation

- 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?

Use Pandas summary statistics functions and generative AI to support the exploratory data analysis in the data and find important observations, correlations or outlier values

PACE: Analyze Stage

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

The dataset has sufficient variables to achieve the goal, but it is necessary to clarify if the initial dataset represents a particular type of driver other than casual drivers. This has important implications since users other than casual riders have different needs and, they are always connected to the road conditions

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

By dividing the dataset in two groups, group with null values and group with non-null values, then I would apply the pandas .describe function to each group. I can divide the set it groups by using the mask function, as used in the attached python notebook of this preliminary analysis.

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

The two tables below shows that there are some fields where the mean is higher than the median, suggesting that the fields indicated in yellow have outlier values for distance and number of trips. This could be indicative of a frequent and intense driving pattern

Rows with null values:

[5]:

	ID	sessions	drives	total_sessions	n_days_after_onboarding	total_navigations_fav1	total_navigations_fav2	driven_km_drives	duration_minutes_drives	activity_days	driving_days
count	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000
mean	7405.584286	80.837143	67.798571	198.483348	1709.295714	118.717143	30.371429	3935.967029	1795.123358	15.382857	12.125714
std	4306.900234	79.987440	65.271926	140.561715	1005.306562	156.308140	46.306984	2443.107121	1419.242246	8.772714	7.626373
min	77.000000	0.000000	0.000000	5.582648	16.000000	0.000000	0.000000	290.119811	66.588493	0.000000	0.000000
25%	3744.500000	23.000000	20.000000	94.056340	869.000000	4.000000	0.000000	2119.344818	779.009271	8.000000	6.000000
50%	7443.000000	56.000000	47.500000	177.255925	1650.500000	62.500000	10.000000	3421.156721	1414.966279	15.000000	12.000000
75%	11007.000000	112.250000	94.000000	266.058022	2508.750000	169.250000	43.000000	5166.097373	2443.955404	23.000000	18.000000
max	14993.000000	556.000000	445.000000	1076.879741	3498.000000	1096.000000	352.000000	15135.391280	9746.253023	31.000000	30.000000

Rows without null values

	ID	sessions	drives	total_sessions	n_days_after_onboarding	total_navigations_fav1	total_navigations_fav2	driven_km_drives	duration_minutes_drives	activity_days	driving_days
count	14299.000000	14299.000000	14299.000000	14299.000000	14299.000000	14299.000000	14299.000000	14299.000000	14299.000000	14299.000000	14299.000000
mean	7503.573117	80.623820	67.255822	189.547409	1751.822505	121.747395	29.638296	4044.401535	1864.199794	15.544653	12.182530
std	4331.207621	80.736502	65.947295	136.189764	1008.663834	147.713428	45.350890	2504.977970	1448.005047	9.016088	7.833835
min	0.000000	0.000000	0.000000	0.220211	4.000000	0.000000	0.000000	60.441250	18.282082	0.000000	0.000000
25%	3749.500000	23.000000	20.000000	90.457733	878.500000	10.000000	0.000000	2217.319909	840.181344	8.000000	5.000000
50%	7504.000000	56.000000	48.000000	158.718571	1749.000000	71.000000	9.000000	3496.545617	1479.394387	16.000000	12.000000
75%	11257.500000	111.000000	93.000000	253.540450	2627.500000	178.000000	43.000000	5299.972162	2466.928876	23.000000	19.000000
max	14998.000000	743.000000	596.000000	1216.154633	3500.000000	1236.000000	415.000000	21183.401890	15851.727160	31.000000	30.000000

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?

I would ask my manager to investigate the initial dataset, how was it extracted from the database, if there were particular filters or parameters in the initial query that might have biased the dataset to a particular driver type.

- What data initially presents as containing anomalies?

The Driven_kilometer_drives field has a maximum value of 21,183 km. More than half the circumference of earth!

- What additional types of data could strengthen this dataset?

User occupation or driver type: Casual, sales person, transportation, delivery, truck driver, Emergency response, etc.

Km driven under heavy traffic, slow condition. If the application could register the number of kilometers the driver got stuck in traffic due to heavy traffic.