

Automatidata PACE Strategy Document III

Exploratory Data Analysis

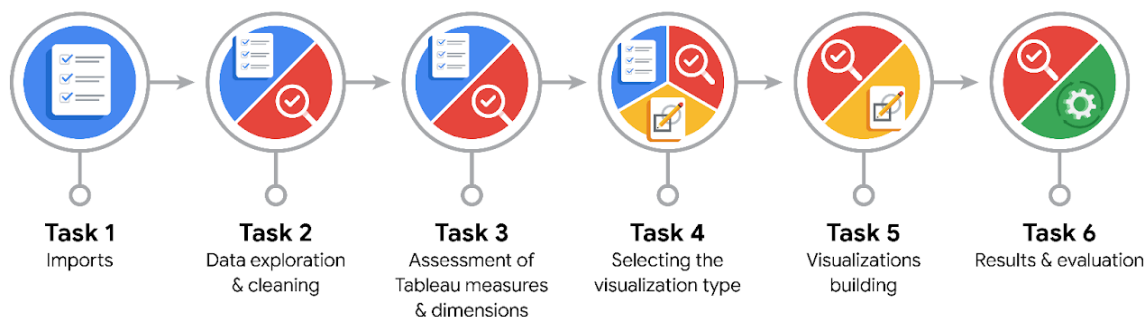
Introduction

PACE stands for Plan, Analyze, Construct and Execute. It is a framework that illustrates the foundation and structure for data analysis projects and each letter represents an actionable stage in a project. The stage “Plan” involves the definition of the project scope, the research of business data and the workflow development. The stage “Analyze” involves data scrubbing, data conversion and database formatting. The stage “Construct” involves building models and machine learning algorithms and selecting a modeling approach. The stage “Execute” involves the presentation of results to decision-makers, stakeholders and others in order to receive feedback. This framework is built upon an iterative cycle where each stage may reveal new insights, requiring the return to earlier stages. A PACE strategy document is used to record decisions and reflections at different stages of the data analytical process. It typically includes the definitions of roles and actions to ensure clarity and accountability.



Purpose

Automatidata is consulting for the New York City Taxi and Limousine Commission (TLC). New York City TLC is an agency responsible for licensing and regulating New York City's taxi cabs and for-hire vehicles. The agency has partnered with Automatidata to develop a regression model that helps estimate taxi fares before the ride, based on data that TLC has gathered. For this stage of the project, we identify six main tasks that are presented in the following visual.





Considerations



PACE: Plan Stage

- What methods are best for identifying outliers?

We use *NumPy* to calculate the mean and median and understand the data range. We create boxplots and histograms to visualize and observe how the data is distributed. We also create scatter plots which allow us to observe trends, patterns and outliers between variables.

- How do we decide whether to keep or remove outliers from future models?

There are three main ways to handle outliers: retaining them, removing them or reassigning values. The choice depends on the nature of the outliers and the model's assumptions. If the outliers are clearly errors like typos and the data will be used in modeling or machine learning, removing them is often best (though this is the least common option). If the dataset is small or used for modeling, reassigning values might be necessary. If the data will be used for EDA or the model is resistant to outliers, it is usually best to keep them.

- Which columns and variables in the dataset are most critical for our analysis and final deliverable?

Relevant variables include trip distance, total amount, tip amount, vendor and passenger count, which are essential for predicting taxi fares and analyzing patterns.

- What are the units of measurement for each variable?

Distances are measured in miles, fare amounts in dollars and time-related variables (like pickup and drop-off) are recorded in date-time format.

- What are our initial assumptions about the data that will guide our exploratory data analysis and what do we expect to confirm or adjust through our findings?

It is assumed that fare prices will correlate strongly with trip distance and duration, while tip amounts might vary by vendor. This will need verification through further analysis.

- Is there any missing or incomplete data in the dataset?

Some data points are incomplete, such as rides with zero passengers or a distance of 0 miles.

- Is the dataset consistent in terms of format?

Most of the data is consistently formatted, but certain anomalies (like 0-mile trips) need to be addressed.

- What exploratory data analysis techniques will we need to use to start this activity?

Key EDA practices include summarizing statistics, identifying outliers (using box plots or scatter plots) and visualizing distributions through histograms.



PACE: Analyze Stage

- What actions should be taken during EDA to ensure the project reaches its objectives?

We begin by cleaning the data, identifying outliers and summarizing the key variables (e.g. using descriptive statistics and visualizations) to detect patterns and outliers.

- Is there a need to combine more datasets as part of the EDA process and what structuring tasks (such as filtering or sorting) are necessary for the current data?

The dataset appears sufficient for now, but any additional external data (e.g. location information) could enhance the analysis. Structuring will involve filtering for outliers and sorting by key variables like trip distance and total cost.

- What initial ideas do we have about which types of visualizations might be most effective for our audience?

Scatter plots and box plots will be ideal for showing the relationship between fare amounts and trip distances, while bar charts can highlight temporal trends (e.g. rides per day or month).



PACe: Construct Stage

- What visualizations, machine learning models or other outputs will be essential to achieve the project's goals?

Box plots, histograms, scatter plots, bar charts and possibly a regression model to predict fare prices.

- What steps are needed to create the necessary visualizations?

Data cleaning, filtering outliers and generating visualizations like scatter plots (trip distance against total fare) and bar charts (distribution of rides by time or vendor).

- Which variables should be prioritized in the data visualizations for this project?

Trip distance and total amount are the most critical for visualizing the relationship between distance and fares.

- Returning to the planning stage, how will we address any missing data that is identified?

Missing data, like zero passenger counts or zero trip distances, will either be removed or investigated for potential correction before modeling.




PACe: Execute Stage

- What key insights have we gathered from your EDA and visualizations?

Outliers like trips with zero miles need addressing and there is a clear correlation between trip distance and fare. Most trips are under two miles, with a right-skewed fare distribution.

- What recommendations would we provide to the organization based on our findings?

Trips with zero distance or zero passengers should be investigated, as these could distort model accuracy. Focusing on high-traffic locations like airports and tourist spots may yield deeper insights.

- 
- Based on our current knowledge of the data and visualizations, what additional research questions could be investigated for the team?

Patterns in ride times, peak periods or vendor-specific behaviors (e.g. higher tips for certain vendors) could be further investigated.

- How do we plan to present these visualizations to different stakeholders or audiences?

Dashboards like Tableau provide a clear summary of key trends and patterns, tailoring the presentation for managers or technical audiences. In general, for non-technical audiences, we should use simplified bar charts and accessible Tableau dashboards with clear explanations. For data-savvy stakeholders, we can present detailed boxplots, histograms and correlation analyses to explain deeper insights into data patterns.