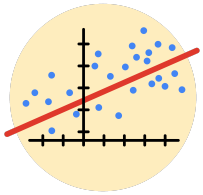
Course Five

Regression Analysis: Simplifying Complex Data Relationships



Instructions

Use this PACE strategy document to record decisions and reflections as you work through this end-of-course project. As a reminder, this document is a resource that you can reference in the future, and a guide to help you consider responses and reflections posed at various points throughout 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 5 PACE strategy document
- ☐ Answer the questions in the Jupyter notebook project file
- ☐ Build a multiple linear regression model
- ☐ Evaluate the model
- ☐ Create an executive summary for external stakeholders

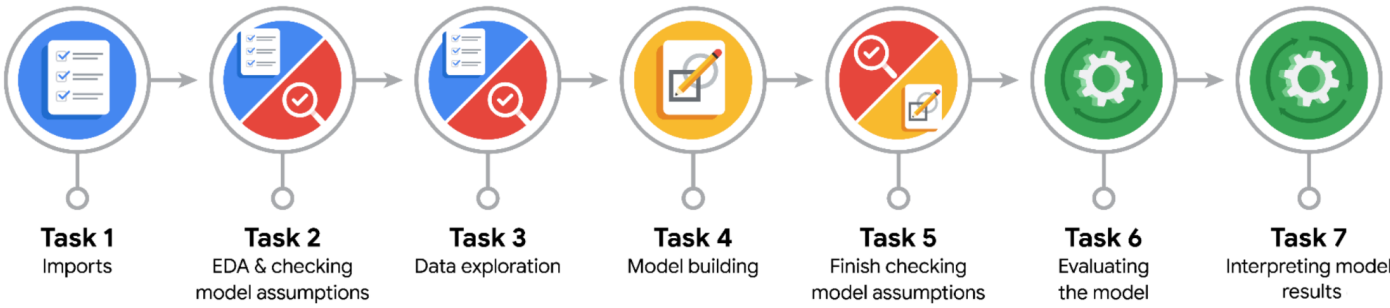
Relevant Interview Questions

Completing the end-of-course project will empower you to respond to the following interview topics:

- Describe the steps you would take to run a regression-based analysis
- List and describe the critical assumptions of linear regression
- What is the primary difference between R^2 and adjusted R^2?
- How do you interpret a Q-Q plot in a linear regression model?
- What is the bias-variance tradeoff? How does it relate to building a multiple linear regression model? Consider variable selection and adjusted R^2.

Reference Guide

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



Data Project Questions & Considerations

PACE: Planning Stage

- What are you trying to solve or accomplish?

The purpose of this task is to develop a multiple linear regression model to predict the trip's duration from other variables contained in the dataset of the company.

- What are your initial observations when you explore the data?

The trip's durations are not explicit in the dataframe, so it will be necessary to calculate it from the hour of pick up and the drop off time. For this purpose I will use the datetime function to perform a calculation with datetype

PACE: Analyzing Stage

- What are some purposes of EDA before constructing a multiple linear regression model?

It is necessary to obtain better performance and precision at the moment of using the model to predict. In addition, it is essential to prevent bias in the distribution of the data.

- Do you have any ethical considerations at this stage?

Do the username and behaviors of the registers express patterns exceptional to each person in the data? Are there variables that can be affected by the predispositions of the interface of the app?

PACE: Constructing Stage

- Do you notice anything odd?

The outliers do not let see the distribution of the data and its behavior with others, so it can affect to the performance of the model

- Can you improve it? Is there anything you would change about the model?

The percentiles included in the model can be evaluated so the distribution is more normal.

- What resources do you find yourself using as you complete this stage?

Boxplots and functions related to quartiles.

PACE: Execute Stage

- What key insights emerged from your model(s)?

The elimination of outputs improved the effectiveness of the model. At the same time, it is important to consider fine tuning.

- What business recommendations do you propose based on the models built?

It is important to collect more data to perform a 2nd stage of reliability analysis for the price and duration.

- Do you think your model could be improved? Why or why not? How?

It would be necessary to show the model's coefficients and evaluation of R adjusted analysis.

- Do you have any ethical considerations at this stage?

It is important to evaluate the methods of data collection, so it cannot cause a possible bias or misunderstanding.