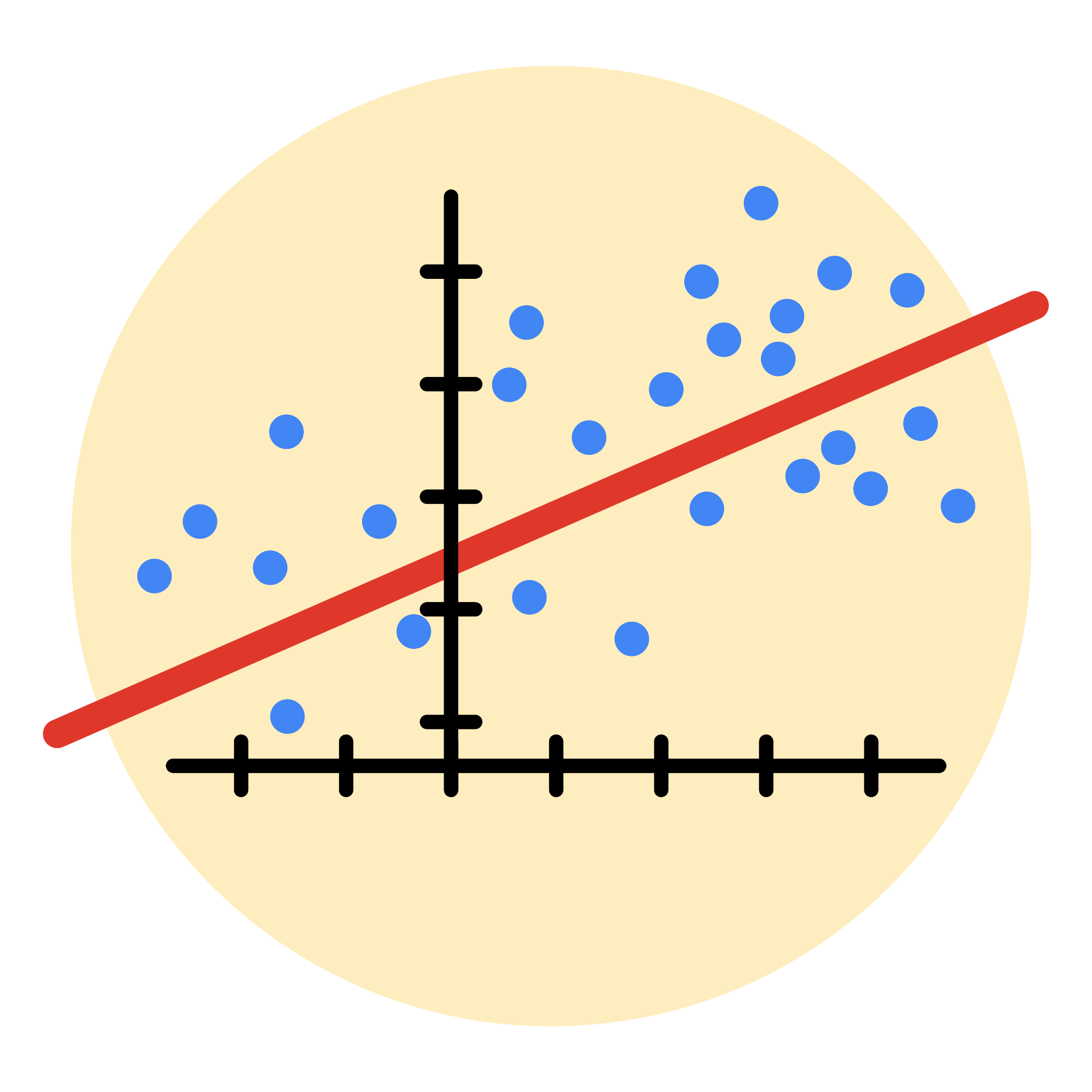
**Course Five**

# 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 team members

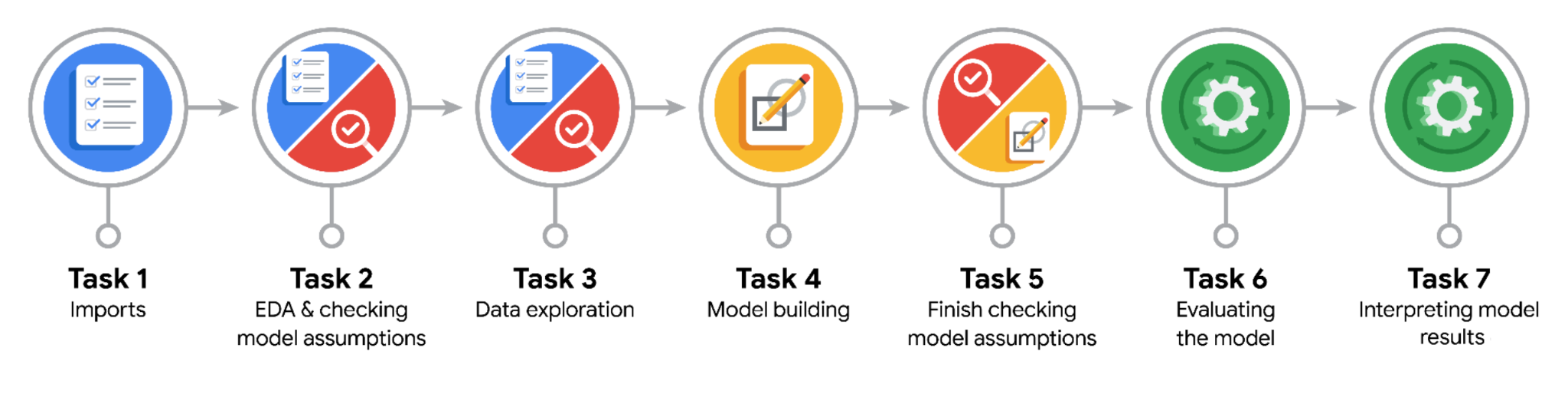
# Relevant Interview Questions

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](https://www.digitalvidya.com/blog/assumptions-of-linear-regression/)
* What is the primary difference between R2 and adjusted R2?
* 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 R2.

**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: Plan Stage**

* Who are your external stakeholders for this project?

New York City Taxi and Limousine Commission

* What are you trying to solve or accomplish?

The New York City Taxi & Limousine Commission contracted Automatidata to predict taxi cab fares. In this part of the project, the Automatidata data team created the deliverable for the original ask from their client: a regression model.

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

Initial observations include the dataset size of 22699 rows and 18 columns with zero missing values. The pickup time and dropoff time fields are not in the right format and after plotting the boxplot get to know there are outliers in the data.

* What resources do you find yourself using as you complete this stage?
* Python libraries like seaborn, matplotlib, sklearn etc.

**PACE: Analyze Stage**

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

EDA helps us to detect outliers, multicollinearity, statistical values like mean, median, mode, quantiles,

And Distribution of the multiple columns.

* Do you have any ethical considerations in this stage?

No.

**PACE: Construct Stage**

* Do you notice anything odd?

Not much as we need to create some fields like duration, mean\_duration, mean\_distance, rush\_hour fields and some outliers in few columns which handled by imputing the IQR values.

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

Yes, we can improve it further by Requesting additional data from under-represented itineraries.

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

Python libraries like seaborn, matplotlib, sklearn etc.

**PACE: Execute Stage**

* What key insights emerged from your model(s)?
* The feature with the greatest effect on fare amount was ride duration, which was not unexpected. The model revealed a mean increase of $7 for each additional minute, however, this is not a reliable benchmark due to high correlation between some features.
* What business recommendations do you propose based on the models built?

The model provides a generally strong and reliable fare prediction that can be used in downstream modeling efforts.

* To interpret model results, why is it important to interpret the beta coefficients?

Because Beta coefficient will let us know on increasing 1 miles how much fare amount will increase.

* What potential recommendations would you make?
* The New York City Taxi and Limousine commission can use these findings to create an app that allows users (TLC riders) to see the estimated fare before their ride begins.
* Do you think your model could be improved? Why or why not? How?
* Yes, we can improve it further by Requesting additional data from under-represented itineraries.
* What business/organizational recommendations would you propose based on the models built?

The model revealed a mean increase of $7 for each additional minute, however, this is not a reliable benchmark due to high correlation between some features.

* Given what you know about the data and the models you were using, what other questions could you address for the team?

How we can further increase accuracy of the model.

* Do you have any ethical considerations at this stage?

No