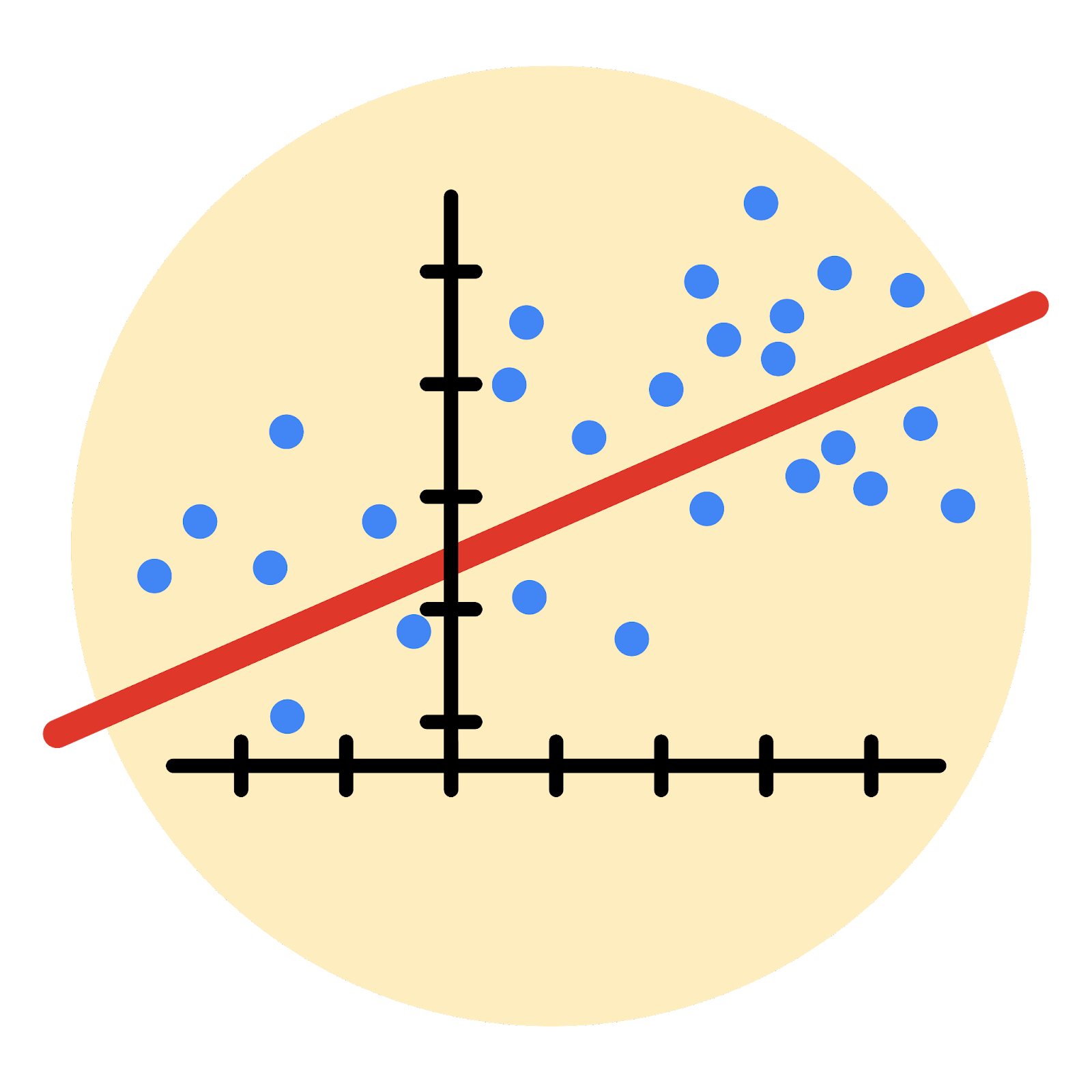
**Course Five**

# **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

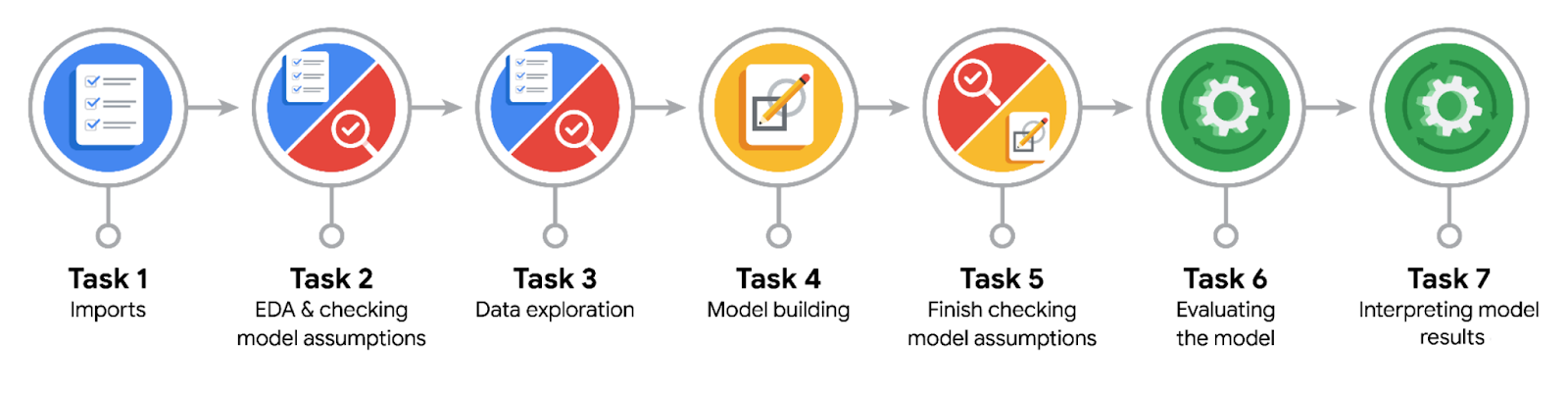
# **Relevant Interview Questions**

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?

The external stakeholders for this project are the customers of the NYC TLC.

* What are you trying to solve or accomplish?

I am trying to create a regression model that details the relationship between ride fare and multiple variables, such as trip distance and number of passengers.

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

There are three important variables: ride fare, trip distance, and trip duration.

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

Jupyter Notebooks and Python libraries.

**PACE: Analyze Stage**

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

Identifying variables that are needed to estimate the value of the dependent variable, fare amount.

* Do you have any ethical considerations at this stage?

Ensuring that the data is anonymous, providing privacy for the customers involved in the data.

**PACE: Construct Stage**

* Do you notice anything odd?

There are zeros in the trip distance column, and negatives in the fare amount column.

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

Yes. Imputation can reset the values of these outliers, removing them from the dataset.

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

The seaborn package for different graphs, such as scatterplots and heatmaps.

**PACE: Execute Stage**

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

The variable with the strongest influence on fare amount is mean distance. This suggests that the farther a taxi travels, the more profits can be made.

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

To increase the average distance customers are driven and potentially reduce trip duration, to improve tips.

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

The coefficients can give you a good idea of which factors have the greatest effect on the dependent variable.

* What potential recommendations would you make?

Determining if reducing trip duration improves profits.

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

Yes. Determining the nature of the mysterious negative outliers could determine whether or not those rows need to be removed, which may change the result of the model.

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

Does mean duration have a positive or negative impact of fare amount?

* Do you have any ethical considerations at this stage?

No. The data has already been anonymized.