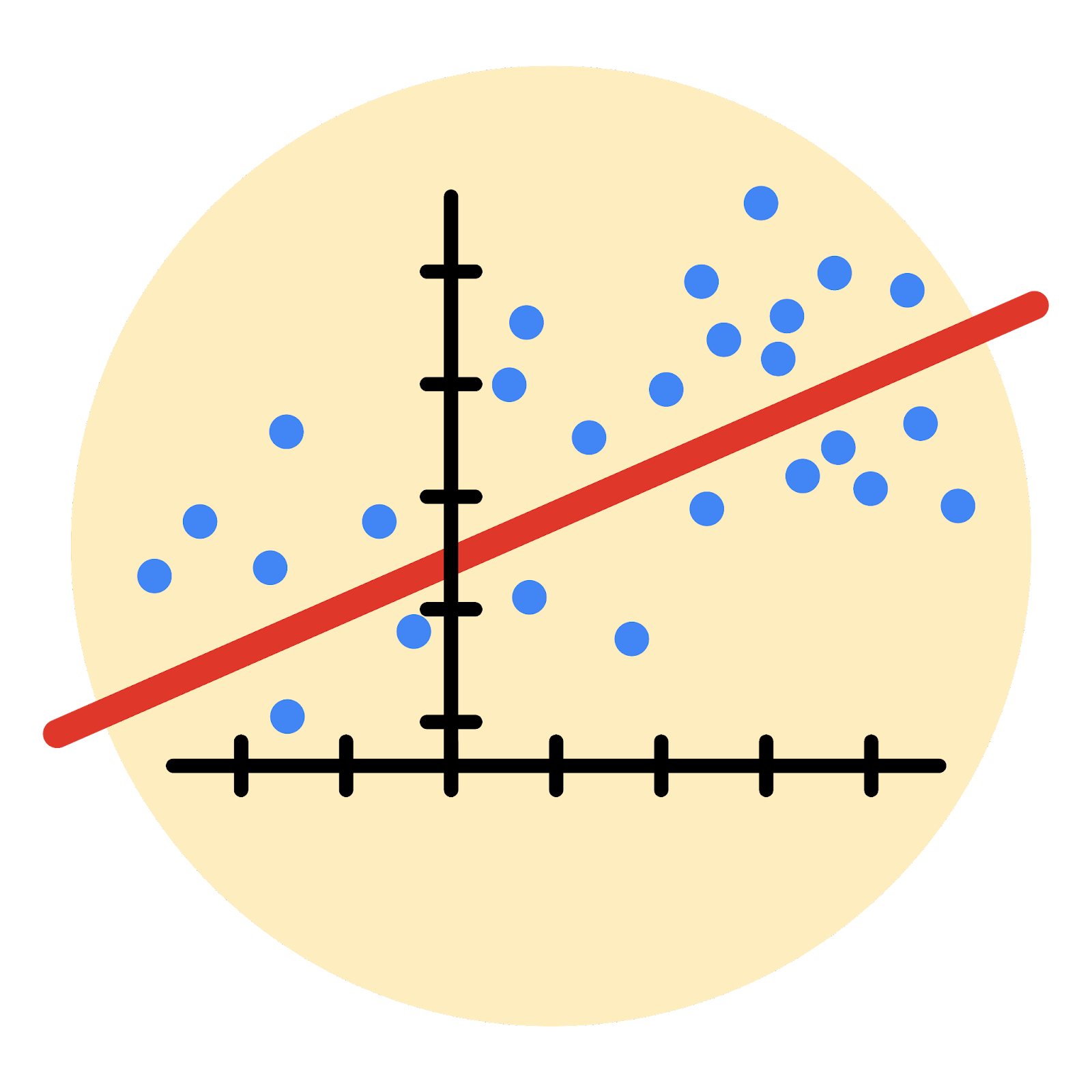
**Course Five**

# **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

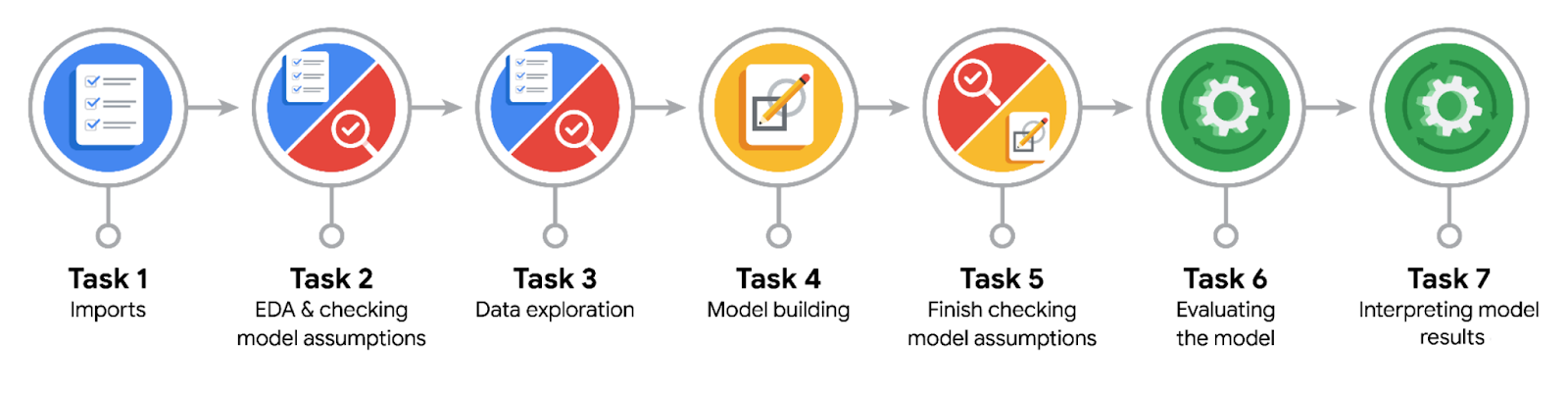
# **Relevant Interview Questions**

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 TLC Team and passengers

* What are you trying to solve or accomplish?

I am working on developing a regression model to predict the fare amount for passengers based on a given dataset that includes several independent variables that might be correlated with the fare amount.

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

My initial observations suggested that distance and the time taken by passengers to reach their destinations might be correlated with fare amounts. By examining the dataset’s statistics, it appears that the fare amount data contains outliers. The mean is 13.03, and the 75th percentile is 14.5, but the maximum value is 999.99.

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

I am using Python libraries, Pandas and Numpy to source and explore the dataset.

**PACE: Analyze Stage**

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

I will gain deeper insights into the data and be able to formulate and test assumptions to develop a robust regression model.

* Do you have any ethical considerations at this stage?

My ethical considerations include the possibility that drivers might decline some passenger requests based on fare amount predictions. Additionally, I am concerned about ensuring that the collected data does not violate passenger privacy.

**PACE: Construct Stage**

* Do you notice anything odd?

Some fare amounts are negative, which doesn’t make sense, and there are instances where the duration is 0, even though it should be greater than 0.

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

I would investigate the reasons behind these anomalies to determine if the odd datasets are valid. If they are not, I can replace them with the median or mean values, or remove them entirely.

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

I am using the Python libraries Pandas, Numpy, and Seaborn to explore and clean dataset.

**PACE: Execute Stage**

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

Key insights from the model indicate that the independent variables, distance and duration, have positive correlations with the dependent variable, fare amount.

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

Longer trips, as measured by distance and duration, tend to generate higher fares. Focusing on attracting customers for longer journeys could be a strategy to increase revenue.

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

Beta coefficients represent the change in the dependent variable for a one-unit change in the independent variable, holding all other independent variables constant. This allows to understand the magnitude and direction of the relationship between each predictor and the outcome.

* What potential recommendations would you make?

I suggest collecting and analyzing passenger distance data to uncover potential correlations that could help optimize fare structures.

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

The inclusion of more variables that are closely tied to fare amount would likely result in a more robust and predictive model.

* What business/organizational recommendations would you propose based on the models built?

Longer trips, as measured by distance and duration, tend to generate higher fares. Focusing on attracting customers for longer journeys could be a strategy to increase revenue.

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

Distance and duration positively correlate with fare amount, but I believe there’s an opportunity to introduce value-added services to further increase revenue.

* Do you have any ethical considerations at this stage?

Potential ethical concerns include the possibility of discriminatory practices, such as prioritizing long-distance passengers, and ensuring the protection of passenger data privacy during the collection process.