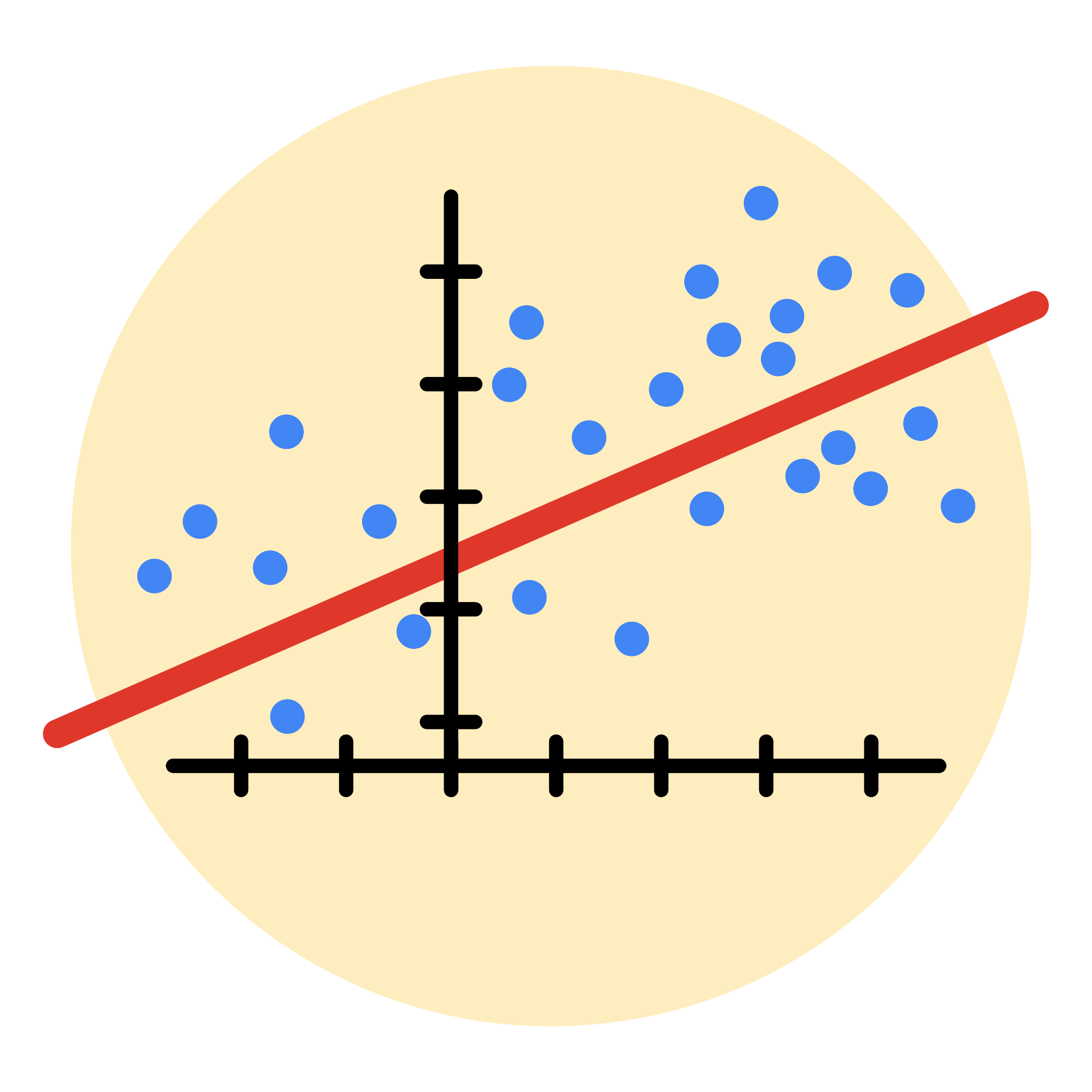
**Course Five**

# 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

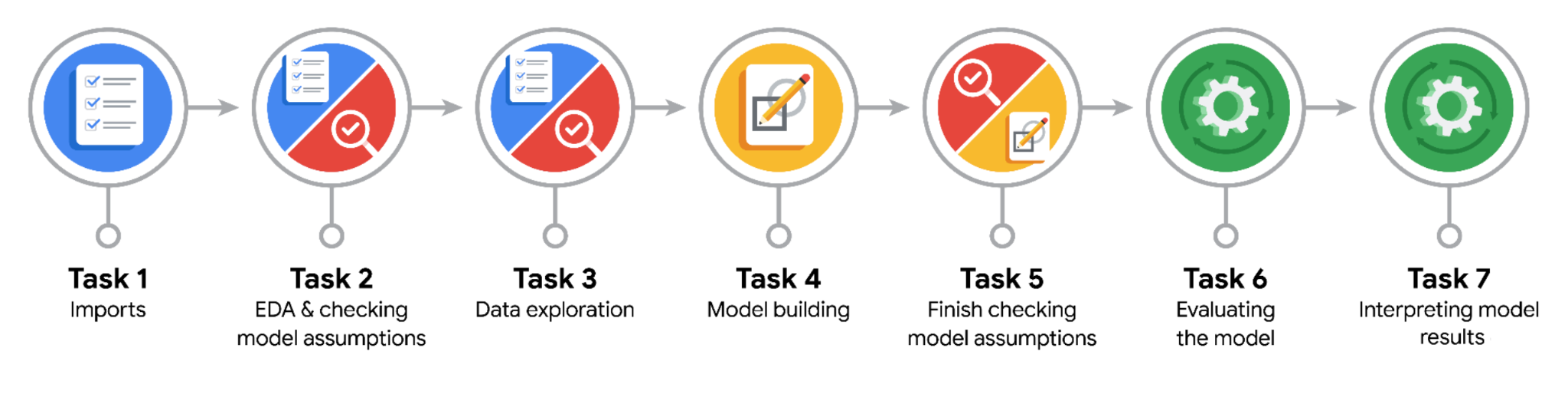
# Relevant Interview Questions

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?

Titus Nelson – Operations Manager, New York TLC (Direct technical stakeholder)

Juliana Soto – Finance and Administration Department

* What are you trying to solve or accomplish?

Building a regression model to estimate fare amount.

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

No null values observed.

No duplicates.

Data has 2269 fields and 18 observations.

Outliers are there with unusually high and negative values in the fields of fare amounts, trip\_distance, duration.

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

Python libraries such as Pandas.

**PACE: Analyze Stage**

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

EDA has the following purposes: -

1. Identifying outliers and investigating them.

2. Finding any missing values which can lead to unbalanced data.

3. Checking for any Multicollinearity among key variables i.e. checking their distributions.

4. Feature engineering of some variables.

* Do you have any ethical considerations in this stage?

Ethical considerations could be doing the necessary checks such as any duplicates, null or missing values. Not to create any anecdotes when either removing the outliers with unusual values. Keeping in consideration the integrity of data while going through the process of feature engineering.

**PACE: Construct Stage**

* Do you notice anything odd?

Variance was 86.8% explained in the test data which meant model was not overfit. Having a MAE of 2 would mean the predicted values are off by 2 than the actual observed values.

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

I would have removed the unusual values of the three key variables considered here: fare\_amount, duration, trip\_distance instead of standardizing them and making the dataset more unbalanced as these unusual values were significant.

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

Python libraries such as sklearn.model\_selection, sklearn.metrics, sklearn.linear\_model, sklearn.preprocessing.

**PACE: Execute Stage**

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

Ride duration has the most influence on fare amount.

This model can be used to predict fares beforehand.

The predictions of the model are reliable.

Correct the outliers which are over and under threshold i.e. the iqr factor.

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

A business recommendation would be setting a minimum ride duration necessary to book a trip at a particular convenience taking into consideration the cost of the ride. Having a particular ride duration, fare amounts would never be negative or 0 which was included in the model.

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

It is important to interpret beta coefficients because they show the direction and strength of a predictor’s effect on an outcome, showing which variables are most significant in predicting the outcome variable.

* What potential recommendations would you make?

Recommendations such as how integral the data is, how the fare amounts were reported and via what systems. Taking into consideration all of the payments were existing for a ride, if there is any distance covered with payments type except then 1 and 2, how is it reported.

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

Yes, it can be improved with a more realistic data set. It was because the dataset had unexplained values, which were often flagged as outliers in the analysis given. It can be improved by conducting a wider analysis and eliminating anti-digital payment types.

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

I could address what payment types could be expected for the upcoming rides and what actions can be taken accordingly.

* Do you have any ethical considerations at this stage?

An ethical consideration I have had were: -

1. Bias and fairness, ensuring no processing leads to unfair outcomes.
2. Transparency and explainability by making the outcome understandable via documenting the findings.
3. Data integrity and purpose by clarifying that the data is only used for assignment purposes.