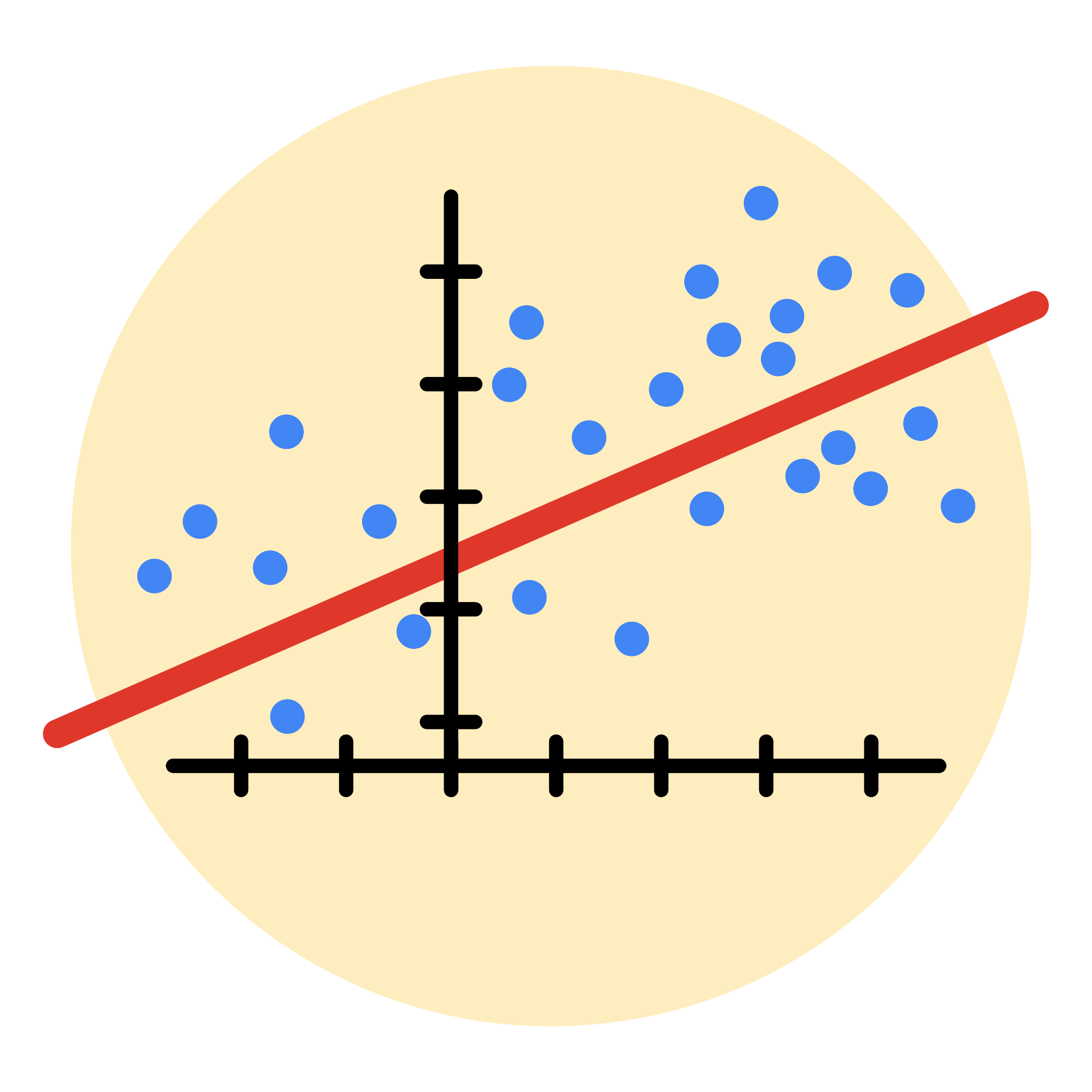
**Course Five**

# 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~~

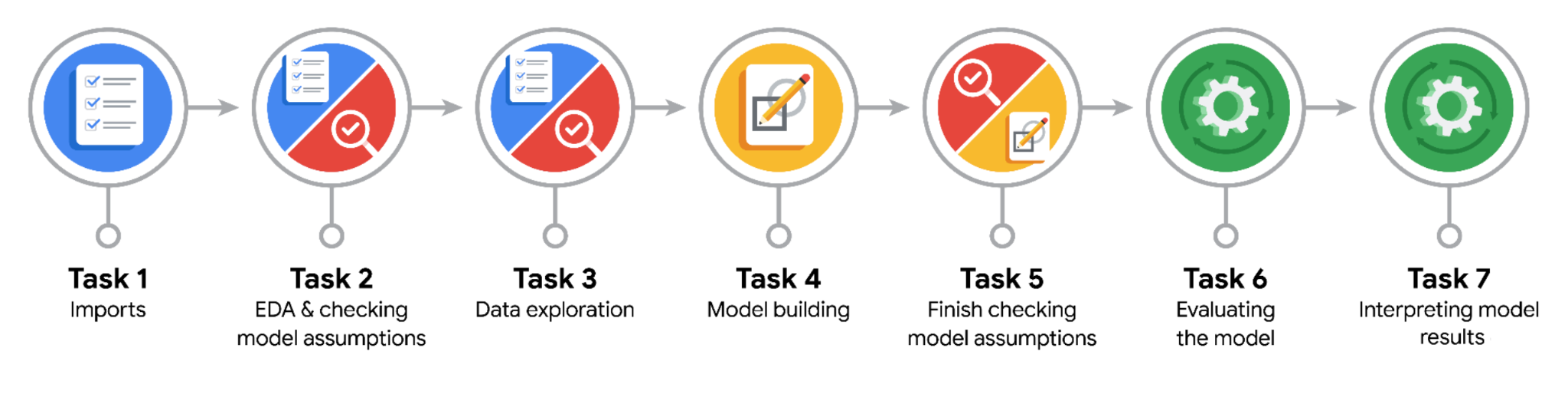
# Relevant Interview Questions

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?

TLC clients

* What are you trying to solve or accomplish?

I will build the accurate linear regression model which predicts fare amounts based on known facts

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

It is clean, however, some extra manipulations, such as normalization and outlier removal will be needed

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

Implementing statistical concepts and pandas functions

**PACE: Analyze Stage**

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

It is important to get the grasp of the data that will be used for model development to establish the workflow to follow on the path to successful completion.

* Do you have any ethical considerations in this stage?

Regarding ethics the data is absolutely fine. It is first party and was collected in unbiased environment. Particularly, it relates to the sample I am working with

**PACE: Construct Stage**

* Do you notice anything odd?

Outliers and minimal values of some columns’ values

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

Using mathematical concepts of interquartile range I am able to replace ridiculously high values with some adequate alternatives. The negative values will be imputed with zeros

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

Python functions, pandas functions

**PACE: Execute Stage**

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

Not all attributes are absolutely important for fare amount’s prediction. Distance and duration predictably play the key role

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

The better capturing of trips’ distance and duration are required. It will make building even more reliable model possible

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

It helps to determine how sensitive the target variable to the change of independent variables.

* What potential recommendations would you make?

Deploy the built model and gather the real data on its performance to determine possible imperfections

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

I do not see some obvious ways to improve it. Results seem already satisfactory. Yet I realize that some issues probably will arise during further stages of the project and my knowledge can be limited, so I do not know the extra ways to improve the model

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

The better capturing of trips’ distance and duration are required. It will make building even more reliable model possible

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

I would ask what is the best way to implement the linear regression model building - via sklearn library’s method or via statsmodel.api’s one

* Do you have any ethical considerations at this stage?

The data which was used is anonymous, no personally-sensitive information was utilized. Therefore, no, there are no ethical considerations