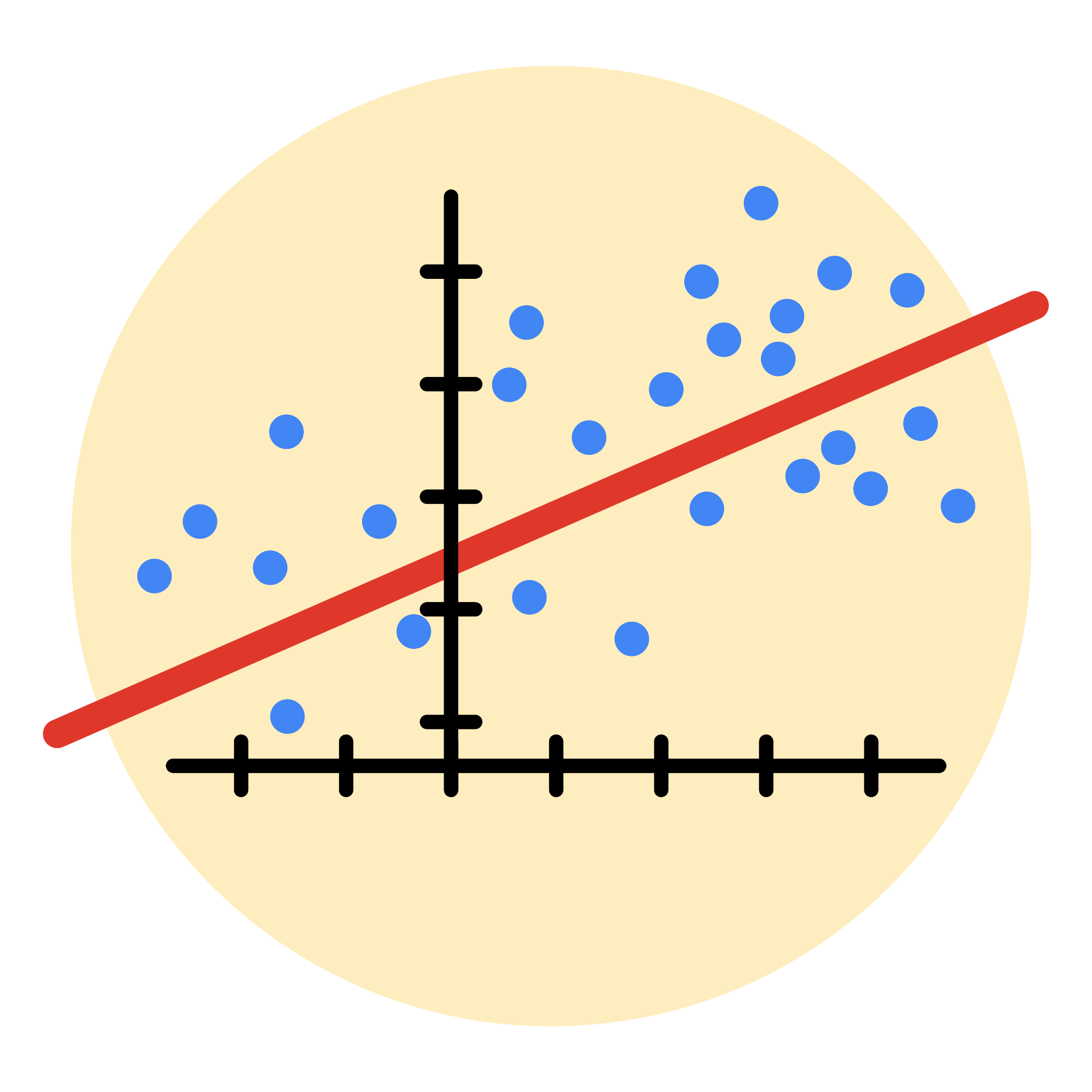
**Course Five**

# 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

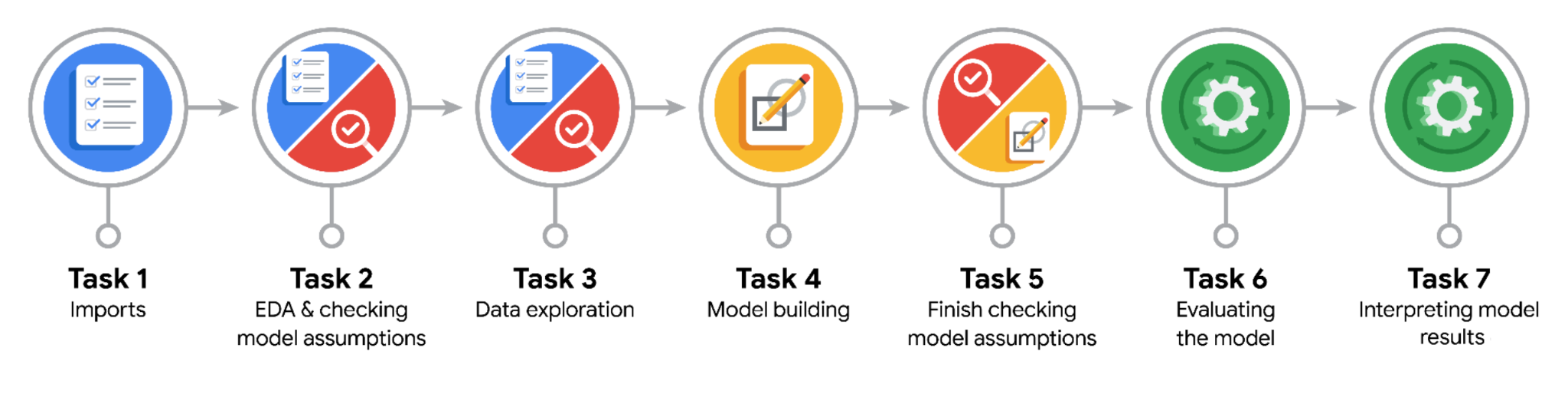
# Relevant Interview Questions

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 from NYC TLC agency. Titus Nelson, Operations Manager.

* What are you trying to solve or accomplish?

I am trying to build a regression model that helps estimate the ride fares before the taxi ride, based on data that TLC has gathered.

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

I found that data has rows of 22699 columns and 18 rows. The data was not indeed clean.

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

I referred to the emails from stakeholders to finalize my plan. I selected few variables, which I thought that I must develop working on them.

**PACE: Analyze Stage**

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

The EDA is performed to make ensure that the data is high quality, and risk-free. Because the data is to be trained by model. It is the task of EDA to understand and familiarize with data. The data was then prepared to be fitted by model.

* Do you have any ethical considerations in this stage?

A data scientist must be impartial to any bias. He must consider all the possibilities of data during EDA. He should not support any claim or phenomenon during the EDA.

**PACE: Construct Stage**

* Do you notice anything odd?

Yes, I found that there were most of the rides related to Rate Code for John F. Kennedy airport, which had fare amounts of $52. Furthermore, I noted that the model was aware of the happenings in test dataset. That’s why the model has R^2 score of 0.83 on training dataset. And R^2 score of 0.86 on test dataset.

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

I can consider creating the separate training and testing datasets. Deriving the means columns from

Training dataset, and then copy to testing dataset. Furthermore, I would like to train the model on dataset not containing rows related to JFK, and then add that data after.

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

I looked for materials like how to write code for Linear Regression from scikit-learn.

**PACE: Execute Stage**

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

The 0.83 of the fare amount can be explained by the model. The coefficients reveal that mean distance feature has the greatest impact on model, that is, for every mile traveled, there is an increase in mean of $7 in fare amount.

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

I would recommend that we should take duration and distance into consideration. Furthermore, we should take rides passing by JFK airport into consideration, as there most of rides having fare of $52.

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

The beta coefficients are important, because they tell how the variables have impact on the model and how much organization should invest into these variables, which will surely provide benefit in future.

* What potential recommendations would you make?

The more data has to be collected from NYC TLC. Additionally, the organization must put efforts to ensure that data being collected is reliable.

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

Yes, the model could be improved. I can consider creating the separate training and testing datasets. Deriving the means columns from Training dataset, and then copy to testing dataset. Furthermore, I would like to train the model on dataset not containing rows related to JFK, and then add that data after.

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

We can make an app based on this model, so that customers can have an estimate fare before the ride begins.

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

The data was indeed not in high quality, we put our best efforts to make it risk free. We used Multiple Linear Regression Model, which is an excellent choice for predicting rides’ fares based on variety of variables.

* Do you have any ethical considerations at this stage?

A data professional must present all the findings he encountered during developing model.