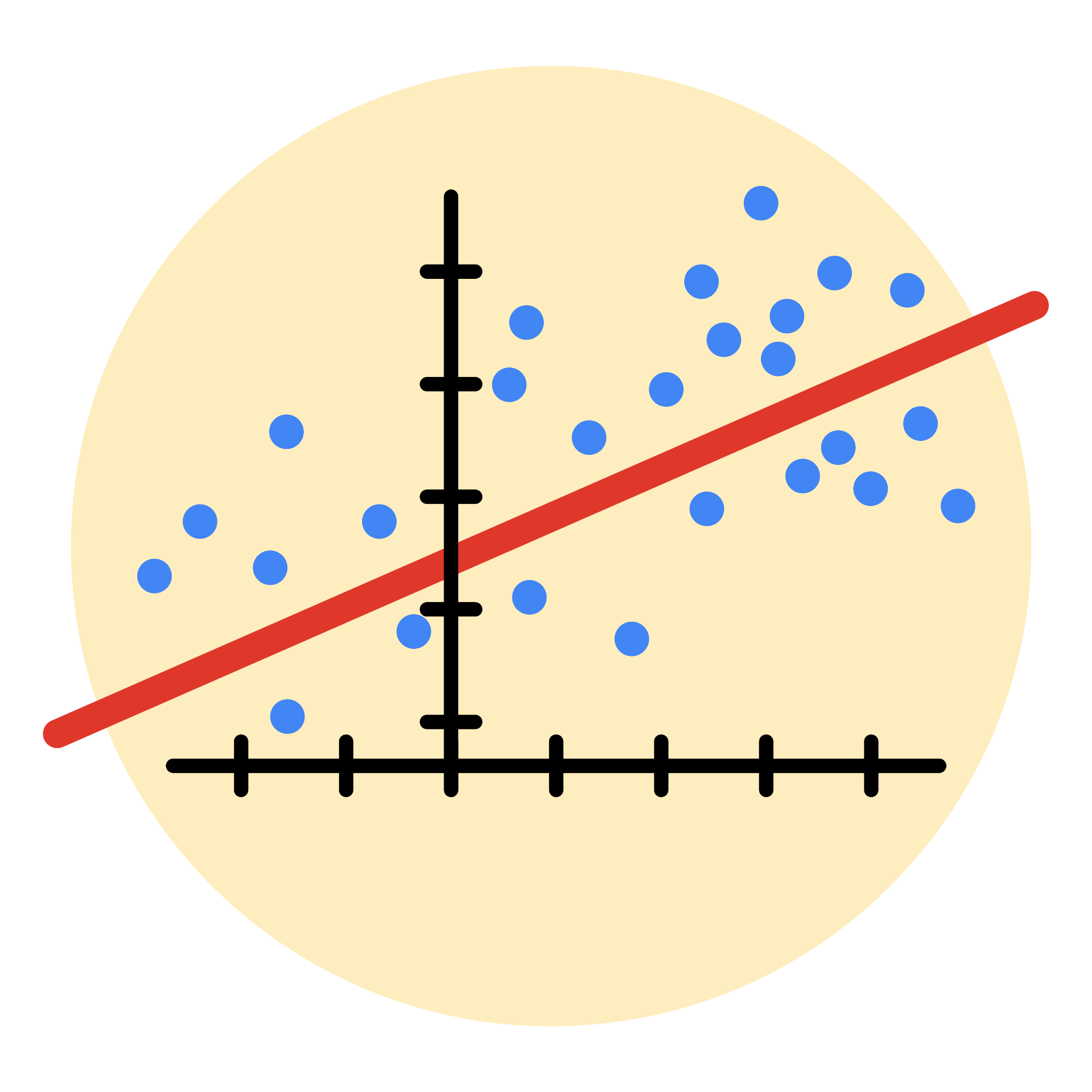
**Course Five**

# 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

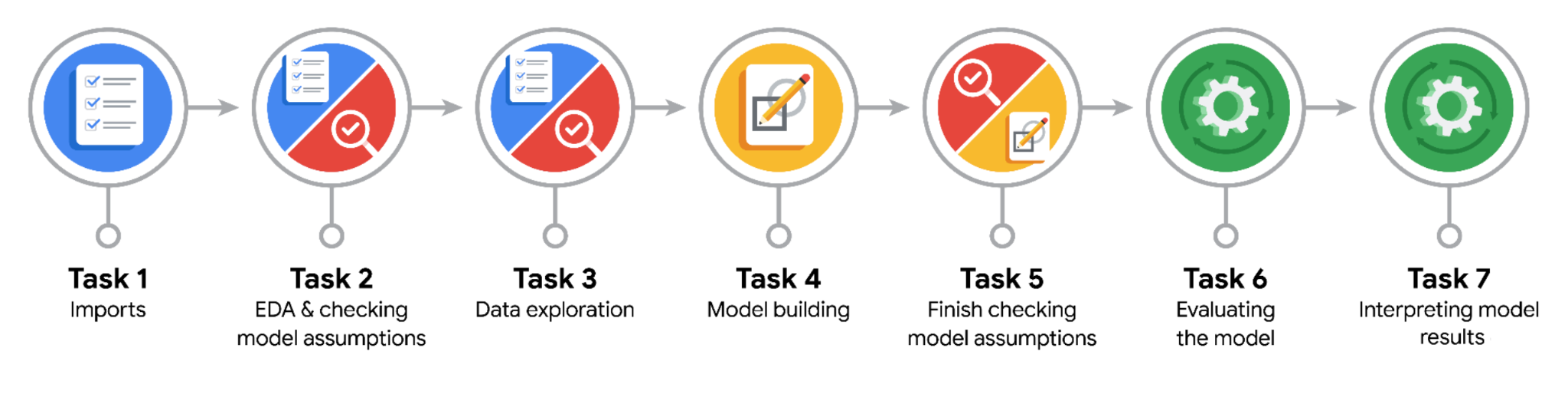
# Relevant Interview Questions

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?

External stakeholders for this project are TLC team members, who are less technically-savvy than automatidata team members.

* What are you trying to solve or accomplish?

We are trying to build a multiple linear regression model that predicts taxi fares.

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

The data doesn’t have null values. Pickup and Dropoff columns need to be transformed to datetime in order to calculate a trip duration. There are outliers in the data that need to be cleaned up.

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

**PACE: Analyze Stage**

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

The purpose of EDA is to understand the data in order to perform any cleaning and preparatory work required before constructing a multiple linear regression.

* Do you have any ethical considerations in this stage?

No

**PACE: Construct Stage**

* Do you notice anything odd?

Mean distance and mean duration are highly correlated with fare amount. They are also highly correlated 0.87 with each other.

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

The model R2 is 0.86 with test data is reasonably good. Other variables could be tried to improve its performance.

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

Visualization is very useful. Data cleaning and preparation are key.

**PACE: Execute Stage**

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

Mean distance and mean durations are the variables with greatest impact on fare amount. For 1 mile travelled, the fare amount increases by $2.

* What business recommendations do you propose based on the models built?
* To interpret model results, why is it important to interpret the beta coefficients?

The Beta coefficients weight the relationship between the independent variables and the dependent variable. In this case, the Beta coefficients input are the scaled independent variables. So to convert to the independent variables, the standard deviation of the independent variables should be involved.

* What potential recommendations would you make?

For this model, the calculated fare would be based on pick-up and drop-off points that are used to derived a mean distance and mean duration for the trip. Vendor ID, passenger count and rush hour have a smaller impact on the fare that the first two variables.

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

Yes. The model can be improved

# NOTES

This notebook was designed for teaching purposes. As such, there are some things to note that differ from best practice or from how tasks are typically performed.

1. When the `mean\_distance` and `mean\_duration` columns were computed, the means were calculated from the entire dataset. These same columns were then used to train a model that was used to predict on a test set. A test set is supposed to represent entirely new data that the model has not seen before, but in this case, some of its predictor variables were derived using data that \*was\* in the test set.

This is known as \*\*data leakage\*\*. Data leakage is when information from your training data contaminates the test data. If your model has unexpectedly high scores, there is a good chance that there was some data leakage.

To avoid data leakage in this modeling process, it would be best to compute the means using only the training set and then copy those into the test set, thus preventing values from the test set from being included in the computation of the means. This would have created some problems because it's very likely that some combinations of pickup-dropoff locations would only appear in the test data (not the train data). This means that there would be NaNs in the test data, and further steps would be required to address this.

In this case, the data leakage improved the R2 score by ~0.03. It turned out to be not that signficant.

* What business/organizational recommendations would you propose based on the models built?
* 2. Imputing the fare amount for `RatecodeID 2` after training the model and then calculating model performance metrics on the post-imputed data is not best practice. It would be better to separate the rides that did \*not\* have rate codes of 2, train the model on that data specifically, and then add the `RatecodeID 2` data (and its imputed rates) \*after\*. This would prevent training the model on data that you don't need a model for, and would likely result in a better final model. However, the steps were combined for simplicity.
* 3. Models that predict values to be used in another downstream model are common in data science workflows. When models are deployed, the data cleaning, imputations, splits, predictions, etc. are done using modeling pipelines. Pandas was used here to granularize and explain the concepts of certain steps, but this process would be streamlined by machine learning engineers. The ideas are the same, but the implementation would differ. Once a modeling workflow has been validated, the entire process can be automated, often with no need for pandas and no need to examine outputs at each step. This entire process would be reduced to a page of code.
* Given what you know about the data and the models you were using, what other questions could you address for the team?
* Do you have any ethical considerations at this stage?