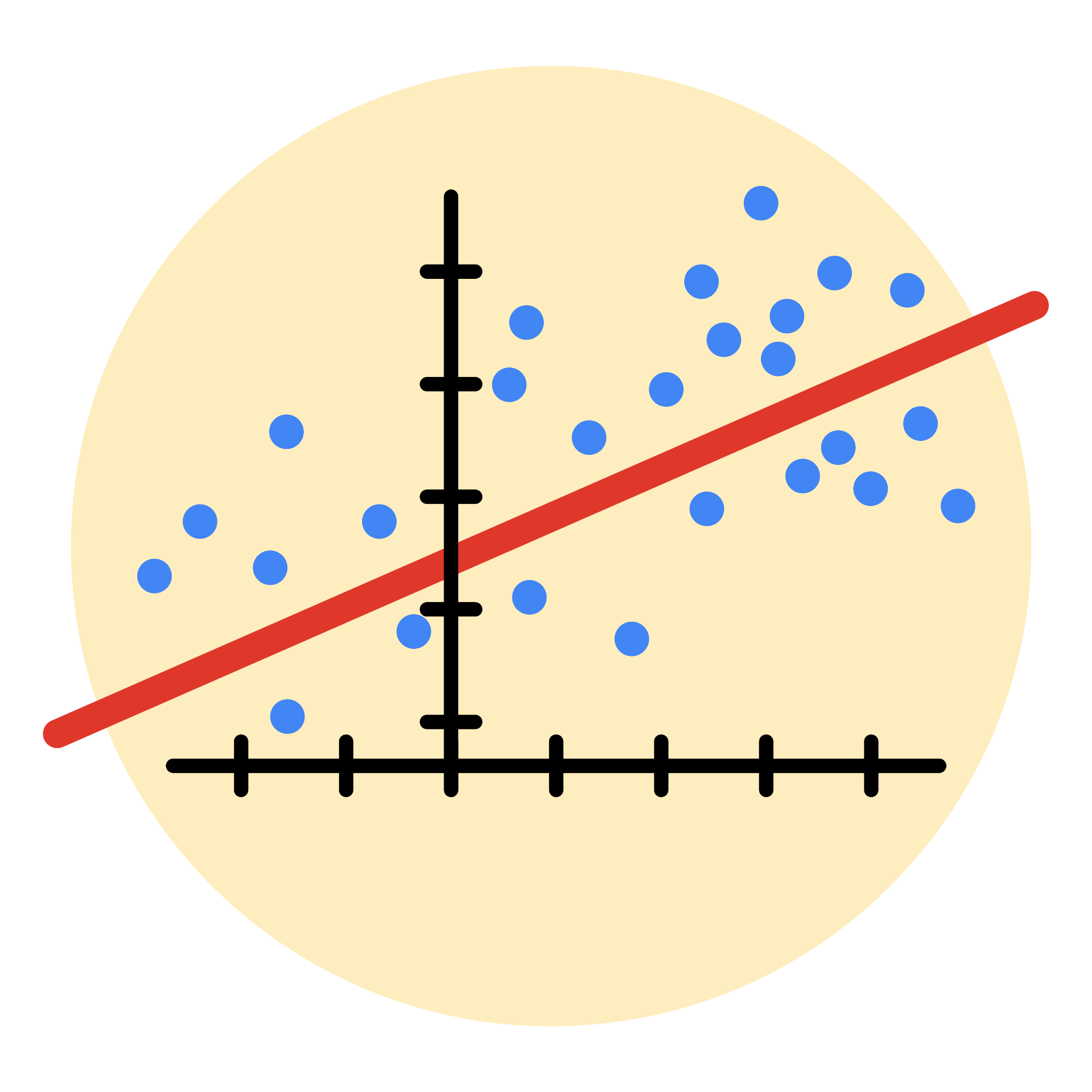
**Course Five**

# 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

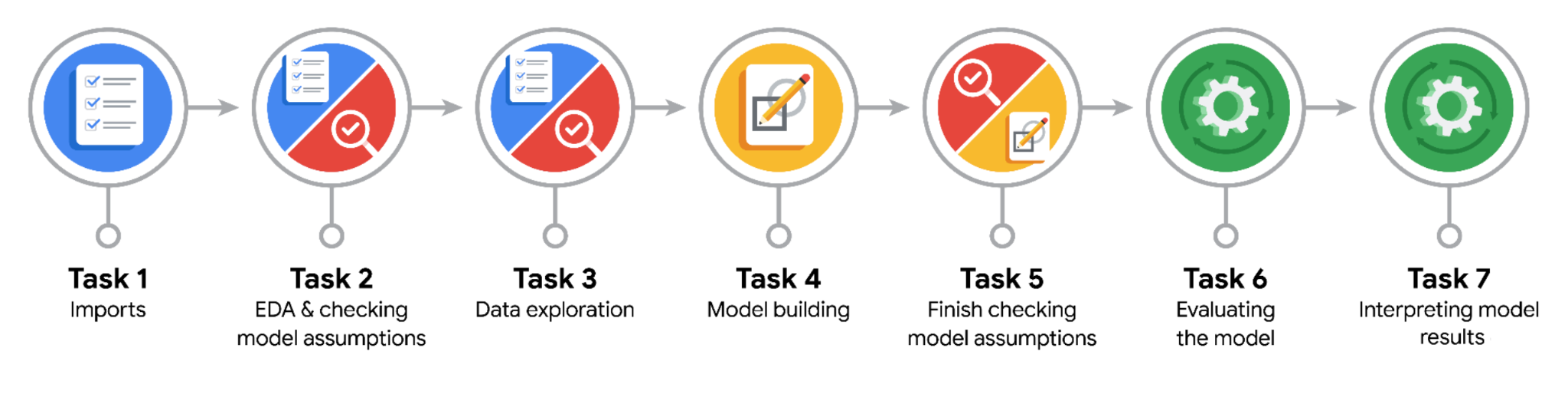
# Relevant Interview Questions

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?

NY TLC program managers Juliana Soto and Titus Nelson

* What are you trying to solve or accomplish?
* Develop a regression model that helps estimate taxi fares before the ride, based on data that TLC has gathered.
* Build a multiple linear regression model to predict taxi fares using existing data that was collected over the course of a year

Following tasks:

* Compute descriptive stats
* Conduct hypothesis test
* Create Execute Summary for Automatidata data team before sharing results with TLC
* What are your initial observations when you explore the data?
* No missing values and no duplicates found exploring data (EDA)
* Several outliers (see tip\_amount and total\_amount) while other fields like mta\_tax (see data dictionary) fairly constant – outlier handling need to be decided on
* Data fields containing dates need to be converted to datetime
* What resources do you find yourself using as you complete this stage?

- Import packages (numerics, dataframes, visualizations, regression models)

- Initial load of TLC data for analysis into dataframe for processing

**PACE: Analyze Stage**

* What are some purposes of EDA before constructing a multiple linear regression model?
* Detect outliers, missing data that would impact linear regression results. For that a plan how to handle these data must be in place (drop, include, exclude?)
* Verify multicollinearity between predictor variables (X) as this impacts statistical modeling
* Calculate additional required fields based on exisiting ones, e.g. duration fields = tpep\_dropoff\_datetime - tpep\_pickup time
* Do you have any ethical considerations in this stage?

None - as the story, all names, characters, and incidents portrayed in this project are fictitious

**PACE: Construct Stage**

* Do you notice anything odd?

Calculated fields ‘mean\_distance’ ( based on ‘PULocationID’ and ‘DOLocationID’ fields) and ‘mean\_duration’ (based on ‘duration’ field --> mean subject to above location fields) were calculated from the entire dataset. Same fields were used to train a model to predict on the test set, i.e. training data contaminates test data which should actually verify the model’s performance (prediction of y ‘fare\_amount’)

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

Train (calculate duration and distance means on the training set only), i.e. test the model with separate (sub) dataset

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

Packages for OLS, MLR, confusion matrix (see sklearn)

**PACE: Execute Stage**

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

Based on the test data metrics MAE, MSE, RMSE show no bias and no model overfit. A coefficient of determination R2 of 0.868 shows 86.8% of the variance in the fare\_amount variable is described by the model

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

The model provides a generally strong and reliable fare prediction that can be used in downstream modeling efforts

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

It is important in order to identify the impact each feature (predictor) has on the outcome variable ‘fare\_amount’

* What potential recommendations would you make?

Request additional data from under-represented itineraries

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

Train and test the model using different datasets

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

Verify the high correlation between several features and adapt model

* Do you have any ethical considerations at this stage?

None, as dataset is not showing personal data (no exact drop off address or names)