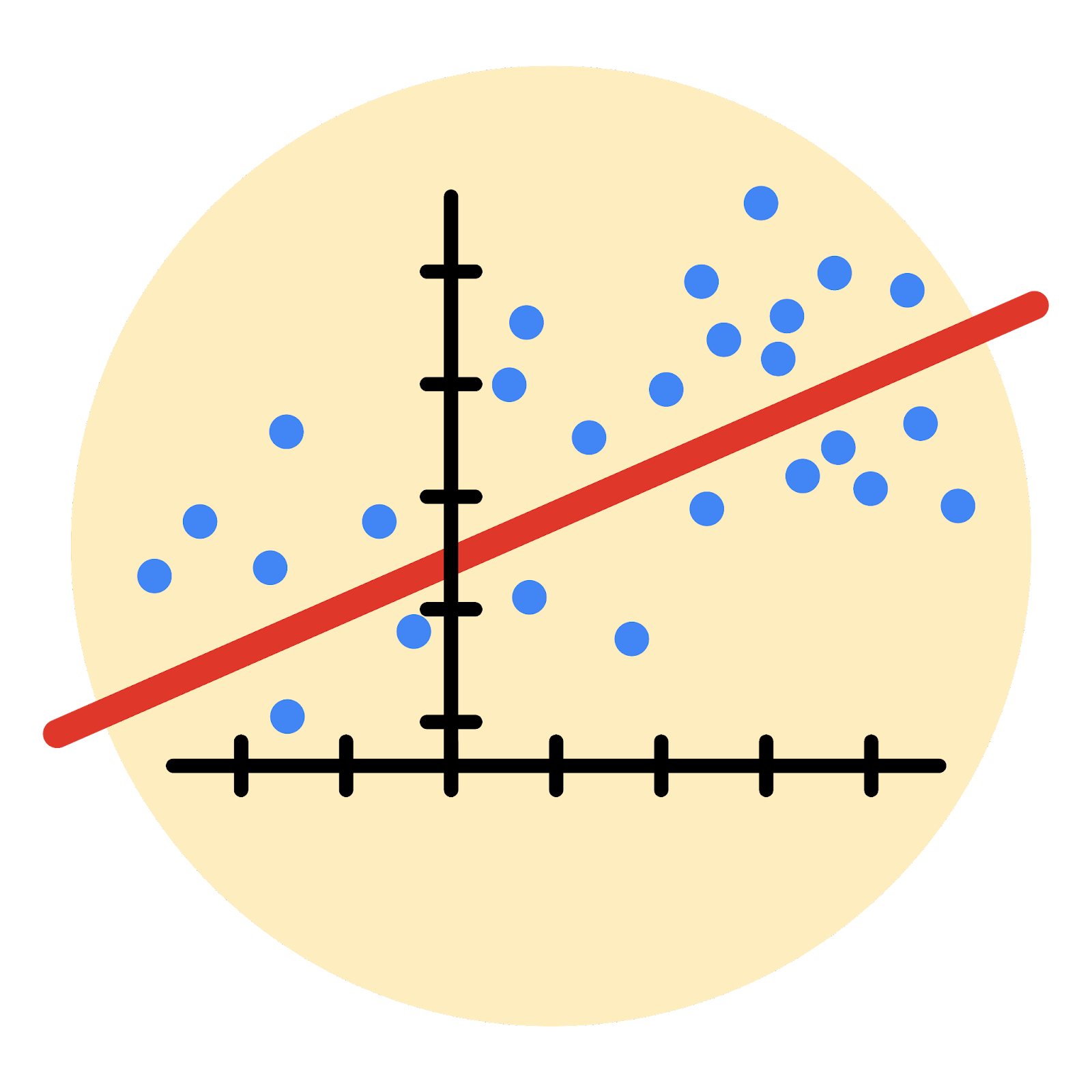
**Course Five**

# Regression Analysis: Simplifying Complex Data Relationships



# Instructions

Use this PACE strategy document to record decisions and reflections as you work through the end-of-course project. As a reminder, this document is a resource that you can reference in the future and a guide to help 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:

* Understand and assess the proposed scenario
* Build a multiple linear regression model
* Evaluate the model
* Articulate findings in an executive summary for external stakeholders

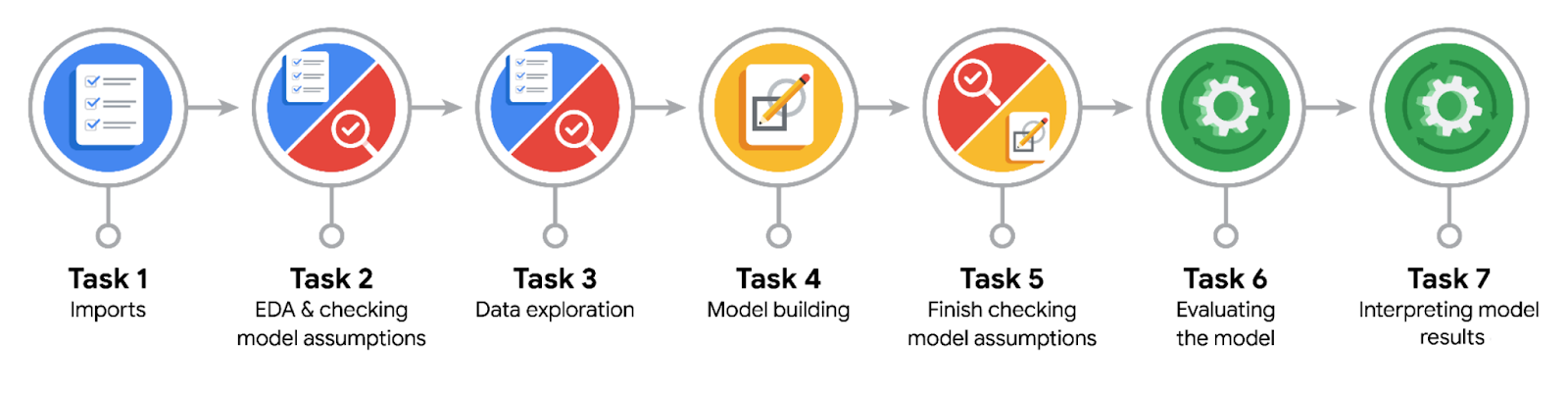
# Relevant Interview Questions

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: Planning Stage**

* Who are your external stakeholders for this project?

- Juliana Soto, Finance and Administration Department Head

- Titus Nelson, Operations Manager

* What are you trying to solve or accomplish?

Build a regression model in Python that predicts that predicts taxi cab ride fares.

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

- There are no duplicates or missing values in the data

- Some things stand out from this table of summary statistics. For instance, there are clearly some outliers in several variables, like tip\_amount ($200) and total\_amount ($1,200). Also, a number of the variables, such as mta\_tax, seem to be almost constant throughout the data, which would imply that they would not be expected to be very predictive.

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

- Exploratory Data Analysis (EDA)

**PACE: Analyzing Stage**

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

- Analyze and discover data, looking for correlations, missing data, outliers, and duplicates.

* Do you have any ethical considerations in this stage?

- Articulating that **correlation is not causation** is part of a data professionals best practices and ethical toolbox

**PACE: Constructing Stage**

* Do you notice anything odd?

mean\_duration and mean\_distance are **both highly correlated** with the target variable of fare\_amount They're also both correlated with each other, with a Pearson correlation of 0.87.

Recall that **highly correlated predictor variables can be bad** for linear regression models when you want to be able to draw statistical inferences about the data from the model. However, correlated predictor variables c**an still be used** to create an accurate predictor if the prediction itself is more important than using the model as a tool to learn about your data.

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

- try modeling with both variables even though they are correlated.

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

- For the test data, an **R2** of 0.868 means that 86.8% of the variance in the fare\_amount variable is described by the model.

The **mean absolute error** is informative here because, for the purposes of the model, an error of two is not more than twice as bad as an error of one.

**PACE: Execute Stage**

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

- The model performance is high on **both** training and test sets, suggesting that **there is no bias** in the model and that the model is not **over-fit**. In fact, the test scores were somewhat better.

- The distribution of the residuals is normal and has a mean of -0.015. The residuals represent the variance in the outcome variable that is not explained by the model. A **normal distribution around zero is good,** as it demonstrates that the models errors are evenly distributed and unbiased.

-The model's residuals are **evenly distributed** above and below zero, with the **exception** of the sloping lines from the upper-left corner to the lower-right corner, which you know are the imputed maximum of $62.50 and the flat rate of $52 for JFK airport trips.

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

- Fitting multiple linear regression models may require **trial and error** to select variables that fit an accurate model while maintaining model assumptions (or not, depending on your use case).

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

- interpreting beta coefficients provides i**nsights into the relationships between** predictor variables and the response variable, **aids** in hypothesis testing, **facilitates** variable comparison, **supports** decision making, and **helps validate** and **improve** the model.

* What potential recommendations would you make?
* Do you think your model could be improved? Why or why not? How?

- 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.

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