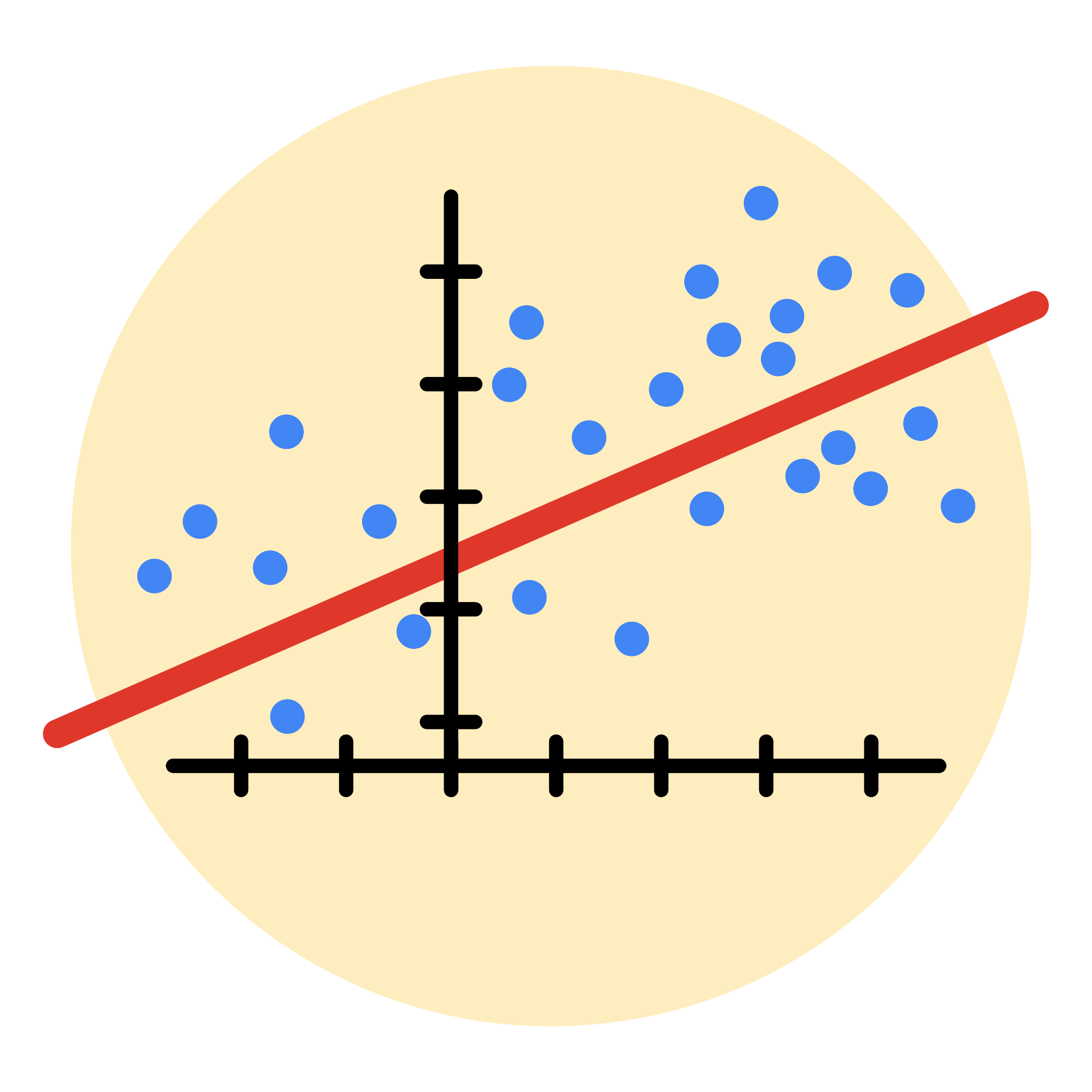
**Course Five**

# 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

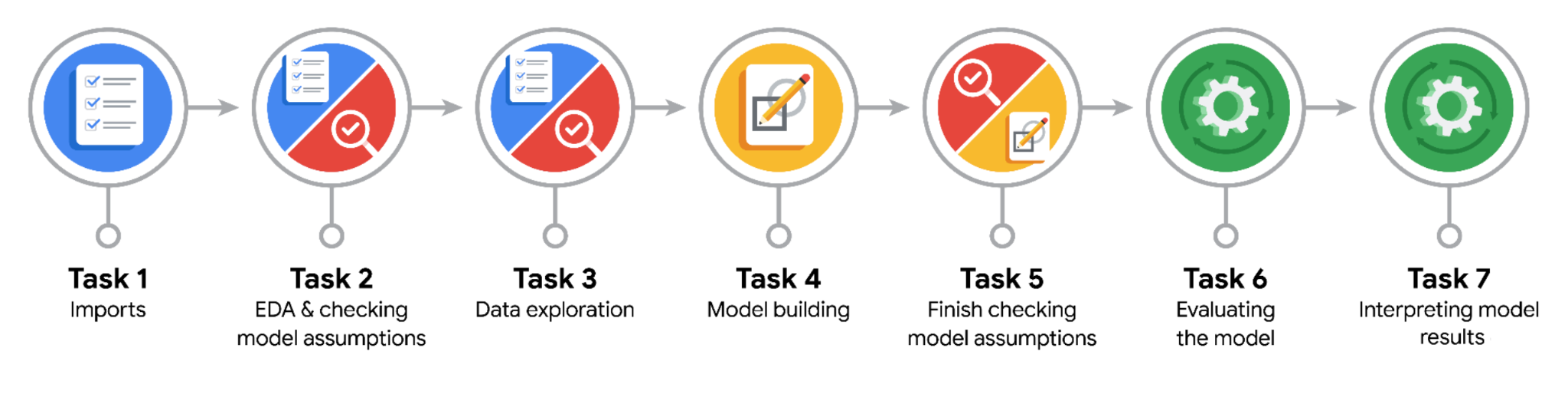
# Relevant Interview Questions

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?
* New York City Taxi and Limousine Commission (TLC) - primary client
* TLC Operations Manager (Titus Nelson) and Finance Department Head (Juliana Soto)
* NYC taxi drivers and passengers who will be affected by fare predictions
* Regulatory bodies overseeing transportation pricing
* What are you trying to solve or accomplish?
* Build a multiple linear regression model to predict taxi fares before the ride begins
* Help TLC estimate fare amounts based on trip characteristics
* Provide accurate fare predictions using historical taxi trip data
* Enable better fare transparency and planning for passengers and operators
* What are your initial observations when you explore the data?
* Dataset contains 22,699 taxi trips with 18 features
* Key variables include trip distance, duration, pickup/dropoff locations, passenger count
* Target variable is fare\_amount (continuous)
* Some outliers present in fare amounts (negative values, extremely high values)
* Missing values only in Arrival\_Delay\_in\_Minutes column (393 missing)
* Data spans different times of day, days of week, and months
* What resources do you find yourself using as you complete this stage?
* Pandas for data exploration and summary statistics
* Data dictionary provided by TLC for understanding column meanings
* Business context from stakeholder emails about regression requirements

**PACE: Analyze Stage**

* What are some purposes of EDA before constructing a multiple linear regression model?
* Identify and handle outliers that could skew model performance
* Check for missing values and determine appropriate handling strategies
* Understand distributions of variables and relationships between predictors and target
* Detect multicollinearity between independent variables
* Validate assumptions for linear regression (linearity, normality, homoscedasticity)
* Do you have any ethical considerations in this stage?
* Ensure fare predictions don't discriminate based on pickup/dropoff locations (avoiding redlining)
* Protect passenger privacy by not using personally identifiable information
* Consider fairness across different neighborhoods and demographics
* Transparency in model decisions for regulatory compliance

**PACE: Construct Stage**

* Do you notice anything odd?
* Horizontal lines at $52 and $62.5 in duration vs. fare plot (fixed rates/outlier cap); high multicollinearity between mean\_distance and mean\_duration (~0.77).
* Can you improve it? Is there anything you would change about the model?
* What resources do you find yourself using as you complete this stage?

**PACE: Execute Stage**

* What key insights emerged from your model(s)?
* Model explains ~88% variance (test R²=0.883); MAE~$2.2 indicates accurate predictions; mean\_distance and mean\_duration are top predictors
* What business recommendations do you propose based on the models built?
* To interpret model results, why is it important to interpret the beta coefficients?

Betas show feature impact (e.g., +$5 fare per std dev in distance); aids prioritization but cautious due to scaling/multicollinearity.

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

Yes; R²=0.883 leaves room. Why: Multicollinearity, potential nonlinearity. How: Ridge for regularization; Random Forest for nonlinearity; cross-validation.

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

TLC: Use for fare auditing; Automatidata: Offer as pricing service; optimize fleet during rush hours.

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

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

Performance by borough/season; impact of surge pricing; integration with traffic APIs.

* Do you have any ethical considerations at this stage?

Monitor for biases (e.g., higher fares in underserved areas); ensure transparency in predictions; avoid over-reliance on model for driver earnings.