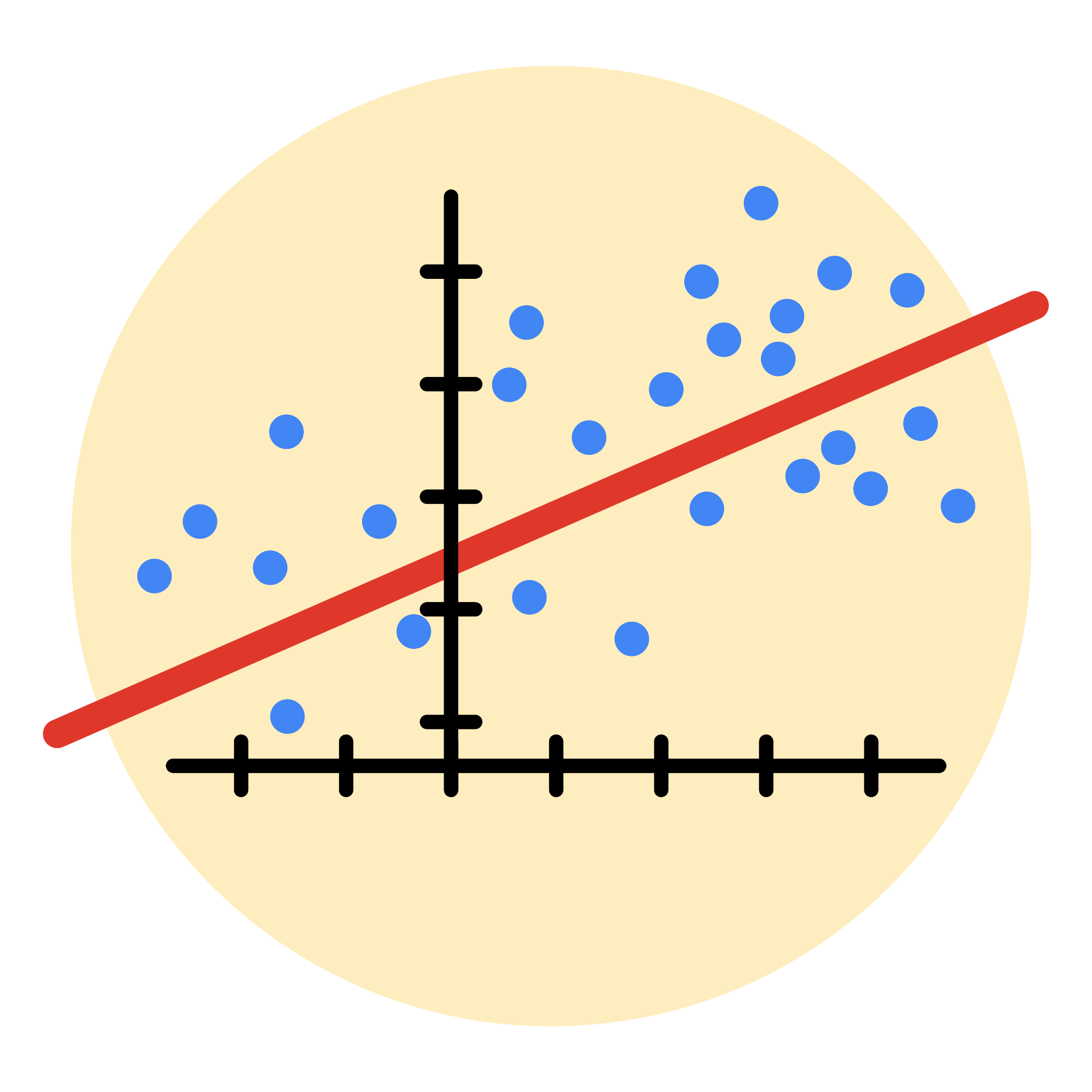
**Course Five**

# 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

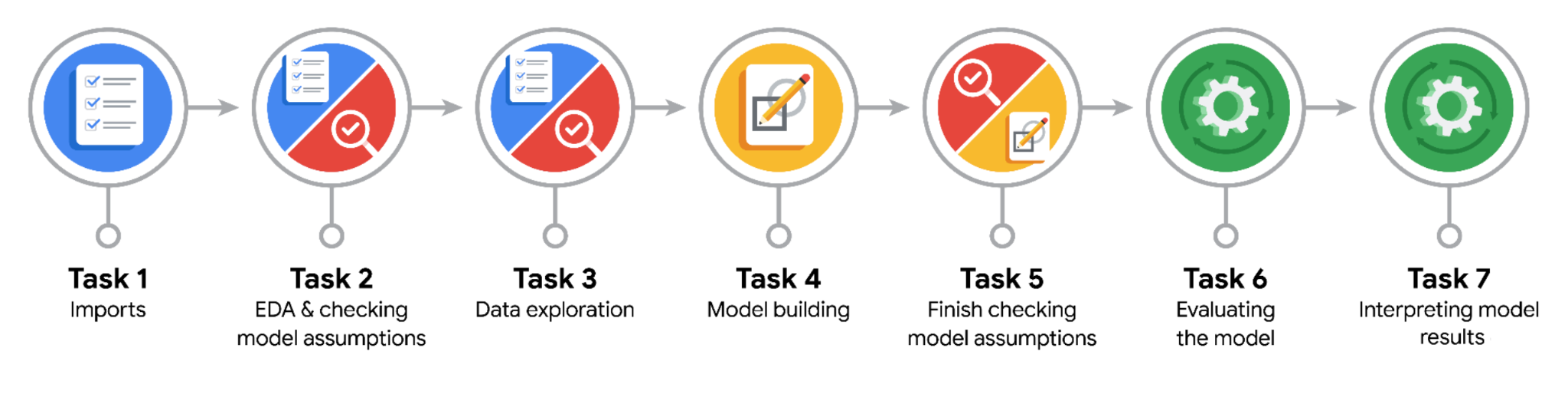
# Relevant Interview Questions

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 primary external stakeholder for this project is the New York City Taxi and Limousine Commission (TLC). Key contacts within the TLC mentioned in the scenario include Juliana Soto (Finance and Administration Department Head) and Titus Nelson (Operations Manager), who are described as program managers with non-highly technical backgrounds, requiring clear, concise communication regarding the project's findings.

* What are you trying to solve or accomplish?

The main objective is to build and evaluate a multiple linear regression (MLR) model using the provided NYC TLC taxi dataset to predict taxi Fare\_amount based on relevant ride characteristics known before or at the start of the trip. This aims to fulfill the TLC's request for a predictive tool that can estimate potential fares, leveraging their collected data to provide useful insights or functionality.

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

Initial exploration confirms the dataset includes the target variable Fare\_amount and key potential predictors like Trip\_distance, Passenger\_count, VendorID, RateCodeID, PULocationID, and DOLocationID. It also contains datetime columns requiring feature engineering (e.g., to extract trip duration, hour, day) and categorical features needing appropriate encoding for regression; basic checks would also focus on identifying potential outliers, impossible values (like negative fares or zero distances with fares), and understanding the distributions of these key variables.

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

The data dictionary for the 2017\_Yellow\_Taxi\_Trip\_Data.csv dataset (essential for understanding potential predictor variables and the target variable), knowledge from Course 5 materials regarding regression analysis principles and model types (like MLR), and the Course 5 PACE strategy document template itself, which guides the planning process.

**PACE: Analyze Stage**

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

Exploratory Data Analysis (EDA) before building a multiple linear regression model is crucial for several reasons: it helps understand the distributions, ranges, and potential outliers of both the dependent variable (Fare\_amount) and potential independent variables; allows for visualizing relationships between predictors and the outcome to check for linearity; aids in detecting multicollinearity (high correlation between predictors); identifies the need for data cleaning (handling missing values, errors) and feature engineering (e.g., extracting information from dates/times); and informs the initial selection of relevant variables, ultimately leading to a more informed and robust model building process.

* Do you have any ethical considerations in this stage?

Yes, ethical considerations during the Analyze stage primarily involve identifying potential biases reflected in the data through exploration; for example, checking if fare patterns or data completeness differ significantly based on location IDs (which might correlate with demographics or socioeconomic status) is important for understanding potential fairness issues. Additionally, decisions made during data cleaning, such as outlier removal or imputation of missing values, should be made transparently and carefully to avoid unintentionally skewing the data or disproportionately affecting certain groups, thereby ensuring responsible preparation for model building.

**PACE: Construct Stage**

* Do you notice anything odd?

During the Construct stage, after building the initial multiple linear regression model, one would critically examine the outputs for potential oddities. This includes checking for unexpected coefficient signs or magnitudes (e.g., Trip\_distance having a negative impact on Fare\_amount), verifying if theoretically important predictors are statistically insignificant (high p-values), assessing if the overall model fit (e.g., R-squared, Adjusted R-squared) is very low, and scrutinizing diagnostic plots like residual plots for patterns suggesting non-linearity or non-constant variance (heteroscedasticity), or Q-Q plots for deviations indicating non-normally distributed residuals, all of which could signal issues with the model.

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

Based on the evaluation and assumption checks performed after the initial build, improvements to the model are often possible. Potential changes could involve refining the set of predictor variables by removing insignificant ones or those causing high multicollinearity (using techniques like VIF analysis or feature selection methods), applying transformations (such as log or square root) to predictors or the dependent variable to better meet linearity or normality assumptions, adding polynomial terms or interaction variables to capture non-linear relationships identified during EDA or residual analysis, or reconsidering the encoding strategy used for categorical features.

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

Resources heavily utilized during the Construct stage primarily include Python libraries like scikit-learn (for building the LinearRegression model, splitting data, calculating evaluation metrics like R², MAE, MSE, RMSE) and statsmodels (often used for more detailed statistical output, assumption checking, and diagnostics like VIF). Visualization libraries such as matplotlib and seaborn are essential for creating residual plots, Q-Q plots, and other diagnostic visualizations. Additionally, Course 5 materials covering MLR implementation, model evaluation, assumption diagnostics, and interpretation, along with the prepared dataset and the Jupyter notebook environment itself, are crucial resources

**PACE: Execute Stage**

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

Key insights typically derived at this stage would focus on identifying the statistically significant predictors of Fare\_amount from your final multiple linear regression model. You would highlight which variables (like Trip\_distance, potentially time-based features, VendorID, etc.) have the strongest impact, the direction of that impact (positive or negative coefficients), and quantify these relationships (e.g., "an extra mile adds $X to the fare, holding other factors constant"), while also summarizing the model's overall predictive power (e.g., using R-squared or Adjusted R-squared) and its limitations.

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

Assuming the model meets acceptable accuracy and reliability standards, recommendations could include deploying it as a pre-ride fare estimation tool for the TLC or licensed apps to enhance transparency for riders. Insights about significant fare drivers (e.g., distance, specific zones) could inform operational strategies or potentially dynamic pricing models (with ethical considerations). It's also crucial to recommend establishing clear guidelines on the model's appropriate use, acknowledging its error margin (e.g., RMSE), and planning for ongoing monitoring and updates.

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

Interpreting the beta coefficients (slopes) in a multiple linear regression model is crucial because each coefficient quantifies the estimated average change in the dependent variable (e.g., Fare\_amount) for a one-unit increase in its corresponding independent variable, while holding all other independent variables in the model constant. This allows stakeholders to understand the unique contribution, direction (positive or negative), and magnitude of each predictor's relationship with the outcome, providing specific, actionable insights beyond just knowing which variables are generally important.

* What potential recommendations would you make?

Potential recommendations could extend to using the model insights for optimizing driver allocation based on predicted fare density in certain areas/times, developing tiered service levels if certain factors strongly predict higher willingness-to-pay or identifying routes/conditions where the model performs poorly, suggesting areas needing further investigation or data enrichment. A key recommendation is always to establish a framework for periodic model retraining and validation to ensure its continued accuracy and relevance over time.

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

Whether the model requires improvement depends on its final performance metrics against the project goals and how well it met regression assumptions. Most models can potentially be improved; possibilities might include incorporating more advanced feature engineering such as interactions between variables like time and distance, trying non-linear models if linear assumptions were poorly met, gathering additional data like real-time traffic information or employing different regularization techniques if overfitting was suspected, always weighing the complexity increase against the performance gain and interpretability needs.

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

Organizational recommendations could involve establishing clear ownership for the model's maintenance and monitoring within the TLC or Automatidata, integrating the fare estimation logic into relevant operational systems or public-facing applications, using the model's findings to inform driver communications or training regarding factors influencing fares, and potentially setting up A/B tests which should be actual experiments this time to validate strategies suggested by the model's correlational insights before full-scale implementation.

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

Beyond predicting Fare\_amount, the dataset and regression techniques could potentially address other questions like predicting trip duration, analyzing factors influencing Tip\_amount (for credit card trips), identifying characteristics of trips with unusually high tolls (Tolls\_amount), exploring geographical patterns in demand or fare discrepancies using LocationIDs, or modeling the likelihood of different Payment\_type usage based on ride characteristics, offering broader operational insights to the team.

* Do you have any ethical considerations at this stage?

Yes, ethical considerations during the Execute stage center on responsible communication and deployment; this includes clearly articulating the model's limitations, potential inaccuracies (error range), and any identified biases to stakeholders. It's crucial to ensure the model isn't deployed in ways that could unfairly disadvantage specific groups of riders or drivers (e.g., due to biases learned from historical data related to location or time) and to maintain data privacy standards in reporting and any application using the model's predictions.