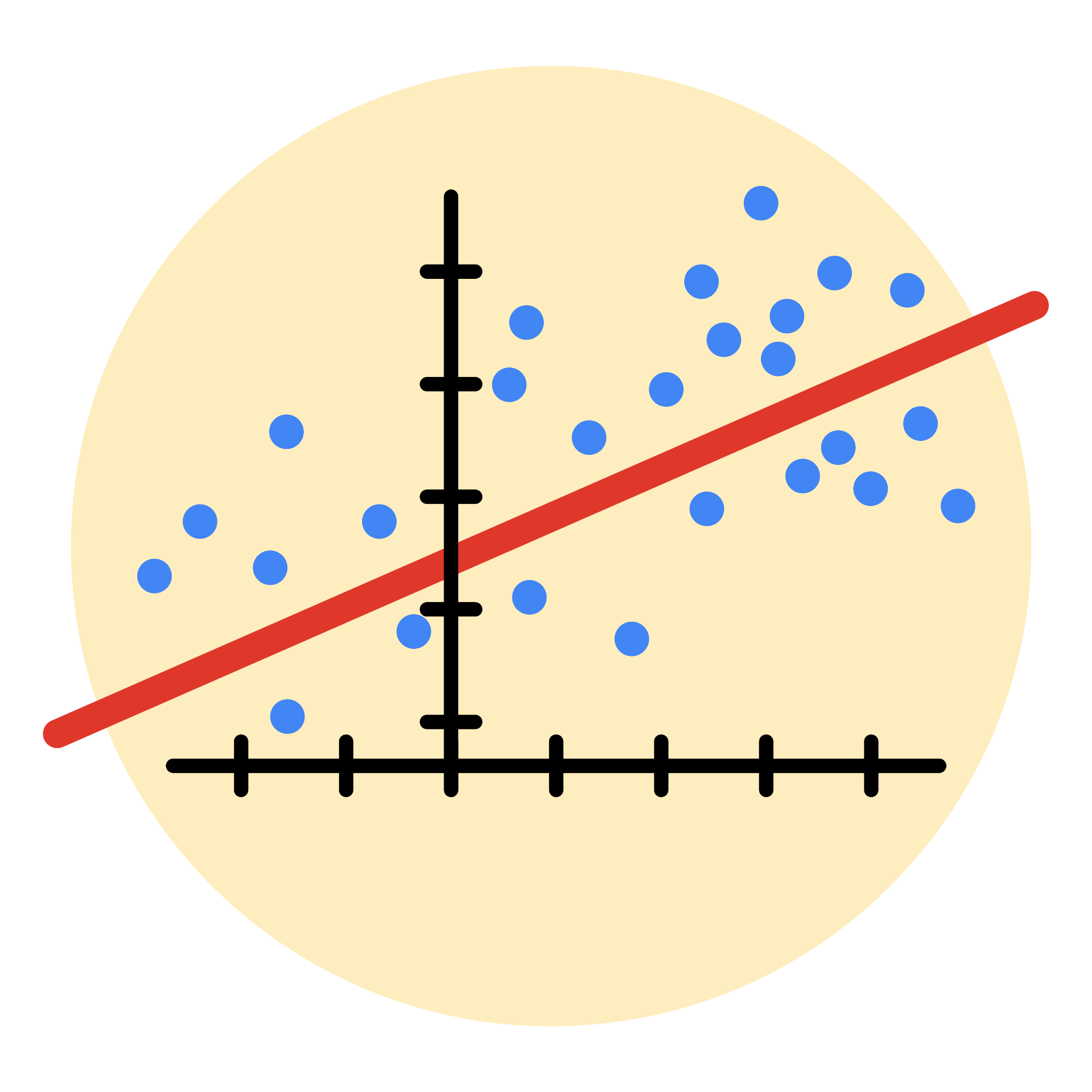
**Course Five**

# 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.

**PACE: Plan Stage**

**Who are your external stakeholders for this project?**

* **Primary Stakeholders:**
  + Juliana Soto (Finance and Administration Department Head, NYC TLC)
  + Titus Nelson (Operations Manager, NYC TLC)

**What are you trying to solve or accomplish?**

* Build a multiple linear regression model to estimate NYC taxi fares based on historical TLC data.
* Provide actionable insights and recommendations to help TLC set accurate fare expectations for customers.

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

* Look for patterns and trends (e.g., correlations between trip distance, duration, time of day, and fares).
* Identify outliers or missing data that could impact the model.
* Assess data distribution to determine if transformations are needed.

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

* Jupyter Notebook for coding and data analysis.
* Course materials (e.g., regression modules, videos, and readings).
* Python libraries: Pandas, NumPy, Matplotlib/Seaborn, and Scikit-learn.

**PACE: Analyze Stage**

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

* Understand the dataset and its structure.
* Identify key variables and their relationships with the target variable (fare).
* Check for missing or incorrect data.
* Test assumptions such as linearity, multicollinearity, and normality.

**Do you have any ethical considerations at this stage?**

* Avoid using sensitive information like passenger identities if present.
* Ensure transparency in handling and interpreting altered/pedagogical datasets.

**PACE: Construct Stage**

**Do you notice anything odd?**

* Outliers or extreme fare values that might distort model predictions.
* Multicollinearity between independent variables.

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

* Address multicollinearity by removing or combining redundant variables.
* Apply transformations if variables don’t meet linearity or normality assumptions.
* Test alternative regression models (e.g., Ridge or Lasso).

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

* Python libraries for model building: Scikit-learn, Statsmodels.
* Documentation and tutorials for troubleshooting.
* Peer-reviewed examples of regression applications.

**PACE: Execute Stage**

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

* The primary factors driving fare predictions (e.g., trip distance, duration, location).
* Areas where the model performs well and where it needs improvement.

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

* Suggest TLC develop tools to estimate fares dynamically based on distance and time, helping customers and drivers plan trips better.
* Recommend fare adjustments during peak hours to optimize income and customer satisfaction.

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

* Beta coefficients indicate the magnitude and direction of each predictor’s impact on fare.
* They help stakeholders understand which factors most influence fare estimates.

**What potential recommendations would you make?**

* Implement predictive fare estimation in customer-facing tools (e.g., apps).
* Use insights to optimize driver routing and minimize idle time.

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

* **Yes:**
  + Incorporate additional features like traffic conditions or weather data.
  + Use advanced techniques (e.g., feature engineering, hyperparameter tuning).

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

* Deploy fare prediction models as part of TLC’s operations to increase transparency.
* Use regression insights to inform strategic fare-setting policies.

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

* Can the model predict peak demand times and locations?
* What is the expected financial impact of changing fare structures?

**Do you have any ethical considerations at this stage?**

* Ensure fairness in fare predictions across different boroughs and times.
* Avoid creating biases that disproportionately impact certain demographics.