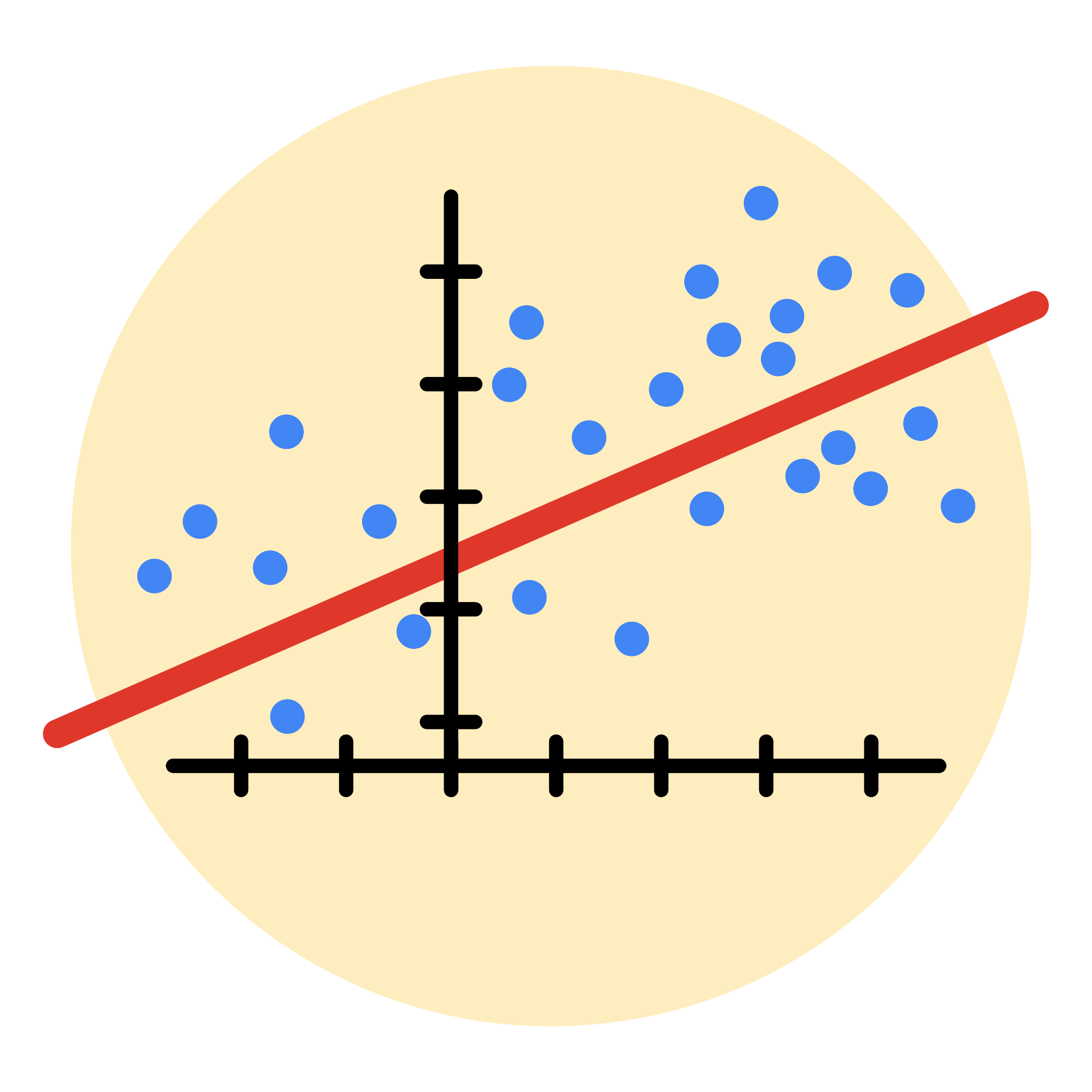
**Course Five**

# 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

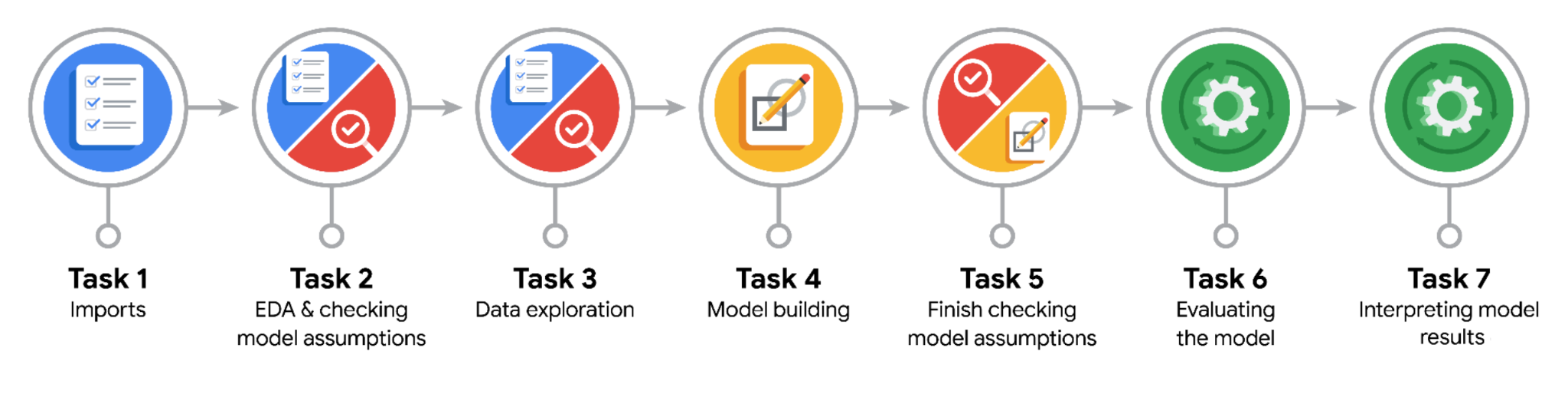
# Relevant Interview Questions

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?

External stakeholders include Waze leadership, Operations Manager (Ursula Sayo), and Director of Data Analysis (Harriet Hadzic). These stakeholders need insights to make data-driven decisions to improve user retention and optimize the Waze app experience

* What are you trying to solve or accomplish?

We aim to build a binomial logistic regression model to predict user churn based on various user behavior metrics. The goal is to identify key predictors of churn and provide actionable insights to reduce churn rates and enhance user engagement.

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

The dataset contains 14,999 rows and 13 columns.  
There are some missing values in the 'label' column.  
Variables such as 'sessions,' 'drives,' 'total\_sessions,' 'total\_navigations\_fav1,' 'total\_navigations\_fav2,' 'driven\_km\_drives,' and 'duration\_minutes\_drives' have potential outliers.  
The dataset includes both categorical and numerical variables.

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

Data dictionary to understand the variables.  
Pandas for data manipulation.  
Initial EDA to understand the data distribution and identify potential issues.

**PACE: Analyze Stage**

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

To understand the distribution of the data and identify any outliers or anomalies.  
To check for missing values and decide on imputation methods if necessary.  
To explore relationships between variables and identify potential predictors for the model.  
To perform correlation analysis to detect multicollinearity.

* Do you have any ethical considerations in this stage?

Ensuring data privacy and confidentiality.  
Avoiding bias in data interpretation and ensuring the model does not unfairly target specific user groups.

**PACE: Construct Stage**

* Do you notice anything odd?

The dataset contains outliers and missing values which need to be addressed before modeling.  
Some variables show high multicollinearity.

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

Impute or remove outliers to improve model performance.  
Address multicollinearity by removing highly correlated variables or using regularization techniques.  
Engineer new features such as 'km\_per\_driving\_day' and 'professional\_driver' to capture additional predictive signals.

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

Pandas for data manipulation and feature engineering.  
Scikit-learn for model building and evaluation.  
Data visualization libraries like Seaborn and Matplotlib for EDA and plotting.

**PACE: Execute Stage**

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

Activity days had a significant negative correlation with user churn, indicating that users who use the app more frequently are less likely to churn.  
Professional drivers had a lower churn rate compared to non-professional drivers.

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

Focus on increasing user engagement and activity within the app to reduce churn rates.  
Develop features specifically tailored for professional drivers to enhance their experience and retain them.

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

Beta coefficients indicate the strength and direction of the relationship between each predictor variable and the target variable. Interpreting them helps understand which factors are most influential in predicting churn.

* What potential recommendations would you make?

Implement targeted engagement strategies to increase app usage.   
Explore additional features that might capture user behavior more accurately.

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

Yes, the model could be improved by addressing multicollinearity, scaling features, and adding new features based on domain knowledge.

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

Enhance user engagement strategies based on the insights from the model. Focus on improving the experience for professional drivers as they have a lower churn rate.

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

How do different types of user engagement impact churn rates? Are there specific times or conditions under which churn is more likely to occur?

* Do you have any ethical considerations at this stage?

Ensure that the model does not inadvertently introduce bias against specific user groups.  
Maintain user privacy and confidentiality while using the data.