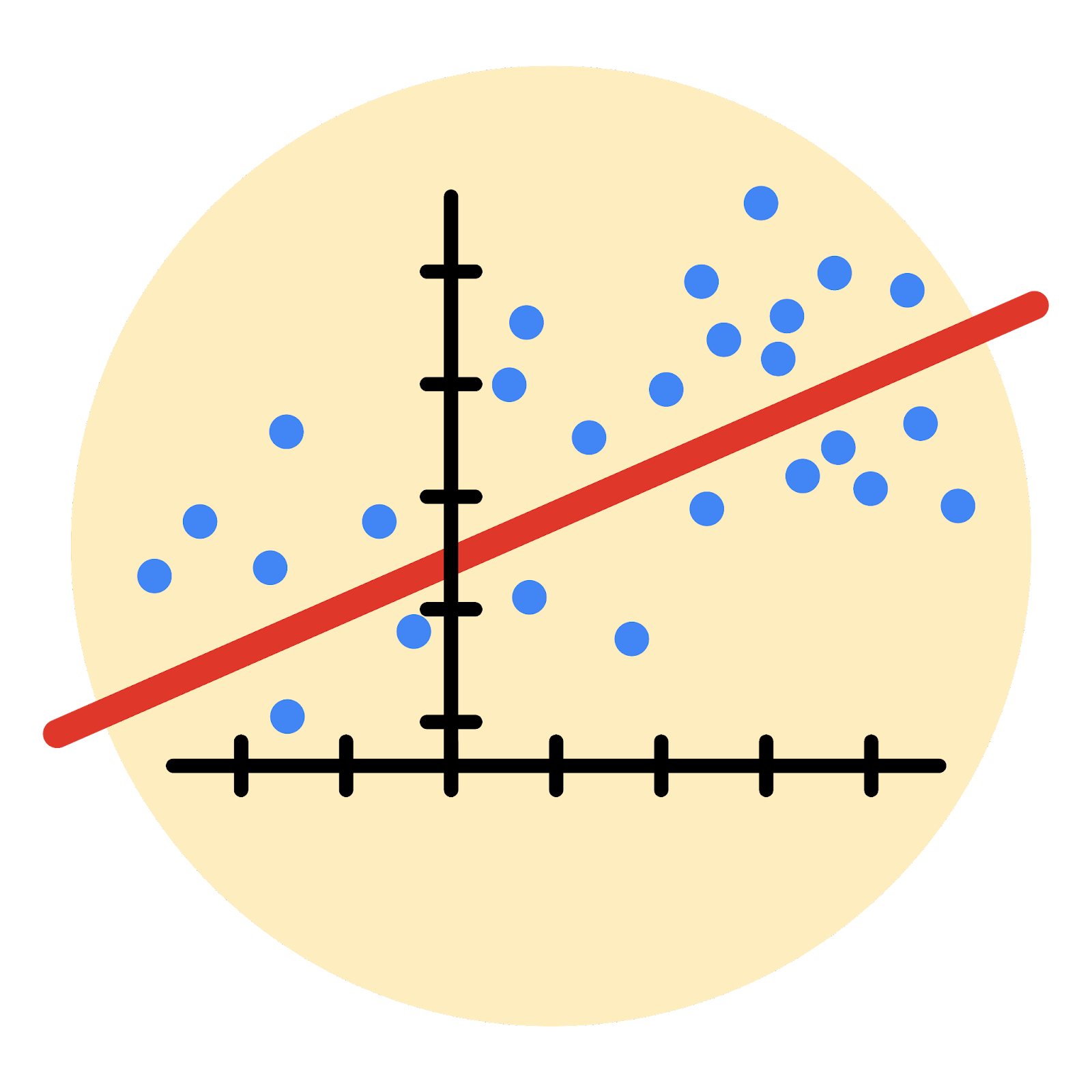
**Course Five**

# **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

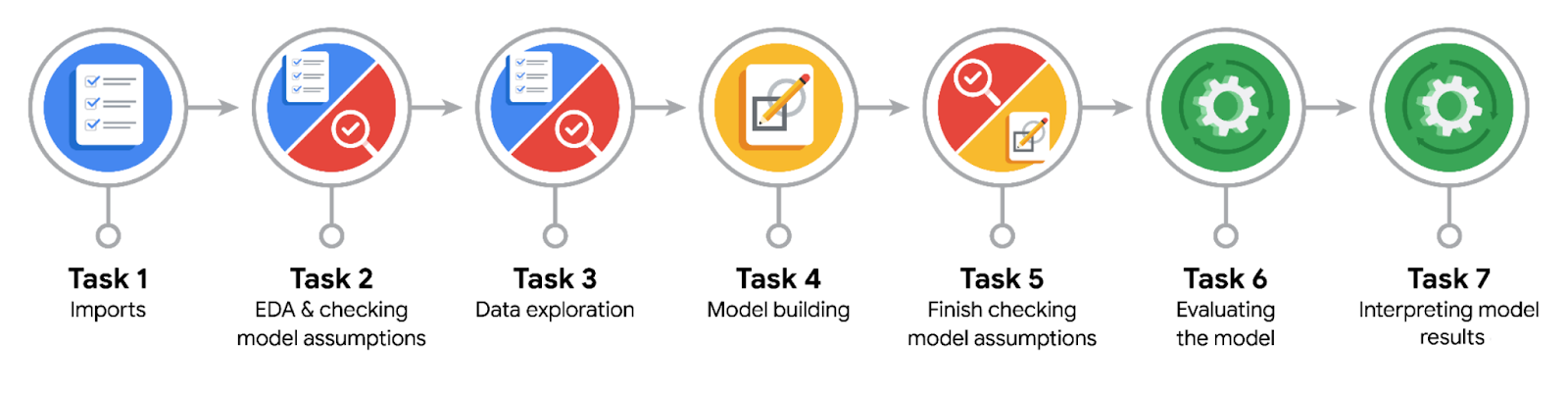
# **Relevant Interview Questions**

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 operations team.

* What are you trying to solve or accomplish?

We are trying to build a multiple logistic regression model capable of predicting user churn rate using the “label” as our dependent variable and some other user characteristics as our independent variable.

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

There were missing rows in the label column, however the dataset appeared to be in a consistent format.

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

Jupyter notebook, python, tiktok dataset, input from cross-functional team.

**PACE: Analyze Stage**

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

We do EDA to familiarize ourselves with the data to understand whether or not it would be able to satisfy the pre-modelling assumptions - to request new data or transform existing data, clean data, remove outliers and missing values as well as duplicates.

* Do you have any ethical considerations at this stage?

Yes, of course. An ethical consideration would be ensuring data privacy and fairness and avoiding bias.

**PACE: Construct Stage**

* Do you notice anything odd?

driving\_days and activity\_days; sessions and drives; these variables are collinear.

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

To satisfy the no multicollinearity assumption, driving days and sessions were removed.

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

Python jupyter notebook, large language artificial intelligence.

**PACE: Execute Stage**

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

activity\_days most influenced the model's prediction, with a negative correlation with user churn.

km\_per\_driving day was expected to be a stronger predictor but it was the second to the least predictor for user churn.

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

I recommend that Waze do not use the model to drive business decisions as it is a weak predictor of user churn but it can be used as a guide for further exploratory efforts.

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

To understand the underlying relationships in the data, which are fundamental to the predictive power of the model.

* What potential recommendations would you make?

Data containing drive-times and geographical locations could help. Much more granular data exploring user interaction with the app.

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

Scaling the predictor variables, model reconstruction with different combinations of predictor variables and/or regularization.

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

I recommend that Waze do not use the model to drive business decisions as it is a weak predictor of user churn but it can be used as a guide for further exploratory efforts.

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

How does being a professional driver contribute to user churn rate?

* Do you have any ethical considerations at this stage?

Bias and fairness

Accountability

Transparency and consent.