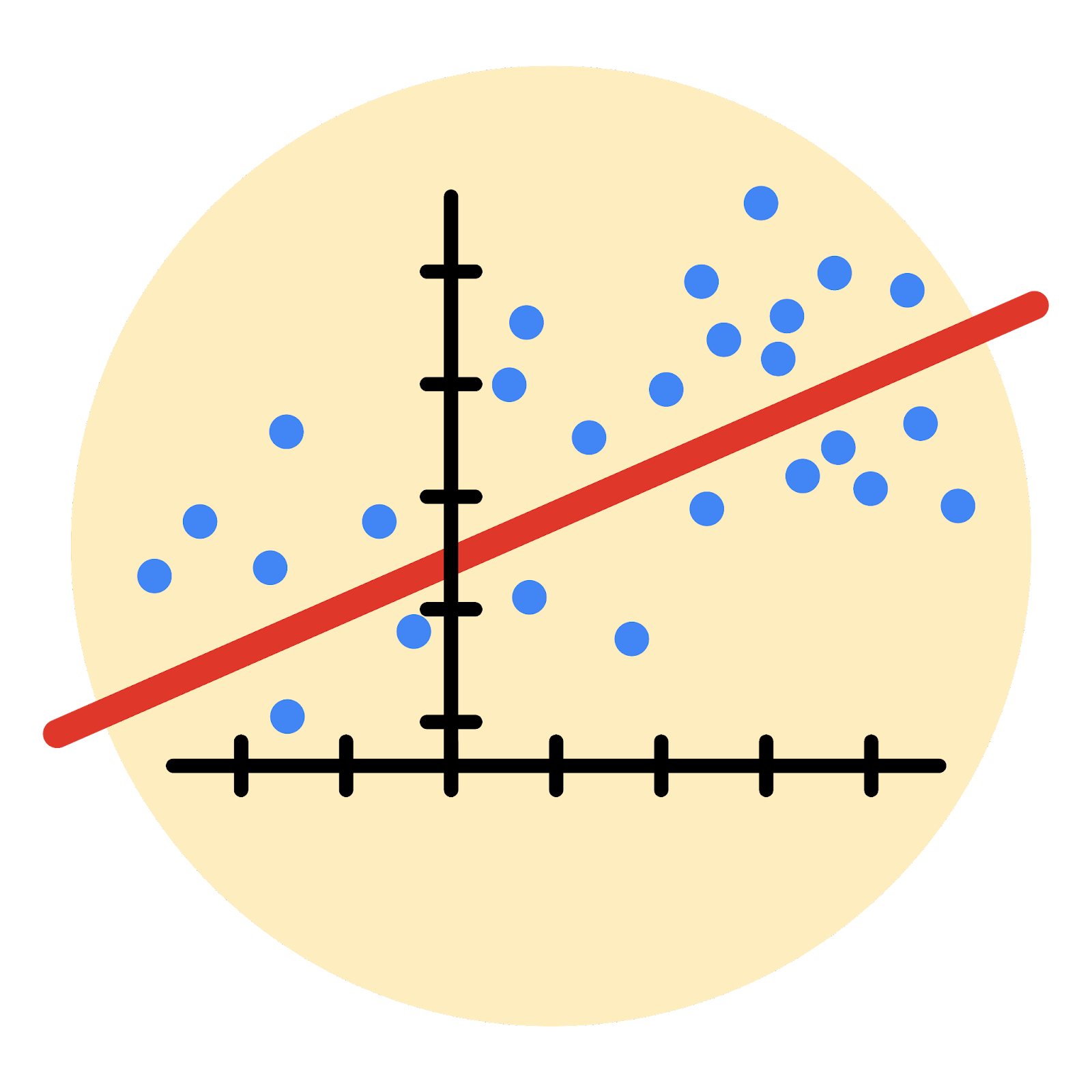
**Course Five**

# **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

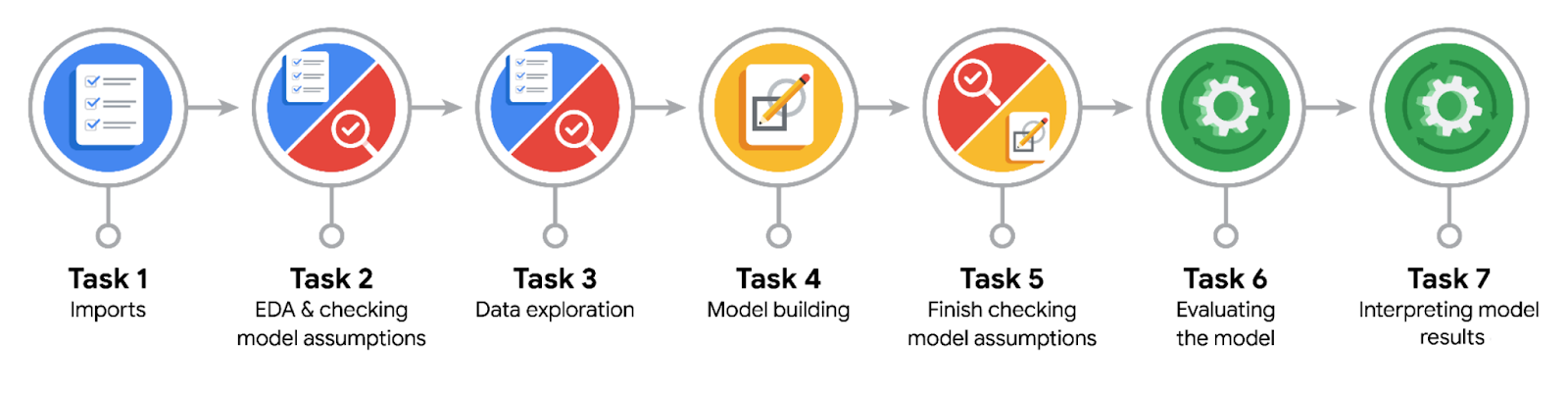
# **Relevant Interview Questions**

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 verification status based on characteristics of videos.

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

There were missing rows in about seven columns in the data, 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 teams.

**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?

The video\_like\_count variable shows high correlation with the other video counts: like, share and comment.

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

Yes, to satisfy the no multicollinearity assumption, the video\_like\_count would be removed.

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

Large language models: phind.

**PACE: Execute Stage**

* What key insights emerged from your model(s)?
* Longer videos are more likely to be posted by verified users.
* View count does not significantly impact whether a user is verified or not.
* Shared videos are slightly more likely to be posted by verified users.
* Downloaded videos do not significantly impact whether a user is verified or not.
* Videos with many comments are less likely to be posted by verified users.
* Opinion reports are more likely to be posted by verified users.
* Banned authors and authors under review are not significantly more likely to post verified content.
* What business recommendations do you propose based on the models built?

**Target Unverified Users**: Since unverified users are more likely to post claims, consider focusing efforts on identifying and analyzing content from unverified users.

**Monitor Author Status**: Keep track of the author ban status. Banned authors and authors under review are not significantly more likely to post verified content, which could potentially increase the likelihood of claims.

* 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?

Building a machine learning model.

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

Yes, the model can be improved because there were a lot of variables that did not have a significant impact on user verification status. Feature selection, particularly a backward selection could improve this model. Also, regularization techniques like Lasso, Ridge or Elastic nets.

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

**Target Unverified Users**: Since unverified users are more likely to post claims, consider focusing efforts on identifying and analyzing content from unverified users.

**Monitor Author Status**: Keep track of the author ban status. Banned authors and authors under review are not significantly more likely to post verified content, which could potentially increase the likelihood of claims.

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

Does length of transcription text significantly affect the user verified status?

* Do you have any ethical considerations at this stage?

Bias and fairness.

Accountability