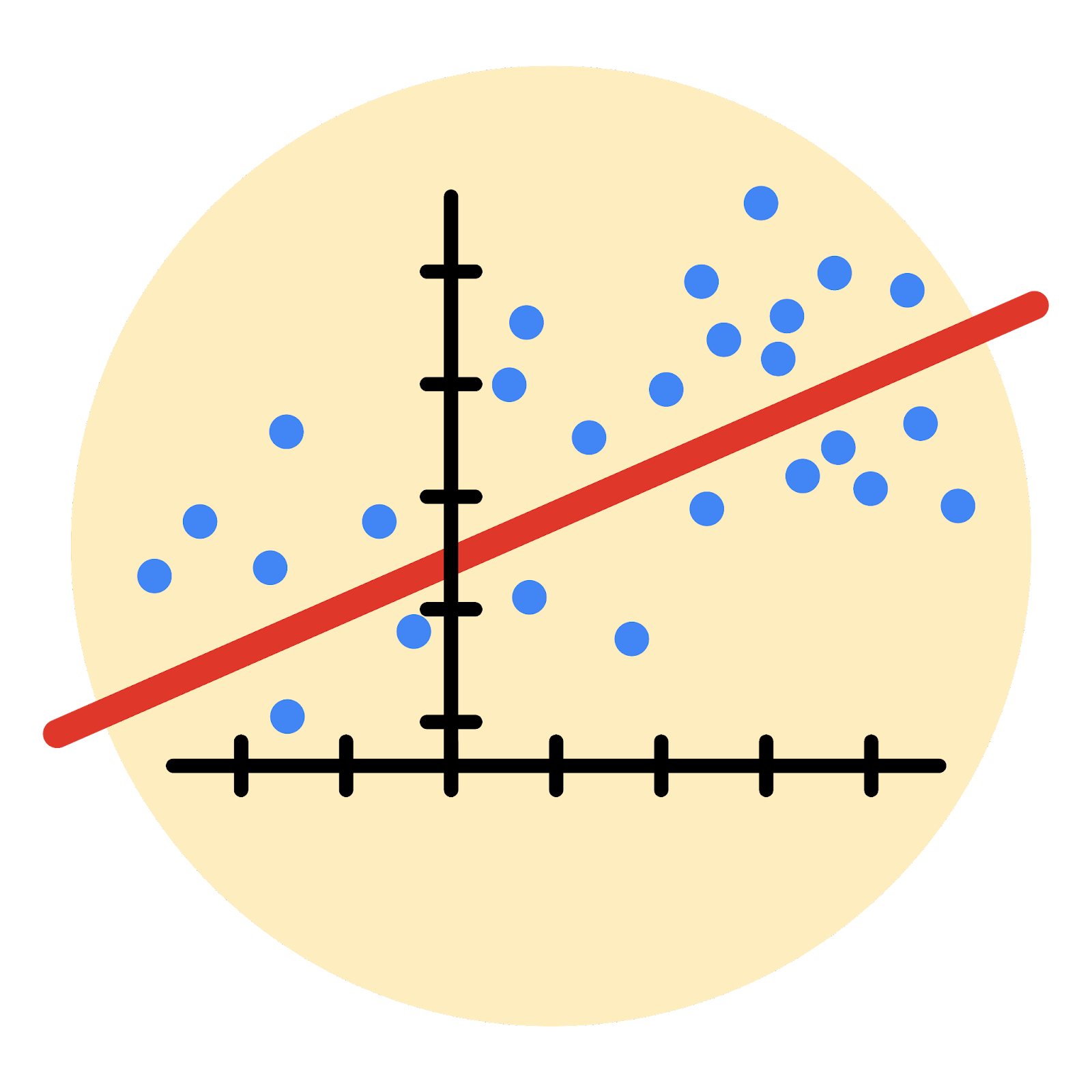
**Course Five**

# **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

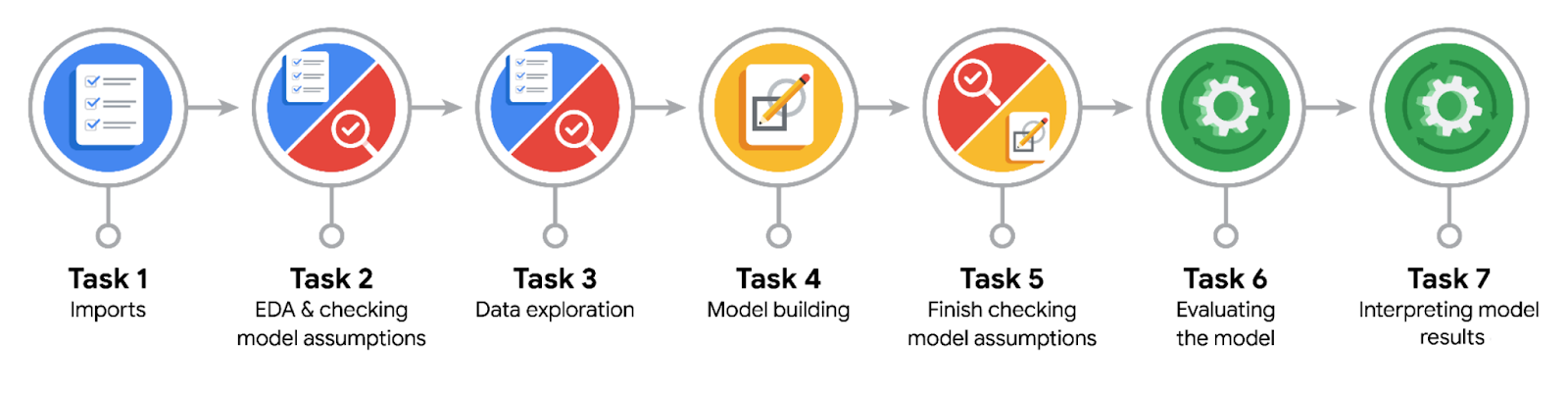
# **Relevant Interview Questions**

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 external stakeholders include TikTok’s content strategy team, marketing managers, social media analysts, and product managers.

These stakeholders are interested in understanding which video features contribute the most to increasing watch time to improve content strategy and platform engagement.

* What are you trying to solve or accomplish?

The objective is to build a multiple linear regression model to predict average watch time for TikTok videos based on various video attributes such as video length, engagement metrics (likes, shares, comments, etc.), and the number of followers.

The goal of this model is to determine which variables significantly impact watch time to inform content optimization strategies.

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

Initial observations indicate that some variables (such as comments, likes, and shares) are highly skewed and may need transformation.

Also, there are strong correlations between some predictors, suggesting potential multicollinearity.

Additionally, some variables (like average watch time and video length) appear to have a linear relationship, which is suitable for regression modeling.

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

The following resources are utilized during this process:

- Coursera course lab notebooks and instructional videos

- Documentation for pandas, seaborn, and statsmodels

- Online forums (e.g., Stack Overflow)

- Articles on interpreting regression outputs and assumptions

**PACE: Analyze Stage**

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

Some of the purposes of EDA before constructing a multiple linear regression model include:

-Understand the structure and distribution of the data

- Identify missing values, outliers, and anomalies

- Explore relationships between variables

- Detect multicollinearity or skewed distributions

- Inform decisions about transformations and feature selection

* Do you have any ethical considerations at this stage?

Some ethical considerations present at this point include:

-Avoiding bias in variable selection (for example., follower count might reflect popularity biases)

-Ensuring data privacy (especially with the user-generated content)

-Transparency in how model results are communicated and used for content decisions

**PACE: Construct Stage**

* Do you notice anything odd?

Yes, there was an odd result that occurred.

The regression model presented evidence of multicollinearity, particularly between likes, comments, and shares.

These variables are all engagement metrics and may be capturing similar signals.

Also, the residual plot suggested slight heteroscedasticity.

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

There are improvements that can be made to the model which include:

-Considering dropping or combining highly correlated variables (e.g., creating a composite engagement score)

-Applying log transformations to skewed variables

-Test interaction terms (e.g., followers × video length)

-Evaluate the model using adjusted R² and RMSE to balance complexity and performance

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

The following resources can be utilized during this point of the stage:

-statsmodels and scikit-learn documentation

-Python packages: pandas, seaborn, matplotlib, numpy

-Online resources for interpreting regression diagnostics

-Course-provided guidance for regression assumptions

**PACE: Execute Stage**

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

Some key insights that emerged from the model include the following:

**Video length** and **number of followers** were statistically significant predictors of the average video watch time.

**Shares** and **comments** also had a positive relationship, though their significance varied depending on multicollinearity adjustments.

The model had a decent R² value, indicating a good fit, though some variance in watch time remains still unexplained.

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

The following business recommendations can be proposed based on the models built:

-Encourage creators to optimize video length within the range presented to increase watch time.

-Promote strategies to increase shares and comments, which appear to positively drive engagement.

-Prioritize growing follower count as a focused long-term strategy for improving watch time.

-Consider further investigation into nonlinear effects or thresholds in video performance.

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

It is important to interpret the beta coefficients when looking at model results because the beta coefficients show the expected change in average watch time for a one-unit increase in a predictor variable, maintaining all other variables constant.

This process helps determine which factors have the greatest impact and informs actionable decisions.

* What potential recommendations would you make?

The following recommendations can be made:

-Use the model to guide content development and creator coaching.

-Develop dashboards to track the key metrics tied to watch time.

-Test changes to content strategy through A/B testing informed by model outputs.

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

The model can be improved by completing the following:

-Refining variable selection to reduce multicollinearity

-Applying nonlinear regression techniques or regularization methods

-Adding new features, such as video category or posting time

-Using cross-validation for more robust performance metrics

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

The following recommendations are suggested based on the models built:

-Provide creators with personalized insights on content optimization

-Enhance recommendation algorithms by prioritizing predictors of high watch time

-Adjust platform incentives (i.e., boosting in-feed placement) based on watch time predictors

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

Based on what is known about the data and models being utilized, the following questions should be addressed for the team:

-What is the predicted average watch time for a given combination of features?

-How do different content categories compare in terms of engagement and watch time?

-Are there diminishing returns to video length or follower count?

-Can creators be segmented into performance tiers based on the model?

* Do you have any ethical considerations at this stage?

One ethical consideration present is that watch time should not be the only focus if it leads to the promotion of addictive or misleading content.

Model outputs should be used to enhance user experience and content quality, not only maximize engagement at all costs.