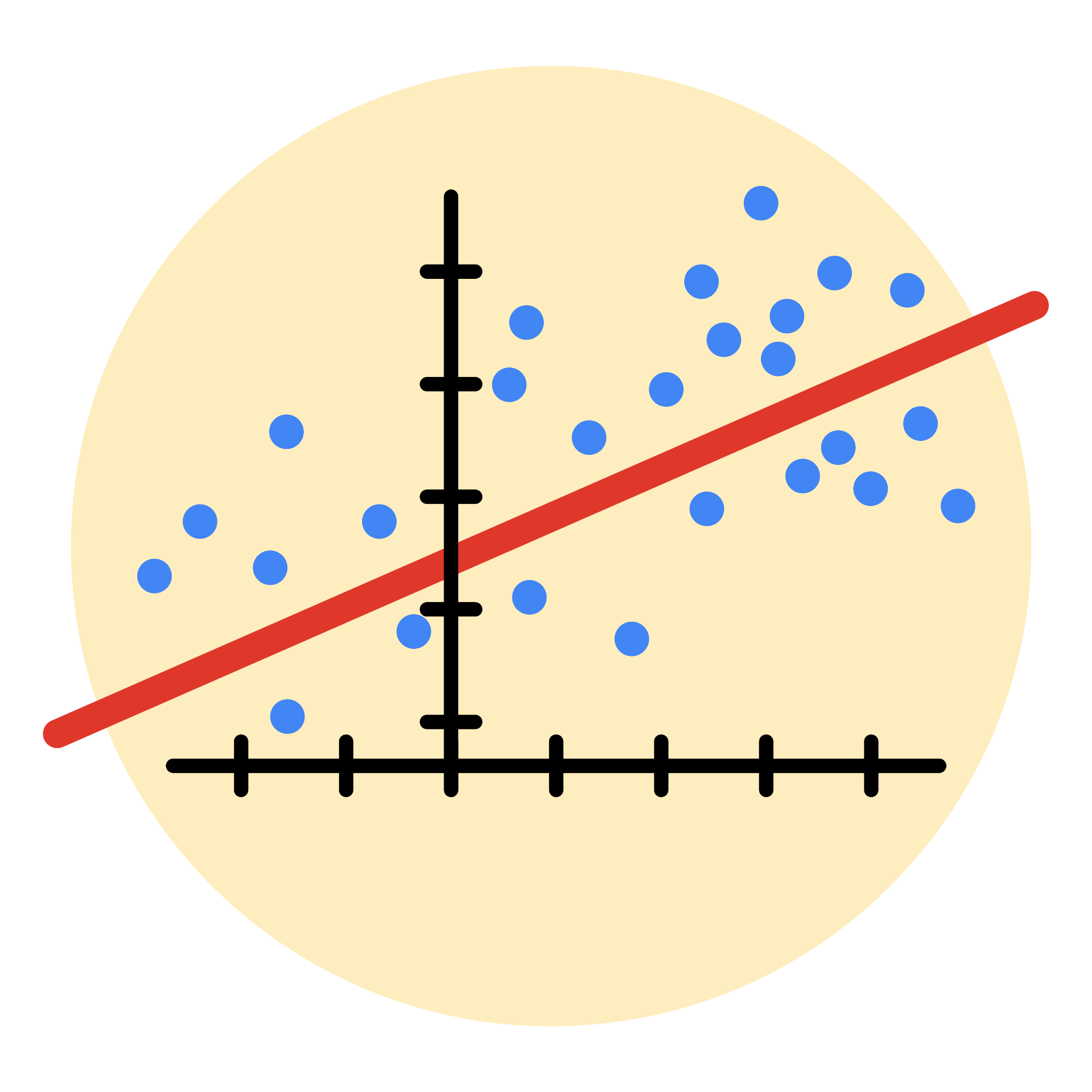
**Course Five**

# 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

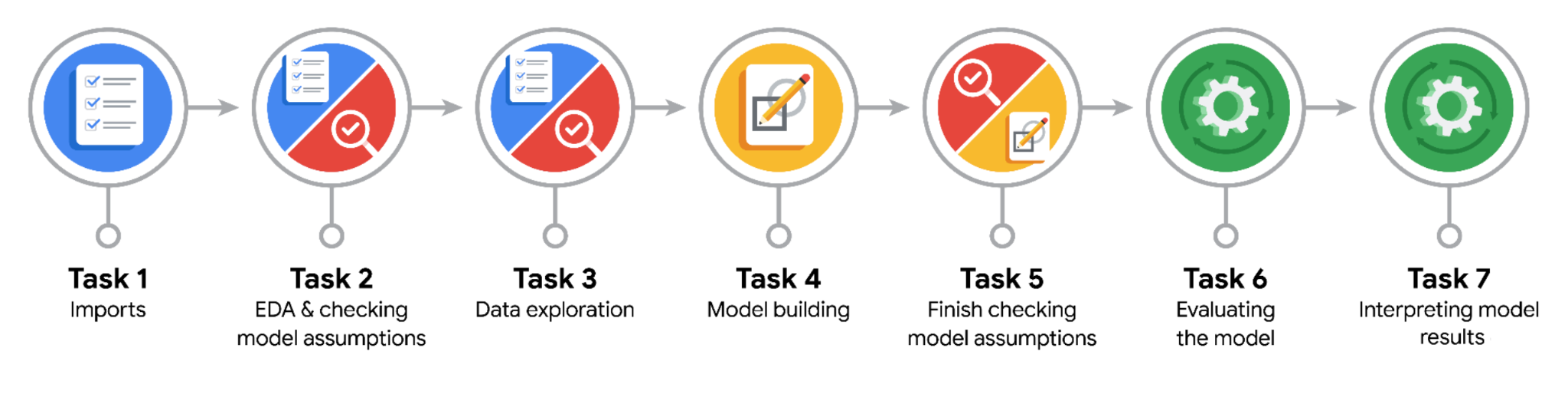
# Relevant Interview Questions

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?

TikTok users

* What are you trying to solve or accomplish?

I am verifying the correlations between variables sourced from given dataset. Then, I am building a model that can predict the status of users who upload contents on TikTok based on the identified correlations with variables

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

My initial observation was that I find so many outliers from the two variables, ‘video\_comment\_count’, and ‘video\_like\_count’

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

Pandas and visualization tool are useful

**PACE: Analyze Stage**

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

To have some sense of how the dataset would look like and to see how should I clean and structure the dataset.

* Do you have any ethical considerations in this stage?

Ethical considerations would be that me constructing a model to predict the result would affect others’ behaviors

**PACE: Construct Stage**

* Do you notice anything odd?

Approximately 94.2% of the dataset represents videos posted by unverified accounts and 5.8% represents videos posted by verified accounts. So the outcome variable is not very balanced.

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

Can utilize resampling to create class balance in the outcome variable.

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

pandas

**PACE: Execute Stage**

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

The dataset has some correlated variables, which might cause an issue with basic assumption in building regression model. I dropped ‘vidoe\_like\_count’ for that matter

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

Based on the logistic regression model, each additional second of the video is associated with 0.009 increase in the log-odds of the user having a verified status.

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

Interpreting beta coefficients is crucial for understanding the relationship between independent and dependent variables in a regression model. 1. Quantifying the impact 2. Identifying key drivers 3. Assessing model performance 4. Making predictions 5. Understanding the underlying mechnisms

* What potential recommendations would you make?

I would recommend determining the optimal video length that balances engagement and verification probability. A longer video might increase verification chances but could also lead to viewer fatigue.

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

Yes, there are other machine learning models available like decision tree or random forest models so I would try to use other models to see if other models show a better performance.

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

I would encourage the creation of high-quality, engaging content that encourages viewers to watch longer videos.

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

How stable is the model’s performance over time? Are there any seasonal or cyclical patterns that might affect the model’s accuracy?

* Do you have any ethical considerations at this stage?

No, I excluded the data that could cause any ethical concerns.