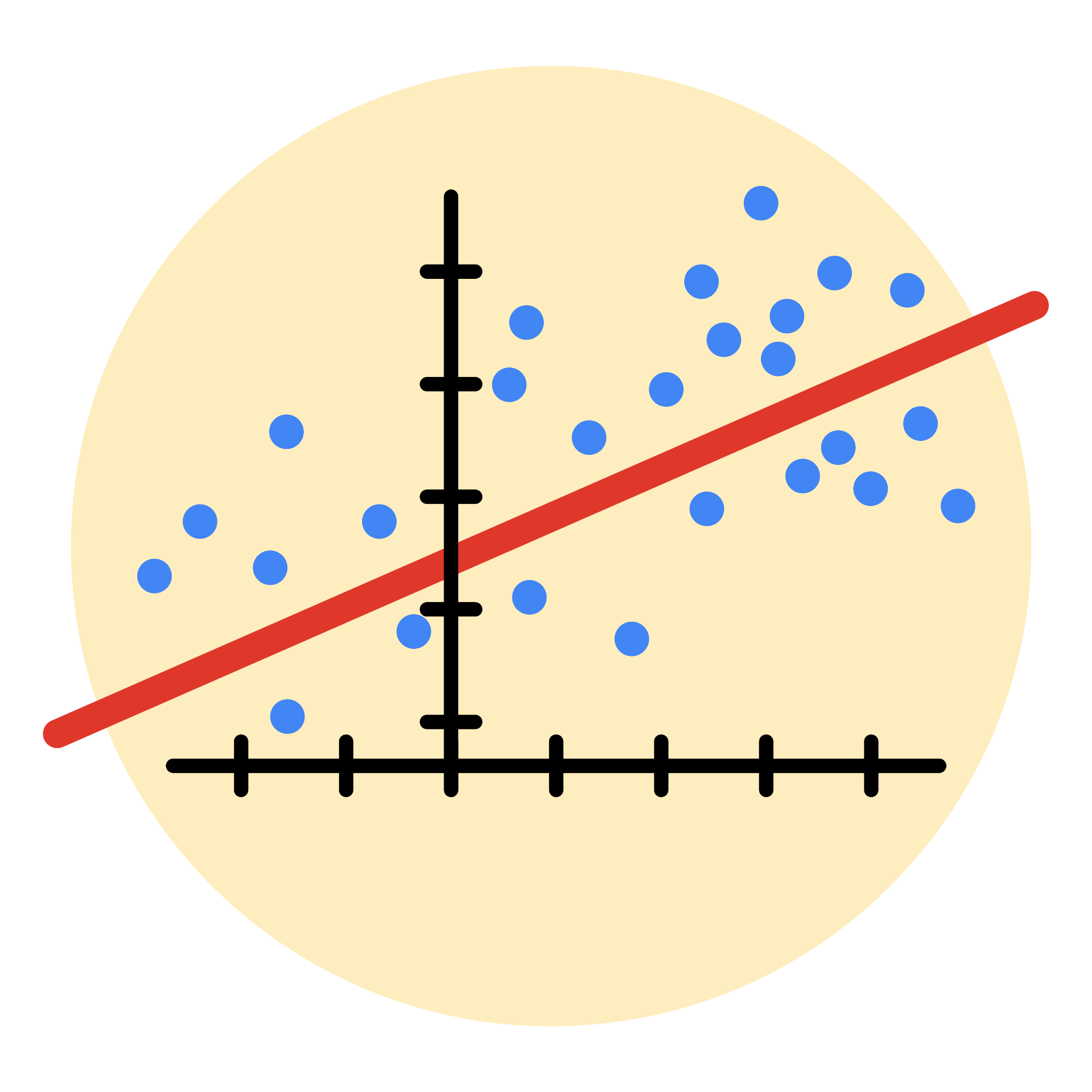
**Course Five**

# 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

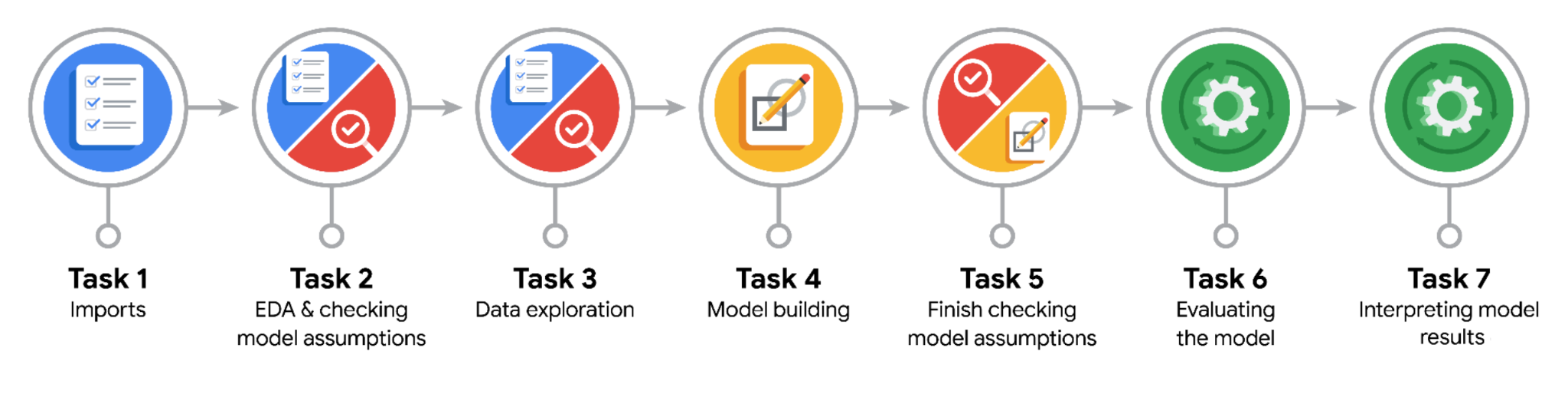
# Relevant Interview Questions

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?

Mary Joanna Rodgers- Project Management Officer

Margery Abebowale- Finance Lead, Americas

Maika Abadi- Operations Lead

* What are you trying to solve or accomplish?

I am trying to find what variables are useful for predicting author verification status.

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

There is a clear relationship between video length and author verification status.

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

Python and numerous packages such as numpy, sklearn, and seaborn.

**PACE: Analyze Stage**

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

It allows me to remove NA values and check for outliers.

* Do you have any ethical considerations in this stage?

Is it ok to remove the NA data? It should be ok because it is a small portion of the overall data set.

**PACE: Construct Stage**

* Do you notice anything odd?

Not all of the engagement metrics are good for predicting author verification status.

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

I would add more independent variables to increase the predictive ability of the model.

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

I found the past exercises very helpful. I also was able to research the process and code needed for the project.

**PACE: Execute Stage**

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

Video length was the key predictor variable for determining author verification status.

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

We should build the model based on the identified predictor variables.

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

Because the beta coefficients show the strength and direction of the relationship between the independent and dependent variables.

* What potential recommendations would you make?

That the machine learning model be built based on the model for identifying author verification status.

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

Yes, the model could be improved by adding additional predictor variables.

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

That this model be used as the basis for the machine learning model.

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

Are there other metrics that we can use to improve the accuracy of the model?

* Do you have any ethical considerations at this stage?

I need to be careful that my recommendations don’t lead to a biased machine learning model. The model is not perfect and so I need to weigh the risks of inaccurate predictions that the model will make.