One of the main objectives of this course is to help you gain hands-on experience in communicating insightful and impactful findings to stakeholders. In this project you will use the tools and techniques you learned throughout this course to train a few linear regressions on a data set that you feel passionate about, select the regression that best suits your needs, and communicate insights you found from your modeling exercise.

After going through some guided steps, you will have insights that either explain or predict your outcome variable. As a main deliverable, you will submit a report that helps you focus on highlighting your analytical skills and thought process.

**Review criteria**

Your peer will review your report from the perspective of a Chief Data Officer or the Head of Analytics and will assess whether your final linear regression went through all the necessary steps to achieve the main objective of your analysis.

Yes, you are expected to leverage a wide variety of tools, but these report should focus on present findings, insights, and next steps. You may include some visuals from your code output, but this report is intended as a summary of your findings, not as a code review. Optionally, you can submit your code as a python notebook or as a print out in the appendix of your document.

The grading will center around 5 main points:

* Does the report include a section describing the data?
* Does the report include a paragraph detailing the main objective(s) of this analysis?
* Does the report include a section with variations of linear regression models and specifies which one is the model that best suits the main objective(s) of this analysis.
* Does the report include a clear and well presented section with key findings related to the main objective(s) of the analysis?
* Does the report highlight possible flaws in the model and a plan of action to revisit this analysis with additional data or different predictive modeling techniques?

**Step-By-Step Assignment Instructions**

**Setup instructions:**

Before you begin, you will need to choose a data set that you feel passionate about. You can brainstorm with your peers about great public data sets using the discussion board in this module.

Please also make sure that you can print your report into a pdf file.

**How to submit:**

The format of your work must adhere to the following guidelines. The report should be submitted as a pdf. Optionally, you can include a python notebook with code.

Make sure to include mainly insights and findings on your report. There is no need to include code, unless you want to.

**Project**

***Optional: find your own data set***

As a suggested first step, spend some time finding a data set that you are really passionate about. This can be a data set similar to the data you have available at work or data you have always wanted to analyze. For some people this will be sports data sets, while some other folks prefer to focus on data from a datathon or data for good.

***Optional: participate in a discussion board***

As an optional step, go into a discussion board and brainstorm with your peers great data sets to analyze. If you prefer to skip this step, feel free to use the Ames housing data set or the Churn phone data set that we used throughout the course.

**Required**

Once you have selected a data set, you will produce the deliverables listed below and submit them to one of your peers for review. Treat this exercise as an opportunity to produce analysis that are ready to highlight your analytical skills for a senior audience, for example, the Chief Data Officer, or the Head of Analytics at your company.

Sections required in your report:

* Main objective of the analysis that specifies whether your model will be focused on prediction or interpretation.
* Brief description of the data set you chose and a summary of its attributes.
* Brief summary of data exploration and actions taken for data cleaning and feature engineering.
* Summary of training at least three linear regression models which should be variations that cover using a simple linear regression as a baseline, adding polynomial effects, and using a regularization regression. Preferably, all use the same training and test splits, or the same cross-validation method.
* A paragraph explaining which of your regressions you recommend as a final model that best fits your needs in terms of accuracy and explainability.
* Summary Key Findings and Insights, which walks your reader through the main drivers of your model and insights from your data derived from your linear regression model.
* Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model adding specific data features to achieve a better explanation or a better prediction.

**Frequently Asked Questions**

Here are frequently asked questions about the assignment and review process. Read these before starting your assignment.

Do I have to come up with my own data set?

* You are highly encouraged to find a data set you feel really passionate about. This will help you showcase analytical work that truly matches your skills. But if you prefer, you can use some of the data sets from this course.

Is it OK to choose the same data set as someone else?

* Yes, more than one person can analyze the same data set. Most likely your insights will be different from your peers and you will still be able to showcase your own talent as a unique solution.

Do I have to train more than 3 variations of a linear regression?

* It is highly recommended that you try at least three variations of a linear regression model to highlight which tool or technique improved your prediction or interpretation.

Is this an individual assignment?

* You can ask for help or assistance on technical issues and general direction of your analysis, but the interpretation of the analytical output and the writing of the report should be your own.

**Examples of Good Feedback**

As a reviewer, you will be required to provide feedback on the work of your peers.

When providing feedback, it is a good practice to first indicate what the learner did well on the assignment. After that, help the learner improve by providing specific, actionable advice.

Here are a few quotes of feedback collected from other learners who have taken this class before.

"It is great that you included scatter plots for the predicted vs actual of the outcome variable across all your regressions. It really showcases the pros and cons of each of your models."

"On this particular regression, you might want to try to remove the outliers. I believe that would improve your predictions".

"The comparison across your regression models could be more exhaustive. It is hard to assess model performance using just one error metric."

"You might want to use encoding for your categorical variables. The current settings do not help you leverage any of your categorial features."

"There were no insights or key takeaways from your analysis. You already put in great work, just needed to interpret your output a bit more into what a stakeholder should retain from your great analysis."

"If you feel really passionate about data related to the health crisis, your next step might be to merge it with data from this repository."