# Instructions:

In this Lab, you will demonstrate the data regression skills you have learned by completing this course. You are expected to leverage a wide variety of tools, but also this 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.

The grading will center around 5 main points:

1. Does the report include a section describing the data?
2. Does the report include a paragraph detailing the main objective(s) of this analysis?
3. 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.
4. Does the report include a clear and well-presented section with key findings related to the main objective(s) of the analysis?
5. 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?

# Import the required libraries

The following required modules are pre-installed in the Skills Network Labs environment. However if you run this notebook commands in a different Jupyter environment (e.g. Watson Studio or Ananconda) you will need to install these libraries by removing the # sign before !mamba in the code cell below.

In [ ]:

```
# All Libraries required for this lab are listed below. The libraries pre-installed on Skills Network Labs are commented.
# !mamba install -qy pandas==1.3.4 numpy==1.21.4 seaborn==0.9.0 matplotlib==3.5.0 scikit-learn==0.20.1
# Note: If your environment doesn't support "!mamba install", use "!pip install"
```

In [ ]:

```
import pandas as pd
```

# Importing the Dataset

Before you begin, you will need to choose a data set that you feel passionate about or you can use data set of **insurance.csv** file.

Read your chosen dataset into pandas dataframe:

In [ ]:

```
#data = pd.read_csv('Data/ insurance.csv')
#data.head()
```

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 explain ability.
- 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.

# 1. Dataset Description

   a. **Dataset Statistics**
   b. **Convert categorical features into numerical features**
   c. **Correlations Study**

# 2. Determining Normality:

Making our target variable normally distributed often will lead to better results If our target is not normally distributed, we can apply a transformation to it and then fit our regression to predict the transformed values. How can we tell if our target is normally distributed?
Transformations techniques to get or approach normal distribution:

   a. Square root
   b. Log
   c. Box cox

# 3. Applying Various Regression Models

# 4. Applying various linear regression models with advanced techniques

Through the following steps for **Vanilla Linear Regression, Lasso Regression, Ridge Regression** and **ElasticNetCV:**

- Chain multiple data processing steps together using Pipeline
- Use the KFolds object to split data into multiple folds.
- Perform cross validation using SciKit Learn with cross_val_predict and GridSearchCV

# 5. Models Flaws and Strength and further suggestions: