# Data Science Essentials

Lab 4 – Visualizing Data

## Overview

In this lab, you will learn how to use R or Python to visualize and explore data. Before creating analytical models, a data scientist must develop an understanding of the properties and relationships in a dataset. There are two goals for data exploration and visualization. First to understand the relationships between the data columns. Second to identify features that may be useful for predicting labels in machine learning projects. Additionally, redundant, collinear features can be identified. Thus, visualization for data exploration is an essential data science skill.

In this lab you will explore two datasets. Your first goal is to explore a dataset that includes information about automobiles, which you want to use to create a solution that predicts the price of an automobile based on its characteristics. This type of predictive modeling, in which you attempt to predict a real numeric value, is known as *regression;* and it will be discussed in more detail later in the course – for now, the focus of this lab is on visually exploring the data to determine which features may be useful in predicting automobile prices.

After exploring the automobile data, you will turn your attention to some adult census data, which you plan to use to classify people as high or low income based information known about them. This technique of predicting whether data entities belong to one class or another is known as *classification*, and will be discussed later in the course.

**Note**: This lab builds on knowledge and skills developed in the preceding labs in this course. If you have little experience with R or Python, and have not completed Lab 3, you are encouraged to do so now.

## What You'll Need

To complete this lab, you will need the following:

- A Web browser
- An Azure Machine Learning workspace.
- The files for this lab

**Note**: To set up the required environment for the lab, follow the instructions in the Setup Guide for this course.

# Upload a Jupyter Notebook

The code for this lab is provided in a Jupyter notebook. Both R and Python versions of the notebook are provided.

1. Browse to https://studio.azureml.net and sign in using the Microsoft account associated with your free Azure ML account.
2. On the **Notebooks** tab, click **+NEW**. Then select the option to upload a notebook from a local file.
3. Select the **Visualizing Data (R).ipynb** or **Visualizing Data (Python).ipynb** file in the **Mod4** folder where you extracted the lab files for this course, accept the default name for the notebook, and select the appropriate language (R or Python 3).
4. After the notebook has been uploaded, open it and follow the instructions it contains.