15.072: The Analytics Edge **Fall 2021**
Short Homework 1: Linear Regression

Out: September 13; Due: September 23, 11:59 pm.

Submission Instructions

Please save your short answers to the questions below in PDF format. Include **all** of your code in PDF format in an Appendix to your short answers. Document your code to indicate which sections correspond to each question. Submit all of your work in a single PDF to Gradescope. You do not need to submit any *.R* files or *.csv* files.

Forecasting Automobile Sales

Almost all companies seek accurate predictions of future sales—so they can better match production with demand, reduce unnecessary inventory costs, and design appropriate pricing strategies.

In this problem, you will aim to predict the monthly US sales of (i) the Jeep Wrangler, a compact SUV (Sports Utility Vehicle) with off-road capability manufactured by Fiat Chrysler Automobiles (FCA), and (ii) the Elantra, a compact sedan manufactured by the Hyundai Motor Company. To this end, you will build a linear regression model using US economic indicators and Google search query volumes.

The data is contained in the file `WranglerElantra2018.csv`. The dataset comprises 108 observations, one for each month from January 2010 to December 2018. It contains 10 columns, described in Table 1.

Table 1: Variables in the dataset `WranglerElantra2018.csv`.

Variable	Description
`Month`	The observation month given as a numerical value (1 = January, 2 = February, 3 = March, etc.).
`Year`	The observation year.
`Wrangler.Sales`	The number of units of the Jeep Wrangler sold in the United States in the given month and year.
`Elantra.Sales`	The number of units of the Hyundai Elantra sold in the United States in the given month and year.
`Unemployment.Rate`	The estimated unemployment rate (given as a percentage) in the United States in the given month and year.
`Wrangler.Queries`	A (normalized) approximation of the number of Google searches for "jeep wrangler" in the United States in the given month and year.
`Elantra.Queries`	A (normalized) approximation of the number of Google searches for "hyundai elantra" in the United States in the given month and year.
`CPI.All`	The consumer price index (CPI) for all products for the given month and year. This is a measure of the prices paid by consumer households for goods and services.
`CPI.Energy`	The monthly consumer price index (CPI) for the energy sector of the US economy for the given month and year.

The data used in this problem were obtained from publicly-available online sources:

- Elantra sales: `https://www.conceptcarz.com/monthly-sales/28079/hyundai-elantra.aspx`
- Wrangler sales: `https://www.conceptcarz.com/monthly-sales/27943/jeep-wrangler.aspx`
- Unemployment rates: `https://data.bls.gov/timeseries/LNS14000000`
- Google search query quantities: `https://www.google.com/trends/explore`
- Consumer Price Index (All): `https://fred.stlouisfed.org/series/CPIAUCSL#`
- Consumer Price Index (Energy): `https://fred.stlouisfed.org/series/CPIENGSL#`

Throughout this problem, you will use a training set comprising all observations in 2010–2017, and a test set comprising all observations in 2018.

a. Build an initial linear regression model to predict monthly Wrangler sales with five independent variables: **`Year`**, **`Unemployment.Rate`**, **`Wrangler.Queries`**, **`CPI.Energy`**, and **`CPI.All`**. [**20 pts**]

   (i) Report the following in-sample and out-of-sample performance metrics: R^2, mean absolute error (MAE) and root-mean-square error (RMSE).

b. Update your model by choosing a subset of these five variables. [**20 pts**]

   (i) Which variables did you choose and why?

   (ii) Report the in-sample and out-of-sample performance metrics (R^2, MAE, RMSE).

   (iii) Compare your results to Question a.

c. Add **`Month`** to the model from Question a. to capture seasonality. [**20 pts**]

   (i) Report the in-sample and out-of-sample performance metrics.

   (ii) Compare your results to Question a.

d. Now, fit the same model as in Question c. to predict the monthly sales of the Hyundai Elantra. [**20 pts**]

   (i) Report the in-sample and out-of-sample performance metrics.

e. Comment on the performance of your regression model with the two car models. [**20 pts**]

   (i) Contrast your results between the Jeep Wrangler and the Hyundai Elantra.

   (ii) Provide two insightful visualizations to support your argument.