DBA4711
Homework Assignment #2

You must post your completed assignment on Canvas.

Problem 1: [INDIVIDUAL] Generalized Regression (30 points)
Consider the SBC.xlsx shared with the problem set. The task is to build a linear regression model to predict the Annual Sales using the remaining data fields.

- Build a regression model on R or excel
- Identify the most significant variables and build the final model.
- Detect an outlier aware linear regression model by excluding 10% of the outlier data. How does the regression model change?
- Build a min-max regression model. How does the regression model change.

Problem 2: [GROUP]Forecasting Jeep Wrangler and Hyundai Elantra Sales (70 points)

Almost all companies seek accurate predictions of future sales of their products. Clearly, if a company can accurately predict sales, it can better match production with customer demand, thus reducing unnecessary inventory costs while being able to satisfy all demand.

In this exercise you are asked to predict the monthly US sales of the Jeep Wrangler (manufactured by Fiat Chrysler Automobiles (FCA)) and Elantra (manufactured by Hyundai Motor Company) automobiles, both of which are sold all over the world. The Wrangler is a compact SUV (Sports Utility Vehicle) with off-road capability made by Jeep – a subsidiary of FCA – and the Elantra is a compact sedan manufactured by Hyundai. Herein you are asked to build a linear regression model to predict monthly US sales of the Wrangler and the Elantra using economic indicators of the United States as well as Google search query volumes. The data for this problem is contained in the file `WranglerElantra2018.csv`, which you will need to download from the Canvas course site under the "Homework" section. Each observation in the file is for a single month, from January 2010 through December 2018. The variables are described in Table 1.

Table 1: Variables in the dataset WranglerElantra2018.csv.

Variable	Description
`Month.Numeric`	The observation month given as a numerical value (1 = January, 2 = February, 3 = March, etc.).
`Month.Factor`	The observation month given as the name of the month (which will be a factor variable in R).
`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 magnitude 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.

a) Start by reading `WranglerElantra2018.csv` into R (do not forget to navigate to the directory on your computer containing `WranglerElantra2018.csv` first). Then split the data into a training set and test set. The training set should contain all observations for 2010–2017. The test set should have all observations for 2018.

Consider the five independent variables **`Year`**, **`Unemployment.Rate`**, **`Wrangler.Queries`**, **`CPI.Energy`**, and **`CPI.All`**. Using your regression skills, you will choose a subset of these five variables and construct a regression model to predict monthly Wrangler sales (**`Wrangler.Sales`**). Use the training set to build your model, and do not add any additional variables beyond the five indicated independent variables.

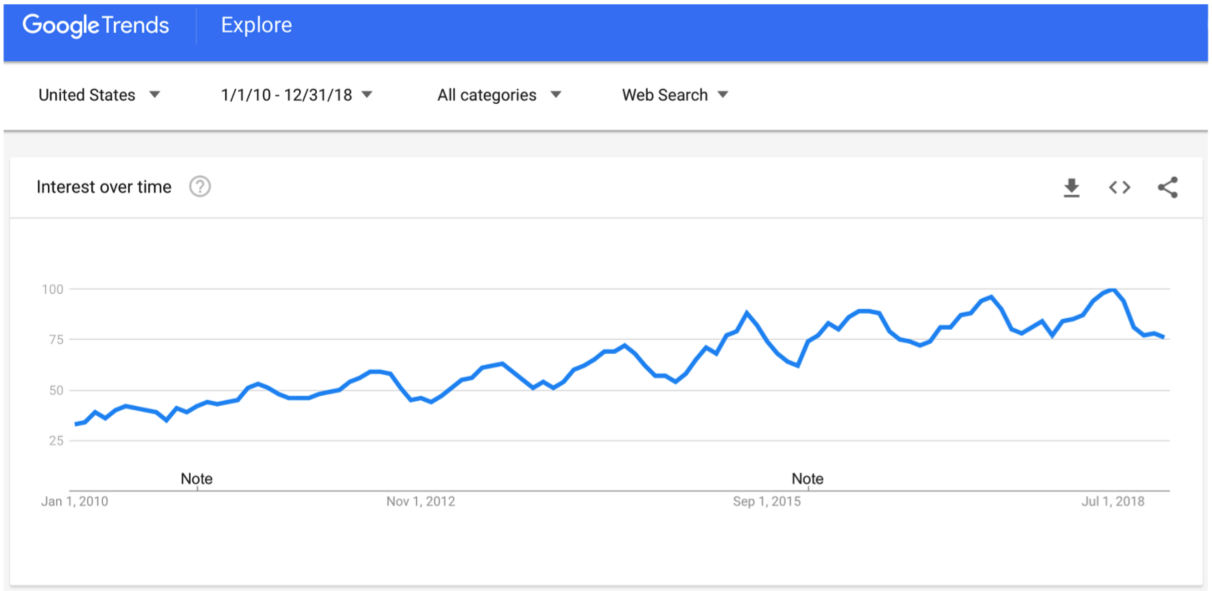
i) (10 points) Build an initial linear model with all five independent variables. Based on model output, which variables are significant, i.e. have at least one "star" in the summary output (or more mathematically, have a p-value less than 0.05)?

ii) (6 points) Choose a subset of the five independent variables to construct a new linear model.

1) (4 points) Justify your choice of variables.
2) (4 points) What is the linear regression equation produced by your new model, and what is your interpretation of the coefficients for the independent variables?
3) (4 points) Do the signs of the model's coefficients make sense?

iii) (4 points) How well does the model predict training-set observations, as captured, for instance, by the R^2 value of the model? In a similar spirit, how well does the model predict test-set observations, as captured, for instance, by the OSR^2 value of the model?

b) **Google Trends.** (4 points) One of our feature variables, `Wrangler.Queries`, was obtained from publicly-available data in Google Trends. In Figure 1, we show a time-series plot of search queries for "jeep wrangler" over the time span from 2010–2018. What trend do you observe? Does this trend make intuitive business sense?



United States ▾ 1/1/10 - 12/31/18 ▾ All categories ▾ Web Search ▾

Figure 1: "jeep wrangler" search queries (normalized) from Jan. 1, 2010 to Dec. 31, 2018. Jourdain says: I updated this picture.

Visit `https://www.google.com/trends/explore` to explore this resource on your own. There are many rich datasets publicly available online; for example, the other features and outcome variables for this regression problem were obtained from:

- Monthly auto sales: `http://www.goodcarbadcar.net/p/sales-stats.html`
- Unemployment rates: `https://data.bls.gov/timeseries/LNS14000000`
- Consumer Price Index (All): `https://fred.stlouisfed.org/series/CPIAUCSL#`

- Consumer Price Index (Energy): `https://fred.stlouisfed.org/series/CPIENGSL#`

It is extremely valuable to be aware of and be able to access these and other such datasets yourself. In fact, publicly-available datasets are likely to be valuable for your Final Project in this course. Even if you have access to proprietary company data, you may still be able to improve the predictive power of your models by adding in publicly-available data as predictive features, as we have done here with Google search frequencies.

c) Drawing on intuition from Figure 1, let us now try to further improve the linear regression model by modeling seasonality. Construct a new linear regression model using the **`Month.Factor`** variable as an independent variable. And as before, construct your model based on the training data.

i) (4 points) Describe your new model. What is the regression equation? (Do not simply copy and paste output from R.) What is your interpretation of the coefficients of each of the **`Month.Factor`** dummy variables?

ii) (6 points) Which variables are significant? What is the training set R^2? Test set OSR^2?

iii) (6 points) Do you think adding the independent variable **`Month.Factor`** has improved the quality of the model? Why or why not?

iv) (4 points) Can you think of a different way that you might use the given data to model seasonality? Do you think your new way would improve on the best model you have constructed so far? (By the way, later in the course we will have a lecture dedicated to *time series* modeling, and we will explore a number of ways to construct models using datasets with an associated time component.)

v) **Elantra Sales.** (6 points) Now, on the same training data set, build a linear regression model to predict the outcome variable **`Elantra.Sales`** (i.e., monthly US Elantra sales) using a subset of the independent variables **`Year`**, **`Unemployment.Rate`**, **`Elantra.Queries`**, **`CPI.Energy`**, and **`CPI.All`**. What is the training set R^2? OSR^2? (*Note:* Your model will probably not look very good.)

vi) (4 points) Compute all correlations among **`Elantra.Sales`** and the five independent variables listed in question d) by running the following command in your R console (assuming you loaded the data into a data frame called **`onedata`**):

```
cor(onedata[,c("Elantra.Sales", "Year", "Unemployment.Rate", "Elantra.Queries",
               "CPI.All", "CPI.Energy")])
```

What do you observe? Does this help explain why the Elantra model is comparatively less predictive?