1. "The http://fueleconomy.gov website, run by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy and the U.S. Environmental Protection Agency, lists different estimates of fuel economy for passenger cars and trucks. For each vehicle, various characteristics are recorded such as the engine displacement or number of cylinders. Along with these values, laboratory measurements are made for the city and highway miles per gallon (MPG) of the car."

   Predictors extracted from the website include: EngDispl, NumCyl, Transmission, AirAspirationMethod, NumGears, TransLockup, TransCreeperGear, DriveDesc, IntakeValvePerCyl, ExhaustValvesPerCyl, CarlineClassDesc, VarValveTiming and VarValveLift. The response variable is FE, which is the unadjusted highway data.

   To access this data set, you will need to install the AppliedPrectiveModeling library. The following code will produce 3 data sets (cars2010, cars2011 and cars2012, which are the training, validation and test data sets, respectively).

```
library(AppliedPredictiveModeling)
data(FuelEconomy)
```

   a. Generate scatter plots and correlations for the variables **EngDispl, NumCyl, ExhaustValvesPerCyl** and the **VarValveTiming** versus the target variable, **FE**.
      - Can linear relationships adequately describe these relationships?
      - Are there any outliers that you should investigate?
      - What variable has the highest correlation with **FE**?

- What is the p-value for that correlation coefficient? Is it statistically significant at the 0.05 level? What can you conclude?

b. Generate correlations among all of the variables in the previously mentioned variables, minus the target, **FE**. Are there any notable relationships?

c. Fit a simple linear regression model with **FE** as the response variable and **EngDispl** as the predictor.
   - What is the value of the F Statistic and the associated p-value? How would you interpret this with regard to the null hypothesis?
   - Write the predicted regression equation.
   - What is the value of R-square? How would you interpret this?

2. The IceCream dataset has two columns, **sales** which gives the total daily sales of a local ice cream shop in hundreds of dollars, and **temperature** which reflects the daily high temperature.

   a. Run a regression analysis predicting daily sales from temperature.
      - Are the errors of your model normally distributed? What evidence would you cite here?
      - Do you see evidence of any relationship between temperature and sales? What statistical evidence (think: p-value) would you cite here?
      - What is the parameter estimate for temperature in the model equation? Interpret this parameter using a sentence.

3. The dataset *MinnTemp* has information for the daily average temperature for a weather station in Minneapolis. The variables **temp** and **time** provide the temperature and time measurements respectively. Time is

measured in hours since the study began.

b. Perform a regression analysis predicting **temperature** using the **time** variable.
   - Are the errors of your model normally distributed? What evidence would you cite here?
   - Do you see violations of our assumptions for simple linear regression? If so, what problems do you see?
   - Is there statistical evidence that **time** is related to **temperature** at the confidence level of 0.05? If so, describe the relationship in a sentence, if not discuss what your next steps in this analysis might be.