**Case Study – Hypothesis Test, Linear Regression**

Submit a report (PDF or Word) that includes:

* Answers to the questions below with relevant visualizations/outputs.
* R/Python code corresponding to each question (attach the code under each answer).

Problem 1 – Hypothesis Testing using Simulation (10 pts)

The dataset iris contains measurements on three species of iris flowers. In this assignment, you’ll focus only on the 50 observations with Species == "versicolor". Complete both of the following tasks and report results and interpretations.

A botanist claims that the average sepal length of her Iris versicolor plants is 5.5 cm. You will test this claim via simulation.

Problem 2 – Linear Regression (25 pts)

The airquality dataset contains daily air quality measurements in New York, May–September 1973:

| **Variable** | **Description** |
| --- | --- |
| Ozone | Daily maximum ozone |
| Solar.R | Solar radiation |
| Wind | Average wind speed |
| Temp | Maximum daily temperature |
| Month | Month (5 = May … 9 = September) |
| Day | Day of month |

.

1. Simple linear regression: investigate if wind speed largely explains ozone variation
   1. Initial exploration - make a scatterplot and calculate correlation coefficient to identify the strength of their relationship.
   2. Fit a model, write down the fitted model and interpret: the coefficient estimates, model performance.
   3. Predict the ozone concentration when wind = 10 mph.
   4. Check model assumptions using diagnostic plots.
2. Multiple Linear regression: include wind speed, temperature and solar radiation to improve predictions
   1. Multicollinearity check using pairwise correlation plot.
   2. Fit a model, write down the fitted model and interpret: the coefficient estimates, model performance.
   3. Check model assumptions using diagnostic plots.