DATS 6101 - Introduction to Data Science

Homework Assignment 5

Linear Regression

Question 1

Read "Pearson vs Spearman.pdf". We will use it next week.

Create a new Rmarkdown file and complete the below tasks. The rest of the exercise uses the bikeshare data (bikedata.csv)

Bikeshare Dataset

Question 2
Import the data, call it "bikeorig". The "Date" variable is probably imported as factor level variable. Copy all other variables except Date, Casual Users, and Registered Users into a new datafame. Call it "bike". How many variables are in "bike"? How many of them are imported as "int"? Feel free to rename longer variable names into shorter ones for convenience.

Question 3
Select only the subset with Hour between 14 and 18 inclusively. These are the afternoon rush hour data. How many observations are there?

Question 4
Find a nice way to present some correlation among different pairs of variables.

Question 5
From your understanding of the variables, which ones should be changed to factor level variables? Change them so that the analysis make sense.

Question 6
Find linear models to predict the "Total Users". Use only linear terms for now. Show us what model you end up with. Remember to check the VIF values for each model.

Question 7
Comment on the coefficient values, p-values, multiple R-squared of the final model.

Question 8
Also find the confidence intervals of the coefficients of your final model.

Question 9
Try at least one more model with interaction terms or non-linear terms. Does the multiple R^2 value improve? What else did you find?