Texas Christian University
COSC 40023 Spring 2025
Assignment 5

Due: April 2nd, Wednesday, at 11:30 PM (Late submission NOT accepted)
Submission (two files, no compression): `assignment5.py` and `assignment5.R` through TCU Online

Download the dataset called "Housing_Data.csv". In the dataset, each observation is corresponding to a house/apartment sold in New Taipei City, Taiwan. Here are the descriptions of the columns.

`X1` = the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)
`X2` = the house age (unit: year)
`X3` = the distance to the nearest subway station (unit: meter)
`X4` = the number of convenience stores in the living circle on foot (integer)
`X5` = the geographic coordinate, latitude. (unit: degree)
`X6` = the geographic coordinate, longitude. (unit: degree)
`Y` = house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

Build an SVR model and a random forest regression model in a Python file named `assignment5.py`. Then, build an SVR model and a random forest regression model in an R file named `assignment5.R`. Here are some additional requirements.

1. `25%` of the data should go to the test set. Do NOT set `random_state` in Python or `seed` in R.
2. When building SVR models, use RBF kernel in Python and epsilon regression with RBF kernel in R. Keep the default values of the other parameters.
3. Calculate adjusted R-squared on the test set for the SVR models. In each language, manually rerun the script of splitting the dataset, building the SVR model and calculating adjusted R-squared for 10 times. Have comments listing 10 calculated adjusted R-squared values. Then, have another comment to clearly show the average adjusted R-squared of the 10 values.
4. When building random forest regression models, set number of trees to 500 for both Python and R. Keep the default values of the other parameters.
5. Calculate adjusted R-squared on the test set for the random forest regression models. In each language, manually rerun the script of splitting the dataset, building the random forest regression model and calculating adjusted R-squared for 10 times. Have comments

listing 10 calculated adjusted R-squared values. Then, have another comment to clearly show the average adjusted R-squared of the 10 values.

6. In each language, compare the average adjusted R-squared values obtained from the SVR model and the random forest regression model. At the end of each file, have a comment to clearly state which model has better performance.
7. Have sufficient single-line and multi-line comments.