DSC 520 Final Project Template

Name: Edris Safari
Date: 02/06/2020
Title: Real Estate Data Analysis

Section 1 – Getting Started

In this section, we will describe three data sets that will be examined for this project. Datasets are related in the Real Estate industry but vary in content. Each dataset will be examined according to their content. A combined report will detail the findings of each dataset.

Dataset 1

Data Source

https://www.kaggle.com/samdeeplearning/vt-nh-real-estate

Description

This dataset contains features of houses in three towns in Vermont, which make up a sizable chunk of the real estate firm's business. The dataset is divided into test, train and validate data sets with test having 24 rows, train 138 and validate with 70 rows. There are 28 column describing features such as number of bedrooms, yard size, etc.

Goal

We will try to cross validate the results between Train, Validate, and, Test. We will select appropriate independent variables after exploratory data analysis and run a regressions model which we will test against the test dataset and then validate.

Dataset 2

Data Source

https://www.kaggle.com/quantbruce/real-estate-price-prediction

Description

Dataset columns are 'transaction_date' , 'house_age' , 'distance_to_the_nearest_MRT_station' , 'number_of_convenience_stores' , 'latitude' , 'longitude' , 'house_price_of_unit_area'. There are 500 records in the dataset.

Goal

Evaluate Correlation between independent variables and make prediction on the dependent variable 'house_price_of_unit_area'

Dataset 3

Data Source

https://www.kaggle.com/c/zillow-prize-1

Description

There are two data sets with over 1 million records each and 58 columns. properties_2016 and properties_2017 datasets contain data for each year.

Goal

Reduce error between actual home price and zillow's estimate. We will perform EDA and correlation analysis and create a model to examine variation in estimates vs. actual home price.

Needed packages

Packages needed for calculation, analysis and plotting the data sets are listed below:

library(data.table)

library(dplyr)

library(ggplot2)

library(stringr)

library(DT)

library(tidyr)

library(corrplot)

library(leaflet)

library(lubridate)

Plots and tables

We will create histograms and density plots, scatter plots and correlation plots to examine and understand the data.

Questions for hypothesis testing

1. Does location determine price?
2. Which affect the price of a home? Number of rooms and the square foot?
3. How Does age affect the price of a home?
4. What are the common factors among certain price range?

Section 2 –Cleaning Data and Exploratory Data Analysis

TBD

Section 3 –Writeups

TBD

Section 4 – Final

TBD

Section 1 – Week 9

- Introduction
- Research questions
- Approach
- How your approach addresses (fully or partially) the problem.
- Data
- Required Packages
- Plots and Table Needs
- Questions for future steps.

Section 2 – Week 10

- How to import and clean my data
- What does the final data set look like?
- Questions for future steps.

Section 3 – Week 11

- What information is not self-evident?
- What are different ways you could look at this data?
- How do you plan to slice and dice the data?
- How could you summarize your data to answer key questions?
- What types of plots and tables will help you to illustrate the findings to your questions?
- Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.
- Questions for future steps.

Section 4 – Week 12

- A story / narrative that emerged from your data. Follow this structure.
  - Introduction.
  - The problem statement you addressed.
  - How you addressed this problem statement
  - Analysis.
  - Implications.
  - Limitations.
  - Concluding Remarks