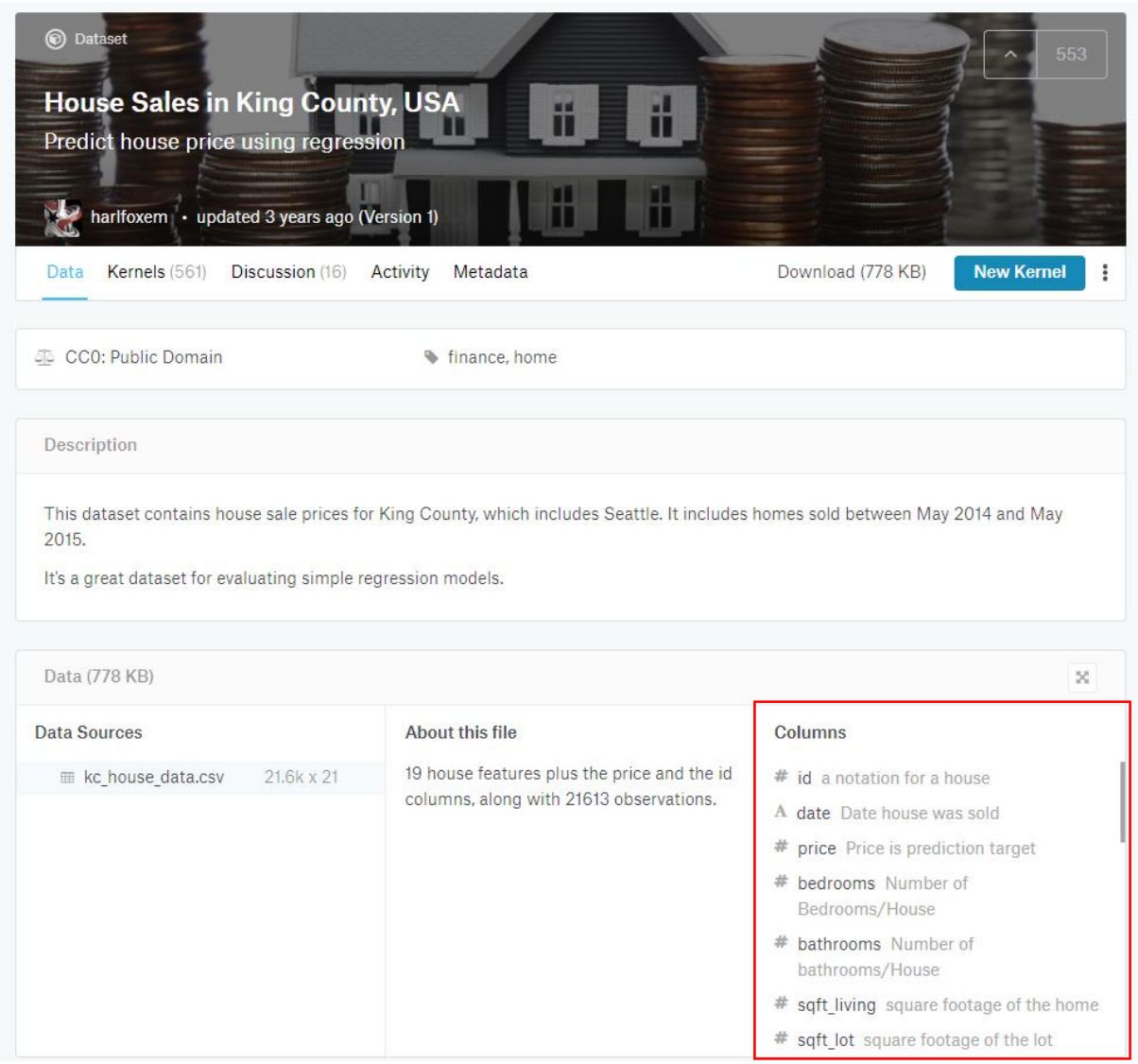
Multivariate Data Analysis Assignment #3

Dataset: House Sales in King County, USA, kc_house_data.csv

(https://www.kaggle.com/harlfoxem/housesalesprediction)

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.



References

http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

http://www.ggplot2-exts.org/gallery/

[Assignment instructions]

Create your own Rmd file with the necessary R script blocks and explanations with mark down syntax by following the instructions below:

[Data Preparation]

[1] Load the dataset using read.csv( ) functions

[2] Create a bar plot with regard to the number of bedrooms

[3] Remove the rows (1) without rooms (number of rooms =1) or (2) with rooms more than 6 (number of rooms > 6)

[4] Randomly sample 5,000 rows for computational ef랴챠두쵸

[Plotting and Interpretation]

[1] Create at least 10 different plots and interpret the results to understand the dataset itself. Visit the recommended reference sites to see what types of graphs can be generated by "ggplot" package. Googling is strongly recommended to create your own plots.

[2] Establish at least 5 hypotheses about the price (ex: The number of rooms is positively related to the price. In other words, houses with more rooms are generally more expensive than houses with fewer rooms)

[3] Create appropriate plots to qualitatively (not quantitatively based on the hypothesis testing method generally used in statistics) verify each hypothesis and determine whether each hypothesis can be accepted based on the generated plot.