

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.

The screenshot shows the Kaggle dataset page for 'House Sales in King County, USA'. The header includes the dataset title, a subtitle 'Predict house price using regression', and the creator 'harlfoxem' with a note 'updated 3 years ago (Version 1)'. Below the header is a navigation bar with tabs for 'Data', 'Kernels (561)', 'Discussion (16)', 'Activity', and 'Metadata'. A 'Download (778 KB)' button and a 'New Kernel' button are also present. The 'Data' tab is selected, showing a table with columns for 'Data Sources', 'About this file', and 'Columns'. The 'Data Sources' table lists 'kc_house_data.csv' with dimensions '21.6k x 21'. The 'About this file' section states: '19 house features plus the price and the id columns, along with 21613 observations.' The 'Columns' section lists the following features: '# id a notation for a house', '# date Date house was sold', '# price Price is prediction target', '# bedrooms Number of Bedrooms/House', '# bathrooms Number of bathrooms/House', '# sqft_living square footage of the home', and '# sqft_lot square footage of the lot'. The 'Columns' section is highlighted with a red border.

Data Sources	About this file	Columns				
<table border="1"><thead><tr><th>File Name</th><th>Dimensions</th></tr></thead><tbody><tr><td>kc_house_data.csv</td><td>21.6k x 21</td></tr></tbody></table>	File Name	Dimensions	kc_house_data.csv	21.6k x 21	19 house features plus the price and the id columns, along with 21613 observations.	<ul style="list-style-type: none"># id a notation for a house# date Date house was sold# price Price is prediction target# bedrooms Number of Bedrooms/House# bathrooms Number of bathrooms/House# sqft_living square footage of the home# sqft_lot square footage of the lot
File Name	Dimensions					
kc_house_data.csv	21.6k x 21					

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 efficiency

[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.