STATS 131 FINAL PROJECT

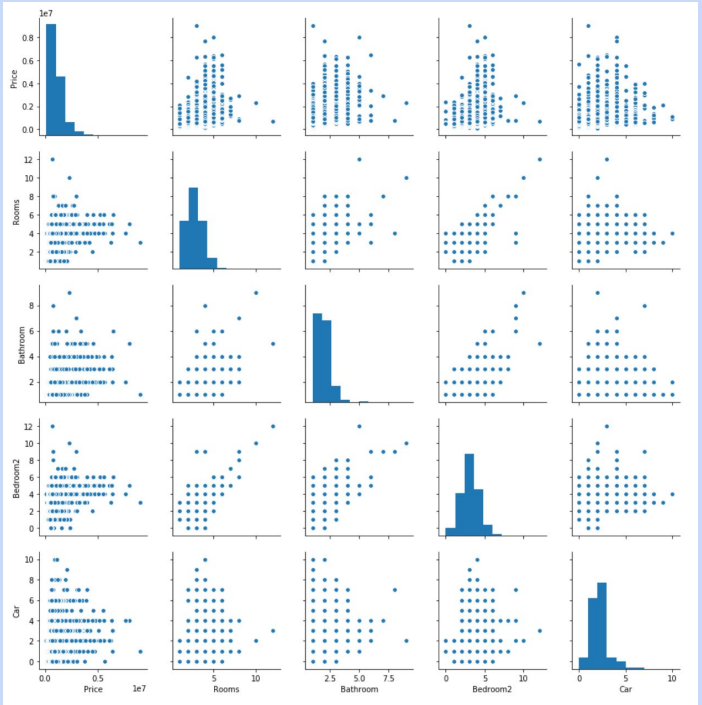
Melbourne Housing Data

-----Jieming Fang, Jin Wang, Huijun Yan

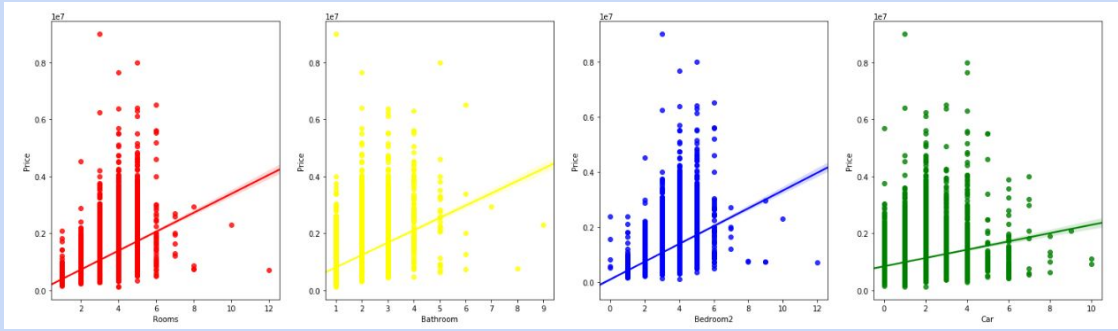
Background information on the subject and field of study.

- Choosing "Melbourne Housing Data" as our data set.
- Information of data collection:
  - Tony Pino collected the data starting from 2016
  - The data was scraped from publicly available resulted posted every week from Domain.com.au
  - The dataset includes Address, Type of Real estate, Suburb, Method of Selling, Rooms, Price, Real Estate Agent, Date of Sale and Distance from C.B.D.
  - 34857 observation & 21 variables
- Research Question:
  - What housing characteristics associate with housing price in Melbourne?

Exploratory Analysis of the Data



--Potential Relationships



- Scatter plots for relationships between numeric data
  - Create a Pairplot and linear regression plots to see the top 4 (“Rooms”; “Bathroom”; “Bedroom2”; “Car”) correlation with “Price”
  - It implies that there are collinearity relationship among “Rooms”; “Bathroom”; and “Bedroom2”

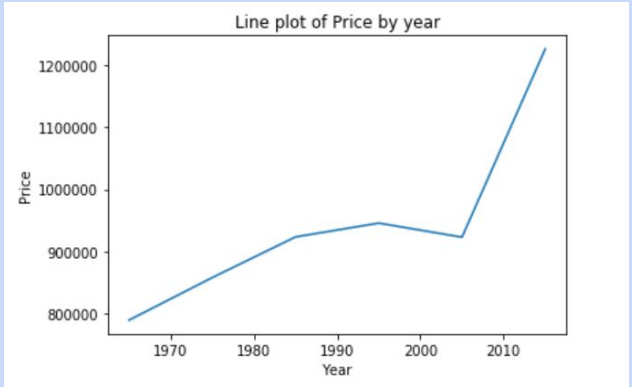
Exploratory Analysis of the Data

--Potential Relationships

- Analysis the Year of Built to create a new variabels (Year Group)
  - Based on the "YearBuilt" statistics (starting with it medium), we divide it into 6 groups
  - Table shows the mean price for each group
  - The line plot shows the trend of mean price between 1965 to 2020 year

yeargroup	Price
[1965, 1975)	7.894192e+05
[1975, 1985)	8.578595e+05
[1985, 1995)	9.233559e+05
[1995, 2005)	9.456095e+05
[2005, 2015)	9.229580e+05
[2015, 2025)	1.225991e+06

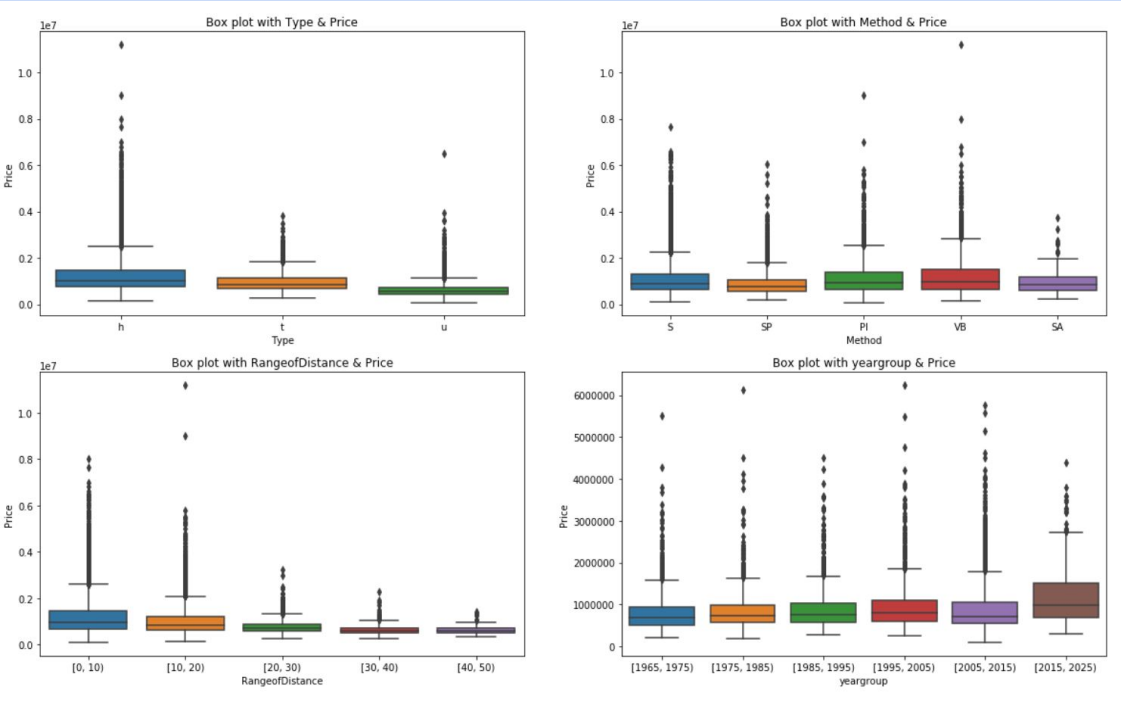
(6, 1)



Exploratory Analysis of the Data

--Potential Relationships

- Create boxplots to see the relationship of each categorical variables ('Type'; 'Method'; 'RangeofDistance'; 'yeargroup') with "Price"
  - "Method" v.s "Price": each level are similar (medium almost on the same line)
  - "Type" v.s "Price"; "RangeofDistance" v.s "Price" and "yeargroup" v.s "Price": different for each level (mediums are not on the same line)



Data Modeling

- Pick High correlation variables and "Price" as a new data
- use test size= 0.2 to split new data as training data and testing data.
- They have the same outcome variables---"Price"
- Model with Numerical & Categorical
  - Rooms + Bathroom + Car + YearBuilt + C(RangeofDistance) + C(Type)
  - Each predictors's P-value are equal to 0
  - MSE : 154504501092.9267
- Model Only with Numerical
  - Rooms + Bathroom + Car + YearBuilt
  - P-value: some of larger than the model 1
  - MSE: 182985781375.0511