2020/10/31

**What do we use to share code?**

Google colab: <https://colab.research.google.com/drive/1qOARCmOA46epX7PvwvvcZPAdSk7m_Fe5?usp=sharing>

**Problem statement**: scope? House type? Which prices do we predict?

**Dataset**

Code:<https://github.com/lijiaxingogo/HousePrice>

<https://www.kaggle.com/faressayah/linear-regression-house-price-prediction#6.-Polynomial-Regression>

Dataset: <https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page>

Dealing with right skew data: <https://cran.r-project.org/web/packages/GlmSimulatoR/vignettes/dealing_with_right_skewed_data.html>

**What to be done before next week?**

1. Look at the code <https://github.com/lijiaxingogo/HousePrice>
2. EDA

**Univariate + bivariate analysis**

1. BOROUGH; NEIGHBORHOOD; BUILDING CLASS CATEGORY; TAX CLASS AT PRESENT - Yi
2. BLOCK; LOT; EASEMENT; BUILDING CLASS AT PRESENT; YEAR BUILT; - Zhuoyuan Xu
   1. pre-war
3. ADDRESS; APARTMENT NUMBER; ZIP CODE; RESIDENTIAL UNITS - Siyu Shen
4. COMMERCIAL UNITS; TOTAL UNITS; LAND SQUARE FEET; GROSS SQUARE FEET - Fan Zhang
   1. How to filter out some of the data
   2. Land square feet/ gross square feet
5. TAX CLASS AT TIME OF SALE; BUILDING CLASS AT TIME OF SALE (Zhuoyuan); SALE PRICE (Target); SALE DATE - Congyun Jin

**Problem**

House types

<https://www1.nyc.gov/assets/finance/jump/hlpbldgcode.html>

1. Deciding Block-lot vs. zip code vs. borough - Zhuoyuan, Siyu
   1. Borough - it is highly correlated with block-lot and zip code
   2. There are more price variation between boroughs than within a borough (according to Siyu’s boxplot)

<https://towardsdatascience.com/stop-using-zip-codes-for-geospatial-analysis-ceacb6e80c38>

https://mindecology.com/blog/visually-target-your-best-customers-block-group-vs-zip-code-level-maps/

1. Tax class at present (13) + tax class at time of sale (3) - Yi, Congyun
   1. Tax class at time of sale
   2. Dropping tax class at present
2. One hot encoding all categorical variables - Congyun, Fan Zhang
   1. 30 classes - one hot encoding
   2. Dropping building class at present
3. Find where the outliers are - Zhuoyuan, Yi
4. Linear Regression, Ridge Regression, LASSO Regression, Elastic Net - Siyu, Fan
   1. Assumptions checking?
5. Robust Regression, Polynomial Regression - Zhuoyuan, Yi
   1. Assumptions checking?

**Email to Professor**

1. PPT or Word document - Fan to Professor
   1. How long?
      1. 5-min presentation + 3-min qa
   2. What is the deliverable format?
      1. Slide + word all in pdf
2. Change data/ problem - Yi
   1. Approved

**Clean data**

1. Identify and Delete outliers (outliers can be analyzed separately at the end)

**Feature engineering**

One hot encoding - Congyun, Fan Zhang

**Modeling**

1. Find where the outliers are - Zhuoyuan, Yi
2. Linear Regression, Ridge Regression, LASSO Regression, Elastic Net - Siyu, Fan
   1. Assumptions checking?
3. Robust Regression, Polynomial Regression - Zhuoyuan, Yi
   1. Assumptions checking?

**Evaluation**

**Deployment**

[**Final Slides**](https://docs.google.com/presentation/d/1qd8A2a0MMFaJcCSxo5ZVJ5m-XCvyPaOhOVGRGoxJLBc/edit#slide=id.p)

[**Final Report**](https://docs.google.com/document/d/1Q9o_IVAURg4TwcutKVHNx73c9HHNbrjw6AnJf9XfDiU/edit)