PROJECT 2 - AMES HOUSING DATA AND KAGGLE CHALLENGE

creating a regression model to determine sale price of a house in the town of Ames, Iowa

Background

- Originally published in Journal of Statistics Education, Volume 19, Number 3(2011)1
- Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression ProjectDean De CockTruman State UniversityJournal of Statistics Education Volume 19, Number3(2011), www.amstat.org/publications/jse/v19n3/decock.pdf
- Released as a Kaggle challenge 10 months ago and had participation from 90 teams

Project objective

- Creating and iteratively refining a regression model

Process:

Regression models covered in class only to be used

Upload prediction values to Kaggle to get a score

EDA and Data cleaning

- Checking for null values and replacement with valid values if necessary
- Conversion of all ordinal and nominal data types into numerical formal
- Dropping of outliers

df_train_alt: Replacing null values

Column 'Lot Frontage' has a mean 69.06 and median 68 being values close to each other.

This indicates that the distribution is almost Normal and hence the mean value is used for the null values

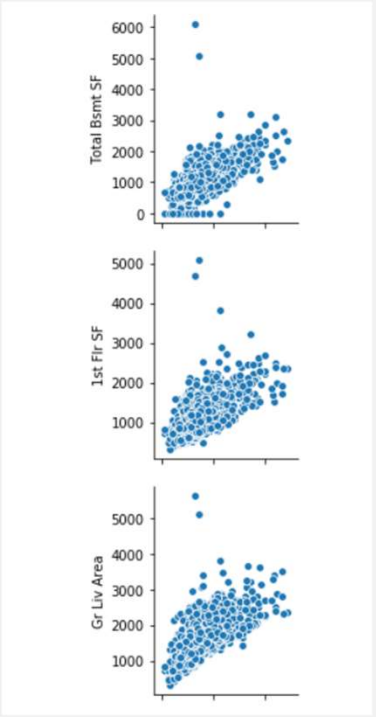
330 null values in 'Lot Frontage' were filled with mean value 69.055 ¶

```
# df_train_alt: Replacing null values in 'Lot Frontage' with mean value

df_train_alt['Lot Frontage']= df_train_alt['Lot Frontage'].fillna(df_train_alt['Lot Frontage'].mean())
df_train_alt['Lot Frontage'].isnull().sum()
```

```
# changing ordinals ratings to rankings - 'Kitchen Qual'

mapping = {np.nan:0, 'Po': 1,'Fa': 2,'TA': 3, 'Gd': 4, 'Ex': 5 }
df_train_alt['Kit_Qual_rk']= df_train_alt['Kitchen Qual'].apply(lambda x : mapping[x])
```

Identifying features for the model

- Techniques used include:
  - Viewing heatmap
  - Calculation of Pearson Correlation coefficient
  - Co-linearity between features
  - Filter method
  - Recursive Feature elimination



Features

Pearson Correlation coefficients

SalePrice	1.000000
Overall Qual	0.800207
Ex_Qual_rk	0.712146
Gr Liv Area	0.697038
Kit_Qual_rk	0.692336
Garage Area	0.649897
Garage Cars	0.647781
Total Bsmt SF	0.629303
1st Flr SF	0.618486
Bsmt_Qual_rk	0.612188
Year Built	0.571849
Gar_Fin_rk	0.557839
Year Remod/Add	0.550370
Firepl_Qu_rk	0.538925
Full Bath	0.537969
FoundationPConc	0.529047
TotRms AbvGrd	0.504014
Mas Vnr Area	0.503579

Target variable

Co-linearity between Gr Liv Area and TotRms AbvGrd
1 : 0.81

Co-linearity between Garage Cars and Garage Area
1 : 0.89

Model designs

Model 1 - correlation (17 features)

- Linear Regression
- Ridge Regression
- Lasso Regression

Overall Qual	Full Bath
Year Built	Kit_Qual_rk
Year Remod/Add',	TotRms AbvGrd
Mas Vnr Area	Firepl_Qu_rk
Ex_Qual_rk	Garage Yr Blt
Bsmt_Qual_rk	Gar_Fin_rk
Total Bsmt SF	Garage Area
1st Flr SF	FoundationPConc
Gr Liv Area	

Model 2 - correlation (16 features)

- Linear Regression
- Ridge Regression
- Lasso Regression

Overall Qual	Full Bath
Year Built	Kit_Qual_rk
Year Remod/Add',	~~TotRms AbvGrd~~
Mas Vnr Area	Firepl_Qu_rk
Ex_Qual_rk	Garage Yr Blt
Bsmt_Qual_rk	Gar_Fin_rk
Total Bsmt SF	Garage Area
1st Flr SF	FoundationPConc
Gr Liv Area	

Model 3 - RFE (13 features)

- Linear Regression
- Ridge Regression
- Lasso Regression

MS SubClass60	Garage Type2Types
MS SubClass75	Garage TypeAttchd
MS SubClass120	GarageTypeBuiltIn
MS SubClass150	Misc FeatureElev
Bldg TypeTwnhs	Misc FeatureGar2
Roof StyleGable	
Roof StyleGambrel	
Roof StyleMansard	

Polynomial (17 features)

Final model selection

Model 1 (17 features) Linear Regression model

Kaggle Public score: 33774.73095

Kaggle Private score: 35055.20712

```
Train R_square: 0.84745

Test R_square : 0.85766

          RMSE : 29536
```

Model 1 (17 features) Ridge Regression model

Kaggle Public score: 33901.40970

Kaggle Private score: 35054.17459

```
Train R_square: 0.84754

Test R_square : 0.85770

          RMSE : 29558
```

Further refining of the model for better fit

- Consider other features
- Feature Engineering
- Polynomial with degree of 3
- Try other models