Milestone 1 Report

➢ Preprocessing techniques

I. Drop all columns with nulls more than 40%.
II. Fill numeric columns nulls with the mean value.
III. Encode string columns.
IV. Predict null encoded string values.

➢ Data analysis

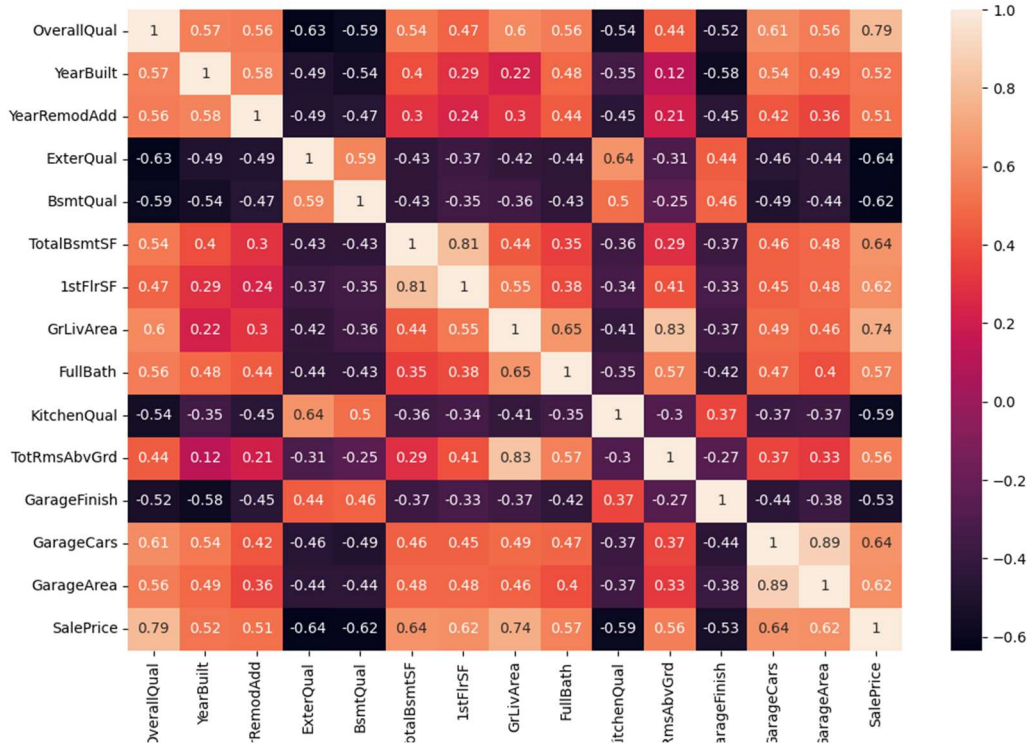

Figure 1.correlation matrix for selected features

➢ Regression techniques

We selected all features that correlate more than 0.4 with target columns then passed them to two regression techniques.

- <u>Multivariable linear regression</u>
  - Mean Square Error→1136767904.4536338
  - Training time→0.005179405212402344
  - Number of features→14

- <u>Polynomial regression</u>

We choose degree by looping on different degrees and choose the best one depends on the RMSE.

  - Training time→ 1.105431318283081
  - Number of features→14

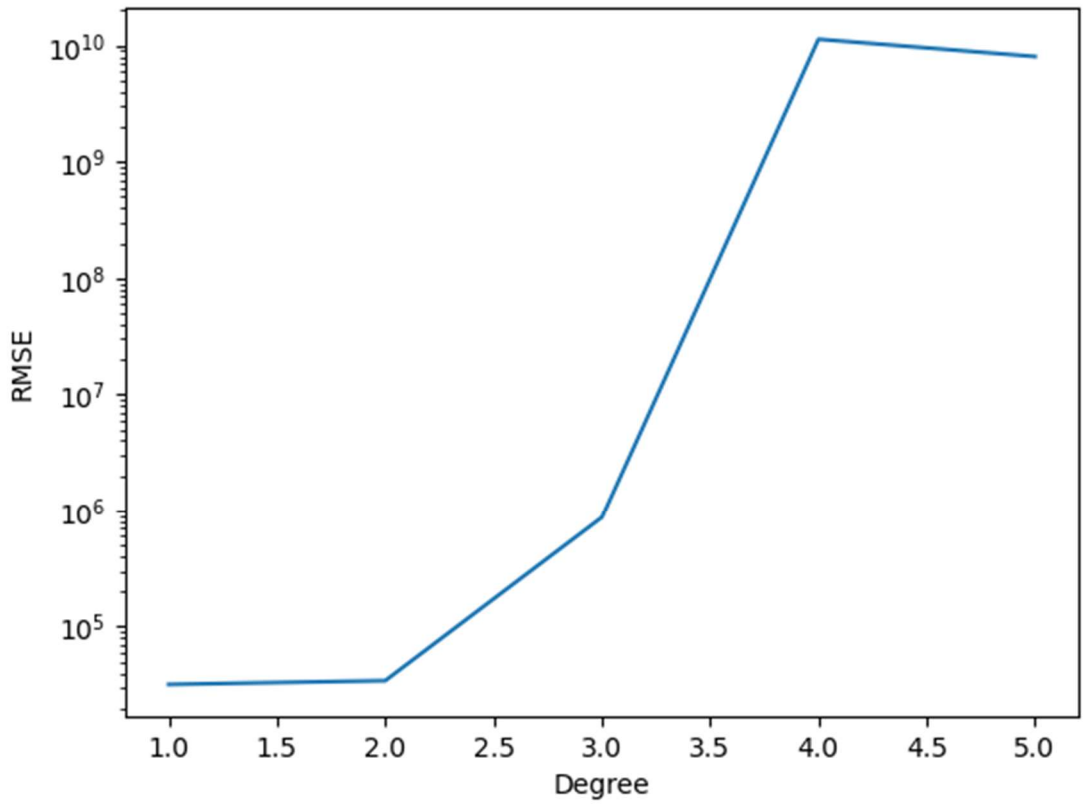

Figure 2.Relation between RMSE and degree of polynomial regression

	Train set	Test set
Size	70%	30%

- **Conclusion**

Columns with null values can has high correlation with the SalePrice(Y) column with out replace the null values (in numerical columns replaced with the mean , in encoded columns replaced with classification model predictions).
Features that has high correlation was 12 features
After replacing null values features that has high correlation was 14 features