Analysis of Ames Dataset

Joseph, Supriya, Mei Lian, Eng Soon

Problem Statement

We are a Data Science team at Propnex pte ltd, a real estate company. We have been with tasked with analysing the Ames Housing Dataset and finding the key factors that influence the Sale Price of a house in Ames city.

Through this analysis, we will develop a regression model to predict the sale prices. We will make recommendations to homeowners on improving their property value and set expectations for their property's Sale Price.

Ames Housing Dataset

Individual residential properties sold in Ames, Iowa from 2006 to 2010.

Train Data Set	2,051 rows, 81 columns
Test Data Set	878 rows, 80 columns

Categorical	Nominal Data	23	MS_SubClass, Garage Type
	Ordinal Data	23	Overall Quality, Overall Condition
Numerical	Discrete Data	20	Year Sold, Bedroom
	Continuous Data	15	Ground Living Area, Lot Area

Data cleaning

Handling missing data

- Check for actual values of NaN values with na_filter=false
- Most NaN values are actually NA (Not applicable)

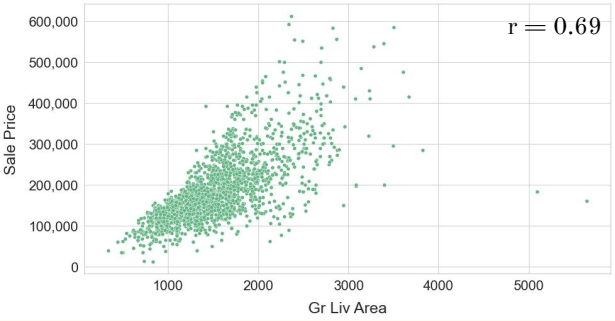
How we imputed the data for some columns

- If values are missing in dataframe, we would try to impute it by enquiring:-

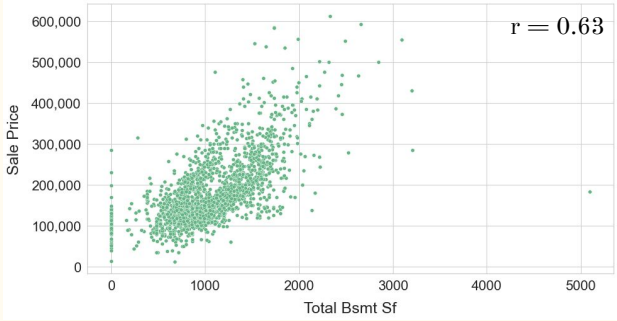
a. Is the amount of data missing significant? (eg. Lot Frontage)
b. What other features would have highest correlation with missing feature
c. If we are unable to identify (b), we can impute with values (mean/median/mode) from the feature with the missing values.
d. If not, we would most probably fill it as NA

Exploratory Data Analysis (Numerical Variables)

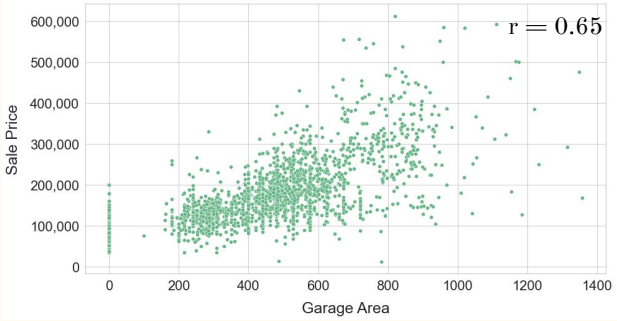
Ground living Area vs Sale Price



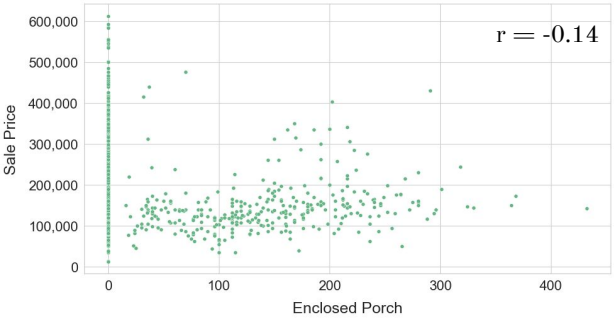
Total Basement Sqft vs Sale Price



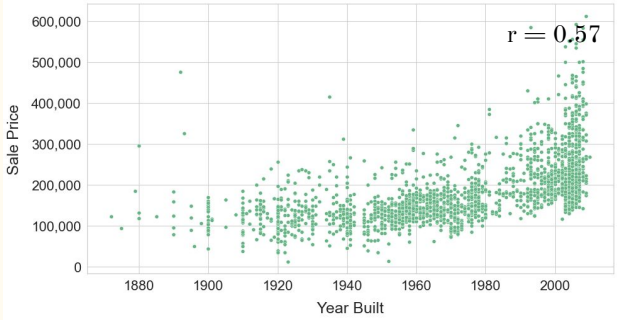
Garage Area vs Sale Price



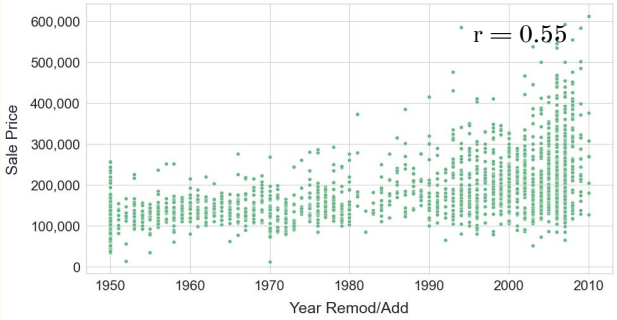
Open Porch Sqft vs Sale Price



Year Built vs Sale Price

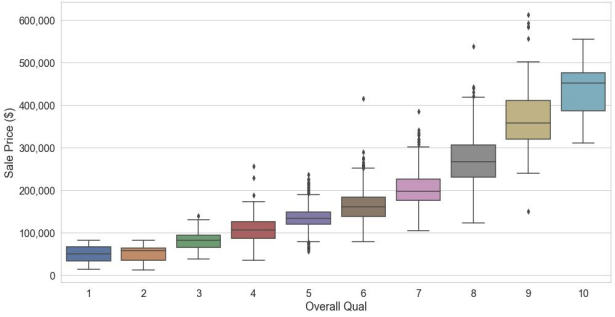


Remodel Year vs Sale Price

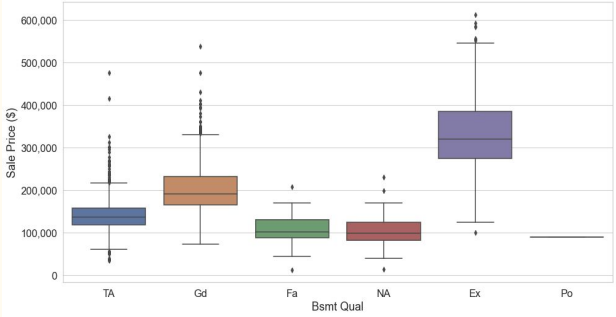


Exploratory Data Analysis (Categorical Variables)

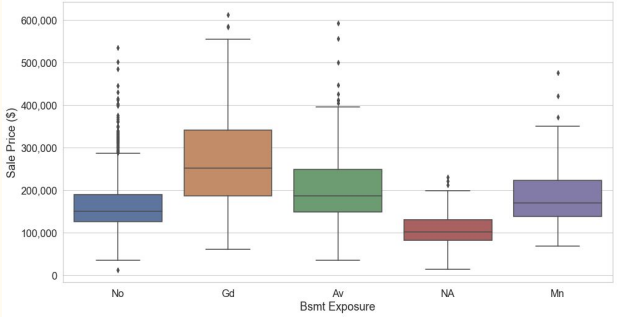
Overall Quality vs Sale Price



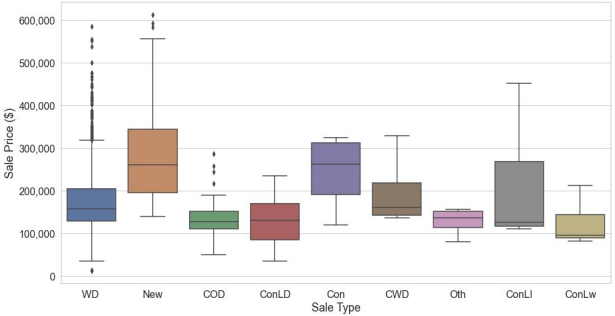
Basement Quality vs Sale Price



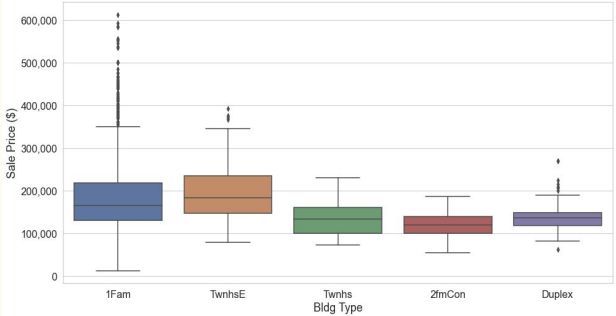
Basement Exposure vs Sale Price



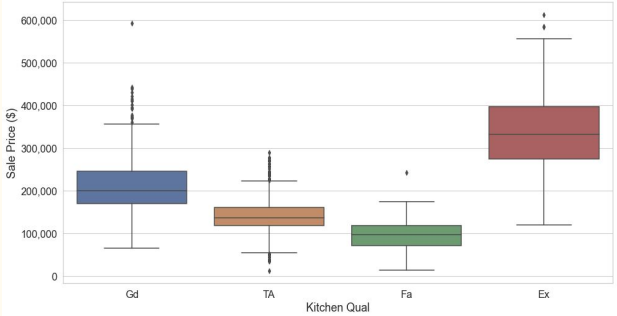
Sale Type vs Sale Price



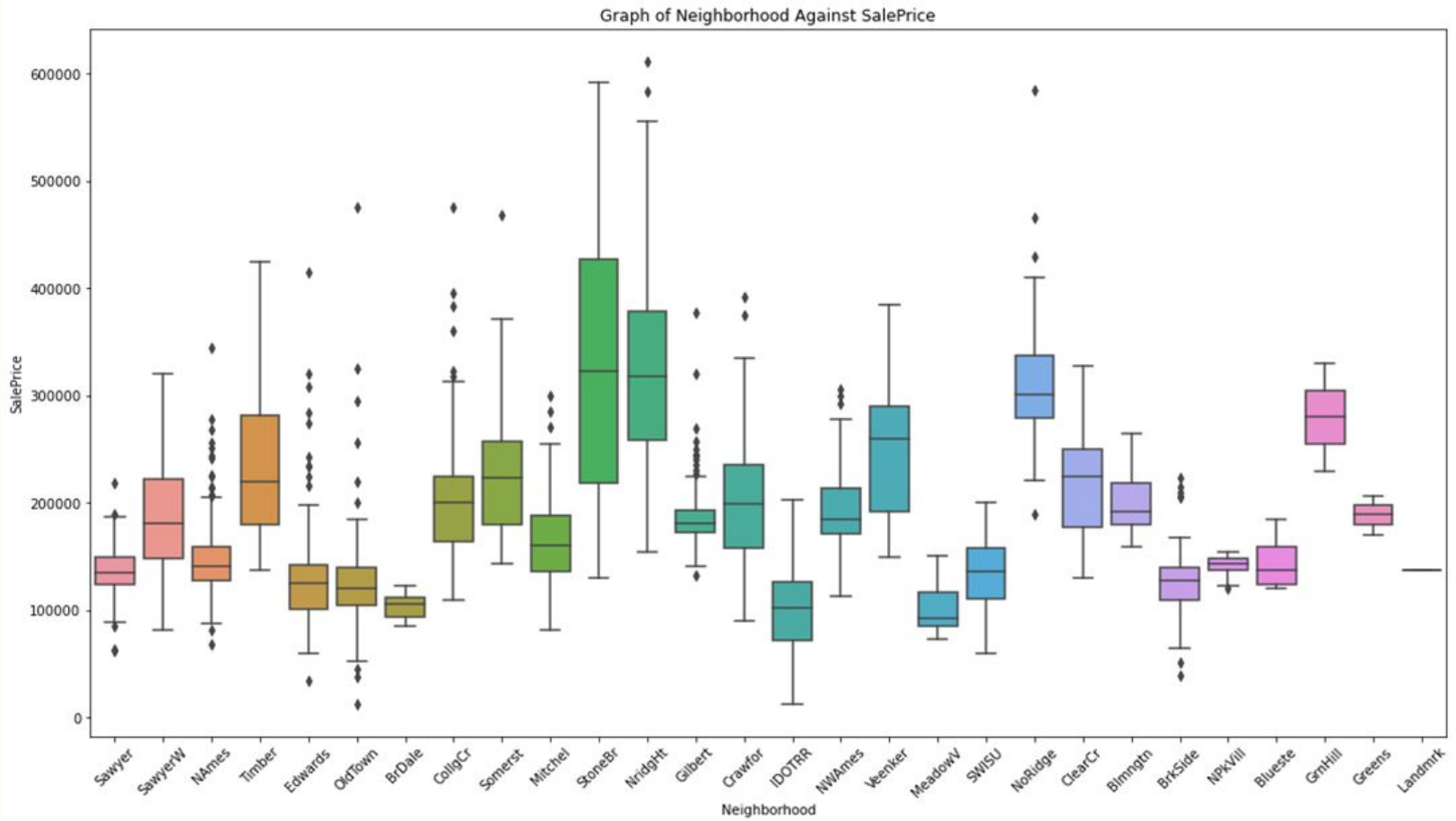
Building Type vs Sale Price



Kitchen Quality vs Sale Price

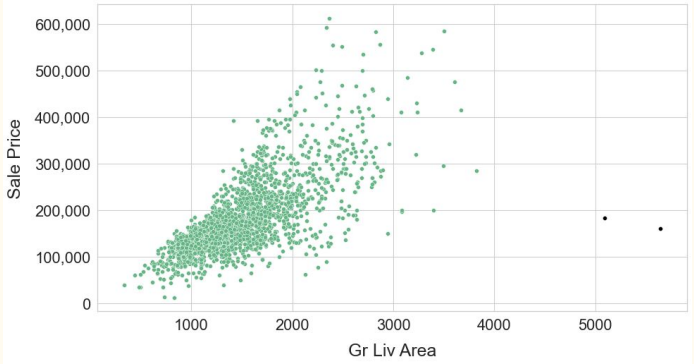


Exploratory Data Analysis (Categorical Variables)

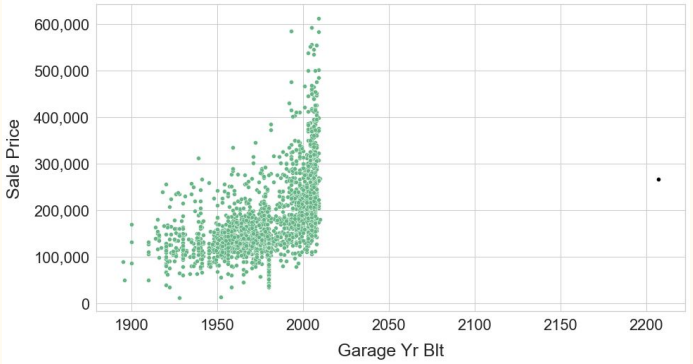


Handling Outliers

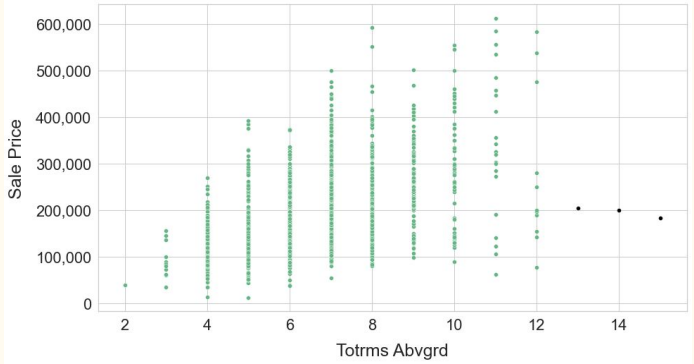
Ground living Area vs Sale Price



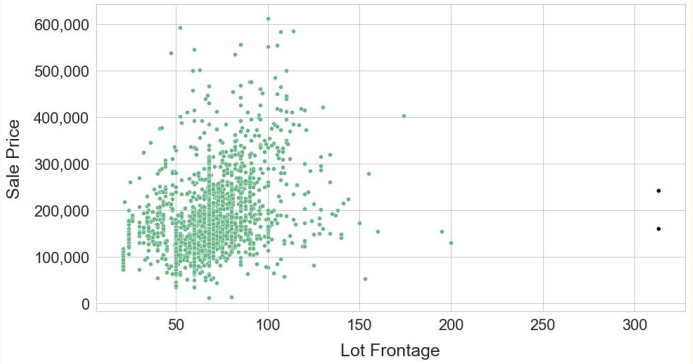
Garage Year Built vs Sale Price



Total Rooms Above ground vs Sale Price



Lot Frontage vs Sale Price



Pre-Processing the Data

- Handling Ordinal Variables
  - Eg: External quality, Basement Condition
    - Before: Excellent, Good, Average/Typical, Fair, Poor
    - After: 5, 4, 3, 2, 1
- One Hot Encoding
  - Eg: Heating, Neighborhood
    - Before: Heating (with sub-categories)
    - After: Heating_Floor, Heating_GasA, Heating_GasW, Heating_Grav, Heating_OthW, Heating_Wall
- Standardisation

Modelling - Lasso regression

R^2 score
Train data: 0.93
Validation data: 0.91

Cross val score: 0.91

Feature Engineering

- Ground Living Area * Overall Quality
- Basement Quality * Basement Exposure
- Garage Cars * Garage Area
- Overall Condition * Remodel date
- Lot Frontage * Lot Area
- Overall Quality * Functionality

Modelling - Lasso & Ridge regression

Lasso

R^2 score
Train data: 0.92
Validation data: 0.90

Cross val score: 0.91

Ridge

R^2 score
Train data: 0.94
Validation data: 0.93

Cross val score: 0.93

Model Coefficients

Variable	Coefficient	Absolute (Coefficient)
gr_liv_area	20,314.41	20,314.41
overall_qual	11,598.11	11,598.11
gr_liv_area*overall_qual	8,678.98	8,678.98
bsmtfin_sf_1	8,284.08	8,284.08
total_bsmt_sf	5,869.06	5,869.06
lot_area	5,822.88	5,822.88
year_built	5,764.04	5,764.04
exter_qual	5,150.10	5,150.10
bsmt_qual*bsmt_exposure	4,986.68	4,986.68
overall_cond	4,557.10	4,557.10
1st_flr_sf	4,293.77	4,293.77
neighborhood_StoneBr	4,282.09	4,282.09
neighborhood_NridgHt	4,087.44	4,087.44
kitchen_qual	4,030.58	4,030.58
mas_vnr_area	3,630.57	3,630.57
year_remod/add	3,445.92	3,445.92
screen_porch	3,385.24	3,385.24
functional	3,356.80	3,356.80
neighborhood_GrnHill	3,275.59	3,275.59
exterior_1st_BrkFace	3,236.05	3,236.05
sale_type_New	3,178.57	3,178.57
roof_style_Mansard	(3,039.32)	3,039.32
garage_cars*garage_area	3,000.79	3,000.79
fireplaces	2,855.81	2,855.81
lot_frontage	2,733.63	2,733.63
garage_type_NA	(2,639.20)	2,639.20
bsmt_qual	2,457.01	2,457.01
ms_zoning_FV	2,380.66	2,380.66
neighborhood_Crawfor	2,291.73	2,291.73
lot_config_CulDSac	2,262.15	2,262.15

Positive Factors

- Ground living area
- Overall quality
- Basement Type 1 finished square feet
- Total square feet of basement area
- Lot area
- Year built
- External quality
- Basement quality with good exposure

Negative Factors

- Roof type - Mansard
- No garage
- Neighborhood College Creek

Conclusions

- Bigger houses fetch better Sale Price
  - High Ground living area
  - High Basement Area
  - High Lot Area
- Better quality houses fetch better Sale Price
  - Overall quality
  - External quality
- Good Basement quality & exposure enhances the Sale Price
- Newer houses positively influence the Sale Price
- Not having a garage lowers the Sale Price

Recommendations to Home Owners

- Home owners can consider undertaking renovation and improving the overall quality of the house by using material of better quality and finish.
  - Avoid roof type Mansard
  - Add a fireplace
  - Improve kitchen quality
  - Exterior covering of the house BrickFace
  - Add a screen porch
- They should focus on specifically improving the external quality using better quality material on the exterior.
- Home owners may set their expectations about their property value according to:
  - Size of their Ground Living Area, Basement, Lot Area
  - Year in which the house was built
  - Basement quality & exposure
  - Neighbourhood (Stone Brook, Northridge Heights, Green Hills have better Sale Price)

Thank You