Housing Project:

In this project, the test and training dataset have been merged and made it into a single dataframe. It contains 1460 rows and 81 columns. There was null values in 19 columns.

Of these the categorical features were:

- MasVnrType
- BsmtQual
- BsmtCond
- BsmtExposure
- BsmtFinType1
- BsmtFinType2
- Electrical
- FireplaceQu
- GarageType
- GarageFinish
- GarageQual
- GarageCond

Of these the numerical features were:

- LotFrontage
- MasVnrArea
- GarageYrBlt

Numerical features containing null values are replaced with mean

Categorical features containing null values are replaced with mode in the features

EXPLORATORY DATA ANALYSIS:

Before we start data analysis or run the data through a machine learning algorithm, we must clean your data and make sure it is in a suitable form. Further, it is essential to know any recurring patterns and significant correlations that might be present in your data. The process of getting to know your data in depth is called Exploratory Data Analysis.

Exploratory Data Analysis is an integral part of working with data. In this tutorial titled 'All the ins and outs of exploratory data analysis,' you will explore how to perform exploratory data analysis on different data types

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations

Steps Involved in Exploratory Data Analysis

1. Data Collection
2. Data Cleaning
3. Univariate Analysis
4. Bivariate Analysis

Data Collection:

- The first step is to import the required libraries
- Read the csv files of both training and testing files merge and convert it to single dataframe df

Data Cleaning:

Numerical features containing null values are replaced with mean. Categorical features containing null values are replaced with mode in the features.

Of these the categorical features were:

- MasVnrType
- BsmtQual
- BsmtCond
- BsmtExposure
- BsmtFinType1
- BsmtFinType2
- Electrical
- FireplaceQu
- GarageType
- GarageFinish
- GarageQual
- GarageCond

Of these the numerical features were:

- LotFrontage
- MasVnrArea
- GarageYrBlt

Univariate Analysis & Bivariate analysis:

In Univariate Analysis, we analyse data of just one variable. A variable in our dataset refers to a single feature/ column.

Here we use Line plots, Box plot, Scatter plot, Violin plot. In this plot we compare each feature with the target feature (Price).

From the plots we come to the following conclusions:

1. MSSubClass (Identifies the type of dwelling involved in the sale)

30 1-STORY 1945 & OLDER is cheap

20 1-STORY 1946 & NEWER ALL STYLES are more in number

60 2-STORY 1946 & NEWER is costly

2. MSZoning(Identifies the general zoning classification of the sale.)

RL Residential Low Density are more in number

FV Floating Village Residential is costly

Commercial houses are cheap

3. LotFrontage (Linear feet of street connected to property)

104 Lot Frontage is costly

45 Lot Frontage is cheap

60 Lot Frontage is more in number

4. Lot Area

Most houses range between 20000 sq.ft

5. Street (Type of road access to property)

Most people prefer Paved

6. LotShape: General shape of property

- Reg Regular - Cheap
- IR1 Slightly irregular - Commonly used
- IR2 Moderately Irregular - Costly
- IR3 Irregular – NA

7. LandContour: Flatness of the property

- Lvl Near Flat/Level -Commonly used
- Bnk Banked - Quick and significant rise from street grade to building - Cheap
- HLS Hillside - Significant slope from side to side - Costly

8. Utilities: Type of utilities available

- AllPub All public Utilities (E,G,W,& S) – Mostly used

9.LotConfig: Lot configuration

- Inside Inside lot – Mostly used
- Corner Corner lot – Cheap
- CulDSac Cul-de-sac - Costly
- FR3 Frontage on 3 sides of property - Coslty

10. LandSlope: Slope of property

- Gtl Gentle slope – Mostly used
- Mod Moderate Slope
- Sev Severe Slope – Less used

11. Neighborhood: Physical locations within Ames city limits

- Names North Ames Mostly preferred

- NoRidge Northridge Costly

12. Condition1: Proximity to various conditions

    - Norm Normal – commonly used
    - PosA Adjacent to postive off-site feature & RRNn Within 200' of North-South Railroad – costly
    - RRAe Adjacent to East-West Railroad cheap

13. Condition2: Proximity to various conditions (if more than one is present)

    - Norm Normal commonly used
    - PosN Near positive off-site feature--park, greenbelt, etc. -costly
    - PosA Adjacent to postive off-site feature – costly

14. BldgType: Type of dwelling

    - 1Fam Single-family Detached price & count is more

15. HouseStyle: Style of dwelling

    - 1Story One story mostly preferred
    - 2 and 2.5 storey is costly

16. OverallQual: Rates the overall material and finish of the house

    - Mostly houses have 5,6,7
    - As quality increases price also increases

17. OverallCond: Rates the overall condition of the house

    - 5^{th} condition is mostly available and costly
    - 9^{th} condition is costly

18. YearBuilt: Original construction date:

    - Newer houses are costly

19. YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

    - 1950 built houses are mostly available
    - Newer remodel houses are costly
    - After 2002 houses are mostly costly

20. RoofStyle: Type of roof

    - Gable Mostly used
    - Hip & Shed costly

21. RoofMatl: Roof material

    - CompShg Standard (Composite) Shingle Mostly used
    - WdShngl Wood Shingles Costly

Since it takes time, we have used pandas and dataframe is created which indicates the Most available,Least available,Costly and Cheap based on the attributes and then replaced it with their description

Features	Mostly Available	Less Available	Costly	Cheap
Id	127	127	692	496
MSSubClass	1-STORY 1946 & NEWER ALL STYLES	1-STORY W/FINISHED ATTIC ALL AGES	2-STORY 1946 & NEWER	1-STORY 1945 & OLDER
MSZoning	Residential Low Density	Commercial	Floating Village Residential	Commercial
LotFrontage	2-STORY 1946 & NEWER	38	2-STORY PUD - 1946 & NEWER	153
LotArea	9600	7407	21535	7879
Street	Paved	Gravel	Paved	Gravel
Alley	Gravel	Paved	Paved	Gravel
LotShape	Regular	Irregular	Moderately Irregular	Regular
LandContour	Near Flat/Level	Depression	Hillside - Significant slope from side to side	Banked - Quick and significant rise from street grade to building
Utilities	All public Utilities (E,G,W,& S)	All public Utilities (E,G,W,& S)	All public Utilities (E,G,W,& S)	All public Utilities (E,G,W,& S)
LotConfig	Inside lot	Frontage on 3 sides of property	Cul-de-sac	Frontage on 2 sides of property
LandSlope	Gentle slope	Severe Slope	Severe Slope	Gentle slope
Neighborhood	North Ames	Bluestem	Northridge	Meadow Village
Condition1	Normal	Within 200' of East-West Railroad	Within 200' of North-South Railroad	Adjacent to arterial street
Condition2	Normal	Adjacent to East-West Railroad	Adjacent to postive off-site feature	Adjacent to arterial street
BldgType	Single-family Detached	Two-family Conversion; originally built as one-family dwelling	Townhouse End Unit	Two-family Conversion; originally built as one-family dwelling
HouseStyle	One story	Two and one-half story: 2nd level finished	Two and one-half story: 2nd level finished	One and one-half story: 2nd

				level unfinished
OverallQual	Average	Very Poor	Very Excellent	Very Poor
OverallCond	Average	Very Poor	Excellent	Very Poor
YearBuilt	2006	1875	1892	1913
YearRemodAdd	1950	1951	2010	1952
RoofStyle	Gable	Shed	Shed	Gabrel (Barn)
RoofMatl	Standard (Composite) Shingle	Roll	Wood Shingles	Roll
Exterior1st	Vinyl Siding	Asphalt Shingles	Imitation Stucco	Brick Common
Exterior2nd	Vinyl Siding	Other	Other	Asbestos Shingles
MasVnrType	None	Brick Common	Stone	Brick Common
MasVnrArea	0	443	1170	381
ExterQual	Average/Typical	Fair	Excellent	Fair
ExterCond	Average/Typical	Poor	Excellent	Poor
Foundation	Cinder Block	Wood	Poured Contrete	Slab
BsmtQual	Average/Typical	Fair	Excellent	Fair
BsmtCond	Average/Typical	Poor	Good	Poor
BsmtExposure	No Exposure	Mimimum Exposure	Good	No Exposure
BsmtFinType1	Unfinshed	Low Quality	Good Living Quarters	Average Rec Room
BsmtFinSF1	0	252	1455	495
BsmtFinType2	Unfinshed	Good Living Quarters	Good Living Quarters	Below Average Living Quarters
BsmtFinSF2	0	123	1474	311
BsmtUnfSF	0	1237	989	225
TotalBsmtSF	0	2392	2444	480
Heating	Gas forced warm air furnace	Floor Furnace	Gas forced warm air furnace	Gravity furnace
HeatingQC	Excellent	Poor	Excellent	Poor
CentralAir	Yes	No	Yes	No
Electrical	Standard Circuit Breakers & Romex	Mixed	Standard Circuit Breakers & Romex	Mixed
1stFlrSF	864	2121	2444	480
2ndFlrSF	0	454	1872	368
LowQualFinSF	0	384	572	156
GrLivArea	864	2353	4316	480

BsmtFullBath	0	Fair	Poor	0
BsmtHalfBath	0	Poor	0	Poor
FullBath	Poor	0	Fair	Very Poor
HalfBath	0	Poor	Very Poor	Poor
BedroomAbvGr	Fair	Very Good	0	Above Average
KitchenAbvGr	Very Poor	Fair	Very Poor	Fair
KitchenQual	Average/Typical	Fair	Excellent	Fair
TotRmsAbvGrd	Above Average	Poor	11	Poor
Functional	Typical Functionality	Severe Slope	Typical Functionality	Major Deductions 2
Fireplaces	0	Fair	Fair	0
FireplaceQu	Good	Poor	Excellent	Poor
GarageType	Attached to home	More than one type of garage	Built-In (Garage part of house - typically has room above garage)	Car Port
GarageYrBlt	2006	1900	2009	1938
GarageFinish	Unfinshed	Finished	Finished	Unfinshed
GarageCars	Poor	Below Average	Fair	0
GarageArea	0	406	832	250
GarageQual	Average/Typical	Excellent	Excellent	Poor
GarageCond	Average/Typical	Excellent	Average/Typical	Poor
PavedDrive	Yes	Partial Pavement	Yes	No
WoodDeckSF	0	303	361	263
OpenPorchSF	0	169	67	523
EnclosedPorch	0	239	291	172
3SsnPorch	0	153	304	140
ScreenPorch	0	260	410	271
PoolArea	0	555	555	480
PoolQC	Good	Excellent	Excellent	Good
Fence	Minimum Privacy	Minimum Wood/Wire	Good Privacy	Good Wood
MiscFeature	Shed	Tennis Court	Tennis Court	Other
MiscVal	0	1150	2500	54
MoSold	Above Average	Poor	Excellent	Below Average
YrSold	2007	2010	2007	2008
SaleType	Warranty Deed - Conventional	Contract 15% Down payment regular terms	Home just constructed and sold	Other

SaleCondition	Normal Sale	Adjoining Land Purchase	Home was not completed when last assessed (associated with New Homes)	Adjoining Land Purchase
SalePrice	140000	198900	755000	34900

Bivariate analysis:

From the univariate analysis and feature importance we find some of the features are taken as important

The most important features are :

1. MSSubClass
2. HeatingQC
3. BldgType
4. LotFrontage
5. Neighborhood
6. BsmtFinType1
7. SaleCondition
8. TotRmsAbvGrd
9. MSZoning
10. BedroomAbvGr
11. OverallCond
12. GarageArea
13. YearRemodAdd
14. 1stFlrSF
15. KitchenQual
16. FullBath
17. Fireplaces
18. YearBuilt
19. BsmtQual
20. GarageArea
21. GarageFinish
22. GrLivArea
23. ExterQual
24. OverallQual

- From the HeatingQC vs SalePrice vs BldgType graph,we find mostly Single Family houses, are costly and more popular
- From the Neighborhood vs SalePrice vs YearBuilt graph,we find newer houses, are costly and more popular especially Northridge Heights
- From the BsmtFinType1vs SalePrice vs BsmtQual graph,we find Good Living Quarters houses, are costly and mostly good and excellent
- From the KitchenQualvs SalePrice vs SaleCondition graph,we find as kitchen quality improves ,house price and preference also increase

- From the GarageFinish vs SalePrice vs ExterQual graph,we find based on garage finish improves ,house price and preference also increase
- From the MSSubClass vs SalePrice vs OverallCond,we find based on MSubclass from 20 to 80 ,house price and preference also increase
- From the TotRmsAbvGrd graph as no of rooms increase, price also increase
- From the BedroomAbvGr graph we find rooms with 1 to 4 are more popular
- From the Garage Area vs Finish,customers prefer garage area of 200 to 850 , finished garages are more costly
- From the years graph, newer homes are more popular and costly
- From the exterior quality graph, better quality are preferred more
- From the first floor area graph, most houses are within 2500 sq.ft
- From the bathroom graph, as the no of bathrooms increase, price also increase
- From the fireplaces graph we find as no of fireplace increase the price also increase with good heat quality

After visualization, categorical features are converted to continuous features.

Correlation:

Correlation is used to check how one or more variables are related to each other. From these variables can be input data features which have been used to predict our target variable.

Correlation, statistical technique which determines how one variables moves/changes in relation with the other variable. It gives the idea about the degree of the relationship of the two variables. It's a bi-variate analysis measure which describes the association between different variables. In most of the business it's useful to express one subject in terms of its relationship with others.

Check correlation and plot heatmap from heat map we find that some of the features are highly correlated. but, we keep all features.

Removing the outliers:

Regarding the outliers, we used two methods, IQR rule and Z-Score

Z-Score is essential to know how many standard deviations away is my actual value from the mean value based on the actual data, you can define the threshold value for the z score to classify a point as an outlier or not in the current scheme of things.

By Z-Score we couldn't remove the outliers, since most of the data is categorical and boxplot was also plotted. Since more than half of the data were present as outliers.

Histogram was plot to check the skewness and the distribution

Training:

Now the model is cleaned, we are training the model for that we are splitting the dataframe into train and test data

Once the splitting is done, the data is scaled using standard scaler.

Testing:

The model is tested using regression and classification algorithms

The regression models used are:

- ❖ Linear Regression
- ❖ Gradient Boosting Regressor
- ❖ AdaBoost Regressor
- ❖ Decision Tree Regressor
- ❖ KNeighbors Regressor,
- ❖ Extra Trees Regressor,
- ❖ Random Forest Regressor

We are taking all these models and looping it from random state 42 to 95 to find the best r2 score,

The training data is 67% and the testing data is 33%

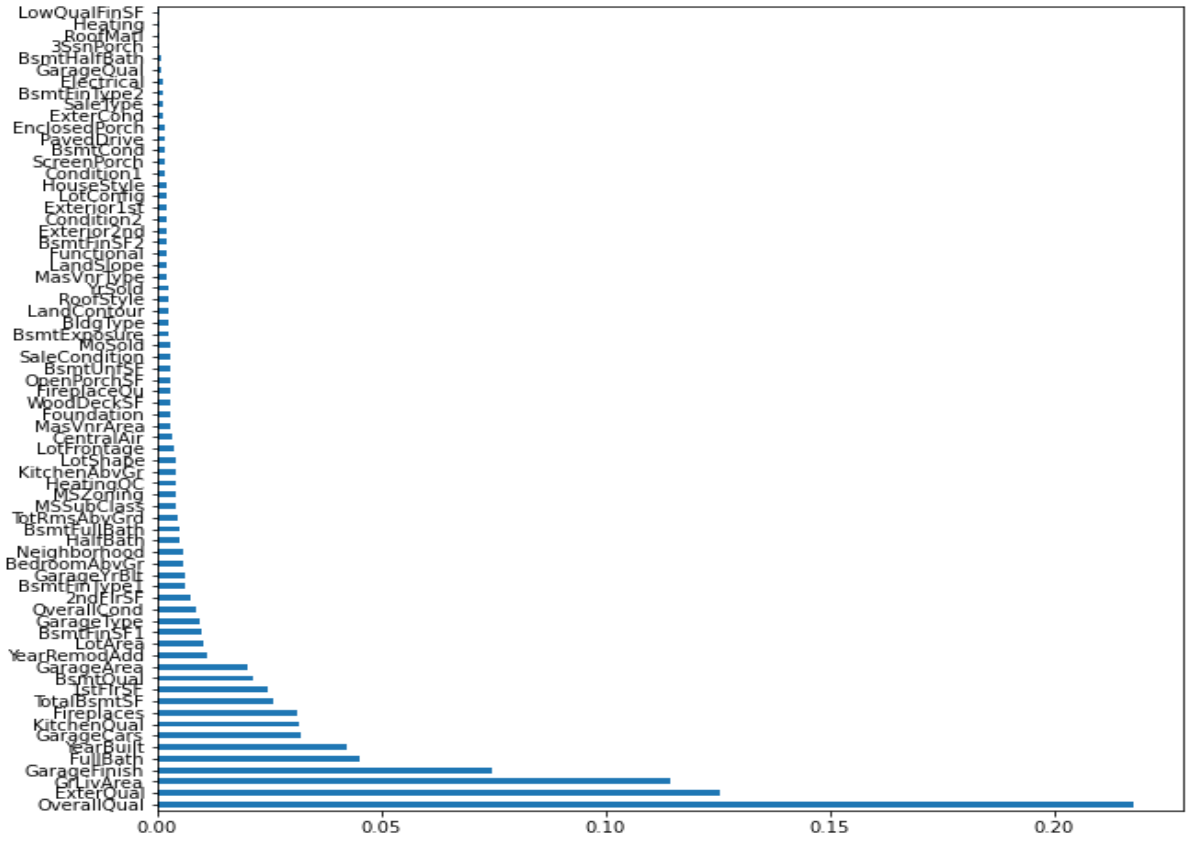
From these iterations we find that the Linear regression gives the highest accuracy of 92% in the 84th random state.

From this we can find the following results:

- R2 Score : 89.1915217715350
- Mean Absolute Error: 34.291056610243736
- Mean squared Error: 2578.755601384853
- Root Mean Absolute Error: 5.855856607725614

Feature Importance:

We have used Extratree Regressor to give the most important features



From this we have selected the most important features

Pipeline:

The dataset is modelled with Standard Scaling,PCA & (LR model,RF model)

- Pipeline with 20 features:

  R2 Score : 86.85245474287953

- Pipeline with 10 features:

  R2 Score : 87.38874715057197

- Pipeline with 20 features:

  R2 Score 87.46100678612534

Hypertuning:

From both regression and classification model, we get linear regression as the best model. But linear regression doesn't have any hyperparameters, so we have used SGD linear regressor and hypertuned the parameters

From this we get the following results:

- R2 Score : 88.91027010973418
- Mean Absolute Error: 34.60455115801892
- Mean squared Error: 2645.858414837134
- Root Mean Absolute Error: 5.882563315257977

We also predicted the test dataset with the hypertuned parameters.

Conclusion:

The house price dataset is collected, cleaned, scaled and test model is prepared. Using this data, it is spitted into train and test data and run with regressor and classifier algorithms

SGR Linear Regressor algorithms gives an r2 score of 88.9 after Hypertuning. This dataset can also be predicted by using pipeline consolidating standard scaling and regression algorithms of 5 features which gives an r2 scores of 87

From this SGR Linear Regressor model, we can predict the Test data and the house price is predicted.

The most important features are :

1. MSSubClass
2. HeatingQC
3. BldgType
4. LotFrontage
5. Neighborhood
6. BsmtFinType1

7. SaleCondition
8. TotRmsAbvGrd
9. MSZoning
10. BedroomAbvGr
11. OverallCond
12. GarageArea
13. YearRemodAdd
14. 1stFlrSF
15. KitchenQual
16. FullBath
17. Fireplaces
18. YearBuilt
19. BsmtQual
20. GarageArea
21. GarageFinish
22. GrLivArea
23. ExterQual
24. OverallQual