FLIP ROBO

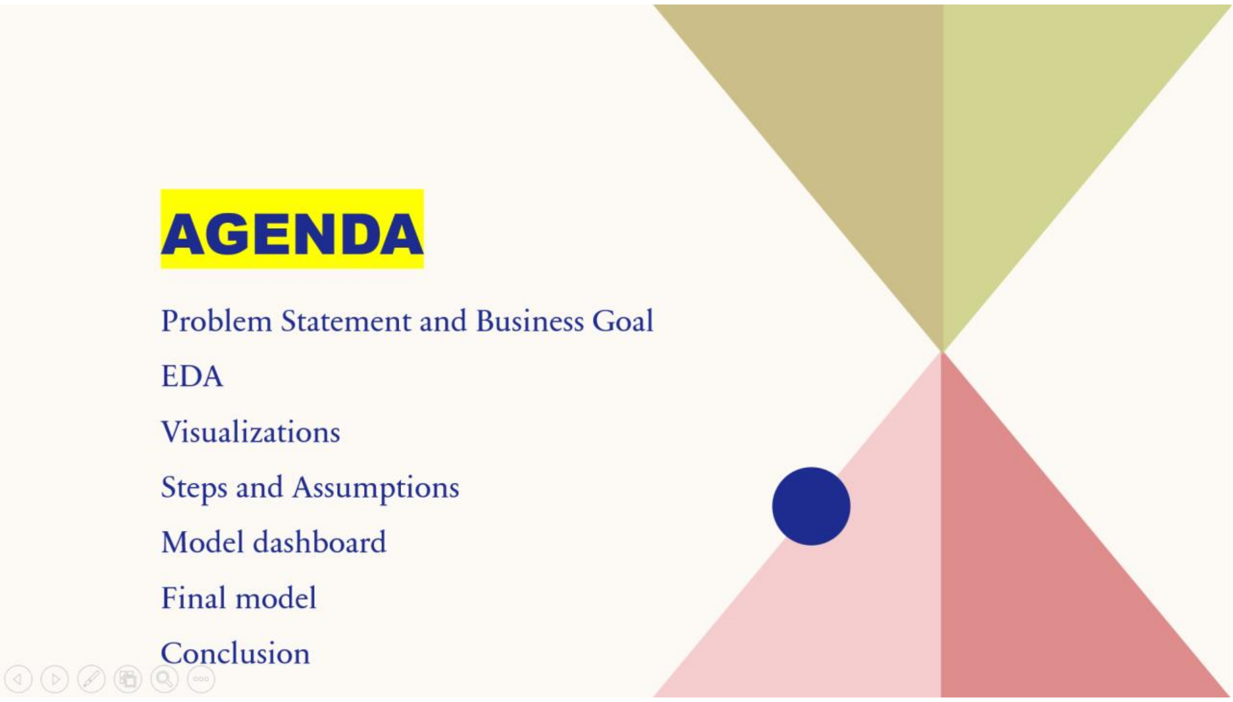
FLIPROBO ASSIGNMENT 6

HOUSING PRICES PREDICTION

Submitted by:

RAMAA DEVI

ACKNOWLEDGMENT



AGENDA

Problem Statement and Business Goal

EDA

Visualizations

Steps and Assumptions

Model dashboard

Final model

Conclusion

A new project is allocated and the project name is Housing Project.

PROVIDED FILE CONTAINED:

Data Description.csv:

This contains the description of data.

train.csv:

This contains the dataset on which you will be working upon

Housing Use case:

This contains the problem statement and business goal.

Sample documentation:

This is a sample report.

test.csv :

Predict the output for these data with your best fit model.

Submission format :

You need to submit three files:

1. The jupyter notebook solution file.

2. The project report in pdf format.

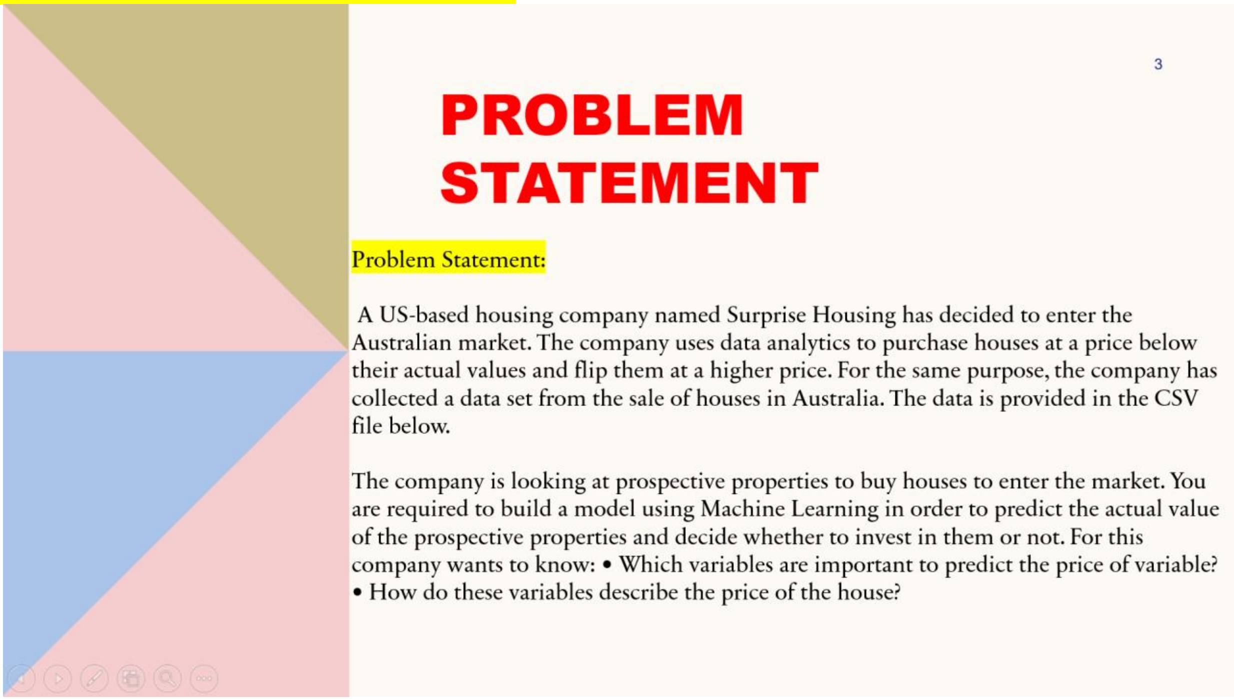
3. A PowerPoint presentation containing problem statement and understanding, EDA steps and visualizations, Steps and assumptions used to complete the project, model dashboard, finalized model, and conclusion.

Create a new Repository, Upload all three files on Github and share the link with me in messages.

USING THE ABOVE RESOURCES GIVEN I STARTED MY PROJECT WORK

INTRODUCTION

- Business Problem Framing



3

PROBLEM STATEMENT

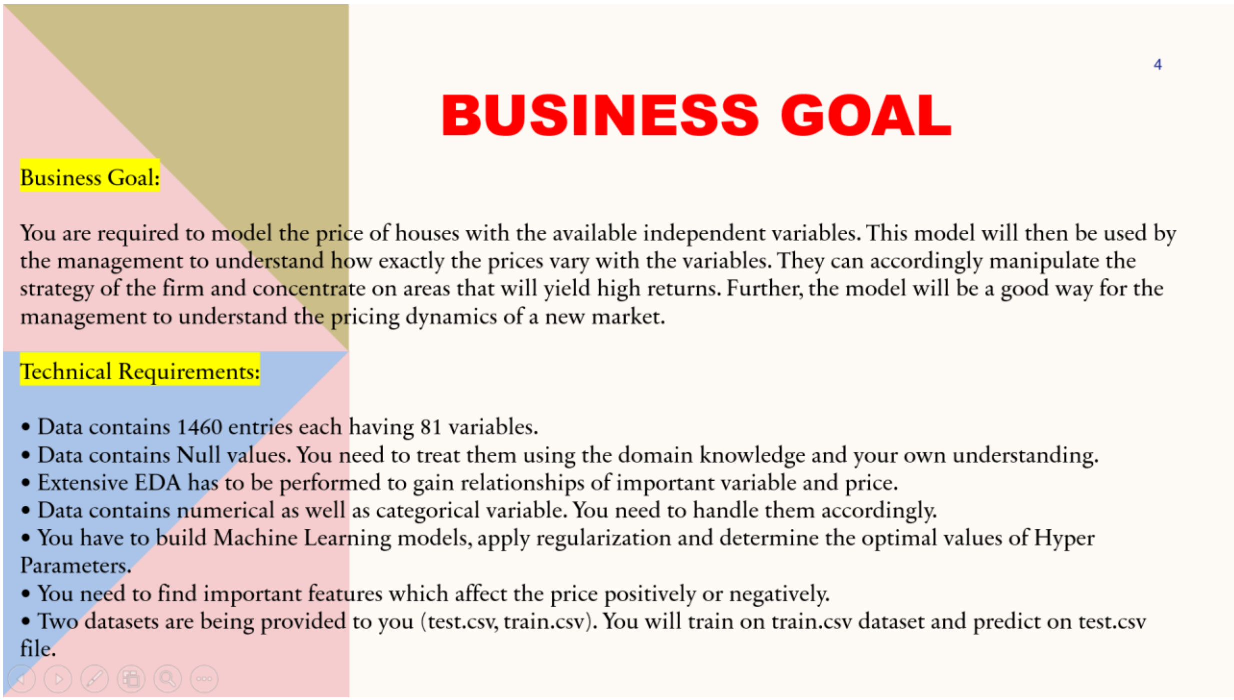
Problem Statement:

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below.

The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know: • Which variables are important to predict the price of variable? • How do these variables describe the price of the house?

- A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

- Conceptual Background of the Domain Problem

- 

4

BUSINESS GOAL

Business Goal:

You are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

Technical Requirements:

• Data contains 1460 entries each having 81 variables.
• Data contains Null values. You need to treat them using the domain knowledge and your own understanding.
• Extensive EDA has to be performed to gain relationships of important variable and price.
• Data contains numerical as well as categorical variable. You need to handle them accordingly.
• You have to build Machine Learning models, apply regularization and determine the optimal values of Hyper Parameters.
• You need to find important features which affect the price positively or negatively.
• Two datasets are being provided to you (test.csv, train.csv). You will train on train.csv dataset and predict on test.csv file.

You are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market

• Review of Literature

• Data contains 1460 entries each having 81 variable

• Data contains Null values. You need to treat them using the domain knowledge and your own understanding.

• Extensive EDA has to be performed to gain relationships of important variable and price.

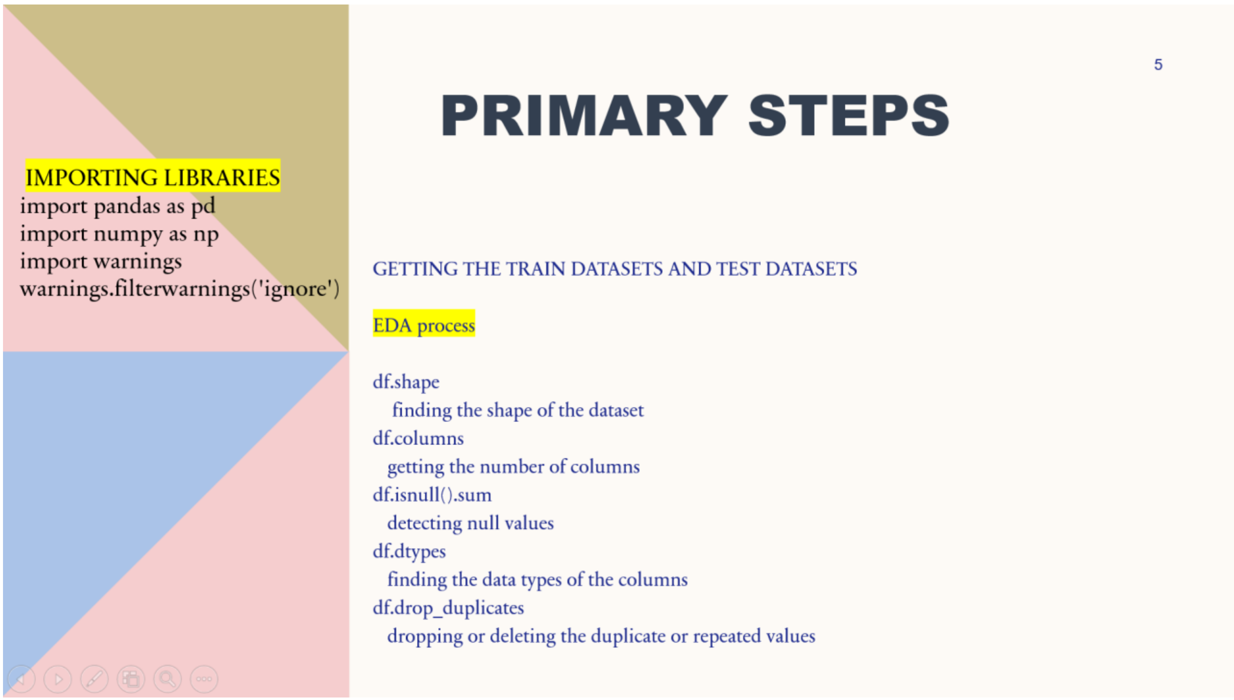
• Data contains numerical as well as categorical variable. You need to handle them accordingly.

• You have to build Machine Learning models, apply regularization and determine the optimal values of Hyper Parameters.

• You need to find important features which affect the price positively or negatively.

. Motivation for the Problem Undertaken

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing company



IMPORTING LIBRARIES

```
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings('ignore')
```

5

PRIMARY STEPS

GETTING THE TRAIN DATASETS AND TEST DATASETS

EDA process

df.shape
  finding the shape of the dataset
df.columns
  getting the number of columns
df.isnull().sum
  detecting null values
df.dtypes
  finding the data types of the columns
df.drop_duplicates
  dropping or deleting the duplicate or repeated values

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

Importing Libraries

Get all necessary libraries.

To get the complete datasets without any spaces in rows and columns

Getting the training dataset

EDA

Removing duplicate datas in this dataset

Data Visualizations

Getting the test dataset

Merging both train and test datasets to get the full data

Dropping few columns as it has more than 3/4 th of null values which may affect the prediction

Getting the dataset after dropping

Converting all object datas into int or float data for better processing

Converting all null datas

Dropping duplicate datas

DATA CLEANSING

Removing outliers

Final dataset for splitting

Training and testing

Since the target variable (y) contains more of float data, we consider Regression for predictionTrying with different alpha parameter

Regularisation

Trying with different alpha parameter

Testing with various models

Hypertuning

Compared to all models RIDGE model has the best r2_score, so hypertuning it.

Comparing the best three models in a dataframe

Getting the best model(RIDGE)

Saving the best fitted model

Reloading and testing the best model

• Data Sources and their formats

PROVIDED FILE CONTAINED:

Data Description.csv:

This contains the description of data.

train.csv:

This contains the dataset on which you will be working upon

Housing Use case:

This contains the problem statement and business goal.

Sample documentation:

This is a sample report.

test.csv :

Predict the output for these data with your best fit model.

• Data Preprocessing Done

Converting all object datas into int or float data for better processing

Converting all null datas

Dropping duplicate datas

DATA CLEANSING

Removing outliers

- Data Inputs- Logic- Output Relationships

EDA

Removing duplicate datas in this dataset

Data Visualizations

- State the set of assumptions (if any) related to the problem under consideration

Since the target variable (y) contains more of float data, we consider Regression for prediction

- Hardware and Software Requirements and Tools Used

Libraries –

Pandas --- used to dataframe a dataset

Numpy --- used to intrepret the data as an array

Scipy --- outliers removal (zscore)

Sklearn --- preprocessing,model_selection ,metrics,linear_model,tree,neighbors,svm and ensemble,

Implearn --- SMOTE

Matplotlib,Seaborn --- data visualizations

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

Since the target variable (y) contains more of float data, we consider Regression for prediction

Trying with different alpha parameter

Regularisation

Trying with different alpha parameter

Testing with various models

Hypertuning

Compared to all models RIDGE model has the best r2_score, so hypertuning it.

Comparing the best three models in a dataframe

- Testing of Identified Approaches (Algorithms)

Regularisation

Trying with different alpha parameter

Testing with various models

Hypertuning

- Run and Evaluate selected models

```
from sklearn.linear_model import Lasso,Ridge,ElasticNet
l=Lasso()
r=Ridge()
en=ElasticNet()
m=[l,r,en]
for i in m:
   i.fit(x_train,y_train)
   predi=i.predict(x_test)
   scorE=r2_score(y_test,predi)
```

```
    print('R2_score of model ' ,i, scorE*100)
R2_score of model  Lasso() 79.60870588513613
R2_score of model  Ridge() 80.0491666720944
R2_score of model  ElasticNet() 79.28370655729428
```

- Key Metrics for success in solving problem under consideration

Compared to all models RIDGE model has the best r2_score, so hypertuning it

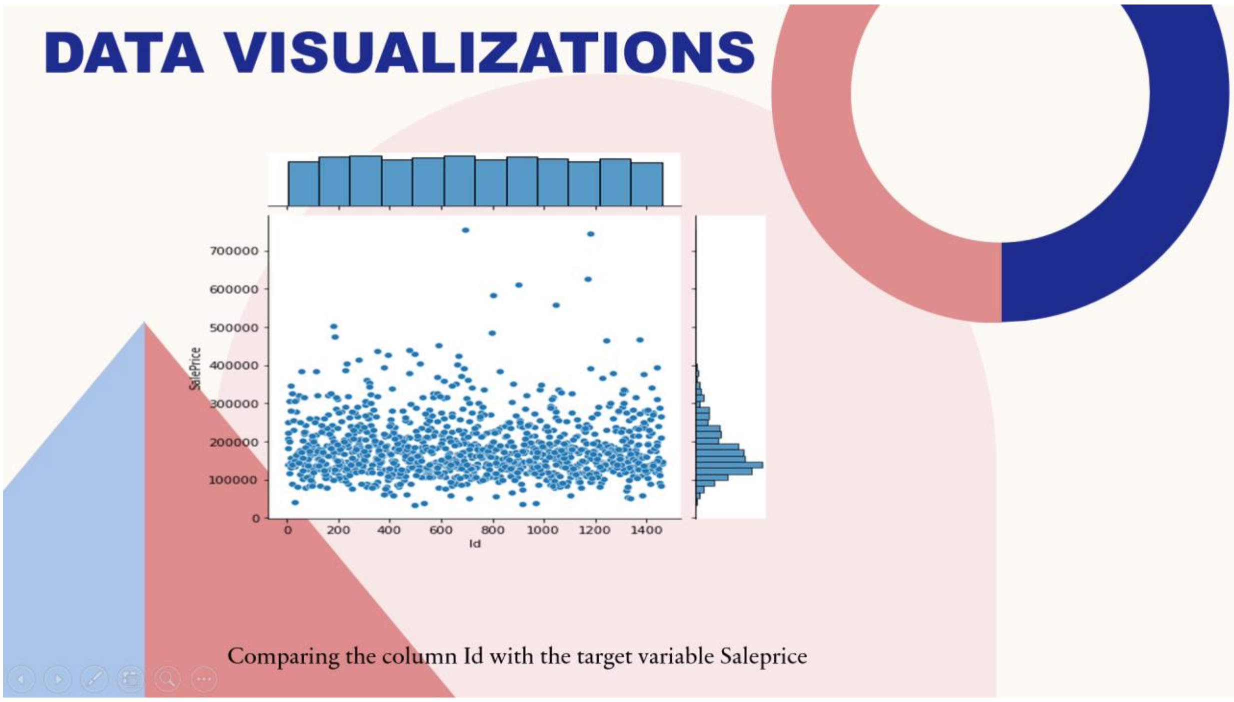
```
parameters={'alpha':[0.1,1.0],
            'fit_intercept':[True,False],
            'copy_X':[True,False],
            'max_iter':[800,900]}
gscv=GridSearchCV(estimator=Ridge(),param_grid=parameters)
gscv.fit(x_train,y_train)
GridSearchCV(estimator=Ridge(),
             param_grid={'alpha': [0.1, 1.0], 'copy_X': [True, False],
                         'fit_intercept': [True, False],
                         'max_iter': [800, 900]})
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.

On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

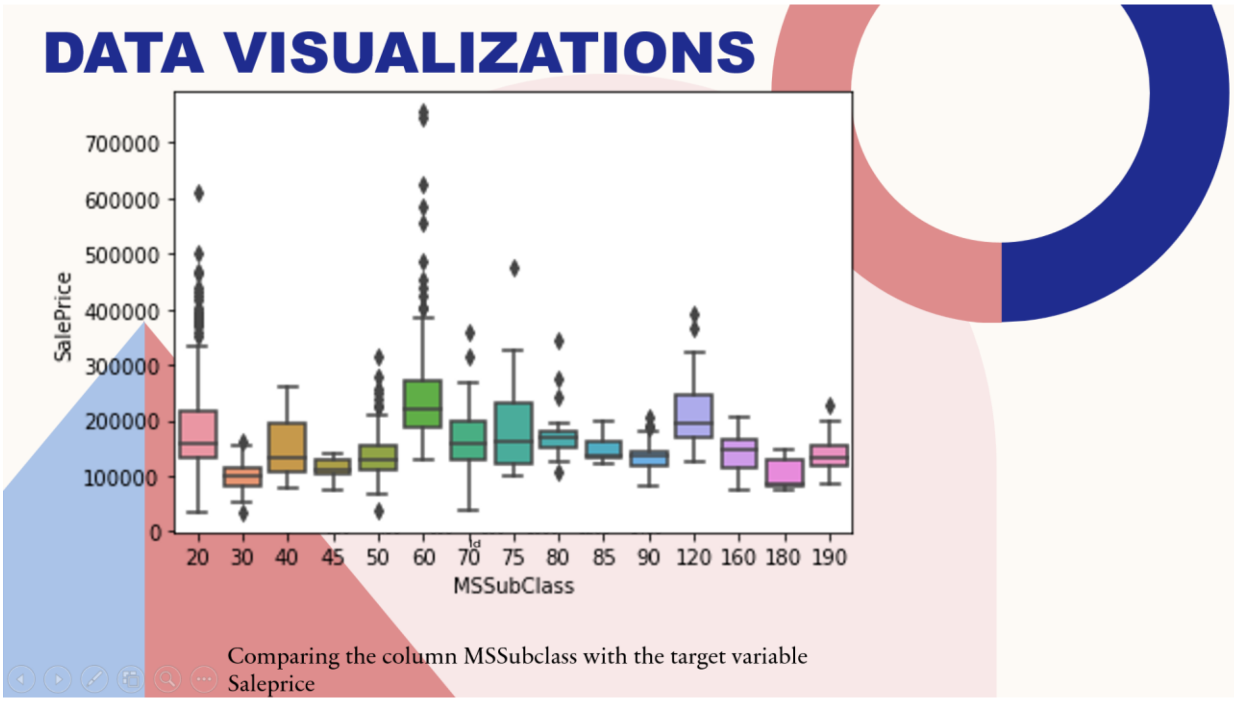
```
predr=gscv.predict(x_test)
print('R2_SCORE OF Ridge' ,r2_score(y_test,predr)*100)
R2_SCORE OF Ridge 80.03871221680038
```

- Visualizations



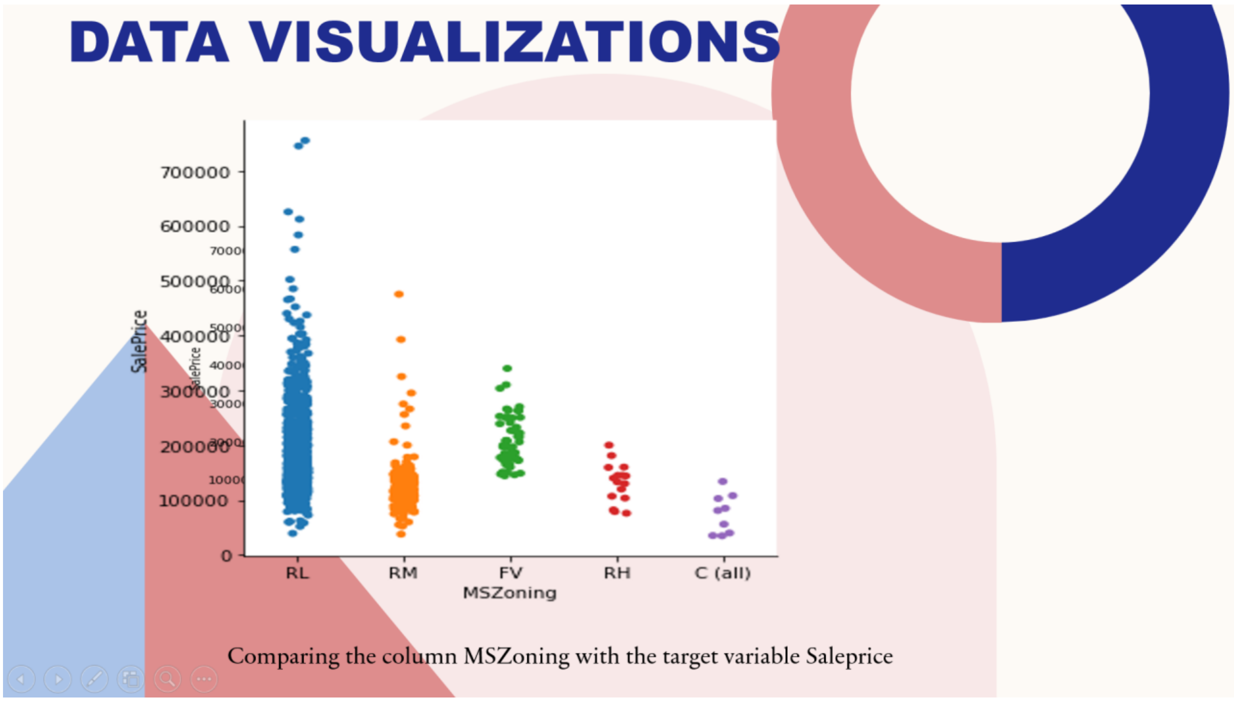
DATA VISUALIZATIONS

Comparing the column Id with the target variable Saleprice



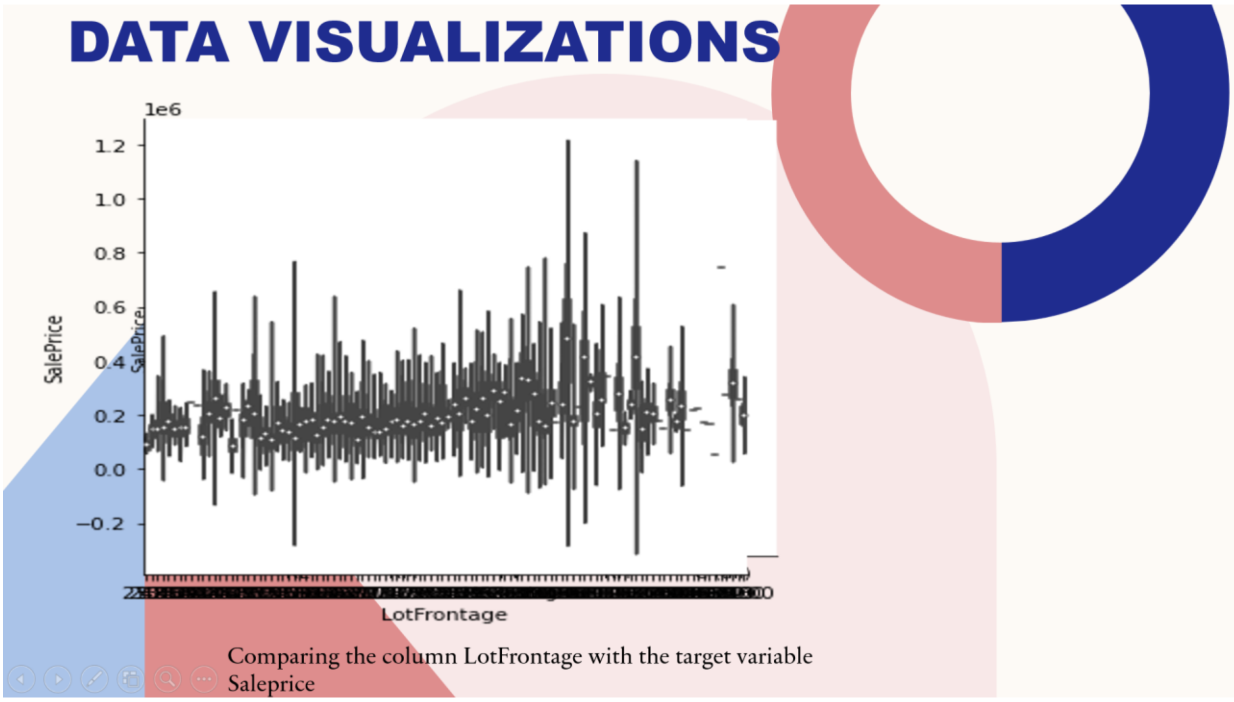
DATA VISUALIZATIONS

Comparing the column MSSubclass with the target variable Saleprice

- 

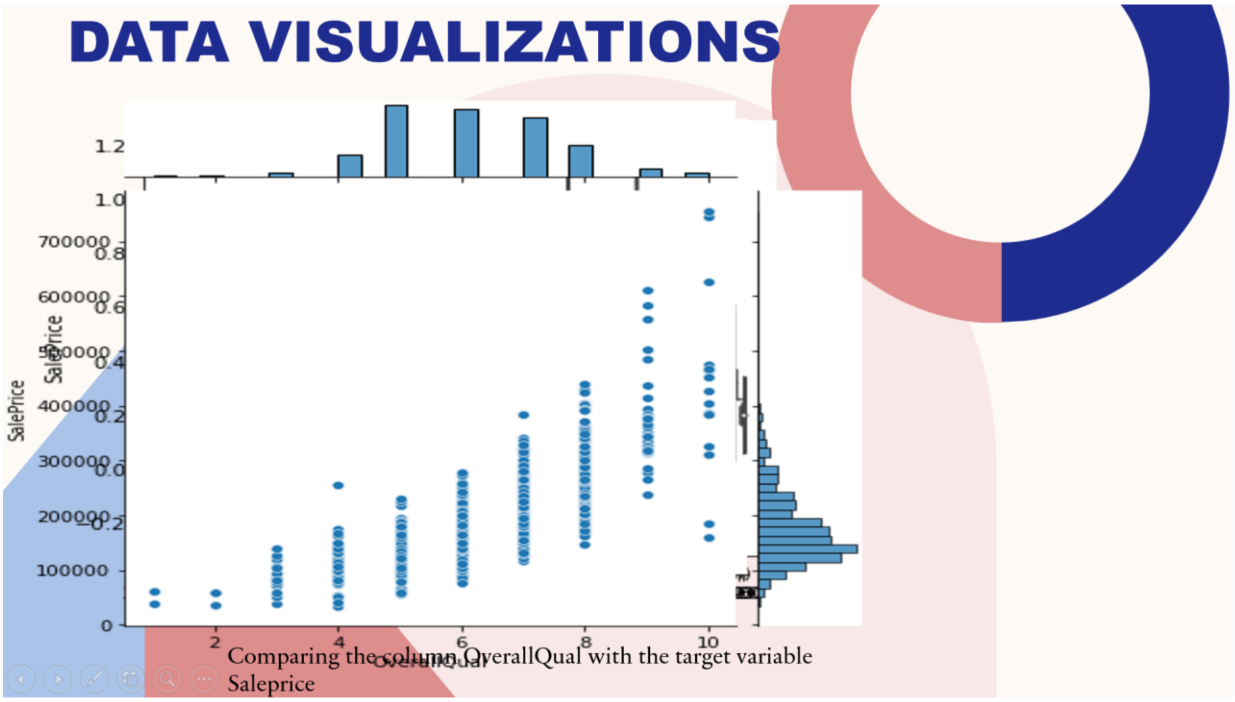
DATA VISUALIZATIONS

Comparing the column MSZoning with the target variable Saleprice

- 

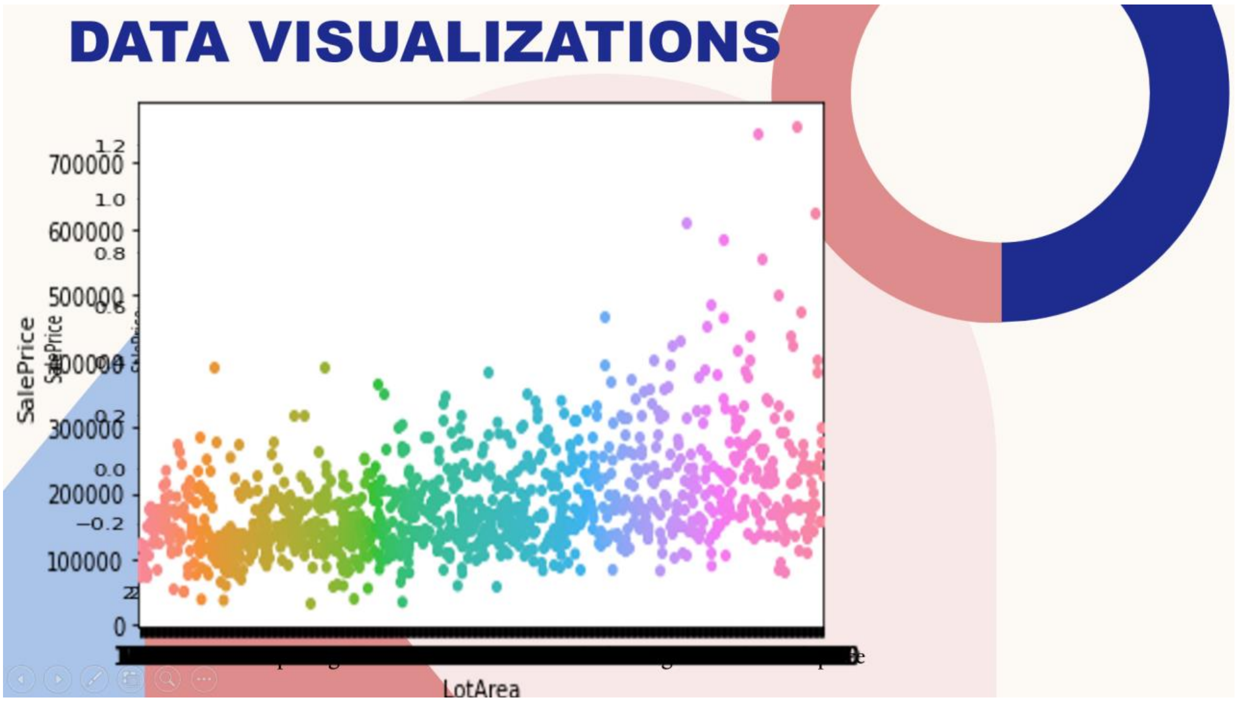
DATA VISUALIZATIONS

Comparing the column LotFrontage with the target variable Saleprice



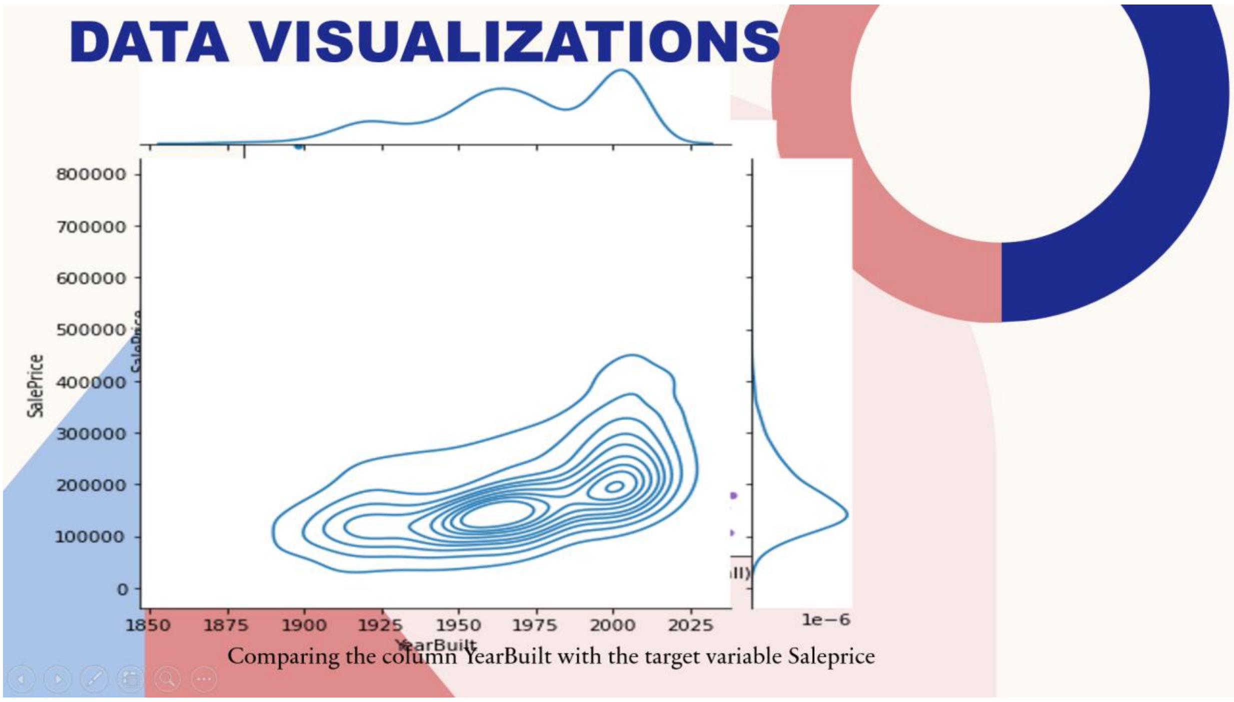
DATA VISUALIZATIONS

- Comparing the column OverallQual with the target variable Saleprice



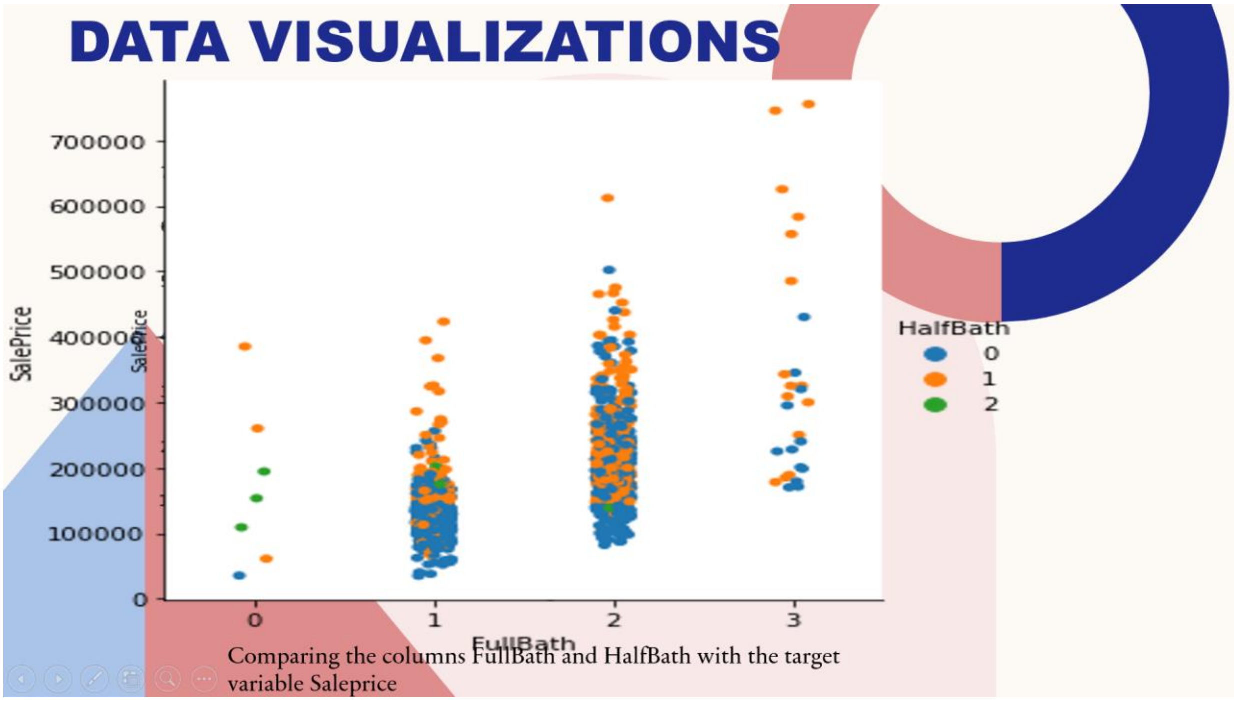
DATA VISUALIZATIONS

- [illegible]

- 

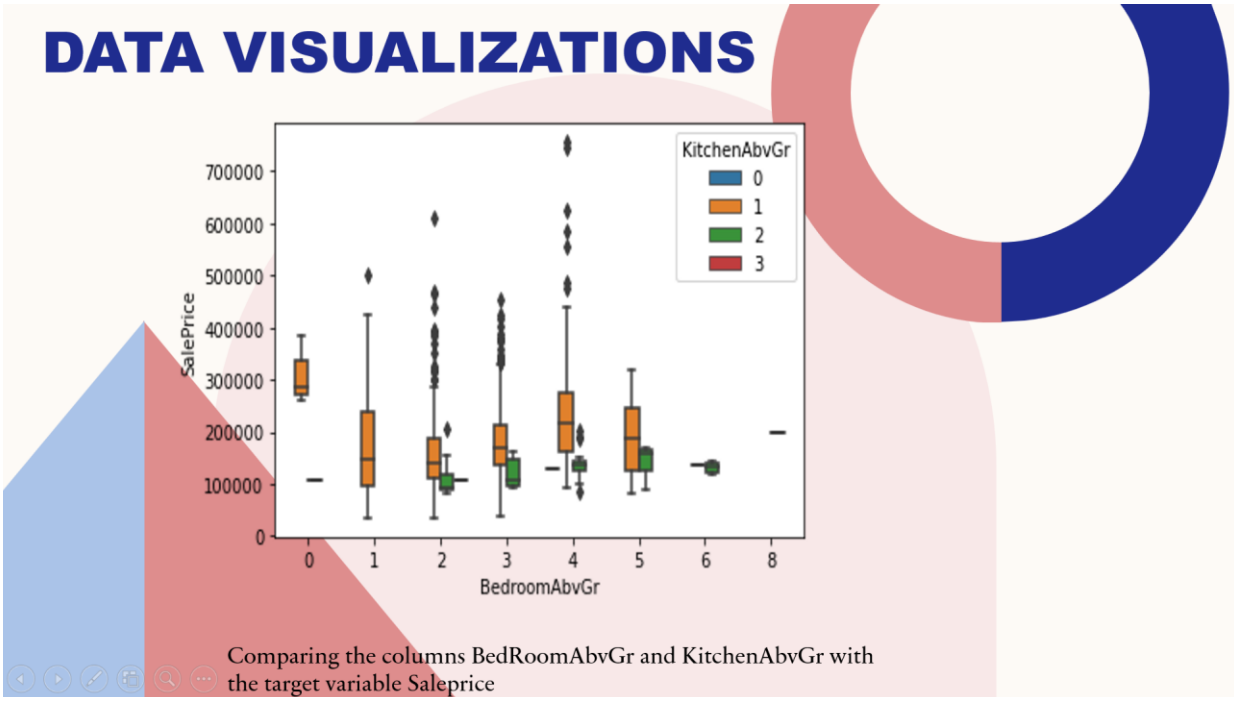
  ## DATA VISUALIZATIONS

  Comparing the column YearBuilt with the target variable Saleprice

- 

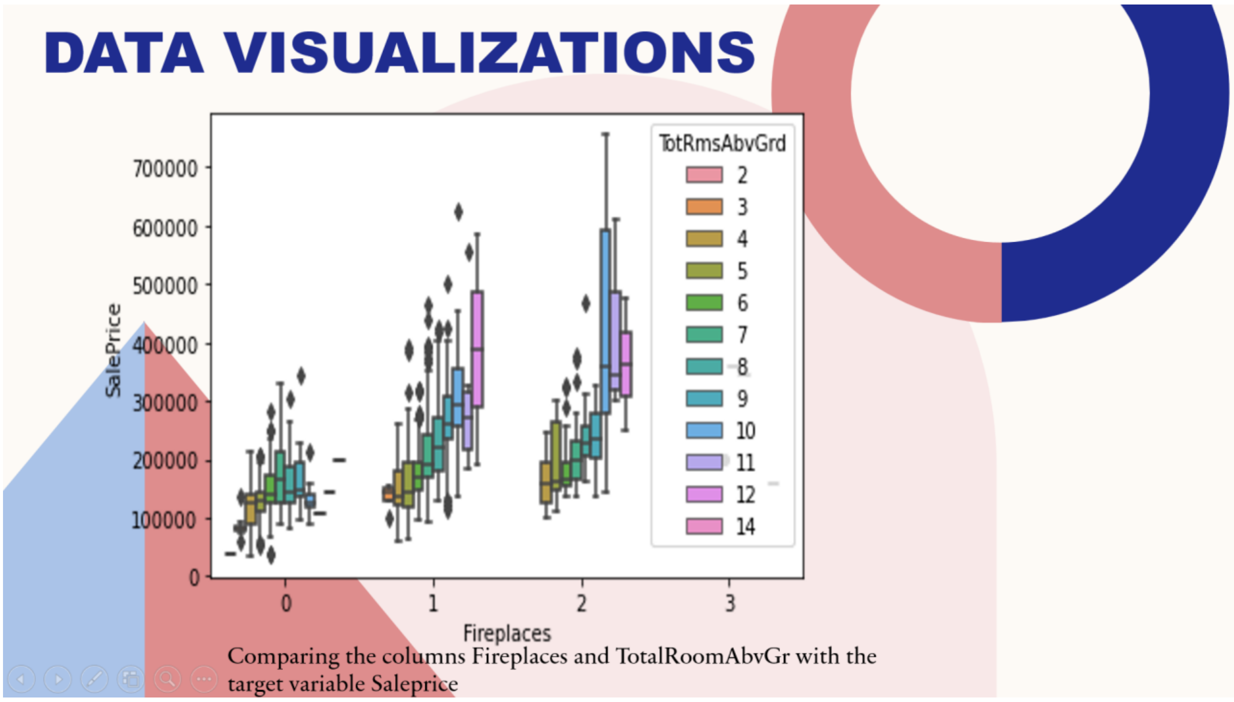
  ## DATA VISUALIZATIONS

  Comparing the columns FullBath and HalfBath with the target variable Saleprice



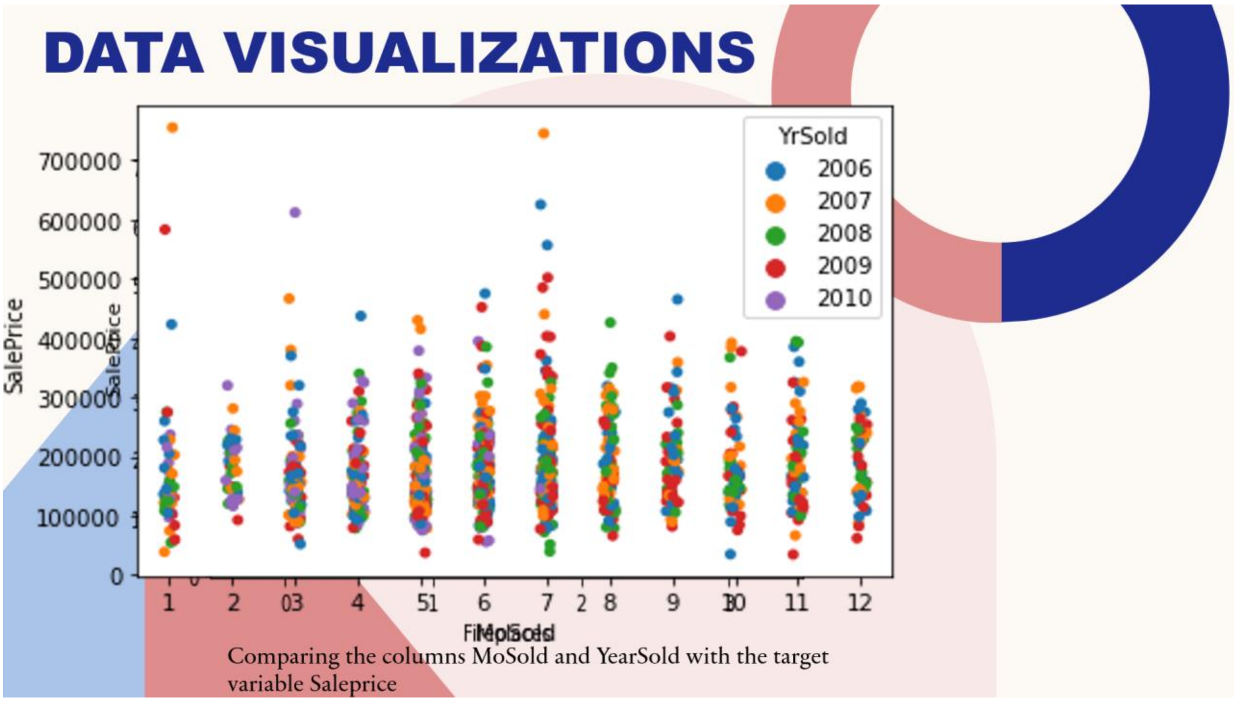
DATA VISUALIZATIONS

Comparing the columns BedRoomAbvGr and KitchenAbvGr with the target variable Saleprice



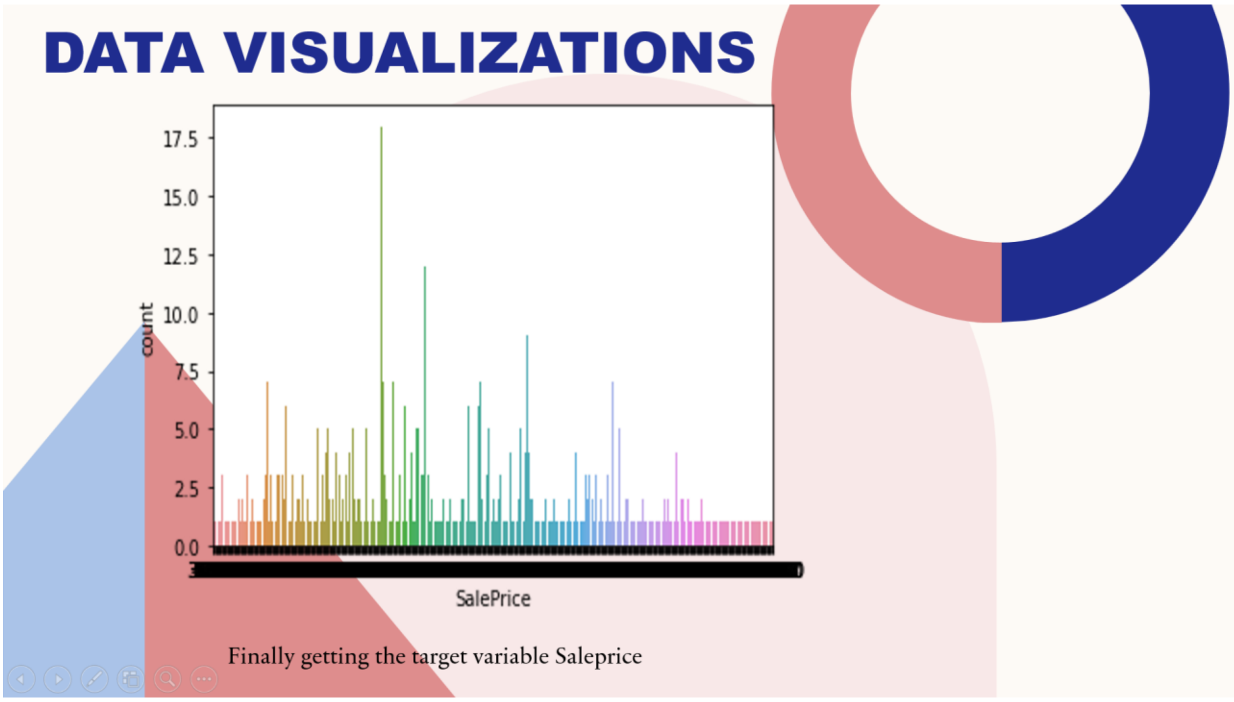
DATA VISUALIZATIONS

Comparing the columns Fireplaces and TotalRoomAbvGr with the target variable Saleprice



DATA VISUALIZATIONS

Comparing the columns MoSold and YearSold with the target variable Saleprice

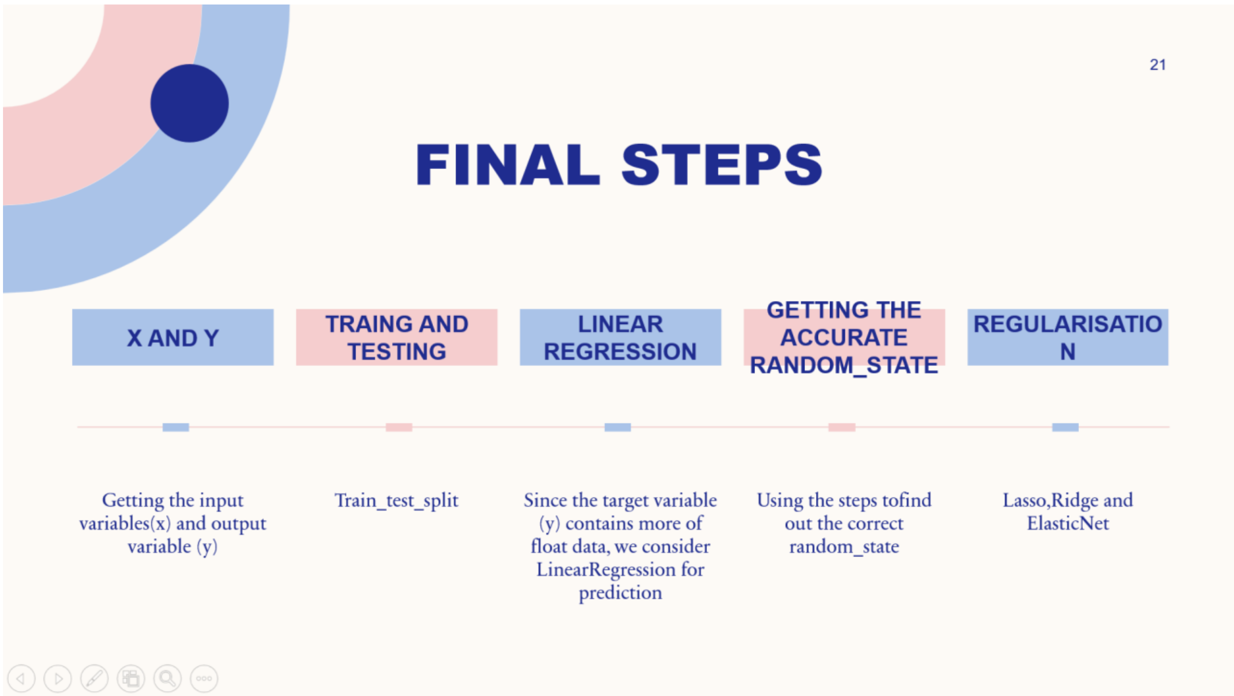


DATA VISUALIZATIONS

Finally getting the target variable Saleprice

CONCLUSION

- Key Findings and Conclusions of the Study



21

FINAL STEPS

X AND Y

Getting the input variables(x) and output variable (y)

TRAING AND TESTING

Train_test_split

LINEAR REGRESSION

Since the target variable (y) contains more of float data, we consider LinearRegression for prediction

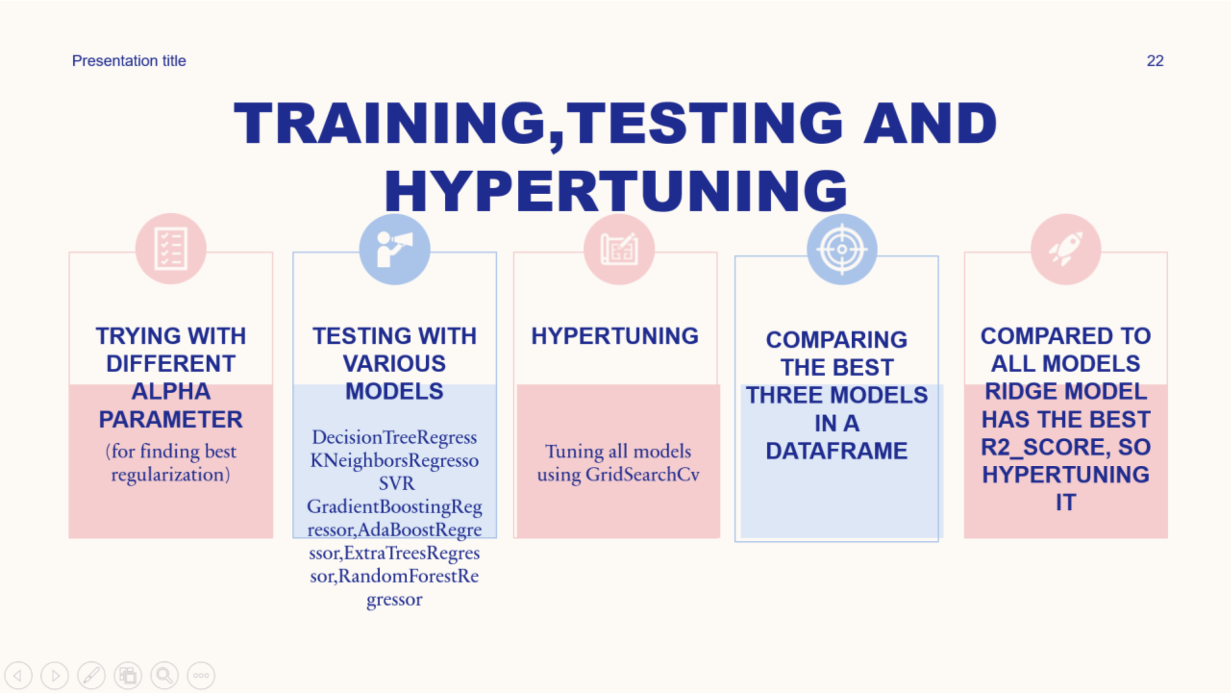
GETTING THE ACCURATE RANDOM_STATE

Using the steps tofind out the correct random_state

REGULARISATIO N

Lasso,Ridge and ElasticNet

- Learning Outcomes of the Study in respect of Data Science



Presentation title

22

TRAINING,TESTING AND HYPERTUNING

TRYING WITH DIFFERENT ALPHA PARAMETER

(for finding best regularization)

TESTING WITH VARIOUS MODELS

DecisionTreeRegress KNeighborsRegresso SVR GradientBoostingRegressor,AdaBoostRegressor,ExtraTreesRegressor,RandomForestRegressor

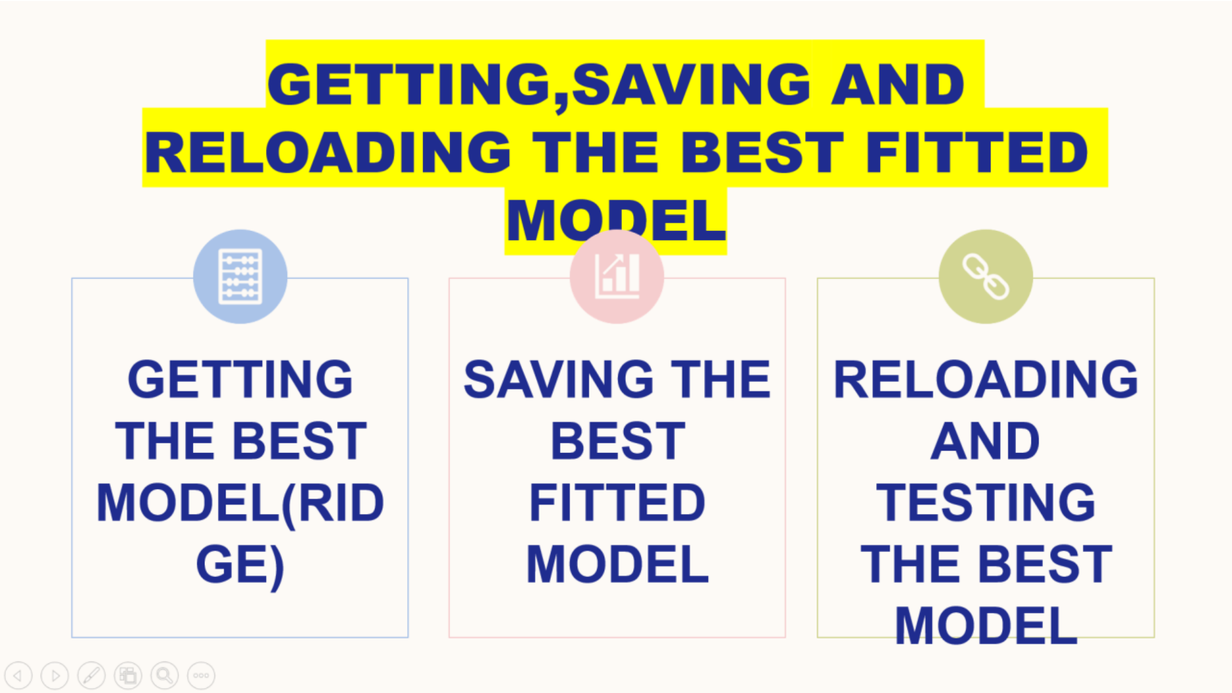
HYPERTUNING

Tuning all models using GridSearchCv

COMPARING THE BEST THREE MODELS IN A DATAFRAME

COMPARED TO ALL MODELS RIDGE MODEL HAS THE BEST R2_SCORE, SO HYPERTUNING IT

- Limitations of this work and Scope for Future

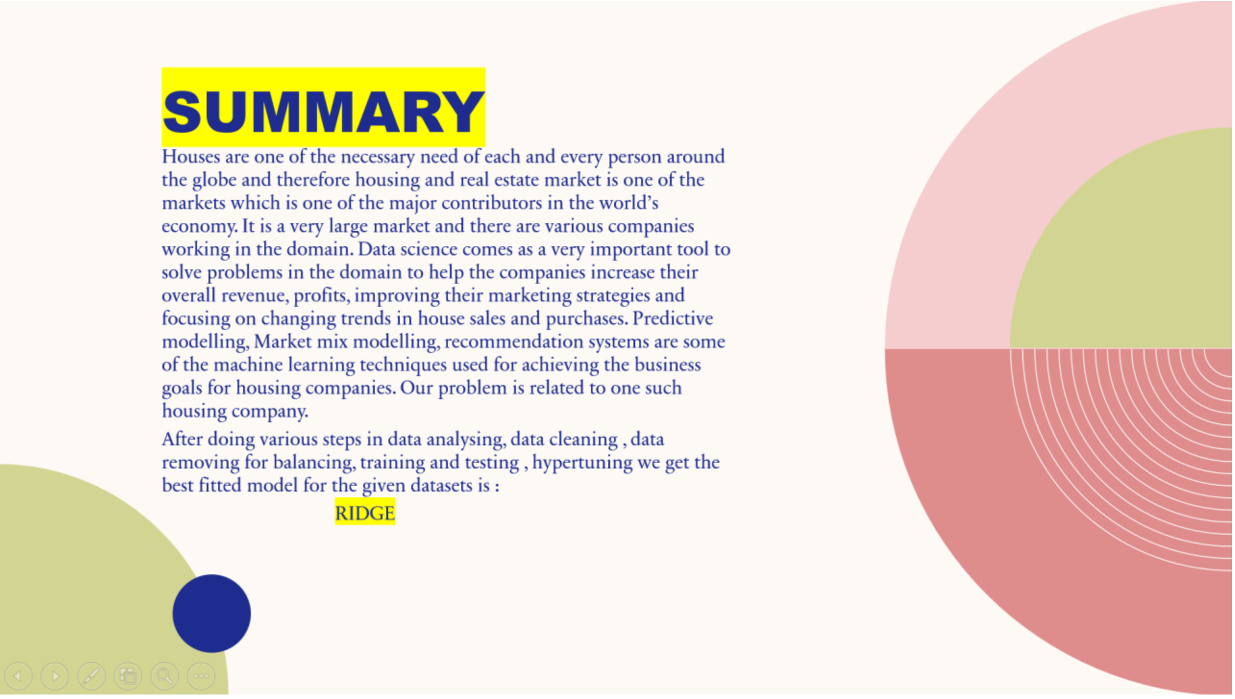


GETTING,SAVING AND RELOADING THE BEST FITTED MODEL

GETTING THE BEST MODEL(RIDGE)

SAVING THE BEST FITTED MODEL

RELOADING AND TESTING THE BEST MODEL



SUMMARY

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

After doing various steps in data analysing, data cleaning , data removing for balancing, training and testing , hypertuning we get the best fitted model for the given datasets is :

RIDGE

-

THANK YOU

from :

Name:

RAMAA DEVI S

Batch No :

INTERNSHIP34