PASH Estates Inc

Authors: Hardi Patel, Aayushi Shah

1. Abstract: Buying a house remains one of the most significant buying decisions that people make in their lifetime. It is difficult to predict the final sale price. There are people who struggle after knowing the expenses they will face in the future. This project will help with the decisions people will come up with. It will help in buying a perfect house which will satisfy their financial needs as well. The readings to examine are the graphs related to sale prices that will provide more context and background. There are a few variables that we would examine like total number of bedrooms, bathrooms; whether remodeling took place, age of the house after and before remodeling , its landscape, new garage quality obtained by multiplying garage quality with the number of cars a garage could fit, also the total square feet and the consolidated porch area. As the dataset from the Kaggle is taken, the training set has 1460 observations (81 variables) and the test set has one less variable besides SalePrice. It is used to examine the house prices in Iowa. We can use Multiple Linear Regression. It is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. Multiple linear regression is used to model the relationship between a continuous response variable and continuous or categorical explanatory variables. The XGBoost algorithm can be used. XGBoost is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Qualitatively we can expect to see higher sale prices when there are many more variables which will affect the prices in the future. Quantitatively we can use

scatter plots and boxplots to evaluate our results between the observations and sale prices. Pash Estates Inc. will give a visual representation of the dataset we have gathered.

2. Introduction: The problem we are working on is related to the change in prices of houses when the features and specifications are separate and updated. For example, if a house is on sale with two bedrooms and one bathroom, its sale price will be lower than a house containing more than two bedrooms and one bathroom. Similarly, many other specifications can change house prices, including garage, pool, schools, fireplace, basement, heater, central air conditioning, parking, area (sq ft), house condition, and the quality of material used in making the house's interior and exterior. Also, the area a house is located in works as a significant factor in the price of the house. It is essential to look into the specifications when buying a house as it drastically impacts its sale price. It is hard to predict the correct and accurate sale price of houses in today's world as there are many variables to take into effect. Let us use one of the specifications: central air conditioning. If a person is buying a house with no central air, then the price should be around $100000, but on the other end, the same house with central air will cost about $160000. One another example is the basement. Without a basement, a house can be sold at a lower price as compared to a house with a basement since the owner has to pay more tax for the house with a finished basement. So PASH Estates brings solutions to all those problems and solutions that a person should know about before buying a house. It will help clarify the concerns by giving precise results that will include bar graphs, scatterplots, and boxplots to compare the sale prices of houses when amenities and specifications are used. The results will consist of all the specifications and the expenses related to it.

3. Related Work: Some of the published work on Kaggle relating to the house price competition include linear regression to show the distribution between specifications and the price of the house. The more the specifications, the more the cost of the house. Some published works have also used XGBoost with XGBRegressor for a more detailed outcome. At the same time, some have used Elastic Net for the model and the final score. Also, Random Forest Regressor, Gradient Boosting Regressor, and AdaBoostRegressor have been used in many of the published works on Kaggle. PASH Estates also uses Random Forest Regressor and Gradient Boosting Regressor that will be used to learn the dataset. Also, Jax is used to doing the coding in python. We have used Multiple Linear Regression and XGBoost algorithms that will be used to show the relationships between multiple variables that are used in deciding the price of houses.

4. Data: The data that we are working on came from the kaggle competition. It is a csv file that includes different features of the house prices. There are a total of 81 columns and 350 rows in the train dataset and 2919 rows and 80 columns in the test dataset which has different data stored. There was no need for preprocessing, filtering, or any other special treatments to use this data in the project. The following is the description of the data stored in the dataset.

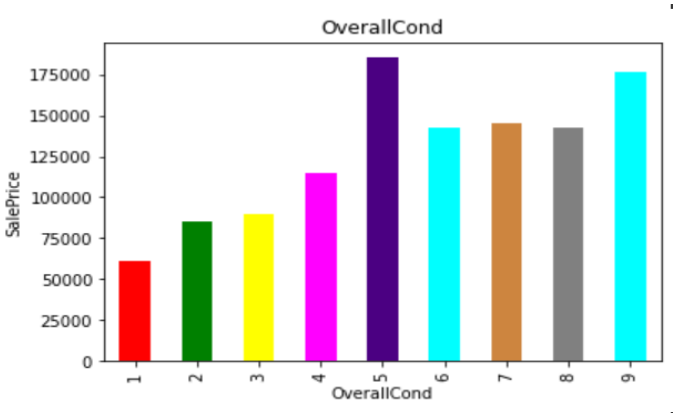
    - SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
    - MSSubClass: The building class which identifies the type of dwelling involved in the sale (1-STORY, 2-STORY, 2-½ STORY)
    - MSZoning: The general zoning classification of the sale (Agriculture, Commercial, Industrial, etc)
    - LotFrontage: Linear feet of street connected to property
    - LotArea: Lot size in square feet
    - Street: Type of road access to property (Gravel, Paved)
    - Alley: Type of alley access to property (Gravel, Paved, No alley access)
    - LotShape: General shape of property (Regular, Irregular)
    - LandContour: Flatness of the property (Near Flat, Banked, Hillside)

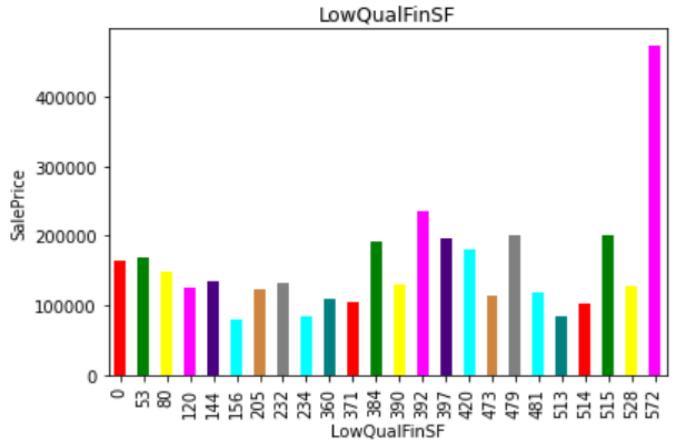
- Utilities: Type of utilities available (All, Electricity/Gas/Water)
- LotConfig: Lot configuration (Inside, Corner, Front)
- LandSlope: Slope of property (Gentle, Moderate, Severe Slope)
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling (Single-family, Duplex, Townhouse)
- HouseStyle: Style of dwelling
- OverallQual: Rates the overall material and finish quality of the house
- OverallCond: Rates the overall condition of the house
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
- RoofStyle: Type of roof (Flat, Gable, Hip, Shed)
- RoofMatl: Roof material (Metal, membrane, Wood)
- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if more than one material)
- MasVnrType: Masonry veneer type (Brick, Cinder, Stone)
- MasVnrArea: Masonry veneer area in square feet
- ExterQual: Evaluates the quality of the material on the exterior
- ExterCond: Evaluates the present condition of the material on the exterior
- Foundation: Type of foundation (Brick, Cinder Block, Slab, Stone, Wood)
- BsmtQual: Evaluates the height of the basement
- BsmtCond: Evaluates the general condition of the basement
- BsmtExposure: Refers to walkout or garden level basement walls
- BsmtFinType1: Rating on quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Rating of quality of second finished area (if present)
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition
- CentralAir: Central air conditioning (Yes, No)
- Electrical: Electrical system
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality

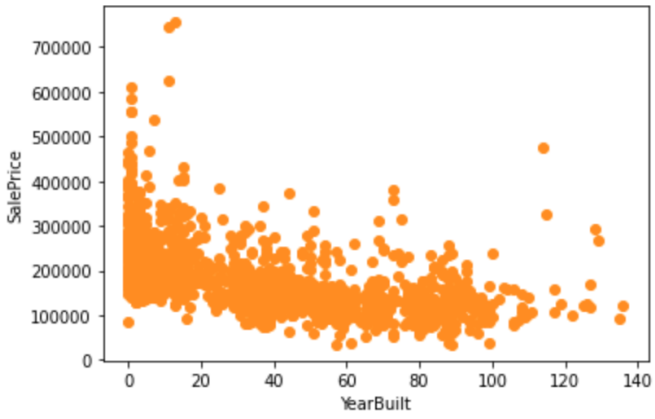
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: $Value of miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale

5. Methods: In order to find a perfect house, the buyer should be aware of all the features of the house that are mentioned in the data. Buying a house has always been a big decision so our approach for solving the problem and bringing the right home to them. The approach that we have chosen clearly explains the features of the houses by comparing them with other houses. The numerical features like LotFrontage, Alley, Area, Condition, Year Built, GarageType, PoolQC, Fence, etc will give the exact price of the house in the current market. The sale prices are relevant or not oro are higher than usual houses should have according to the statistics. The overall rating quality from scale of 1 - 10 according to the sale prices also help to clarify. We also consider other alternatives where the house prices were concluded by the not asking neighbors rating of the area. So, our approach

includes that feature where the house price is concluded after going over all the features including the ratings. Our idea was basically to draw conclusions by plotting graphs of the different features versus the different sale prices.

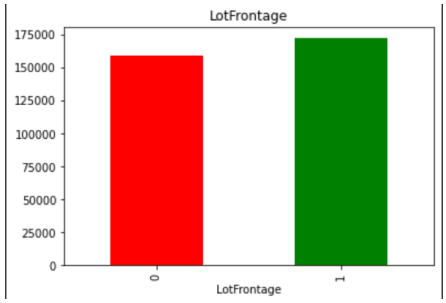






In the above graphs we can see that different houses have different sales price values based on the overall condition of the house. The second graph shows the low quality of the finished square feet a house has and its prices accordingly. The third graph shows a scatter plot of year built versus sale price to help predict the correct price of the house.

6. Experiments: Different experiments have been done in order to figure out the exact price of a house. At first, the data set is loaded, from that dataset the features that are missing in a data of houses are listed and named missing values. Then the variables are made that indicate 1 if the observations were missing and if not 0. Then we calculated the mean Sale Price where the information is missing or present. Which resulted us to the graphs of different features and they are following.

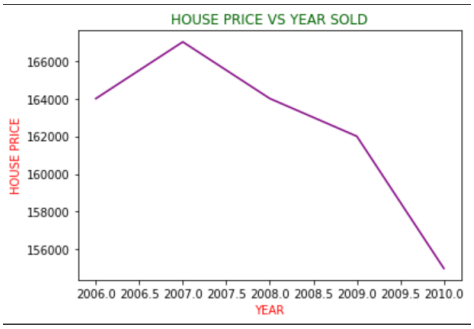


Then we visualized the numerical Features which turned out to be 38 for different features of the houses.

	ID	MSSubClass	LotFrontage	LotArea	OverallQual
0	1	60	65.0	8450	7

1	2	20	80.0	9600	6
2	3	60	68.0	11250	7
3	4	70	60.0	9550	7
4	5	60	84.0	14260	8

The above data shows the information of the houses and its lot front yard, total area, overall quality and its condition. Then we further listed the variables that contained the year information. Later, we plot a graph showing the house price and the year sold below:



Then we captured the difference between the year variable and the year the house was sold. There were some limitations to this prediction since there were many ups and downs in the prices of the houses in those years. However, at last the proper predictions were able to be made using all the features of the houses.

7. Conclusion: From these predictions, one can conclude the price of the house having full knowledge of it. Not only exterior, but also interior costs and its worth can be known of a house by PASH Estates Inc. We learned that the more the facilities the more the price of the house. Jax is able to differentiate through sorts of python and numpy functions. It is useful for deep learning as we can run back propagation effortlessly. Jax takes less time

to run on GPU as compared to numpy. It can differentiate through a large subset of python's features including loops and recursion. Currently, Jax is used for research purposes due to its cool features. In the future, we can improve our model by showing the results of the best facility with less money in a 3D environment so that way we can take a closer look at all the installed features in a house. This will really help for the people in the future at the house without taking an in-person look if they are not living in a commutable distance.