COMP309 Assignment 2

By Viy Moodley (300565283)

Part 1: Business and Data Understanding

1) EDA

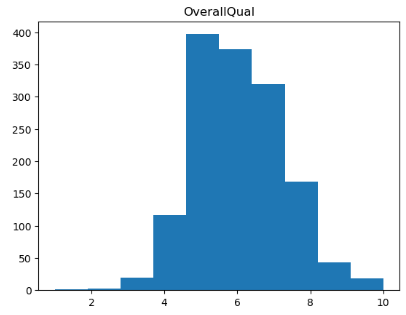
a) The summary statistics of the House Price data had 1460 instances and 81 features. Of the 81 features, 43 were categorical and 38 were numerical. It should be noted that one of the numerical features is the "Id" column, which will be excluded since it doesn't have any real analytical value, additionally '*SalesPrice'* is the target feature, so there are actually 79 explanatory variables.

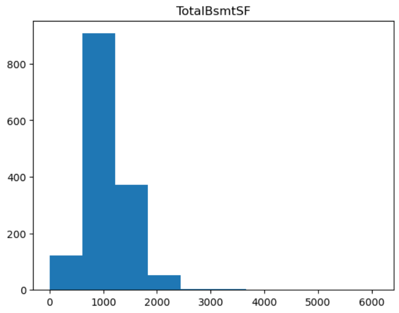
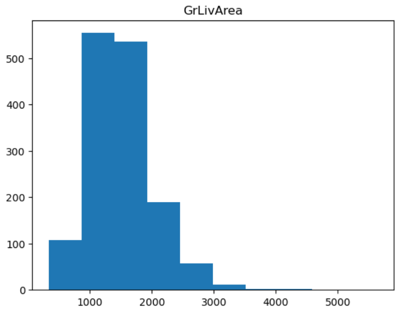
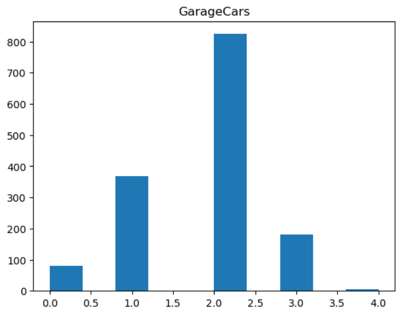
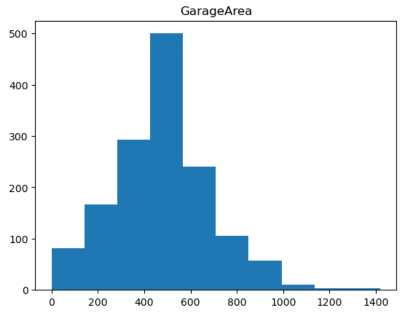
b) According to Pearson's Correlation, the top 5 numerical features that are most highly correlated with the target value of *SalePrice* are:

- *OverallQual*: 0.790982
- *GrLivArea*: 0.708624
- *GarageCars*: 0.640409
- *GarageArea*: 0.623431
- *TotalBsmtSF*: 0.613581

Note: While the data for 'OverallQual' represents a numeric scale with a value range from 1 to 10 – the data description has labels for each numeric value, which suggests the variable may be ordinal categorical. After conducting research, I have decided to include "OverallQual" as a numerical feature because it's values quantitatively measure a house's overall quality. Overall quality had the highest correlation(0.79) with the target feature SalePrice, which indicated that higher quality houses are strongly associated with higher sale prices and further justifies its inclusion in a predictive model.

c) ***Table 1: Analysis of Distributions of Most Highly Correlated Features***

Histogram	Shape (4 d.p.)	Comments
	Skewness: 0.2167	The distribution of the histogram is slightly right skewed, as indicated by the small skewness value. Small kurtosis value suggesting a more normal shape, not too flat and not too sharp – not many outliers.
	Kurtosis 0.0919	

	Skewness: 1.5227 Kurtosis: 13.2010	The histogram appears to be distributed toward the left of the x axis and presents a moderately right skewed distribution, the skewness value supports this observation. Additionally the graph looks quite “sharp” compared to a normal distribution and has an extremely high kurtosis value. Significant departures from normality with outliers contributing to a heavy tail.
	Skewness: 1.3652 Kurtosis: 4.8743	The distribution presents a moderately right skewed spread, with a tight spread of bars along the left of the x axis. The kurtosis value is high and the shape is “sharper” than a normal distribution and more ‘heavy tailed’ suggesting some outliers that are causing ‘heavy tailed-ness’.
	Skewness: -0.3422 Kurtosis: 0.2161	Slightly left skewed distribution, as indicated by the negative skewness value and the data being spread more toward the right of the x axis. Moderate kurtosis value, shape is close to that of a normal distribution, outliers are a minor concern.
	Skewness: 0.1798 Kurtosis: 0.9098	Histogram bars appear to be slightly more spread out over the left of the graph, skewness value suggests a slight right skew. The shape of the distribution is sharp and the kurtosis value suggests a moderate peak.

Key patterns: Firstly, features like ‘GrLivArea’, ‘TotalBsmtSF’, and ‘GarageArea’ exhibit right skewed distributions, which indicates that the majority of houses are clustered in the lower values of these variables’ data ranges with evidence of some extreme values on the higher end. The second pattern is that ‘OverallQual’ and ‘GarageCars’ appear to have relatively normally distributed spreads with little skewness, this suggests these features are more evenly distributed across the dataset.

d) The dataset shows that 19 features have missing values, with the proportion of missing data varying widely. For some features, like ***Electrical***, only one value is missing, representing less than 0.1% of the dataset. Other features, such as ***MasVnrType***, ***MasVnrArea***, and basement related features like ***BsmtQual*** and ***BsmtCond***, have between 8 and 38 missing values, which is roughly under 3% of the data. In contrast, features such as ***LotFrontage*** have 17.7% missing, while ***FireplaceQu*** is missing in 47.3% of cases. The more extreme cases are ***Fence***, ***Alley***, ***MiscFeature***, and ***PoolQC***, where over 80% of the data is missing, with ***PoolQC*** having the largest portion of missing data with 99.5% of its values being null. These gaps in the data will likely require different handling approaches, such as imputation or exclusion, depending on their impact.

2) Business Understanding

a) Data Mining Goal 1 – Identify the most important features that highly influence the price of a house. Data Mining Goal 2 – Model the relationships between these influential features to understand the impact and effects these features have on house prices.

b) Paradigms:

- Regression: Since the goal is to predict a continuous target variable (house price), regression is an appropriate choice. It will give insight and quantify how each feature impacts the house price
- Dimensionality Reduction: Dimensionality reduction techniques (e.g., PCA) can help identify the most significant variables and reduce the complexity of the model, improving interpretability and performance.

3) Further EDA

The results of further EDA using hierarchical clustering suggest that house prices do vary by neighbourhood. The analysis of the dendrogram and boxplot provides a clear understanding of the variation in house prices across neighbourhoods. The dendrogram was generated through hierarchical agglomerative clustering and groups houses based on five key features highly correlated with sale prices: *Overall Quality, Gr Liv Area, Garage Cars, Garage Area*, and *Total Basement SF*. This hierarchical structure reveals how neighbourhoods with similar structural attributes cluster together, suggesting that certain neighbourhoods do share similar housing characteristics, which could potentially reflect similar price ranges.

Figure 1: Dendrogram of Hierarchical Clustered Data and Neighbourhoods

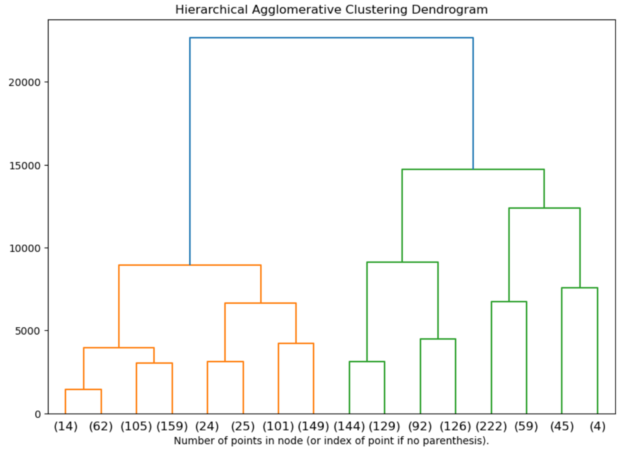
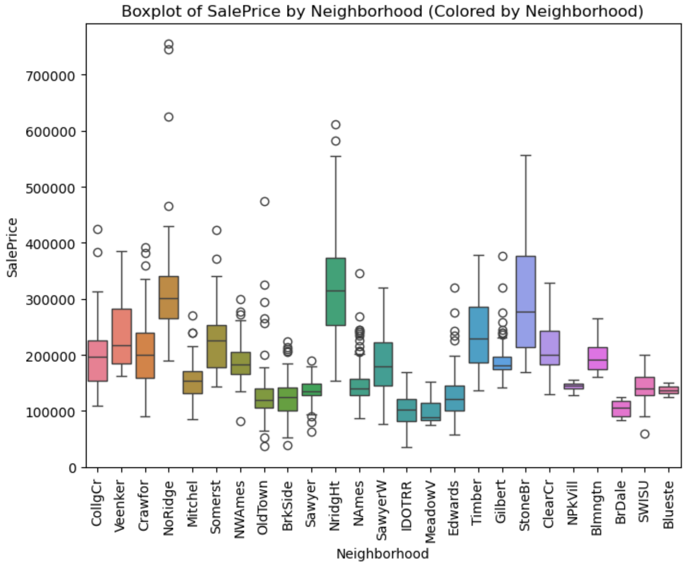
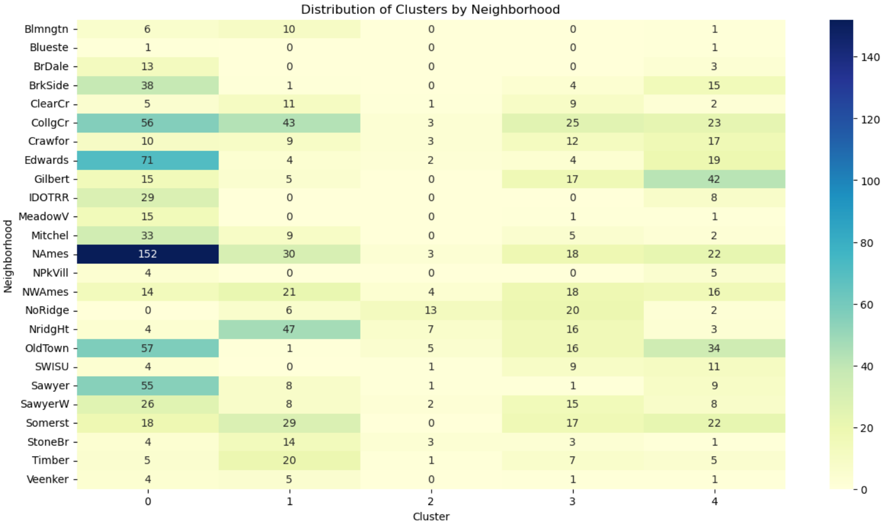


Figure 2: Boxplots of Neighborhood Distributions across House Prices



The boxplot of *SalePrice* by *Neighborhood* further reinforces this insight by displaying the distribution of sale prices across various neighbourhoods. The results show that median sale prices vary significantly, with some neighbourhoods, such as "*NoRidge*" and "StoneBr," exhibiting consistently higher prices, while others, like "BrDale" and "MeadowV," have lower median prices. The spread of prices within each neighborhood also differs, indicating that some areas have more consistent price ranges, while others, like "NridgHt," display a broader range of house prices.

Figure 3: Heatmap Showing Neighbourhoods and Clusters



The heatmap reveals a clear relationship between neighbourhoods and house prices, as determined by agglomerative clustering based on the top five features correlated with price. Neighbourhoods such as *NAmes* and *OldTown* show a strong concentration in specific clusters, indicating that houses in these areas share similar features and price ranges. In contrast, neighbourhoods like *CollgCr* are spread across multiple clusters, suggesting more variation in house characteristics and prices. This analysis highlights that neighborhood location significantly influences house price patterns, with some areas exhibiting more homogeneity in pricing than others.

All of this makes perfect sense in a real world context. Neighbourhoods are often shaped by a variety of factors, including location, access to good schools, safety, and proximity to amenities like or shopping centres. These factors create demand and drive up house prices in certain areas, while other neighbourhoods might be less desirable in this aspect and thus see lower prices. The clustering in the dendrogram and the spread of values in the boxplots reflect the natural variation in what people are willing to pay for a home in different parts of a city. It reflects what you'd expect to see in any housing market, where different neighbourhoods cater to different needs, preferences, and price points.

Part Two: Data Preparation and Machine Learning

1) Preprocessing Overview:

a) Missing Data:

- Numeric Columns: Missing values were imputed using the median, this also helped address cases of outliers. For *LotFrontage*, neighbourhood specific medians were used based on the assumption that frontages would be similar across neighbourhoods.

- Categorical Columns: Missing values were filled with "NA", since null values represented cases where data was not applicable. For example, all cases of *FireplaceQu* with missing data, were directly related to houses with 0 *Fireplaces* and therefore a value for the *FireplaceQu* feature was not entered.
- I contemplated deleting columns with more than 10 percent of its values missing, but ultimately decided against it since I will be carrying our dimensionality reduction anyway.

b) Categorical Data Encoding:

- Used ordinal encoding to assign unique integers to each category. Unseen categories in the test set were handled with a default value of -1.

c) Normalisation:

- StandardScaler was used to standardise numerical features to have a mean of 0 and a standard deviation of 1.

d) The steps are explained further in my code notebook for Part 2.

2) Dimension Reduction

In this analysis, I used Principal Component Analysis (PCA) and SelectKBest – to help identify and remove irrelevant or redundant features when predicting house prices. The first technique I applied was PCA, with the goal of reducing the feature space while still retaining 95% of the variance present in the original data. After fitting the PCA model to the training data and applying the same transformation to the test data, I successfully reduced the number of features to 44 principal components. This allowed me to significantly simplify the dataset while still capturing most of the important information. However, since these principal components are linear combinations of the original features, they lack direct interpretability.

For the second technique, I used SelectKBest, which focuses on selecting the top features based on their statistical relevance to the target variable – house prices. By ranking the features using the f_regression scoring function, I was able to identify the most important features, such as *'OverallQual*,' *'YearBuilt*,' and *'GrLivArea*.' Unlike PCA, SelectKBest preserved the original features, which provided clear insight into which specific variables are most strongly correlated with house prices. Additionally, SelectKBest's top features aligned with widely accepted real estate knowledge. For example, higher quality homes, larger living areas, and bigger basements tend to be associated with higher prices.

Both methods had their advantages. PCA effectively reduced the dimensionality of the dataset, albeit at the expense of interpretability, while SelectKBest provided a straightforward way to isolate the key drivers of house prices. After running these techniques, I removed redundant and irrelevant features from the dataset according to each method, leaving a two subsets of features for building predictive models. Moving forward, I will use reduced versions of the data using 44 components from PCA as well as the selected features from SelectKBest. Both techniques helped to simplify the data and focus attention on the most important variables.

3) Machine Learning

a) Table 2: Results of Linear and Ridge Regression Models on PCA and SelectKBest Data

	Linear		Ridge	
PCA	**Train MSE:**	1207651130.0447474	**Train MSE:**	1207651206.596266
	Test MSE:	837237981.061347	**Test MSE:**	837278424.8034133
SelectKBest	**Train MSE:**	1416411378.7681773	**Train MSE:**	1416411799.555341
	Test MSE:	1113792942.308731	**Test MSE:**	1113901764.9324772

In my analysis of linear and ridge regression models using PCA and SelectKBest for feature selection, several key insights emerged based on the mean squared errors (MSE) on both the training and test sets.

For linear regression with PCA, the model had an MSE of 1.21 billion on the training set and 837 million on the test set, indicating good generalisation. Ridge regression with PCA produced nearly identical results, suggesting that regularisation provided little additional benefit, likely because the principal components did not suffer from multicollinearity. Both models using PCA effectively retained the variance necessary for predicting house prices, with minimal overfitting.

In contrast, when using SelectKBest, both linear and ridge regression models showed higher errors, with linear regression yielding 1.42 billion on the training set and 1.11 billion on the test set. The increased MSE suggests that although SelectKBest selected relevant features, the reduced feature set did not capture as much variance as PCA, leading to weaker generalisation. Ridge regression again showed minimal improvement over linear regression, indicating that regularisation was not particularly useful.

The analysis revealed that PCA outperformed SelectKBest in both linear and ridge regression models, leading to better generalisation and lower test errors. PCA's ability to retain a larger portion of the data's variance resulted in more accurate predictions, while the feature selection approach of SelectKBest, though effective at identifying key features, reduced the feature set in a way that led to higher errors and poorer generalisation. Additionally, regularisation through ridge regression had minimal impact, suggesting that overfitting and multicollinearity were not significant issues in the models. Ultimately, the PCA based models demonstrated a stronger ability to preserve the complexity of the data, making them more effective for predicting house prices compared to the SelectKBest models.

b) Random Forest Regression

	Random Forest Regression	
PCA	**Train MSE:**	190033663.85781825
	Test MSE:	1040741294.2819452
SelectKBest	**Train MSE:**	201323091.84153545
	Test MSE:	790028489.9397485

The Random Forest model consistently outperformed both the Linear and Ridge Regression models, especially when paired with features selected by SelectKBest. The test MSE for Random Forest with SelectKBest was the lowest, at 790 million, showing that Random Forest was more effective at handling high dimensional feature spaces and capturing complex, nonlinear relationships that linear models might miss. The model's performance gap was more noticeable with PCA data where the training MSE was much lower than the test, which could suggest some overfitting. In contrast, when using SelectKBest, the model generalised better, likely because it focused on the most relevant features while ignoring less important ones.

The difference in how PCA and SelectKBest handle feature selection may have influenced these outcomes. PCA reduces dimensionality by transforming the data into components that explain the most variance, which can compress the feature space in a way that benefits linear models. However, this transformation might have caused Random Forest to lose some of the more subtle, nonlinear patterns it would normally excel at identifying. On the other hand, SelectKBest directly chooses the top features based on their statistical significance, which seems to have helped Random Forest zero in on the most impactful variables, improving its performance on the test set.

In the end, while PCA was useful for simplifying the feature space and worked well for the linear models, SelectKBest better aligned with Random Forest's strengths in capturing complex patterns. This highlights how different approaches to feature selection can have a significant effect on model performance, depending on the model's nature and the data's complexity.