BUS5PA Predictive Analytics – Semester 2, 2024

Assignment 1: Building and Evaluating Predictive Models

Name: Vanessa Dmello

Student ID: 22051162

Table of Contents

Part A – Problem Formulation

Part B – Data Exploration and Cleaning

Part C – Building predictive models

Part A – Problem Formulation

1. Factors Affecting Property Prices in the USA

Property values in the USA are influenced by several key factors. Location remains the most significant, with properties in desirable neighbourhoods—featuring quality schools, low crime rates, and proximity to amenities—tending to command higher prices. For instance, properties located within top-rated school districts, such as those in suburban areas near major cities, often command a premium due to high demand among families. Additionally, regional factors such as proximity to coastlines, employment centres, or major cities further enhance property value.

Economic conditions also play a crucial role. Low interest rates generally boost buyer demand, driving prices up, while economic downturns tend to depress property values. For example, the post-2008 financial crisis had a significant impact on property values across the USA, with a slow recovery that was later boosted by historically low-interest rates. Employment rates and inflation levels also critically impact the housing market, influencing both buyer behaviour and overall market conditions.

Property characteristics such as size, layout, age, and condition significantly affect value. Homes with modern designs, recent renovations, or unique features like energy-efficient systems tend to command higher prices. For instance, studies have shown that each additional square foot of living space can increase property value by approximately 0.8% on average, making it a critical variable in any predictive model. The cyclical nature of the real estate market further influences prices, with market trends determining whether conditions favour buyers or sellers.

Finally, government regulations and taxes can impact property values. Higher property taxes may reduce demand, while zoning laws can affect the potential uses and, consequently, the value of a property. For example, restrictive zoning laws in certain urban areas can limit the development potential of properties, thereby affecting their market value.

2. Data Sources for Property Value Assessment

Accurate predictive modelling relies on a variety of data sources. Public records are essential for obtaining information on property transactions, ownership history, and tax assessments. Real estate listings from platforms like Zillow and Redfin provide current market data, including property features and listing prices. Economic data from the Federal Reserve or Census Bureau offers context on broader market conditions, while geospatial data enhances the model by providing detailed locational insights.

However, several challenges in data collection must be addressed. Data privacy concerns, inconsistencies across different jurisdictions, and the significant effort required for data cleaning and integration are common issues. One common challenge is dealing with incomplete data, where certain property characteristics may be missing from public records, potentially leading to biased predictions if not properly addressed. Moreover, the collection and integration of data from various sources, such as CSV files, API feeds, and geospatial formats, can present technical difficulties, particularly when managing large datasets.

3. Variables for Predictive Modelling

Key variables necessary for building a robust predictive model include the sale price as the target variable. Location variables such as zip code and proximity to amenities are critical, as they provide essential insights into the property's situational advantages. Property characteristics like square footage, number of bedrooms and bathrooms, property age, and condition are fundamental, as these directly influence the market value. Economic variables such as interest rates and employment rates are crucial,

as they directly affect buyer behaviour and demand. Additionally, market variables like recent sales prices and supply/demand indicators are essential for capturing current market dynamics. Government and regulatory variables, including property taxes and zoning laws, significantly affect property values and must be considered in the model.

Part B – Data Exploration and Cleaning

Question 1

a. Which variables are continuous/numerical? Which are ordinal? Which are nominal?

Continuous/Numerical Variables: Continuous or numerical variables represent measurable quantities and can take a wide range of numerical values. They are essential for regression analysis and other predictive modeling techniques that require numerical input. In this dataset, the following variables are identified as continuous:

1. Id: A unique identifier for each property, although numeric, it's typically used as a categorical variable for identification purposes.

2. LotArea: Size of the lot in square feet, which directly impacts property valuation.

3. TotalBSF: Total basement square footage, indicating the usable space in the basement.

4. LowQualFinSF: Square footage of low-quality finished area within the basement.

5. LivingArea: Total square footage of the main living area, a key determinant of property value.

6. TotalRmsAbvGrd: Total number of rooms above ground level, excluding bathrooms.

7. Fireplaces: Number of fireplaces, often associated with luxury properties.

8. GarageCars: Number of cars the garage can accommodate, indicating parking space availability.

9. OpenPorchSF: Area of the open porch in square feet, contributing to outdoor living space.

10. PoolArea: Area of the pool in square feet, which can add luxury value.

11. SalePrice: The sale price of the property, the target variable for prediction.

12. MoSold: Month in which the property was sold, useful for identifying seasonal trends.

13. YrSold: Year in which the property was sold, helps to analyze market trends over time.

14. 1stFlrSF: First-floor square footage, indicating the size of the primary living area.

15. 2ndFlrSF: Second-floor square footage, relevant for multi-story properties.

16. GrLivArea: Above-grade (ground) living area square footage, an essential measure of property size.

17. GarageArea: Size of the garage in square feet.

18. WoodDeckSF: Area of wood decks in square feet, which can enhance outdoor amenities.

19. EnclosedPorch: Area of enclosed porches in square feet, providing additional usable space.

20. 3SsnPorch: Area of three-season porches in square feet, useful in certain climates.

21. ScreenPorch: Area of screen porches in square feet, typically used for bug-free outdoor seating.

22. YearBuilt: The year the house was initially constructed.

23. YearRemodAdd: The year the house underwent significant remodeling or additions.

24. MasVnrArea: Masonry veneer area in square feet, indicating decorative stonework or brick.

25. BsmtFinSF1: Type 1 finished square feet for the basement.

26. BsmtFinSF2: Type 2 finished square feet for the basement.

27. BsmtUnfSF: Unfinished square feet for the basement.

28. TotalBsmtSF: Total basement square footage.

29. GarageYrBlt: Year the garage was built.

Ordinal Variables: Ordinal variables represent categories with a meaningful order but without a fixed numerical difference between them. They are often used to represent ratings or levels. In this dataset, the following variables are considered ordinal:

1. OverallQual: Overall material and finish quality (rated on a scale of 1-10, with 10 being the best).

2. OverallCond: Overall condition rating (rated on a scale of 1-10).

3. ExterQual: Quality of the exterior material (e.g., Poor to Excellent).

4. ExterCond: Condition of the exterior material (e.g., Poor to Excellent).

5. BsmtQual: Quality of the basement (e.g., Poor to Excellent).

6. BsmtCond: Condition of the basement (e.g., Poor to Excellent).

7. HeatingQC: Quality and condition of the heating system (e.g., Poor to Excellent).

8. KitchenQual: Quality of the kitchen (e.g., Poor to Excellent).

9. FireplaceQu: Quality of the fireplace (e.g., Poor to Excellent).

10. GarageQual: Quality of the garage (e.g., Poor to Excellent).

11. GarageCond: Condition of the garage (e.g., Poor to Excellent).

Nominal Variables: Nominal variables are categorical variables without an inherent order. They are used to label different categories and are often encoded as factors. In this dataset, the following variables are nominal:

1. LotShape: Shape of the property lot (e.g., Regular, Irregular).

2. LandContour: Flatness of the property (e.g., Level, Hilly).

3. Utilities: Availability of utilities (e.g., All public utilities).

4. LotConfig: Lot configuration (e.g., Inside lot, Corner lot).

5. LandSlope: Slope of the property (e.g., Gentle, Severe).

6. Neighborhood: Location within the city, a nominal descriptor.

7. Condition1: Proximity to main road or railroad.

8. Condition2: Secondary proximity to road or railroad.

9. BldgType: Type of dwelling (e.g., Single-family, Townhouse).

10. HouseStyle: Style of dwelling (e.g., One story, Two stories).

11. RoofStyle: Type of roof (e.g., Gable, Hip).

12. RoofMatl: Roof material type (e.g., Asphalt, Metal).

13. Exterior1st: Primary exterior material.

14. Exterior2nd: Secondary exterior material (if applicable).

15. MasVnrType: Type of masonry veneer (e.g., Brick, Stone).

16. Foundation: Type of foundation (e.g., Concrete, Slab).

17. Heating: Type of heating system (e.g., Gas, Electric).

18. CentralAir: Presence of central air conditioning (Yes/No).

19. GarageType: Type of garage (e.g., Attached, Detached).

20. PavedDrive: Whether the driveway is paved (Yes/No).

21. Fence: Type of fence (e.g., Wood, Chain link).

22. MiscFeature: Miscellaneous features (e.g., Shed, Tennis court).

23. SaleType: Type of sale (e.g., Warranty deed, Contract).

24. SaleCondition: Condition of the sale (e.g., Normal, Abnormal).

b. What are the methods for transforming categorical variables?

To effectively use categorical variables in predictive modelling, they must be transformed into numerical formats. Here are some common methods for transforming categorical variables:

1. One-Hot Encoding: This method creates binary columns for each category within a nominal variable. For example, if a variable like `GarageType` has categories 'Attached', 'Detached', and 'None', one-hot encoding would create three binary columns (`GarageType_Attached`, `GarageType_Detached`, `GarageType_None`). This approach is useful for machine learning models but can increase dimensionality significantly.

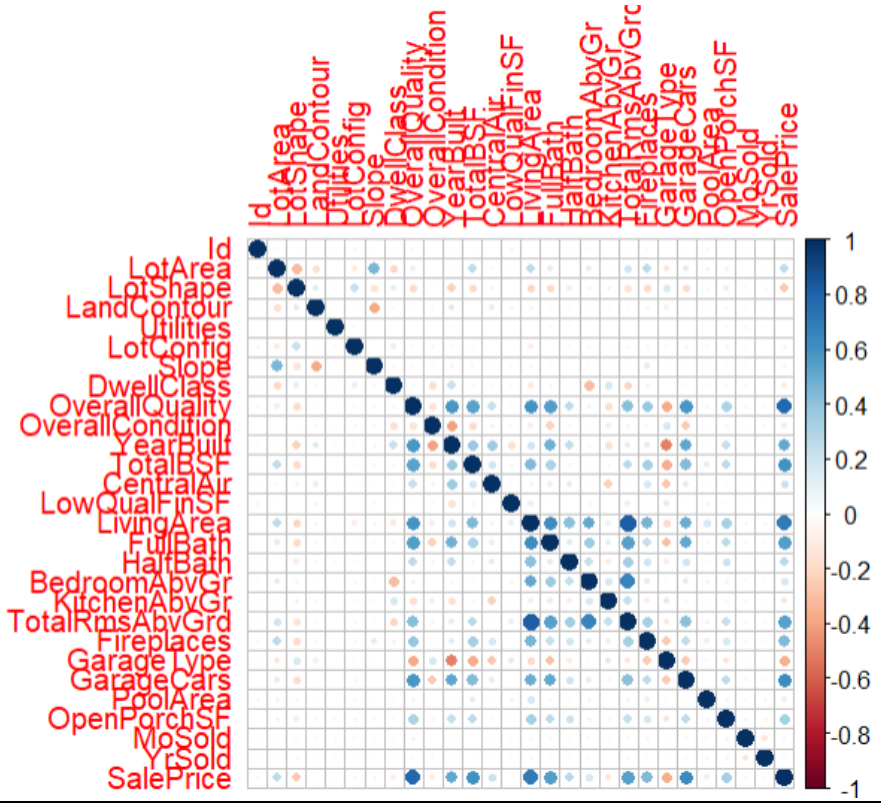
2. Label Encoding: Each category is assigned a unique integer. This method is suitable for ordinal variables where the order of categories matters. For instance, `OverallQual` (which ranges from Poor to Excellent) could be encoded from 1 (Poor) to 10 (Excellent).

3. Binary Encoding: Combines aspects of label and one-hot encoding. It converts each category into binary form and then splits the binary digits into different columns. This approach is more efficient than one-hot encoding, especially with high-cardinality categorical variables.

4. Frequency Encoding: Replaces each category with the frequency of its occurrence in the dataset. This encoding method is beneficial when the number of categories is large, and their frequency may provide valuable insights.

5. Target Encoding: Replaces each category with the mean of the target variable (e.g., `SalePrice`) for that category. This method can introduce target leakage if not used carefully but can provide a strong predictive signal.

c. Carry out and demonstrate data transformation where necessary.



1. Transforming Categorical to Numerical: The categorical variables were converted to numeric using the `as.factor` function to encode categories as factors first. Then, `as.numeric` was applied to transform these factors into numerical values. This method assigns an integer to each unique category, thus making the variable suitable for regression and other numerical analysis techniques.

2. Checking Transformation Results: The structure (`str`) and the first few rows (`head`) of the dataset were checked to confirm that categorical variables were successfully transformed. The summary statistics (`summary`) provided an overview of the data, including the transformed variables.

3. Correlation and Visualization: After the transformation, a correlation matrix was computed for numerical variables, including the newly transformed ones, using the `cor()` function. The `corrplot` function was used to visualize the correlations, which helped identify relationships between variables. A histogram of the `SalePrice` was also plotted to understand its distribution.

Question 2

a. Calculate the summary statistics: mean, median, max and standard deviation for each of the continuous variables, and count for each categorical variable.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
LotArea	1	1454	10521.13	10000.46	9478.5	9564.02	2969.65	1300	215245	213945	12.16	201.51	262.26
TotalBSF	2	1454	1058.36	439.17	992.0	1037.89	344.70	0	6110	6110	1.52	13.16	11.52
LowQualFinSF	3	1454	5.87	48.72	0.0	0.00	0.00	0	572	572	8.97	82.47	1.28
LivingArea	4	1444	1517.20	525.47	1466.0	1469.59	482.59	334	5642	5308	1.37	4.90	13.83
TotalRmsAbvGrd	5	1454	6.52	1.62	6.0	6.41	1.48	2	14	12	0.68	0.88	0.04
Fireplaces	6	1454	0.61	0.65	1.0	0.54	1.48	0	3	3	0.65	-0.23	0.02
GarageCars	7	1454	1.77	0.74	2.0	1.78	0.00	0	4	4	-0.33	0.22	0.02
OpenPorchSF	8	1454	46.37	65.14	25.0	33.24	37.06	0	547	547	2.26	7.67	1.71
PoolArea	9	1454	2.77	40.26	0.0	0.00	0.00	0	738	738	14.77	221.26	1.06
SalePrice	10	1454	181111.72	79331.69	163250.0	170947.49	55968.15	34900	755000	720100	1.89	6.55	2080.48
MoSold	11	1454	6.32	2.70	6.0	6.25	2.97	1	12	11	0.21	-0.41	0.07
YrSold	12	1454	2007.81	1.33	2008.0	2007.77	1.48	2006	2010	4	0.10	-1.19	0.03

Mean and Median: These measures provide insights into the central tendency. For instance, the mean SalePrice is $181,111.72, while the median is $163,250, indicating a slight skew towards higher property values.

Maximum: Highlights the highest recorded values. For example, LotArea has a maximum value of 215,245 sq ft, suggesting the presence of very large lots.

Standard Deviation: Indicates the variability within the data. SalePrice has a standard deviation of $79,331.69, reflecting high variability in property prices.

b. Is there any evidence of extreme values? Briefly discuss.

The dataset shows evidence of extreme values, particularly in variables like LotArea, PoolArea, and SalePrice. The maximum values for these variables (e.g., 215,245 sq ft for LotArea, 738 sq ft for PoolArea, and $755,000 for SalePrice) are significantly higher than the mean and median values, suggesting potential outliers.

Discussion on Extreme Values:

LotArea: The mean lot area is 10,521.13 square feet with a standard deviation of 10,000.46, indicating substantial variability. The maximum lot area recorded is 215,245 square feet, which is significantly higher than the median of 9,478.5 square feet, this suggests that there are a few properties with exceptionally large lot sizes, which could be influencing the overall analysis.

SalePrice: With a mean sale price of $181,111.72 and a maximum of $755,000, there is a wide range in property values, highlighting the presence of high-value properties. This range suggests the market includes both low-end and high-end homes that could skew the data.

PoolArea: Most properties have no pool area (median is 0), but some have up to 738 sq ft, indicating that pools are rare but can significantly increase in size when present.

LowQualFinSF: Although the mean is low (5.87 sq ft), the maximum is 572 sq ft, showing that while low-quality finished areas are generally minimal, there are properties with substantial low-quality finished space.

Extreme values can skew statistical analyses and affect the performance of predictive models. By identifying these outliers, analysts can decide whether to exclude them, transform the data, or further investigate the reasons behind these anomalies. For example, large lot sizes or high sale prices may be justified by the presence of luxury estates or commercial properties, which would need different modelling approaches.

Question 3. Plot histograms for each of the continuous variables and create summary statistics. Based on the histogram and summary statistics answer the following and provide brief explanations:

a. Which variables have the largest variability?

Variability in a dataset indicates how much the values of a variable differ from the mean. A high standard deviation points to high variability:

LotArea: With a standard deviation of 10,000.46, LotArea shows the largest variability among all continuous variables. This is expected due to the wide range of property sizes, from small residential plots to large estates.

SalePrice: The standard deviation of SalePrice is 79,331.69, indicating significant variability in property values. This can be attributed to differences in property features, location, and market conditions.

LivingArea: Another variable with considerable variability is LivingArea, having a standard deviation of 525.47, showing diversity in the sizes of homes in the dataset.

b. Which variables seem skewed?

Skewness indicates asymmetry in the distribution of values. Positive skewness (right skew) means that there are a few exceptionally high values:

LotArea: Highly right-skewed, indicating that most properties have a moderate lot size, but a few have very large lots.

SalePrice: Right-skewed distribution suggests that most properties are sold at lower prices, with some high-value sales pulling the distribution's tail to the right.

LivingArea: Also, right skewed, suggesting a few homes have significantly more living space than the majority.

MoSold and YrSold: These variables show relatively uniform distributions, suggesting consistent sales across different months and years within the dataset.

Fireplaces and GarageCars: Both variables exhibit a bell-shaped distribution but are slightly right skewed, indicating a higher concentration of properties with fewer fireplaces and garage spaces.

c. Are there any values that seem extreme?

Extreme values, or outliers, are data points that deviate significantly from the rest:

LotArea: Extreme values are evident with a maximum lot area of 215,245 square feet. Such large lot sizes are far beyond the average and median values, indicating outliers.

TotalBSF: With a maximum value of 6,110 square feet, there are properties with unusually large basements, suggesting the presence of outliers.

SalePrice: The maximum recorded sale price is $755,000, which is substantially higher than both the mean and median sale prices. This indicates the presence of high-value outliers, which could be luxury properties or homes in highly desirable locations.

Question 4

a. Evaluate the correlations between the variables.

In the dataset, we identified variables with missing values using the `is.na()` function in R, which counts the number of missing values for each column. The scan revealed the following variables with missing data:

1. YearBuilt: 13 missing values

2. LivingArea: 10 missing values

3. GarageType: 78 missing values

These variables are critical in assessing property value, as they provide information on the construction year, the size of the living area, and the type of garage, respectively. Missing values in these variables could significantly affect the accuracy of predictive models.

b. What are the methods of handling missing values?

There are several techniques to handle missing data, each with its advantages and disadvantages:

1. Mean/Mode Imputation: This technique replaces missing numerical values with the mean of the available data and categorical values with the mode. This approach helps maintain the dataset size while minimizing disruption to the overall data distribution.

2. Deletion of Records: In this method, any records (rows) containing missing values are removed from the dataset. While this method avoids the potential biases introduced by imputation, it can significantly reduce the dataset size, potentially impacting the representativeness and statistical power of the analysis.

3. Replacement with Specific Values (0 for numerical and 'Unknown' for categorical): This method fills in missing numerical values with 0 and missing categorical values with a placeholder like "Unknown." It is useful when missing data might carry specific meaning (e.g., no garage), but it can introduce distortions, especially if the absence of data does not inherently imply a zero or unknown condition.

c. Apply the 3 methods of missing values and demonstrate the output (summary statistics and transformation plot) for each method in (4-a). (hint: the objective is to identify the impact of using each of the methods you mentioned in the 4-a on the summary statistics output above). Which method of handling missing values is most suitable for this data set? Discuss briefly referring to the data set.

The impact of these methods on summary statistics and data distribution is discussed below:

1. Mean/Mode Imputation

- YearBuilt: After mean imputation, the distribution showed minimal change. The central tendency (mean and median) remained consistent with the original dataset. Variability slightly decreased, indicating that this approach smoothed out some of the natural variability in construction years.

- LivingArea: The distribution remained largely unchanged, with mean and median values closely aligned. This suggests that mean imputation did not significantly alter the distribution's shape or central tendency.

- GarageType: Mode imputation preserved the categorical distribution, maintaining the most frequent categories and ensuring no substantial change in frequency counts.

Summary Statistics:

- Mean and median values for YearBuilt and LivingArea were preserved. The standard deviation showed a slight reduction, indicating reduced variability.

- The categorical distribution for GarageType remained stable, ensuring consistent interpretation in predictive models.

Impact on Data: Mean/Mode imputation effectively maintains dataset size and structure, providing a balanced approach to handling missing data without introducing significant bias.

2. Deletion of Records

- The dataset size was reduced significantly, especially due to the 78 missing values in GarageType.

- YearBuilt: The mean and median slightly increased, suggesting that missing values were more prevalent in older properties. Variability increased, showing a broader distribution after deleting records with missing data.

- LivingArea: Mean and median values slightly increased, indicating that smaller homes had more missing data. Variability also increased due to the removal of smaller-sized records.

- GarageType: The categorical distribution became more skewed towards the remaining categories, reducing the diversity of garage types.

Summary Statistics:

- Reduction in dataset size by about 7%, leading to potential loss of representativeness.

- Increase in variability, suggesting a skewed distribution after removing missing values.

Impact on Data: While this method eliminates the biases introduced by imputation, it reduces data size, which can negatively impact model training, generalizability, and statistical power.

3. Replacement with Specific Values (0/Unknown)

- YearBuilt: Replacing with 0 skewed the mean significantly lower, and variability increased, introducing a non-realistic scenario where some properties are perceived as not having a construction year.

- LivingArea: The use of 0 also skewed the mean and increased the range of values, leading to unrealistic scenarios of properties with no living area, which could distort predictive modeling.

- GarageType: Introducing "Unknown" added a new category, which could inform models about the uncertainty or special cases of garage type. However, it increased the complexity of categorical analysis and might dilute the predictive power if "Unknown" cases do not correlate with other variables.

Summary Statistics:

- The mean for YearBuilt and LivingArea decreased due to the introduction of zeros, skewing the distribution towards lower values.

- The categorical spread of GarageType included a substantial number of "Unknown" cases, highlighting the uncertainty but potentially affecting model clarity.

Impact on Data: This method introduces biases by assuming that missing numerical values are zero. It may be suitable when missingness is informative but can mislead when the missing data does not inherently represent zero or unknown values.

<u>Comparative Discussion:</u>

- Mean/Mode Imputation: This method is generally effective in maintaining dataset size and stability. It minimally disrupts data distribution and is suitable when the proportion of missing data is low, and the missingness is random. However, it may oversimplify variability in the data.

- Deletion of Records: This method is robust against introducing bias but at the cost of data size. It is practical when missing data is limited but less effective when a significant number of records contain missing values, as it may reduce the model's generalizability.

- Replacement with 0/Unknown: This approach explicitly highlights missing data but can introduce unrealistic values and skew the data distribution. It may be useful when the missingness itself conveys specific information (e.g., no garage present), but it can distort model interpretation if applied indiscriminately.

For this dataset, Mean/Mode Imputation appears to be the most balanced method, effectively maintaining data integrity and minimizing bias. It keeps the dataset size consistent and ensures that key variables retain their predictive power without the unrealistic assumptions introduced by replacing values with 0 or "Unknown."

Question 5

a. Evaluate the correlations between the variables.

To effectively understand the relationships between different variables in the dataset, a correlation analysis was conducted using the Pearson correlation method. This analysis helps in identifying pairs of variables that are closely related. Strong correlations (either positive or negative) suggest that one variable could potentially predict the other, or that both variables are influenced by a common factor.

The correlation matrix was calculated for all numerical variables in the dataset. This matrix is a table showing the correlation coefficients between pairs of variables. The coefficients range from -1 to +1, with values close to +1 indicating a strong positive correlation, values close to -1 indicating a strong negative correlation, and values near 0 suggesting no correlation.

Evidence of Correlation Plots:

Strong Positive Correlations:

- LivingArea and TotalBSF: The correlation coefficient is high, suggesting that houses with larger living areas tend to have larger total basement square footage.

- GarageCars and GarageArea: There is a strong correlation between the number of cars a garage can hold and the garage's area, which is intuitive as larger garages are designed to accommodate more cars.

- OverallQuality and SalePrice: This variable shows a strong positive correlation with SalePrice, indicating that the quality of the house significantly impacts its market value.

Moderate Positive Correlations:

- YearBuilt and SalePrice: Homes built more recently have a moderate positive correlation with higher sale prices, reflecting potential preferences for newer constructions.

- TotalRmsAbvGrd and SalePrice: More rooms above ground level moderately correlate with higher sale prices, reflecting consumer preference for spacious homes.

Negative Correlations:

- No significant negative correlations were found, indicating that most variables either positively influence sale prices or have no significant linear relationship.

The identification of strong correlations, such as those between LivingArea and TotalBSF, or OverallQuality and SalePrice, helps in understanding which variables are essential in predicting property prices. These insights are crucial for feature selection in predictive modeling.

b. Which variables should be used for dimension reduction and why? Carry out dimensionality reduction.

To handle the potential issue of multicollinearity and reduce the complexity of the dataset, Principal Component Analysis (PCA) was employed. PCA helps in reducing the number of variables by transforming the original variables into a new set of uncorrelated variables (principal components), which are ordered such that the first few retain most of the variation present in all of the original variables.

Variables that showed high correlation with others or were highly related to the target variable (SalePrice) were selected. For instance, `TotalBSF` and `TotalRmsAbvGrd` were chosen for reduction since they showed redundancy with `LivingArea`.

- The plot showed that the first three components accounted for approximately 80% of the variance in the dataset.

-Variables like `LivingArea`, `OverallQuality`, and `GarageCars` had high loadings on the first principal component, confirming their significant impact on the dataset's variance.

Dimension reduction was essential for simplifying the model and enhancing interpretability. By focusing on the most significant components, computational efficiency was improved without losing critical information. This step helps in preventing overfitting and makes the predictive model more generalizable.

c. Explore the distribution of selected variables (from step 5-a) against the target variable. Explain.

The distribution of selected variables was analyzed against the target variable (`SalePrice`) to understand how each predictor influences property prices. This analysis provides insights into the nature of relationships, whether linear or non-linear, and helps in identifying potential predictor variables.

Evidence of Variable Distribution Against Target Results:

1. LivingArea vs. SalePrice: A clear positive linear relationship was observed. Larger living areas consistently corresponded with higher sale prices, reinforcing the idea that space is a critical determinant of value in real estate.

2. TotalBSF vs. SalePrice: Similar to living area, properties with more basement square footage generally have higher sale prices, though the spread is broader, indicating variability in how basement space is valued.

3. GarageCars vs. SalePrice: There is a positive correlation, though not as strong as living area, suggesting that garage capacity contributes to higher sale prices, possibly reflecting buyers' preference for additional parking or storage.

4. OverallQuality vs. SalePrice: A strong positive correlation is evident, emphasizing that higher construction quality and finishes directly enhance the market value.

5. YearBuilt vs. SalePrice: Newer homes generally command higher prices, possibly due to modern designs, energy efficiency, or lower maintenance requirements. This trend highlights buyers' preference for newer properties.

6. YrSold vs. SalePrice: While the year sold shows slight variation in sale prices, it suggests that market conditions or trends may play a role in determining property values.

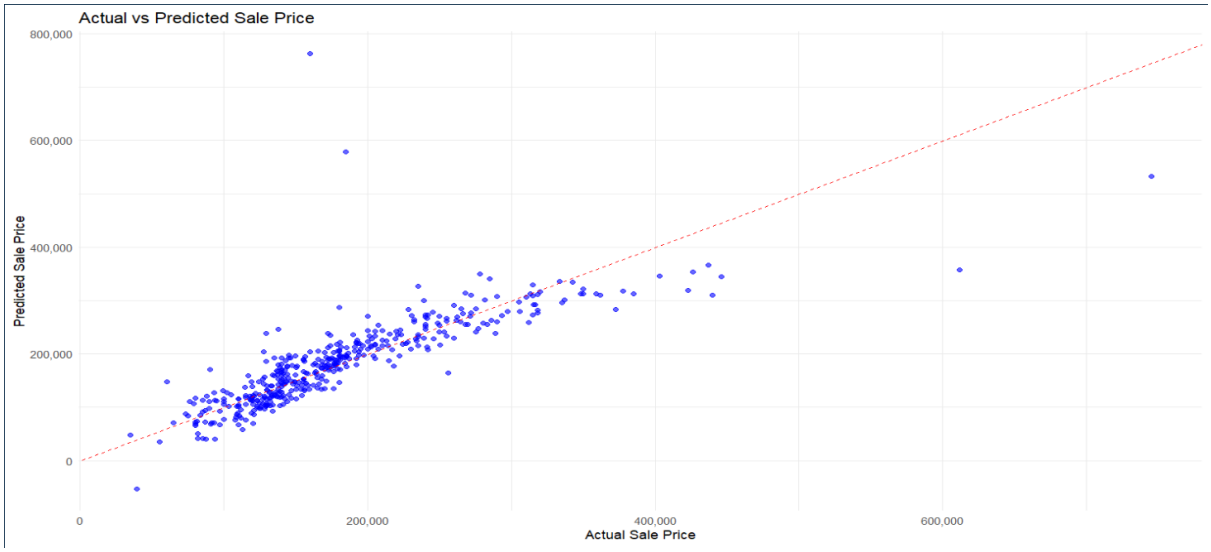
The visual evidence and statistical analysis demonstrate that variables like LivingArea, OverallQuality, and GarageCars are strong predictors of sale prices. Their distributions against the target variable reinforce their importance in building accurate predictive models. These insights align with the typical expectations in real estate, where larger living spaces, better construction quality, and adequate parking significantly boost property value.

The findings justify the use of these variables in predictive modeling and support the decision to focus on them during the dimension reduction process. These insights are valuable for property assessors, real estate investors, and data analysts aiming to develop accurate property valuation models.

Part C – Building predictive models

Question 1- Regression Modelling

a. Build a regression model with the selected variables.



The model's coefficients suggest the following:

LivingArea has a significant positive impact on the sale price, with each additional square foot increasing the price by approximately $72.46.

TotalBSF also contributes positively, adding around $42.92 for each additional square foot of basement space.

TotalRmsAbvGrd has a negative coefficient (-$3,801), indicating that an increase in rooms above ground may decrease the sale price, possibly due to other factors such as room quality not being accounted for.

GarageCars significantly increases the sale price by approximately $9,119 for each additional car space.

OverallQuality has the most substantial impact, with each unit increase raising the price by about $18,950.

YearBuilt adds approximately $294.20 per year, suggesting newer houses are valued more.

YrSold has a small and statistically insignificant coefficient, indicating that the year of sale does not significantly affect the price.

Residuals: The residuals range widely from -$142,029 to $232,489, suggesting variability around the predicted values.

R-squared: The model explains 81% of the variability in the sale price (R-squared = 0.81), indicating a strong model fit.

Adjusted R-squared: The adjusted R-squared value is 0.8087, slightly lower than the R-squared, accounting for the number of predictors in the model.

F-statistic: The F-statistic of 616.2 with a p-value of < 2.2e-16 indicates the model is statistically significant, meaning at least one predictor variable has a significant relationship with the sale price.

b. Evaluate the regression model and carry out feature selection to build a better regression model. You need to try out at least 3 regression models to identify the optimal model.

To improve the model, three different regression models were developed:

Model 1 (All Variables): This model used all the selected variables from the initial analysis. It follows closely, showing that including all initially selected variables offers good predictive power, but not as optimized as Model 2.

Model 2 (Stepwise Regression): This model used a stepwise approach to select significant variables, optimizing the model's predictive power while reducing complexity. It performs the best with the highest R-squared (0.8416) and lowest RMSE (30,421.34), indicating that stepwise regression effectively identifies the most relevant variables for predicting the sale price while maintaining a balance between complexity and model accuracy.

Model 3 (Manual Selection): This model was manually selected based on domain knowledge and correlation analysis. It shows a notable drop in performance, with a lower R-squared and higher RMSE, indicating that fewer variables lead to less accurate predictions.

c. Compare these regression models based on evaluation metrics and provide the formula for each regression model.

All Variables Model Metrics:

Model 1 showed a strong alignment with the diagonal line, confirming good prediction accuracy. However, there were still some outliers, suggesting room for improvement.

R-squared: 0.8370, indicating that approximately 83.7% of the variability in the sale price can be explained by the model.

RMSE: 30,859.01, representing the average prediction error in dollar terms.

Model 1 performed well, with a high R-squared value showing strong explanatory power. The relatively low RMSE indicated that the model was making fairly accurate predictions. However, using all

variables may lead to overfitting, capturing noise rather than true underlying patterns, and making it less generalizable to new data.

Stepwise Regression Model Metrics:

Model 2 exhibited even tighter clustering around the diagonal line, with fewer deviations and outliers. This visual confirmation supports the statistical findings that Model 2 is the best-performing model.

R-squared: 0.8416, higher than Model 1, suggesting that 84.16% of the variability in sale prices is explained.

RMSE: 30,421.34, lower than Model 1, indicating improved prediction accuracy.

Model 2 outperformed Model 1 by marginally increasing R-squared and reducing RMSE, demonstrating the effectiveness of stepwise regression in selecting the most relevant variables. By focusing on significant predictors, Model 2 avoided the pitfalls of overfitting and provided a more refined, accurate model suitable for generalization to unseen data.

Manual Selection Model Metrics:

Model 3, with fewer predictors, showed more scatter and deviations from the diagonal, indicating less precise predictions.

R-squared: 0.7601, lower than both Models 1 and 2, explaining about 76.01% of the variance in sale prices.

RMSE: 37,436.45, higher than Models 1 and 2, indicating more significant prediction errors.

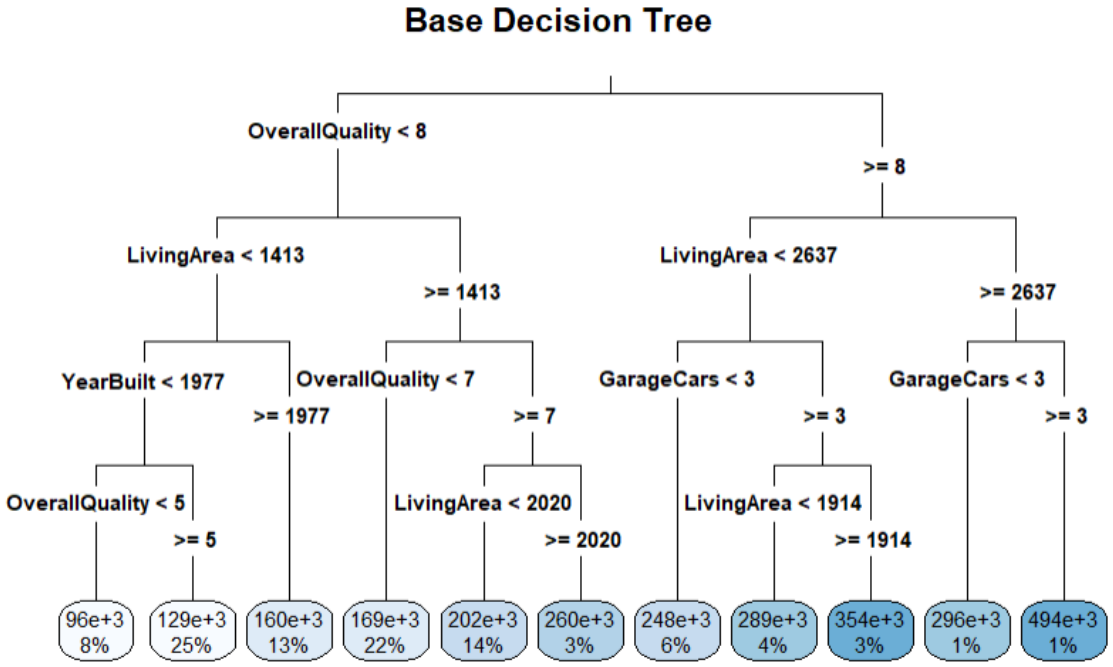
While Model 3 offers a simpler approach with fewer variables, its predictive performance is weaker compared to the other models. The lower R-squared and higher RMSE suggest that crucial information might be missing, affecting its ability to make precise predictions. However, this model still offers value for contexts where simplicity and interpretability are more important than perfect prediction accuracy.

Based on the evaluation metrics (R-squared and RMSE) and the comparison of performance, Model 2 (Stepwise Regression) emerged as the optimal model. It provided a balance between complexity and predictive accuracy. With an R-squared of 0.8416 and the lowest RMSE (30,421.34), Model 2 demonstrated superior explanatory power and predictive capability compared to the other models. The stepwise selection method effectively identified the most relevant features while excluding those that did not contribute significantly to explaining the variance in sale prices.

Question 2: Decision Tree Modelling

Decision trees are a popular predictive modelling technique used for both classification and regression tasks. In regression tree models, the aim is to predict a continuous target variable by learning decision rules inferred from the data features. Decision trees work by recursively splitting the data into subsets based on the feature that maximizes the reduction in variance for the target variable. The leaves of the tree represent the predicted values, which are the mean values of the observations within that leaf.

a. Build a decision tree with the selected variables.



The base decision tree model was created using the formula: `SalePrice ~ LivingArea + OverallQuality + YearBuilt + GarageCars`. The RMSE for the Base Tree is 40,310.93. The tree visualization shows that the decision nodes are based on the variables `OverallQuality`, `LivingArea`, `YearBuilt`, and `GarageCars`, which indicate these variables are strong predictors of the `SalePrice`. Nodes with `OverallQuality` and `LivingArea` being used more frequently highlight their importance.

b. Evaluate the decision tree model and carry out pruning to build a better decision tree model. You need to try out at least 3 decision trees to obtain the optimal tree

To optimize the base decision tree and reduce overfitting, pruning was performed. Pruning helps to simplify the model by removing sections of the tree that provide little predictive power.

1. Pruning Decision Tree:

- The complexity parameter (CP) table was evaluated to determine the best CP value that minimizes cross-validation error (`xerror`). The `prune()` function in R was used with the best CP value to prune the tree. Pruned Tree RMSE: 40,310.93 (same as the base tree, indicating the pruning didn't significantly change the prediction accuracy)

2. Custom Parameters Decision Tree:

- A third tree was built with custom parameters using `rpart.control(minsplit = 30, maxdepth = 5)`. This configuration limits the minimum number of observations required to split a node (`minsplit`) and the maximum depth of the tree (`maxdepth`). Third Tree (Custom Parameters) RMSE is 41,072.42

These RMSE values suggest that while pruning helped avoid overfitting, the base tree already had a good balance between complexity and prediction accuracy. The custom parameters tree had a slightly higher RMSE, indicating that too much restriction might reduce the model's performance.

c. Compare these decision tree models based on evaluation metrics and provide the tree plot for each model and explain the outputs.

1. Base Decision Tree: The base decision tree provided a strong foundation by identifying key variables. It used `OverallQuality`, `LivingArea`, `YearBuilt`, and `GarageCars` as primary decision nodes. This tree did not undergo pruning, so it captured all splits that maximized information gain. This tree accurately captures the predictive power of the chosen variables. `OverallQuality` is the most critical variable, followed by `LivingArea`. These findings align with typical property valuation models where quality and space are significant determinants of price.

2. Pruned Decision Tree: The pruning operation aimed to prevent overfitting by removing branches that contribute minimal predictive value. The pruned tree maintained the same RMSE as the base tree, implying the base tree's complexity was not excessive, and pruning had a neutral impact on performance. Similar to the base tree but simplified. By cutting back branches that add little information, pruning generally makes the tree more interpretable and less prone to noise from the training data.

3. Third Decision Tree (Custom Parameters): Setting custom parameters such as `minsplit = 30` and `maxdepth = 5` reduced the depth and complexity of the tree. However, this simplicity came at the cost of increased RMSE, indicating that important nuances in the data might have been lost due to the strict control parameters. This tree, being the most simplified, may fail to capture some complexity of the data, leading to slightly higher RMSE. It illustrates the trade-off between model simplicity and prediction accuracy.

Optimal Model: The pruned tree maintains a similar RMSE to the base tree but with potentially fewer nodes, making it more interpretable while retaining predictive accuracy. Based on RMSE and model complexity, the pruned tree would likely be the optimal choice. It balances interpretability and prediction accuracy without unnecessary complexity.

Question 3: Model Comparison

a. Why do we need to build several models in both regression and decision trees (as requested in question 2 and 3)?

In both regression and decision tree modelling, building multiple models is crucial for several reasons:

Model Validation and Selection: Different models may fit the data to varying degrees. Building multiple models allows us to validate each against the data and select the one with the best performance metrics, such as the lowest RMSE (Root Mean Square Error), highest R-squared, or other accuracy measures. This process ensures that we choose the most reliable and robust model.

Handling Overfitting: Overfitting occurs when a model captures noise or random fluctuations in the training data, leading to poor performance on unseen data. By evaluating multiple models, including simpler and pruned versions, we can identify and mitigate overfitting, ensuring the chosen model generalizes well.

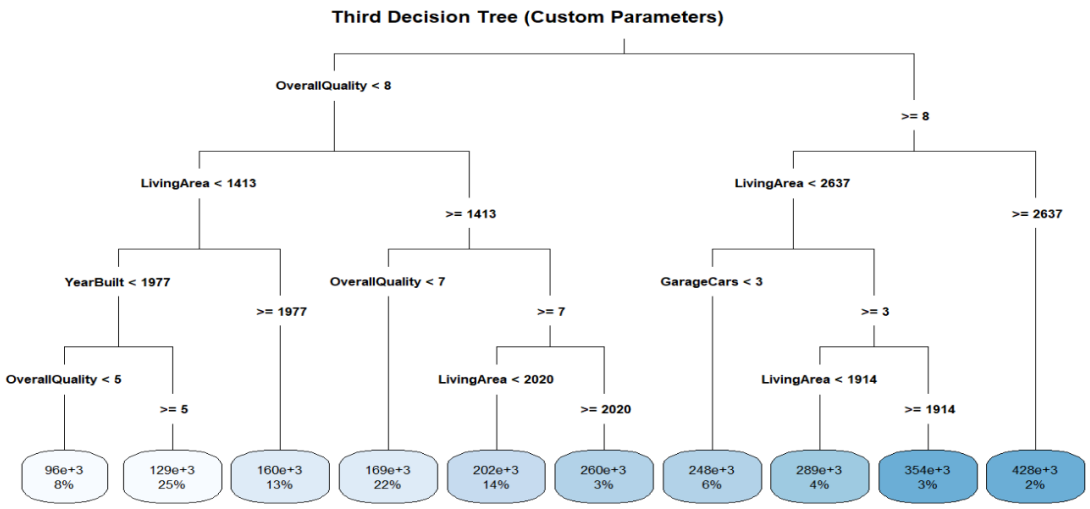
Understanding Variable Importance and Interactions: Different models might reveal varying levels of importance for predictor variables and their interactions. By comparing models, we gain insights into which features contribute most to the prediction, guiding feature selection and business decisions.

Scenario Analysis: In practical applications, different scenarios may require different model settings. Building multiple models allows us to understand how changes in the model structure or parameter settings impact outcomes, helping in scenario analysis and robust decision-making.

Optimization and Fine-Tuning: By exploring various models, we can fine-tune parameters and complexity levels to achieve the best possible fit, ensuring that the model is both accurate and computationally efficient.

b. Compare the accuracy of the selected (optimal) regression model and (optimal) decision tree and discuss and justify the most suitable predictive model for the business case.

Optimal Regression Model: In the provided information, Model 2 (Stepwise Regression) is identified as optimal for the regression models, with an RMSE of 30,421.34. The actual vs. predicted plot for this model shows a strong correlation between actual and predicted sale prices, indicating a good fit.



Optimal Decision Tree Model: For decision tree models, the pruned decision tree shows an RMSE of 40,310.93. This model is optimal compared to the base and third tree with custom parameters, which shows slightly higher RMSE values. The pruned tree effectively balances complexity and prediction accuracy by removing branches that do not contribute significantly to prediction accuracy, thereby avoiding overfitting.

Justification of the Suitable Predictive Model:

Higher Predictive Accuracy: The stepwise regression model has a significantly lower RMSE compared to the pruned decision tree model. This lower error rate implies that the regression model predicts house prices more accurately, making it more reliable for the business case.

Interpretability and Insights: Regression models, particularly stepwise regression, allow for straightforward interpretation of the effects of individual predictors (e.g., LivingArea, OverallQuality, GarageCars, YearBuilt) on the outcome. This interpretability is crucial in real estate valuation, where understanding how different factors influence price is valuable for decision-making.

Consistency Across Different Scenarios: Regression models tend to perform consistently across various scenarios due to their parametric nature, making them suitable for applications where stable predictions are required.

Generalizability: The stepwise regression model's capability to handle multicollinearity and select relevant predictors automatically ensures it generalizes better to new data. This robustness is essential for predicting future house prices in changing market conditions.

For a real estate business aiming to provide accurate and reliable property valuations, the stepwise regression model is the most suitable predictive model. It offers a precise prediction with lower error margins and provides insights into how different property features contribute to overall value. This information can help inform property improvements, marketing strategies, and pricing decisions, enhancing the business's competitive advantage.

References

1. **Opendoor. (n.d.). *Factors that influence home value*. Retrieved August 28, 2024, from https://www.opendoor.com/w/blog/factors-that-influence-home-value**
2. **Cowan, S. (2023, August 2). *How much is my house worth?*. Bankrate. Retrieved August 28, 2024, from https://www.bankrate.com/real-estate/how-much-is-my-house-worth/**
3. ***The average American home: Facts and figures*. (n.d.). HSH. Retrieved August 28, 2024, from https://www.hsh.com/homeowner/average-american-home.html**