The Housing Prices Prediction Analysis Report

Assignment 4: Creating Reports and Dashboards for Predictive and Prescriptive Analysis

Dayanara Torres Macalino (NF1001047),

Miko L. Tan (NF1008647),

Dev D. Rabadia (NF1005560),

Stephy Marvin Christi (NF1003839)

Master of Data Analytics, University of Niagara Falls Canada

CPSC-510-5: Winter 2025 Data Warehousing/Visualization

Professor Mehdi (Matt) Mostofi

March 23, 2025

Overview of Workflow and Methodology

Data Cleaning and Preparation

The Ames Housing dataset was loaded using pandas in Jupyter Lab from a CSV file (Housing Prices Dataset.csv), with "NA" specified as the missing value indicator. The dataset was inspected by displaying the first five rows and summary statistics to understand its structure and identify missing values. It contains 2,919 properties with 52 columns, including features like SalePrice, LotArea, OverallQual, YearBuilt, and PavedDrive.

SalePrice had 49.98% missing values (1,459 out of 2,919 rows). A Linear Regression model was trained on rows with known SalePrice using features (LotArea, OverallQual, GrLivArea, YearBuilt, TotalBsmtSF) to predict missing values. The model achieved an R^2 of 0.83 on training data, and predictions were applied to the missing rows. Other Features such as GarageYrBlt values of 0 were replaced with YearBuilt, and MasVnrArea values of 0 were replaced with the median to correct for unrealistic zeros. No other columns had missing values.

Outliers in SalePrice, LotArea, and GrLivArea were capped using the Interquartile Range (IQR) method (clipping values below Q1 - 1.5*IQR and above Q3 + 1.5*IQR) to reduce the impact of extreme values on analysis and modeling.

After cleaning, the dataset was verified to have no missing values, and final summary statistics were computed to confirm the cleaned data's integrity. The cleaned dataset was saved as *housing_prices_cleaned.csv.*

Predictive Modeling

Two models—Linear Regression and Decision Tree (with max_depth=4)—were trained to predict SalePrice using features like OverallQual, YearBuilt, GrLivArea, TotalBsmtSF, BsmtFinSF1, and BsmtFinSF2. Models were developed in Jupyter Lab using scikit-learn. The dataset was split into training and testing sets (though the split ratio isn't specified in the OCR). Performance metrics (MSE, RMSE, R^2) were calculated, and predictions were exported to housing_prices_with_predictions.csv for visualization in Power BI. The predictions were imported into Power BI to create the Predictive Analysis page, enabling interactive visualization of model performance.

Dashboard Creation in Power BI

The Overview Page focused on property characteristics using features like PavedDrive, BedroomAbvGr, OverallCond, YearBuilt, and YearRemodAdd. Visualizations included bar charts, pie charts, and line charts to show distributions and trends.

The Predictive Analysis Page visualized model performance with scatter plots (actual vs. predicted SalePrice) and line charts (average SalePrice by YearBuilt). Slicers for YearBuilt were added for interactivity.

The Summary Statistics Page pesented statistical summaries and visualizations (box plots, scatter plots) for numerical features (SalePrice, GrLivArea, BsmtFinSF1, BsmtFinSF2, OverallQual, YearBuilt, TotalBsmtSF) and additional features (GarageCars, GarageArea, Fence, LotArea).

Analysis and Insight Generation

Descriptive statistics, distributions, and relationships were analyzed to derive insights about the housing market, property conditions, and predictive model performance. Visualizations were used to identify trends, outliers, and correlations, such as the relationship between SalePrice and features like GarageCars or LotArea.

Summary Report

Overview of the Datasets and Their Attributes

The Ames Housing dataset includes 2,919 properties across 52 columns, capturing a wide range of property characteristics. After cleaning, the dataset (housing_prices_cleaned.csv) contains no missing values, with outliers capped to improve analysis accuracy. Key attributes include SalePrice, the target variable, which ranges from $12,141 to $242,444 (post-capping), with a mean of $176,311.23 and a standard deviation of $64,600.97, indicating significant price variability. LotArea, representing lot size in square feet, averages 9,576.51 sq ft (post-capping), ranging from 1,340 to 10,708 sq ft, reflecting diverse property sizes. OverallQual, a quality rating on a 1–10 scale, averages 6.09, with most homes rated 5–7, suggesting average to above-average quality. YearBuilt spans from 1872 to 2010, with a mean of 1971.31, showing a mix of older and newer homes. GrLivArea, the above-ground living area, averages 1,490.62 sq ft, with a pre-capping maximum of 2,869.75 sq ft, indicating varied home sizes. BsmtFinSF2, the Type 2 finished basement area, is mostly 0 (median and 75th percentile at 0), with a maximum of 1,526 sq ft, highlighting its rarity. Additional features like TotalBsmtSF, GarageCars, GarageArea, and Fence were analyzed for their influence on SalePrice, showing diverse distributions and relationships.

Visualizations Created

Visualizations were created to explore the dataset comprehensively. On the Overview Page, a pie chart of Paved Drive Status reveals that 92.6% of properties lack a paved driveway, while 7.4% have one. A bar chart of Bedroom Distribution highlights that most properties have 2–4 bedrooms, with 1,616 having 3 bedrooms, and a Floor Distribution bar chart shows 1,704 single-story and 1,214 two-story homes. A Property Condition bar chart classifies properties by OverallCond, with 19.7% rated "Very Good," 74.5% "Good," and 5.76% "Bad." A line chart of Construction Trends tracks YearBuilt from 1900 to 2020, noting a peak in 2006 with 31 properties. The Predictive Analysis Page includes scatter plots of actual vs. predicted prices, where Linear Regression shows strong alignment along the diagonal with some outliers, while the Decision Tree exhibits more scatter, especially for higher SalePrice values. Line charts of Average SalePrice by YearBuilt indicate an upward trend for Linear Regression with peaks in the 1920s and post-2000, while the Decision Tree shows erratic trends, including sharp drops around 1950. The Summary Statistics Page features box plots for SalePrice, GrLivArea, LotArea, and BsmtFinSF2, confirming outliers (e.g., SalePrice above $300,000 pre-capping), scatter plots of SalePrice vs. GarageArea and LotArea showing positive but scattered relationships, a bar chart of Average SalePrice by GarageCars (0-car: $110K, 3-car: $274K), and Average SalePrice by Fence ("Good Privacy" at $183K vs. "No Fence" at $172K).

Key Insights Derived from Visualizations and Advanced Analysis

The analysis uncovers critical insights into property characteristics and predictive modeling. Paved driveways are rare (7.4% of properties), suggesting they may be a premium feature increasing property value, while 68.14% of properties have been renovated, likely commanding higher prices due to modern upgrades. The market favors family-sized homes, with 2–4 bedrooms and 1–2 floors, as 3-bedroom homes dominate (1,616 properties), and most properties are in "Good" condition (74.5%), though 5.76% are in "Bad" condition, presenting renovation opportunities. Construction peaked in 2006 (31 properties), reflecting a housing boom, with newer homes prominent post-2000. In predictive modeling, Linear Regression outperforms the Decision Tree, with a higher R^2 (0.9232 vs. 0.8263), lower RMSE ($18,428 vs. $27,719), and lower MAE ($12,688 vs. $19,530), making it more reliable for price predictions. The Decision Tree, with a depth of 5, captures non-linear relationships but underfits, needing tuning to improve performance, and both models struggle with outliers in high-value homes, often underpredicting them. Statistically, SalePrice is right-skewed (mean $176,311.23, median $165,400), with high variability (std $64,600.97), contributing to prediction errors. Larger garages (3-car at $274K vs. 0-car at $110K) and high-quality fences ("Good Privacy" at $183K) increase property value, while LotArea shows significant variability (std 3,611.99 post-capping), with large lots not always correlating with higher prices, likely due to location or condition factors.

Recommendations for Further Exploration or Applications of the Insights

The insights provide actionable recommendations for real estate stakeholders. For market targeting, developers and realtors should focus on 3-bedroom, two-story homes in "Good" condition to meet demand for family housing, emphasizing properties with paved driveways and recent renovations in marketing to support premium pricing. Investment strategies should target the 5.76% of properties in "Bad" condition and the 31.86% unrenovated for renovation projects, particularly those without paved driveways, to enhance value, and investigate outliers in SalePrice and LotArea to identify undervalued properties, possibly in less desirable areas. Model improvement involves enhancing the Decision Tree by increasing its depth or using a Random Forest to capture non-linear patterns, applying log-transformation to SalePrice and LotArea to reduce skewness and improve prediction accuracy, and including features like PavedDrive, BedroomAbvGr, and OverallCond in models to better predict high-value properties. Further analysis should examine location-based features (e.g., neighborhood data) to explain SalePrice vs. LotArea outliers, analyze economic conditions (e.g., interest rates, GDP growth) to understand construction trends like the 2006 peak, and explore the interaction between YearRemodAdd and OverallQual to quantify the price premium of renovations across quality levels, refining market strategies and predictive models.