**Portfolio Project Documentation**

**Correlation One**

**Tech for jobs 7**

**Title:**

Ames Housing Dataset Analysis

*A project submitted**in fulfillment of the requirements for the capstone project*

**by:**

Osama Mohammed Abuzraiq

January, 2025

**Abstract**

This project aims to provide data-driven insights using the Ames Housing dataset from Kaggle. The goal is to uncover valuable insights that can inform decisions on house purchases or help analyse the factors influencing changes in house prices in Ames, Iowa, USA. The analysis will utilize one or more of the following tools and languages: Excel, SQL, Tableau, and Python. Given that the dataset contains 80 features, a feature selection process will be conducted to identify and focus on the most relevant columns, as using all 80 features would be time-consuming and not all features equally impact house prices. Potential insights include, but not limited to, determining the key factors affecting house prices, exploring the relationship between house prices and neighbourhood locations, and identifying trends in house prices over time or across different property types. Lastly, a dashboard will be created to visualize the most important insights, providing an interactive and comprehensive overview of the findings to support decision-making.

**Table of Contents**

[**Chapter 1: Project Description** 1](#_Toc191827293)

[**1.1. Problem Statement** 1](#_Toc191827294)

[**1.2. Possible impact of my analysis** 1](#_Toc191827295)

[**1.3. Overview of the dataset** 1](#_Toc191827296)

[**1.4. Review of Related Literature and Systems** 2](#_Toc191827297)

[**Chapter 2: Project Scoping** 3](#_Toc191827298)

[**2.1. Business Problem** 3](#_Toc191827299)

[**2.2. Business Impact** 3](#_Toc191827300)

[**2.3. Dataset(s)** 3](#_Toc191827301)

[**2.4. Methods** 4](#_Toc191827302)

[**2.5. Dashboard** 5](#_Toc191827303)

[**2.6. Milestones** 6](#_Toc191827304)

[**2.7. Timeline** 6](#_Toc191827305)

[**Chapter 3: Data Analysis & Computation** 8](#_Toc191827306)

[**3.1. Data Curation:** 8](#_Toc191827307)

[**3.2. Data Wrangling:** 12](#_Toc191827308)

[**3.3. Exploratory Data Analysis:** 13](#_Toc191827309)

[**3.4. Statistical Analysis:** 15](#_Toc191827310)

[**Chapter 4: Challenges and Solutions** 21](#_Toc191827311)

[**Chapter 5: Description of Dashboard** 22](#_Toc191827312)

[**Chapter 6: Conclusions and Future Work** 23](#_Toc191827313)

[**6.1. Conclusions:** 23](#_Toc191827314)

[**6.2. Future Work:** 23](#_Toc191827315)

**List of Figures**

[**Figure 1, Mockup Dashboard (Made with Tableau)** 5](#_Toc188421093)

[**Figure 2, Gantt Chart of the Project TimeLine** 7](#_Toc188421094)

[**Figure 3, Sale Price Distribution** 11](#_Toc188421095)

**List of Tables**

[**Table 1, Dataset Description Table** 1](#_Toc188421096)

[**Table 2, TimeLine Table** 6](#_Toc188421097)

[**Table 3, Data Table Schema (Before Feature Selection)** 8](#_Toc188421098)

# **Chapter 1: Project Description**

**1.1. Problem Statement**

The real estate industry plays a crucial role in economic growth, and understanding the factors that influence house prices is essential for buyers, sellers, investors, and financial institutions. House prices are impacted by various factors such as location, property size, age and many more, making it challenging to identify the key drivers of price fluctuations. This project aims to analyze the Ames Housing dataset to **uncover the most important features affecting house prices**, providing valuable insights to guide decision-making in the residential housing market in Ames, Iowa, USA.

**1.2. Possible impact of my analysis**

Accurately predicting house prices can lead to better investment decisions, more effective pricing strategies, and optimized property management. Also, identifying trends and key features that drive price fluctuations can help improve market forecasts, allowing for more informed decisions in the housing market. Therefore, Understanding the factors that influence house prices is useful for a variety of stakeholders, including homebuyers, sellers, and real estate investors.

## **1.3. Overview of the dataset**

The Ames Housing dataset, which describes the sale of individual residential property in Ames, Iowa from 2006 to 2010, was compiled by Dean De Cock is publicly available for use in data science education. [[1]](#Ref1)

Here is a brief description of the dataset’s variables:

**Table 1, Dataset Description Table**

|  |  |
| --- | --- |
| **Title of the Dataset** | Ames Housing Dataset |
| **Link to the Dataset** | <https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset/data> |
| **Number of Samples** | 2930 |
| **Number of Attributes** | 82 |
| **Explanatory variables** | 80 (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) [[1]](#Ref1) |
| **Sample Identifiers** | Order (Sample Number), PID (Parcel identification number) |
| **Target Variable** | SalePrice |
| **Data types** | Both categorical and numerical variables |

**Potential key columns to consider after an initial review of the dataset include:**

**LotArea:** Lot size in square feet.   
**YearBuilt:** Year the house was built.  
**Bedrooms:** Number of bedrooms.  
**Bathrooms:** Number of bathrooms.

**GarageType:** Type of garage (e.g., attached, detached).   
**Neighborhood:** Name of the neighborhood.

**HouseStyle:** Architectural style of the house.

**Condition:** Overall condition rating.

**PoolQC:** Pool quality.

**FireplaceQu:** Fireplace quality.

**LotFrontage:** Linear feet of street connected to property.

**SalePrice:** Final price of the house.

(Additional columns may be identified as key after further analysis of the dataset, column descriptions were gathered from an online document.) [[2]](#Ref2)

**1.4. Review of Related Literature and Systems**

**Gradient Boosting for Ames Housing Price Prediction**

A project on the Ames Housing Dataset utilized Scikit-Learn’s GradientBoostingRegressor to predict house prices in Ames, Iowa. The study emphasized feature engineering, creating new metrics like total square footage, finished square footage, and total bathrooms, which showed strong correlations with sale prices. Data preprocessing included handling missing values, one-hot encoding categorical features, and standardizing numerical ones.

GradientBoostingRegressor outperformed RandomForestRegressor, achieving a lower RMSE of 29,932.58 after hyperparameter tuning. The study concluded that gradient boosting effectively handled the dataset's complexity and produced predictions consistent with the expected skew-right distribution of housing prices. [[3]](#Ref3)

This project will focus on gathering insights and analyzing the features of the Ames Housing Dataset to identify those that have the most impact on house prices. Unlike Michael May's approach, which utilized gradient boosting, this project will not involve machine learning algorithms. However, machine learning techniques could be applied in future work to enhance the analysis. By the end of this project, I will have identified the most important features influencing house prices, making them ready for use in machine learning models if needed.

**Chapter 2: Project Scoping**

## **2.1. Business Problem**

For businesses in the real estate sector, failure to correctly assess house values can lead to missed opportunities, overpricing, or under-pricing properties. For investors, understanding the key factors that affect property prices is important for maximizing returns. For financial institutions, accurate house price predictions are essential for assessing loan risk and setting appropriate mortgage rates.

Uncovering the most important features influencing house prices through detailed data analysis, providing insights that can help guide decisions in the real estate market. By understanding these factors, stakeholders can optimize pricing strategies, improve investment decisions, and navigate the housing market more effectively.

## **2.2. Business Impact**

Accurate predictions of house prices can greatly improve investment decisions, pricing strategies, and property management. By identifying the main factors that affect price changes, the analysis will help stakeholders better understand market trends, make more accurate forecasts, and reduce uncertainty. This will allow homebuyers, sellers, and real estate investors to make smarter, more strategic choices, leading to better financial outcomes and a more efficient real estate market. In addition, a dashboard will be created to display key insights, providing an easy-to-use tool for stakeholders to explore the results and make informed decisions. **Although the dataset is not real-time, the insights gained can still offer valuable guidance for understanding historical trends and making future predictions.**

## **2.3. Dataset(s)**

Currently, I am working with the Ames Housing dataset from Kaggle. This dataset contains 80 features related to residential properties in Ames, Iowa, including details about house prices, property characteristics, and neighborhood attributes.

While this dataset is the primary example for this project, it may not be the final one. Depending on the project's evolution or further data exploration, other datasets could be considered to supplement or enhance the analysis.

**Strengths:** -

* **Comprehensive Data:** The dataset contains a wide range of features that provide a detailed picture of residential properties.
* **Real-World Relevance:** The Ames Housing dataset reflects real-world data, making it useful for understanding actual market trends and property values.
* **Rich in Features:** With numerous features, the dataset allows for the exploration of multiple factors influencing house prices, leading to an in-depth analysis.

**Weaknesses:** -

* **Outdated Data:** The dataset contains historical data from 2010, which may not fully capture current market trends.
* **Limited Geographic Scope:** The dataset focuses on a specific region (Ames, Iowa), which may limit its applicability to other geographic locations with different housing market dynamics.

This mix of strengths and weaknesses highlights both the value and limitations of the Ames Housing dataset in providing insights into the real estate market.

## **2.4. Methods**

The primary variables in the **Ames Housing dataset** include:

**Target Variable: -**

* **SalePrice**: The final price of the house, which will be the dependent variable for analysis.

**Key Independent Variables (Features): -**

* **LotArea**: Lot size in square feet.
* **YearBuilt**: Year the house was built.
* **Bedrooms**: Number of bedrooms.
* **Bathrooms**: Number of bathrooms.
* **GarageType**: Type of garage (e.g., attached, detached).
* **Neighborhood**: Name of the neighborhood.
* **HouseStyle**: Architectural style of the house.
* **Condition**: Overall condition rating.
* **PoolQC**: Pool quality.
* **FireplaceQu**: Fireplace quality.
* **LotFrontage**: Linear feet of street connected to property.

These are just some examples, and further feature selection will be performed to narrow down or expand upon the most significant predictors of house prices.

**Data Analysis Plan:**

**Data Cleaning:** -

* Handle missing data using imputation techniques, or by removing irrelevant rows or columns.
* Detect and handle outliers to ensure they don’t skew the analysis results.
* Ensure data type consistency in columns.

**Feature Selection:** -

Given that the dataset contains 80 features, it’s important to perform feature selection to identify the most impactful variables that influence house prices.

**Exploratory Data Analysis (EDA):** -

* Visualize key relationships between our target variable (SalePrice) and other features. Using scatter plots, box plots, box and whisker plot, and histograms.
* Analyse the impact of internal factors, such as Bedrooms and Bathrooms, on house prices, and compare them with external factors, like LotArea and Neighborhood, to determine which has a greater influence on house prices.
* Additional insights and analyzes will be explored as the EDA progresses and will be included in the EDA chapter.

**Datafolio & Dashboard:** -

Create an interactive dashboard using either Tableau or Looker Studio to visualize key insights and model predictions. The dashboard will provide an accessible, visual representation of the analysis.

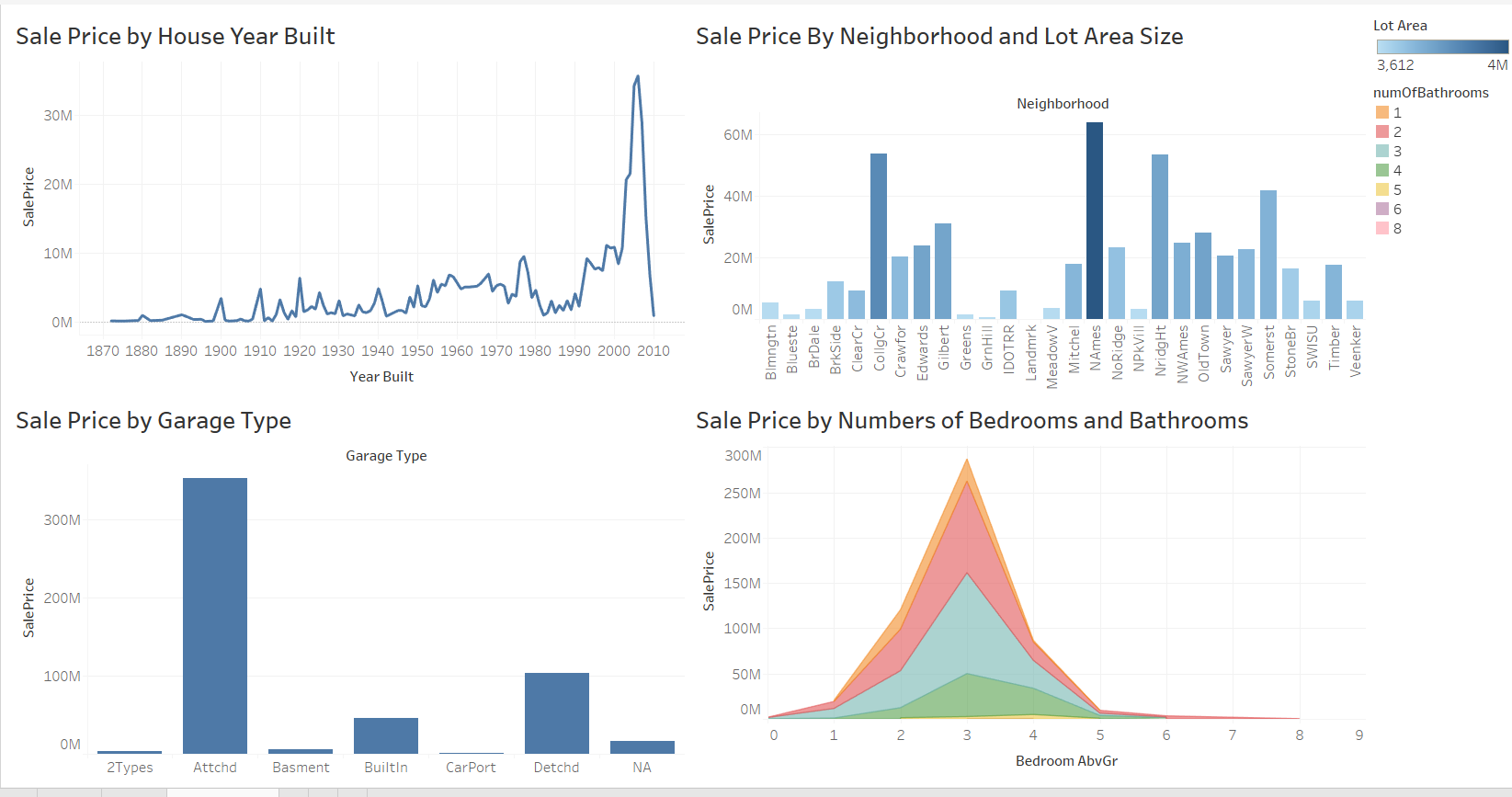
**Final Report:** -

Compile all findings, including feature selection and key insights from the EDA into a comprehensive final report.

The report will summarize the methodologies used, and highlight the most impactful factors influencing house prices in Ames, Iowa.

## **2.5. Dashboard**

Below are some examples of the visualizations that will be included in the dashboard (please note that the final version of the dashboard will likely be more refined by the conclusion of the project's "[Description of Dashboard](#_Chapter_5:_Description)" section):



**Figure 1, Mockup Dashboard (Made with Tableau)**

## **2.6. Milestones**

**Project Scoping and Dataset Familiarization:** -

* Define the project’s goals, scope, and deliverables.
* Understand the structure, variables, and challenges of the Ames Housing dataset.

**Data Cleaning and Preparation:** -

* Address missing values, outliers, and inconsistencies.
* Ensure data type consistency and prepare the dataset for analysis.

**Exploratory Data Analysis (EDA):** -

* Analyze the dataset to uncover patterns and relationships between variables.
* Visualize key insights using scatter plots, box plots, histograms, and more.
* Investigate the influence of internal and external factors on house prices.

**Feature Selection and Engineering: -**

* Narrow down the 80 features to the most impactful variables for house price prediction.
* Create new variables if necessary to enhance the dataset’s predictive capabilities.

**Dashboard and Datafolio Development: -**

* Design and develop an interactive dashboard using Tableau or Looker Studio (TBD).
* Compile the project’s findings, methodologies, and insights into a datafolio.

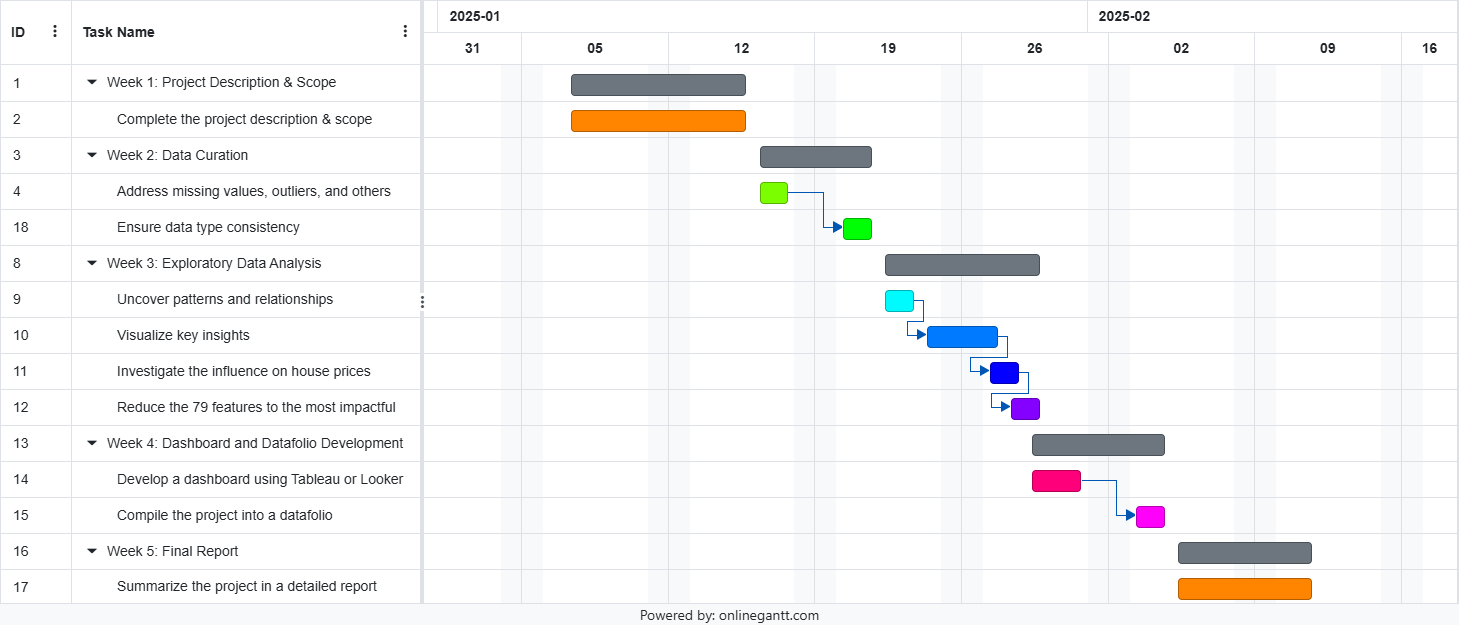
**Final Report and Presentation: -**

* Summarize the project in a detailed report, including the methodologies and key findings.
* Present the dashboard, datafolio, and report to tech for jobs 7 submission for review.

## **2.7. Timeline**

**Table 2, TimeLine Table**

|  |  |
| --- | --- |
| Week | Tasks |
| Week 1 | - Define project scope and objectives - Outline project deliverables and timeline - Document project description |
| Week 2 | - Collect and prepare the Ames Housing Dataset - Handle missing values, data cleaning, and feature selection |
| Week 3 | - Conduct EDA to understand data patterns - Visualize relationships between features and target variable (house prices) |
| Week 4 | - Develop a dashboard for visualizing key insights using Tableau/Looker - Integrate findings from EDA into the dashboard |
| Week 5 | - Summarize findings and insights from the analysis - Write the final report documenting the analysis process and results |



**Figure 2, Gantt Chart of the Project TimeLine**

# **Chapter 3: Data Analysis & Computation**

## **3.1. Data Curation**

* **Data profile:** 
  + **Dataset Name:** Ames housing.
  + **Dataset details:** 2930 samples, 82 variables.
  + **Dataset file size:** 936KB.
  + **Description:** Data set contains information from the Ames Assessor’s Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010.
  + **Dataset Source:** <https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset/data>
  + **Dataset Excel File Link:**

[AmesHousing Portfolio Project.xlsx](https://1drv.ms/x/c/63a6cabba5e66308/EUrN_2VQzzJKsbMBbb3YWEgBpk40euXZw6GFkl5-TmF7XQ?e=3vgMjt)

* + **Dataset Documentation Source:** <https://jse.amstat.org/v19n3/decock/DataDocumentation.txt>
  + **Data Table Schema:**

The table below displays the data schema prior to feature selection. Feature selection will be applied to reduce the number of fields, as 82 features represent a very high dimensionality.

**Table 3, Data Table Schema (Before Feature Selection)**

|  |  |  |
| --- | --- | --- |
| Field | Type | Description |
| Order | Discrete | Observation number |
| PID | Nominal | Parcel identification number |
| MS SubClass | Nominal | Identifies the type of dwelling involved in the sale. |
| MS Zoning | Nominal | Identifies the general zoning classification of the sale. |
| Lot Frontage | Continuous | Lot size in square feet |
| Street | Nominal | Type of road access to property |
| Alley | Nominal | Type of alley access to property |
| Lot Shape | Ordinal | General shape of property |
| Land Contour | Nominal | Flatness of the property |
| Utilities | Ordinal | Type of utilities available |
| Lot Config | Nominal | Lot configuration |
| Land Slope | Ordinal | Slope of property |
| Neighborhood | Nominal | Physical locations within Ames city limits |
| Condition 1 | Nominal | Proximity to various conditions |
| Condition 2 | Nominal | Proximity to various conditions (if more than one is present) |
| Bldg Type | Nominal | Type of dwelling |
| House Style | Nominal | Style of dwelling |
| Overall Qual | Ordinal | Rates the overall material and finish of the house |
| Overall Cond | Ordinal | Rates the overall condition of the house |
| Year Built | Discrete | Original construction date |
| Year Remod/Add | Discrete | Remodel date (same as construction date if no remodeling or additions) |
| Roof Style | Nominal | Type of roof |
| Roof Matl | Nominal | Roof material |
| Exterior 1 | Nominal | Exterior covering on house |
| Exterior 2 | Nominal | Exterior covering on house (if more than one material) |
| Mas Vnr Type | Nominal | Masonry veneer type |
| Mas Vnr Area | Continuous | Masonry veneer area in square feet |
| Exter Qual | Ordinal | Evaluates the quality of the material on the exterior |
| Exter Cond | Ordinal | Evaluates the present condition of the material on the exterior |
| Foundation | Nominal | Type of foundation |
| Bsmt Qual | Ordinal | Evaluates the height of the basement |
| Bsmt Cond | Ordinal | Evaluates the general condition of the basement |
| Bsmt Exposure | Ordinal | Refers to walkout or garden level walls |
| BsmtFin Type 1 | Ordinal | Rating of basement finished area |
| BsmtFin SF 1 | Continuous | Type 1 finished square feet |
| BsmtFinType 2 | Ordinal | Rating of basement finished area (if multiple types) |
| BsmtFin SF 2 | Continuous | Type 2 finished square feet |
| Bsmt Unf SF | Continuous | Unfinished square feet of basement area |
| Total Bsmt SF | Continuous | Total square feet of basement area |
| Heating | Nominal | Type of heating |
| HeatingQC | Ordinal | Heating quality and condition |
| Central Air | Nominal | Central air conditioning |
| Electrical | Ordinal | Electrical system |
| 1st Flr SF | Continuous | First Floor square feet |
| 2nd Flr SF | Continuous | Second Floor square feet |
| Low Qual Fin SF | Continuous | Low quality finished square feet (all floors) |
| Gr Liv Area | Continuous | Above grade (ground) living area square feet |
| Bsmt Full Bath | Discrete | Basement full bathrooms |
| Bsmt Half Bath | Discrete | Basement half bathrooms |
| Full Bath | Discrete | Full bathrooms above grade |
| Half Bath | Discrete | Half baths above grade |
| Bedroom | Discrete | Bedrooms above grade (does NOT include basement bedrooms) |
| Kitchen | Discrete | Kitchens above grade |
| KitchenQual | Ordinal | Kitchen quality |
| TotRmsAbvGrd | Discrete | Total rooms above grade (does not include bathrooms) |
| Functional | Ordinal | Home functionality (Assume typical unless deductions are warranted) |
| Fireplaces | Discrete | Number of fireplaces |
| FireplaceQu | Ordinal | Fireplace quality |
| Garage Type | Nominal | Garage location |
| Garage Yr Blt | Discrete | Year garage was built |
| Garage Finish | Ordinal | Interior finish of the garage |
| Garage Cars | Discrete | Size of garage in car capacity |
| Garage Area | Continuous | Size of garage in square feet |
| Garage Qual | Ordinal | Garage quality |
| Garage Cond | Ordinal | Garage condition |
| Paved Drive | Ordinal | Paved driveway |
| Wood Deck SF | Continuous | Wood deck area in square feet |
| Open Porch SF | Continuous | Open porch area in square feet |
| Enclosed Porch | Continuous | Enclosed porch area in square feet |
| 3-Ssn Porch | Continuous | Three season porch area in square feet |
| Screen Porch | Continuous | Screen porch area in square feet |
| Pool Area | Continuous | Pool area in square feet |
| Pool QC | Ordinal | Pool quality |
| Fence | Ordinal | Fence quality |
| Misc Feature | Nominal | Miscellaneous feature not covered in other categories |
| Misc Val | Continuous | $Value of miscellaneous feature |
| Mo Sold | Discrete | Month Sold (MM) |
| Yr Sold | Discrete | Year Sold (YYYY) |
| Sale Type | Nominal | Type of sale |
| Sale Condition | Nominal | Condition of sale |
| SalePrice | Continuous | Sale price $$ |

* **Initial Observations:**
* **Missing values**: Used conditional formatting to check which cells contain blanks, which turned out to be mainly in the “Garage Yr Blt” and “Lot Frontage” columns, the dataset contains 661 samples with missing values, out of 2930 samples.
* **Duplicates:** Checked the identifier columns “Order” and “PID” for duplicates using conditional formatting, no duplicates were found.
* **Outliers:** Considering this is housing data, there is bound to be outliers. (when comparing a Duplex building type with a Townhouse End Unit, for example).
* **Standardized Categorical Entries:** Consistent Formatting (e.g., "Two" vs. "2").
* **Statistical summary of Target Variable (SalePrice):**
  + **Mean:** 180796.1
  + **Median:** 160000
  + **Mode:** 135000
  + **Standard Deviation:** 79886.69
  + **Distribution:** The figure below displays that the Sale Price’s distribution is right-skewed.



**Figure 3, Sale Price Distribution**

## **3.2. Data Wrangling**

* **Data Cleaning:**
  + **Handling missing data:** 
    - **Total number of rows with missing values:** 661 rows
    - **Garage Yr Blt:** There are 139 missing values in this variable, the missing values in the "Garage Yr Blt" column are linked to the "Garage Type" column. When the garage type is "NA," the "Garage Yr Blt" is intentionally left blank, signifying the absence of a garage and thus no applicable year. Since my analysis doesn’t involve machine learning or other techniques that require complete data, I will leave these missing values unchanged.
    - **Lot Frontage (Linear feet of street connected to property):** There are 490 missing values in this variable, and no other variable explains the missing values. We assume lot frontage is similar across properties in the same neighborhood, so we'll impute the missing values with the median and reassess whether to include this variable in our EDA later. Doing this step lowered the total number of rows with missing values from 661 to 171.

If we exclude the missing values from the “Garage Yr Blt” column, there are 32 rows remaining with missing values. I will drop these rows, as they will not significantly impact the overall dataset.

**Overall results:** 490 rows cleaned, 139 rows left as is, and 32 rows dropped.

* + **Addressing Outliers:**

Given the large number of columns in the dataset, it would be challenging to determine which ones to use for outlier detection. Therefore, I will focus on the target variable, **SalePrice**, for this purpose. While both Z-score and Interquartile Range (IQR) methods can be used, IQR is more suitable in this case because the SalePrice distribution is skewed (as shown in Figure 3), and Z-score assumes a normal distribution.

**Quartile 1:** 129000 **[**=QUARTILE(data\_range, 1)**]**

**Quartile 3:** 213000 **[**=QUARTILE(data\_range, 3)**]**

**IQR:** 84000 **[**=Q3−Q1**]**

**Lower Bound:** 3000 **[**=Q1−1.5×IQR**]**

**Upper Bound:** 339000 **[**=Q3+1.5×IQR**]**

**Sum of data points outside of lower & upper bounds:** 135 **[**=SUM(CL:CL)**]**

**Percentage of outliers: [**(135 / 2930) \* 100 ~= 4.61%**]**

The percentage of outliers, approximately 4.61%, is low and falls within the expected range for real-world datasets. This indicates that the presence of outliers is not excessive and aligns with common data characteristics.

These outliers could represent luxury or low-end houses, and the decision to keep or remove them depends on the analysis goal: are we focusing on the majority or aiming to reflect real-world scenarios? Removing these outliers might lead to losing valuable insights, such as understanding why some houses are significantly more expensive or cheaper than others. Therefore, they will be kept in the dataset.

* + **Data Reduction:** 
    - **Duplicates:** The data contains no duplicates to be removed.
    - **Merged columns:** The data contains no columns that could be merged together seamlessly.

## **3.3. Exploratory Data Analysis**

My exploratory data analysis will focus solely on numerical values and statistical insights. Charts and diagrams will be presented in the next section, **“Project Datafolio & Dashboard.”**

* **Relationship with Target Variable ("SalePrice"):**

Since the data contains 79 variables (excluding the 2 ID columns and 1 target variable), we will compute correlation of all features with the target variable.

* + **Python Code:**

Link to Jupyter Notebook in which correlation analysis was conducted:

[Ames Housing Dataset.ipynb](https://colab.research.google.com/drive/1T-D29seV--w8xCU-ynngJ56q2MEBc5iJ?usp=sharing)

* + **List of the variables with a strong correlation with the target variable (> 0.3 correlation):**

These are the variables that we’ll focus on for the rest of the analysis.

|  |  |
| --- | --- |
| **Name** | **Correlation** |
| Overall Qual | 0.79926 |
| Gr Liv Area | 0.70678 |
| Garage Cars | 0.64788 |
| Garage Area | 0.64040 |
| Total Bsmt SF | 0.63228 |
| 1st Flr SF | 0.62168 |
| Year Built | 0.55843 |
| Full Bath | 0.54560 |
| Year Remod/Add | 0.53297 |
| Garage Yr Blt | 0.52697 |
| Mas Vnr Area | 0.50829 |
| TotRms AbvGrd | 0.49547 |
| Fireplaces | 0.47456 |
| BsmtFin SF 1 | 0.43291 |
| Lot Frontage | 0.35732 |
| Wood Deck SF | 0.32714 |
| Open Porch SF | 0.31295 |

* **Checking for multicollinearity:**

Check to see if some of the strong correlations are highly correlated with each other (>0.8 correlation), which may indicate redundancy.

* List of variable pairs with high correlation:

|  |  |  |
| --- | --- | --- |
| **1st Variable** | **2nd Variable** | **Correlation** |
| Garage Area | Garage Cars | 0.88968 |
| 1st Flr SF | Total Bsmt SF | 0.80072 |
| Garage Yr Blt | Year Built | 0.83485 |
| TotRms AbvGrd | Gr Liv Area | 0.80777 |

|  |  |  |
| --- | --- | --- |
| **Which features to drop and why:** | | |
| **Garage Area vs. Garage Cars (0.89)** |  | **1st Flr SF vs. Total Bsmt SF (0.80)** |
| **Drop:** |  | **Drop:** |
| Garage Cars |  | Total Bsmt SF |
| **Reason:** |  | **Reason:** |
| ‘Garage Area’ is more precise because it represents continuous space, whereas ‘Garage Cars’ is categorical (number of cars). |  | ‘1st Flr SF’ represents livable above-ground space, which is generally more impactful for pricing. |
|  |  |  |
| **Garage Yr Blt vs. Year Built (0.83)** |  | **TotRms AbvGrd vs. Gr Liv Area (0.81)** |
| **Drop:** |  | **Drop:** |
| Garage Yr Blt |  | TotRms AbvGrd |
| **Reason:** |  | **Reason:** |
| ‘Year Built’ is more general and covers the entire house construction, making it more relevant. |  | ‘Gr Liv Area’ accounts for total livable space in square feet, which is more continuous and useful than room count. |

* **Feature Selection Summary**

**Features Kept:**

* **Garage Area:** More precise, represents continuous space.
* **1st Flr SF:** Represents livable above-ground space, more impactful for pricing.
* **Year Built:** Covers the entire house construction, making it more relevant.
* **Gr Liv Area:** Accounts for total livable space in square feet, more useful than room count.

**Features Dropped:**

* **Garage Cars:** Categorical (num of cars), less precise than Garage Area.
* **Total Bsmt SF:** Redundant with 1st Flr SF.
* **Garage Yr Blt:** Redundant with Year Built.
* **TotRms AbvGrd**: Redundant with Gr Liv Area, which is a continuous measurement.
* **Added features to the dataset:**
* **Remodeled:** Indicates whether the house has been remodeled.

It's 1 if Year Remod/Add is different from Year Built, otherwise 0.

**FORMULA:** "=IF(@[Year Remod/Add] <> @[Year Built], 1, 0)"

* **Age:** The number of years since the house was built.

Calculated as Yr Sold - Year Built.

**FORMULA:** "=@[Yr Sold] - @[Year Built]"

## **3.4. Statistical Analysis**

Further detailed statistics can be found in the [**Excel File**](https://1drv.ms/x/c/63a6cabba5e66308/EUrN_2VQzzJKsbMBbb3YWEgBpk40euXZw6GFkl5-TmF7XQ?e=3vgMjt) related to this project under the “EDA” sheet.

* **Numerical features statistics:**

1. **Gr Liv Area:**

* **General Distribution:** The data is fairly **evenly distributed,** as the **mean (1498.2) and median (1440) are close**.
* **Spread of Data:** The **standard deviation (506.6)** indicates a moderate spread around the mean, suggesting some variation in house sizes but no extreme outliers.
* **Most Frequent Value:** The **mode (864)** shows the most common house size in the dataset.
* **Range & Binning:** The **range (5308)** suggests a wide variety of house sizes, but most houses fall in the first two bins(**≤2821 sqft**).
* **Potential Outliers:** A small number of houses (**34 in bin 2822-4231.5 and 5 in bin 4232.5-5642**) have significantly larger living areas, which could be considered **outliers**.

**Key Takeaway:** Most homes are **around 1400-1500 sqft**, with a few significantly larger ones. The distribution is fairly balanced.

1. **Garage Area:**

* **General Distribution:** The mean (471.8) and median (478) are close, indicating a fairly even distribution of garage sizes.
* **Spread of Data:** The standard deviation (215.2) suggests a moderate spread, meaning garage sizes vary but are not extremely scattered.
* **Most Frequent Value:** The mode (0) shows that many houses do not have garages, which is important to note.
* **Range & Binning:** The range (1488) indicates significant variation in garage sizes, but the majority fall within 372-744 sqft.
* **Potential Outliers:** Very few houses (17 in bin 1117-1488 sqft) have extremely large garages, which may be outliers or high-end properties.

**Key Takeaway:** Most houses have a garage around 470-480 sqft, but a considerable number lack a garage entirely. The distribution is fairly balanced.

1. **1st Flr SF (1st Floor Square Footage):**

* **General Distribution:** The mean (1158.86) and median (1083) are close, indicating a fairly even distribution of first-floor sizes.
* **Spread of Data:** The standard deviation (391.36) shows that floor sizes vary, but most values are within a reasonable range.
* **Most Frequent Value:** The mode (864) suggests that a significant number of houses have this first-floor size.
* **Range & Binning:**

1. Most homes (1950 observations) have a first-floor size ≤ 1273.75 sqft.
2. A smaller portion (942 observations) falls between 1274.75 - 2547.5 sqft.
3. Very few houses (9 observations in 2548.5 - 3821.25 sqft, 2 observations above 3822.25 sqft) have exceptionally large first floors, which could be considered outliers.

**Key Takeaway:** The majority of houses have a first-floor size around 1000-1200 sqft. with a few **large properties** that might impact price analysis.

1. **Mas Vnr Area (Masonry Veneer Area):**

* **Skewed Distribution:** The mean (101.81) is greater than the median (0), indicating a right-skewed distribution where many houses have little to no masonry veneer, but a few have large values.
* **Most Frequent Value:** The mode is 0, meaning a significant portion of houses have no masonry veneer.
* **Spread of Data:** The standard deviation (179.17) suggests a high variance, with some homes having very large masonry areas.
* **Range & Binning:**

1. 2696 observations (majority) have ≤ 400 sqft of masonry veneer.
2. A small number of houses fall in the 401-800 sqft (181 observations) and 801-1200 sqft (21 observations**)** bins.
3. Only 5 houses have extremely large masonry veneer areas (1201-1600sqft), likely outliers.

**Key Takeaway:** Most houses have little or no masonry veneer, but a few with large veneer areas might affect price correlations.

1. **BsmtFin SF 1 (Conclusion for Basement Finished Square Feet):**

* **Skewed Distribution:** The mean (442.54) is greater than the median (370), indicating a right-skewed distribution where most houses have smallornofinishedbasementareas, but a few have very large values.
* **Most Frequent Value:** The modeis0, meaning a large number of houseshave no finished basement.
* **Spread of Data:** The standard deviation (455.09) suggests high variability, indicating significant differences between houses.
* **Range & Binning:**
  + 1. 2823 observations (majority) have ≤ 1411 sqft of finished basement space.
    2. Only 78 houses fall in the 1412-2822 sqft range, and just two extreme cases exceed 2823 sqft, likely outliers.

**Key Takeaway:** Most houses have small or no finished basements, but a fewlarge outliers could heavily influence analysis.

1. **Lot Frontage:**

* **Even Distribution:** The mean (68.98) and median (68) are approximately equal, suggesting a fairly symmetrical distribution with no strong skewness.
* **Most Frequent Value:** The mode is 68, meaning many properties share this lot frontage value.
* **Spread of Data:** The standard deviation (21.35) indicates moderate variability, meaning lot frontages do not vary drastically for most properties.
* **Range & Binning:**
  1. 2185 observations (majority) have ≤ 78.25 feet of frontage.
  2. 707 houses fall within 79.25-156.5 feet, and only a few extreme cases exceed 157.5 feet, making them potential outliers.

**Key Takeaway:** Lot frontage is fairly evenly distributed, with most properties having a frontage of around 68 feet and only a few with very large lot frontages.

1. **Wood Deck SF:**

* **Right-Skewed Distribution:** The mean (94) is greater than the median (0), indicating that most homes have little to no deck space, while a few have significantly large decks.
* **Most Frequent Value:** The mode is 0, meaning many houses have no wood deck at all.
* **Spread of Data:** The standard deviation (126.64) shows high variability, meaning deck sizes differ greatly between properties.
* **Range & Binning:**
  1. 2789 homes (majority) have ≤ 356 SF of deck space, indicating that small decks or no decks at all are common.
  2. Only a few homes (1) have decks larger than 1069 SF, making them extreme cases.

**Key Takeaway:** Wood deck sizes are heavily skewed, with most properties having no deck or a small deck, while only a few have large decks.

1. **Open Porch SF:**

* **Right-Skewed Distribution:** The mean (47.26) is greater than the median (26**)**, indicating that most homes have small or no open porches, while a few have much larger porches.
* **Most Frequent Value:** The mode is 0, meaning many homes have no open porch at all.
* **Spread of Data:** The standard deviation (67.31) suggests high variability, meaning porch sizes differ significantly.
* **Range & Binning:**
  1. 2766 homes (majority) have ≤ 185.5 SF of porch space, showing that small porches are common.
  2. Only 2 homes have porches larger than 557.5 SF, making them extreme cases.

**Key Takeaway:** Open porch sizes are heavily skewed, with most properties having small or no porches, while only a few have large porches.

1. **Age:**

* **Even Distribution:** The mean (36.73) and median (35) are approximately equal, suggesting a fairly uniform distribution rather than a strong skew.
* **Most Frequent Value:** The mode is 1, meaning many homes were recently built or remodeled.
* **Spread of Data:** The standard deviation (30.26) indicates moderate variation in building ages.
* **Range & Binning:**
  1. 1,442 homes (majority) are ≤ 34 years old, meaning many homes are relatively new.
  2. 973 homes fall between 35 and 68 years old, indicating a sizable portion of mid-aged properties.
  3. Only 62 homes exceed 103 years old, making them relatively rare.

**Key Takeaway:** The dataset contains a mix of new, mid-aged, and old homes, with a slight concentration toward newer houses.

* **Categorical Features Statistics:**

1. **Overall Qual (Overall Quality):**

* **Distinct Categories:** There are 10 unique values (ranging from 1 to 10), representing different levels of overall building quality.
* **Most Frequent Value (Mode):** The most common rating is 5, accounting for 28.38% of the total dataset, meaning many homes are of average quality.
* **Distribution:**
  1. The majority of homes (53.46%) fall between 5 and 6, indicating that most houses are of average to slightly above-average quality.
  2. Higher-quality homes (7-10) make up 36.79%, with only 1.03% being rated 10 (best quality).
  3. Lower-quality homes (1-4) are rare, comprising only 9.76% of the dataset.

**Key Takeaway:** The dataset is skewed toward average to above-average quality homes, with only a small proportion of very low- or very high-quality houses.

1. **Year Built:**

* **Distinct Categories:** There are 118 unique years, making it impractical to analyze with a pivot table. A histogram is used instead to show the distribution.
* **Most Frequent Year (Mode):** The most common year is 2005, meaning many houses were built in this year.
* **Distribution Trend:**
  1. Few houses were built before 1900.
  2. The number of houses gradually increases from 1900 to 1955.
  3. A significant rise is seen from 1955 onward, with the highest concentration of houses built after 1995.
  4. The most recent decades (1996-2010) have the highest number of houses built.
* **No Missing Values:** There are no blank entries in this column.

**Key Takeaway:** The dataset is skewed towards newer houses, with most homes built after 1955 and a peak in recent years.

1. **Full bath:**

* **Distinct Categories:** There are 5 unique values (0, 1, 2, 3, 4), making it a categorical numerical feature.
* **Most Frequent Value (Mode):** 2 full bathrooms is the most common, found in 66.65% of houses.
* **Distribution Trend:**
  1. Few houses (0.00%) have 0 full baths.
  2. A moderate number (29.03%) have 1 full bath.
  3. The majority (66.65%) have 2 full baths.
  4. Very few houses (3.97%) have 3 full baths.
  5. Only 0.35% have 4 full baths, making them rare.
* **No Missing Values:** Every house has a recorded value for full baths.

**Key Takeaway:** Most houses have 2 full bathrooms, making it the standard. Houses with 3 or more full baths are uncommon, and houses with no full bath are extremely rare.

1. **Year Remod/Add:**

* **Distinct Categories:** There are 61 unique values, making it a widely spread feature.
* **Most Frequent Year (Mode):** 1950 had the highest number of remodels or additions.
* **Distribution Trend:**
  1. There was a significant number of remodels in the 1950s.
  2. The number remained relatively stable until 1990, with occasional peaks.
  3. A sharp increase occurred after 200**0**, with a peak between 2001-2006.
  4. 2006-2011 also had a high number of remodels, though slightly lower than the 2001-2006 period.
* **No Missing Values:** Every house has a recorded remodel year.

**Key Takeaway:** There was a significant remodeling trend in the 1950s, followed by another surge after 2000. Houses that haven't been remodeled since the 1950s-1970s may need upgrades**,** while those remodeled in the 2000s are likely modernized.

1. **Fireplaces:**

* **Distinct Categories:** The feature has **5 unique values** (0, 1, 2, 3, 4 fireplaces).
* **Most Frequent Value (Mode):** **0 fireplaces** is the most common, with **1,412 houses (48.6%)** having none.
* **Distribution Trend:**
  1. **72.48% of houses have at least one fireplace.**
  2. **25.22% have two fireplaces**, which is a considerable percentage.
  3. **Only 2.3% of houses have 3 or more fireplaces**, making them rare.
* **No Missing Values:** Every house has a recorded number of fireplaces.

**Key Takeaway**: Nearly half of the houses do not have fireplaces. However, for homes that do, most have one or two fireplaces, suggesting that fireplaces are common but not a dominant feature in all properties.

1. **Remodeled:**

* **Distinct Categories:** There are only two values (0 = Not Remodeled, 1 = Remodeled).
* **Most Frequent Value (Mode):** 0 (Not Remodeled), with 53.46% of houses having no remodeling.
* **Distribution Trend:**
  1. 46.54% of houses have been remodeled, meaning nearly half of the properties had some form of renovation.
  2. The split between remodeled and non-remodeled homes is fairly balanced.
* **No Missing Values:** Every house has a recorded remodeling status.

**Key Takeaway:** A significant portion of houses (almost half) have undergone remodeling**,** indicating that renovations are common in this dataset. However, slightly more houses remain in their original condition.

# **Chapter 4: Challenges and Solutions**

During the course of this project, several challenges arose, requiring careful problem-solving and adaptation. Below are some of the key difficulties encountered and how they were addressed:

1. **Handling Missing Values:**

**Challenge:** The Ames Housing dataset contains several missing values across different features, such as LotFrontage, GarageYrBlt, and PoolQC. Simply removing missing data could lead to loss of valuable information.

**Solution:** Imputation techniques were applied based on the nature of the missing data. For numerical variables, median imputation was used, while categorical variables were filled with the most frequent category or a placeholder like "None" when applicable.

1. **High Dimensionality of the Dataset:**

**Challenge:** With 80 features available, selecting the most relevant variables for analysis was a challenge. Analyzing all features would be computationally expensive and could lead to redundant information.

**Solution:** Feature selection techniques, including correlation analysis, domain knowledge, and exploratory data analysis (EDA), were used to identify the most significant predictors of house prices.

1. **Dashboard Development Complexity**

**Challenge:** Designing an informative yet user-friendly dashboard required careful selection of visualizations that effectively communicate key insights.

**Solution:** Multiple iterations were tested to refine the dashboard layout, ensuring that relevant trends and patterns were highlighted. The final dashboard was developed using Tableae with interactive elements to enhance usability.

1. **Historical Data Limitations**

**Challenge:** The dataset only covers sales from 2006 to 2010, meaning it may not reflect current real estate trends.

**Solution:** While acknowledging this limitation, the analysis focused on historical trends and factors that remain relevant in housing markets. The insights extracted can still provide valuable knowledge applicable to other real estate datasets.

By overcoming these challenges, the project successfully identified key features impacting house prices and provided a structured approach for real estate market analysis.

# **Chapter 5: Description of Dashboard**

* **Use Case:**

The “Understanding House Prices in Ames, Iowa” dashboard provides an interactive visual analysis of housing prices in Ames, helping users explore how the most notable features highlighted in the previous [EDA section](#_3.3_Exploratory_Data) influence sale prices. It is designed for real estate analysts, homeowners, and potential buyers who want to gain insights into market trends and property values.

**Link to dashboard:** [Understanding House Prices in Ames, Iowa](https://public.tableau.com/app/profile/osama.abuzraiq/viz/UnderstandingHousePricesinAmesIowa/Dashboard#1)

* **Users can interact with the dashboard in several ways:**
* **Filtering by Minimum Sale Price:** A slider allows users to set a minimum sale price threshold, helping them focus on properties within a specific price range.
* **Remodeling Filter:** A toggle option lets users choose between remodeled and non-remodeled houses, enabling a comparison of their impact on sale prices.
* **Visualizations and their purposes:**
* **Neighborhood-Based Insights:** A horizontal bar chart compares average sale prices across different neighborhoods, with color coding distinguishing remodeled vs. non-remodeled homes.
* **Year Built Analysis:** A stacked area chart shows how sale prices have changed over time based on the year of construction, highlighting trends in housing demand.
* **Overall Quality vs. Total Sales Price:** A bar chart groups houses by quality ratings, displaying total sales volume for each category and revealing the correlation between quality and pricing.
* **Sale Price by Age of Home:** A scatter plot visualizes the relationship between a house’s age and its sale price, showing a clear pattern where older homes tend to have lower prices, particularly when not remodeled.

# **Chapter 6: Conclusions and Future Work**

## **6.1. Conclusions**

This project analyzed the **Ames Housing Dataset** to uncover key factors influencing house prices. Through data curation, exploratory data analysis (EDA), and visualization, several important findings emerged:

* **Strongest Predictors of Sale Price:** Features such **as Overall Quality, Gr Liv Area (Above-Ground Living Area), Garage Area, and Total Basement Square Footage** showed the highest positive correlations with house prices.
* **Neighborhood Impact:** House prices vary significantly by **neighborhood**, with areas like **NoRidge and StoneBr** commanding higher values.
* **Effect of Renovation:** Remodeled houses tend to have higher prices than non-remodeled ones, suggesting that **upgrading homes increases value**.
* **Age vs. Sale Price Relationship:** Older homes generally sell for lower prices unless they have been remodeled.
* **Multicollinearity in Features:** Some highly correlated variables, such as **Garage Area & Garage Cars** and **Total Basement SF & 1st Floor SF**, were identified, leading to feature selection to **remove redundancy**.

These findings provide valuable insights for **real estate investors, homeowners, and analysts** looking to understand property valuation trends in Ames, Iowa.

## **6.2. Future Work**

To build upon the findings from this analysis, several areas of improvement and expansion can be considered:

1. **Predictive Modeling:**
   * Implement machine learning algorithms (e.g., **Linear Regression, XGBoost, Random Forest)** to develop a **house price prediction model**.
   * Conduct **feature engineering** to improve model performance.
2. **Interactive Dashboards & Web Applications:**
   * Expand the existing **Tableau dashboard** or develop an interactive **web-based tool** using **Streamlit, Power BI, or Dash.**
   * Allow users to **input house details** and receive an estimated price based on data insights.
3. **Enhancing Data Analysis:**
   * Include additional factors such as **lot size, basement condition, and exterior quality** for deeper insights.

By incorporating these improvements, this project can evolve into a more comprehensive **real estate analytics tool**, providing deeper insights and aiding data-driven decision-making in the housing market.

**References**

[1] Dean De Cock. 2011, Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project, Page 1. Retrieved 15 Jan 2025 from: <https://jse.amstat.org/v19n3/decock.pdf>

[2] Max Kuhn. 2022, The Ames Iowa Housing Data, Page 2. Retrieved 15 Jan 2025 from:

<https://cran.r-project.org/web/packages/AmesHousing/AmesHousing.pdf>

[3] Michael May. 2023, Tackling the Ames Housing Dataset, Retrieved 15 Jan 2025 from:

<https://medium.com/@mskmay66/tackling-the-ames-housing-dataset-c63e7947908a>