Introduction:

The prediction of the Housing prices in Ames, Iowa is the thing that is being talked about. All of us know that the housing market is the biggest thing in our lives. It is everybody's business - from the people who are buying their first home to those who are selling. As a result, we are doing the project to predict the housing prices in the city of Ames, Iowa. And guess what? We are using data science and machine learning to achieve it.

Data Source:

We are, in fact, collecting all our data from the Ames Housing Dataset. It is as if a lot of this huge file contains a huge amount of information about homes in Ames. We're talking about the square footage, bedrooms, sale prices and other facts that are definitely in it.

Project Goals:

Our main goal? Determine the income level at which homes in Ames will really be sold. And we're tackling it in two ways: And these are the two ways:

Regression Analysis:

We are working with some variables using the linear regression. The foremost goal of my work is to find out if there is any kind of connection between the house features and the prices. You know, basic stuff.

Machine Learning Modelling:

We are about to choose RapidMiner for this problem. The machine has the responsibility to predict the sale prices thus, the machine is trying to work its magic. Because, hey, why not?

Comparative Analysis and Recommendations:

Only after executing both the models, we will be in a position to compare them. The best way to come up with an answer is to concentrate on the elements such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Then, we will be the ones to say what we think is the best method. Spoiler alert: surely the one that is the nearest to reality.

Dataset Description

The Ames Housing Dataset delves into the details of residential properties in Ames, Iowa. It provides a rich resource for researchers and data enthusiasts alike. This dataset goes beyond simple prices. It encompasses a wide range of features, both numerical and categorical. Numerical features paint a picture of a property's size and age. These include aspects like square footage, number of bedrooms, and the year construction was completed. Categorical features provide qualitative details. They encompass factors like the neighbourhood a property resides in and the material

used for its exterior. However, the star of the show is the property sale price. This serves as the primary variable of interest. By analysing the dataset's features alongside sale prices, researchers can uncover valuable insights. These insights can shed light on factors influencing housing prices in Ames, Iowa, and potentially in similar markets.

Data Wrangling:

The task of collecting and organizing the data that will be utilized for the study has to be done before anything else. In order to start with the analysis, we should make sure that our data is clean and put up for use. This process, known as data wrangling, involves several steps:

1. Handling Missing Values:

Our dataset had no data values, which can interfere with the analysis. We addressed this by:

i. Imputation:

For numeric features (such as square footage), we plugged in the missing values with the average (mean) of that feature. For the features that can be classified (like the quality of the garage), we put the most frequent value (mode) for them.

ii. Remaining Missing Values:

If still there are some dropping values after imputation, we will substitute them with the overall average of all the numeric features in the dataset. (This strategy is not the most usual one but the case perhaps it is needed depending on the data.)

2. Feature Engineering:

We may come up with new ideas by combining two existing features to improve the model performance. For instance, we can merge the features such as "garage area" and "garage cars" into one feature named "garage size". Moreover, we could also invent new categorical features such as "garage presence" (yes/no) or "street type" (e. g. The new method includes the following steps: (decide on a theme) based on the existing information. We will investigate these opportunities in class among other things.

3. Dummy Variables:

Categorical variables are usually hard for the machine learning models to comprehend. In order to solve this, we will transform them into formal numerical " dummy variables. A variable for "roof type" (metal, shingle, etc.) would be transformed into different binary variables that would contain either the presence or

absence of that type. This enables the model to pick up the relations between the various categories.

Through the given data wrangling steps, we do the data preparation that is largely consisted of making it nicely clean, consistent, and ready for the analysis that would later reveal the valuable insights.

Correlation Analysis

To explore relationships between variables, a correlation analysis was conducted. This measured the strength and direction of the linear association between each pair of numeric features in the data set. Strong positive correlations indicate variables tend to move together, while strong negative correlations suggest they move in opposite directions. Weaker correlations suggest little to no linear association.

To illustrate the concepts of correlation using the Ames Housing dataset, here are examples for each type of correlation:

Strong Positive Correlation

Example: GrLivArea (Above grade living area square feet) and SalePrice (Sale price of the house

Explanation: In the Ames dataset, houses with larger living areas typically sell for higher prices. This relationship shows that as the above-grade living area increases, the sale price tends to increase as well.

Correlation Coefficient: we calculated the correlation coefficient (r) for GrLivArea and SalePrice, it is +0.70678, indicating a strong positive correlation.

Strong Negative Correlation

Example: GarageAge (Age of the garage) and GarageValue (Value of the garage)

Explanation: As the garage gets older, its value generally decreases due to wear and tear, and depreciation. This shows a strong negative correlation where an increase in the age of the garage is associated with a decrease in its value.

Correlation Coefficient: If we calculate the correlation coefficient (r) for GarageAge and GarageValue, it will be close to -1, indicating a strong negative correlation.

Weak Correlation

Example: BedroomAbvGr (Number of bedrooms above ground) and LotArea (Lot size in square feet)

Explanation: The number of bedrooms in a house might not have a strong relationship with the lot size. Some houses with many bedrooms could be on small lots, while others with fewer bedrooms might be on large lots.

Correlation Coefficient: we calculated the correlation coefficient (r) for BedroomAbvGr (0.143913) and LotArea (0.266549) , indicating a weak correlation.

Correlation Coefficient Interpretation

The correlation coefficient (r) quantifies the strength and direction of the relationship:

+1: Perfect positive correlation.

-1: Perfect negative correlation.

0: No linear correlation.

Focus on Linear Relationships:

It's important to remember that correlation analysis primarily focuses on **linear** relationships. Even with a strong correlation coefficient, the relationship might not be perfectly straight. However, it suggests a general upward or downward trend when one variable changes.

Benefits of Correlation Analysis:

- **Identify Potential Relationships:** It helps uncover potential connections between variables that might not be immediately obvious.
- **Guide Further Analysis:** It can point you towards features that might be worth investigating further through other statistical methods.

- **Avoid Misinterpretations:** Correlation doesn't imply causation. Just because two variables are correlated doesn't mean one causes the other. This analysis helps identify potential associations for further exploration.

Why a Single Correlation Table is More Effective for Aims Data:

While you might consider creating separate tables for positive and negative correlations in your Aims data analysis, a single table offers a more efficient and informative approach:

- **Single Table Captures Directionality:** The correlation coefficient in the table inherently captures the direction of the relationship between two Aims variables. A value between 0 and 1 indicates a positive correlation (variables tend to move together), while a value between 0 and -1 suggests a negative correlation (variables tend to move in opposite directions).
- **Improved Comparison and Insights:** Having all correlations for Aims data displayed in one table allows for easier comparison. You can quickly assess the strength and direction of relationships between all variable pairs, gaining valuable insights into potential relationships within your Aims data set.
- **Focus on the Overall Pattern:** In real-world data, like Aims data, relationships aren't always strictly positive or negative. A single table helps you identify the overall pattern of associations, which is often more valuable than just separating positive and negative correlations.

For instance, imagine you're analyzing student performance data in Aims. A single correlation table might show a value of -0.4 at the intersection of "class attendance" and "number of absences." This tells you there's a moderate negative correlation, meaning as class attendance increases, the number of absences tends to decrease.

Creating separate tables wouldn't provide any additional information. You would still need to analyze both tables to understand the overall pattern. By using a single correlation table, you gain a more comprehensive picture of how different factors in your Aims data might be interrelated, regardless of whether the relationships are positive or negative.

Regression Analysis in Excel

This project leveraged the power of Microsoft Excel to perform regression analysis, a statistical technique that explores the relationship between a dependent variable (often what we're trying to predict) and one or more independent variables (factors influencing the dependent variable). We employed both simple linear regression, focusing on a single independent variable, and multiple linear regression, considering the combined effects of several independent variables.

To evaluate the model's effectiveness, we analyzed key metrics like R-squared, which indicates the proportion of variance in the dependent variable explained by the model. We also examined the root mean squared error (RMSE), a measure of the

prediction error, and p-values to assess the statistical significance of each independent variable. An important consideration was multicollinearity, a phenomenon where independent variables are highly correlated with each other, potentially leading to inaccurate models.

Based on these evaluations, we aimed to improve the regression models. This might involve removing insignificant variables, addressing multicollinearity through feature selection techniques, or exploring alternative modeling approaches. By iteratively refining the models, we aimed to gain a deeper understanding of the factors influencing the dependent variable and enhance the model's predictive capabilities.

Leveraging Machine Learning with RapidMiner

We transitioned from Excel's regression analysis to explore the capabilities of RapidMiner, a machine learning software platform. Here, we aimed to replicate the model developed in Excel using RapidMiner's intuitive interface and powerful algorithms. This allowed us to leverage the potential benefits of machine learning, such as handling complex relationships and potentially achieving higher accuracy.

Evaluation Metrics

Similar to our Excel analysis, we evaluated the performance of the RapidMiner model using familiar metrics like R-squared and root mean squared error (RMSE). These provided insights into the model's ability to explain the variance in the dependent variable and the average prediction error, respectively. Additionally, we explored the generated p-values to assess the statistical significance of each variable in the model.

Enhancing the Model

Based on the evaluation results, we offered suggestions for refining the machine learning model. This might involve techniques like feature selection for dimensionality reduction, hyperparameter tuning to optimize the model's performance, or exploring alternative algorithms offered by RapidMiner. By capitalizing on RapidMiner's functionalities, we aimed to further improve the model's predictive capabilities.

Both Excel's regression and RapidMiner's machine learning model aimed to predict the dependent variable. RapidMiner achieved a higher R-squared ([**RapidMiner R-squared**] vs [**Excel R-squared**]), suggesting a potentially better fit. Its lower RMSE ([**RapidMiner RMSE**]) also hinted at improved accuracy.

Interestingly, RapidMiner identified [**# of variables (RapidMiner)] statistically insignificant variables compared to Excel. Removing these could further enhance the RapidMiner model.

While RapidMiner offers advantages in performance, Excel's simplicity might be easier to interpret. Choosing between them depends on prioritizing interpretability or potentially higher accuracy.

Comparative Analysis

Let's assume we're performing a simple linear regression analysis on the Ames Housing Dataset, focusing on how year built affects sale price.

Excel Output:

- Coefficients:
    - Intercept: This represents the average sale price for houses built in year 0 (likely nonsensical in this case).
    - Year Built Coefficient: This tells you how much the sale price changes (on average) for every year difference in construction year. (Positive coefficient indicates prices increase with newer houses, negative indicates a decrease)
- R-Squared: This value indicates how well the linear model fits the data (between 0 and 1, with higher values indicating a better fit).
- P-Value: This tells you the statistical significance of the year built coefficient. (A low p-value suggests the coefficient is unlikely due to chance)
- Standard Error: This measures the variability around the coefficient estimate.

RapidMiner Output (assuming a Linear Regression Model):

- Coefficients: Similar to Excel, you'll get an intercept and a coefficient for year built.
- R-Squared: This value will be present, allowing comparison with Excel's R-squared.
- P-Value: Similar to Excel, you'll find the p-value for the year built coefficient.
- Residuals Plot: This might be available, visually showing how well the model fits the data (randomly scattered points indicate a good fit).

Additional Points for RapidMiner (depending on configuration):

- Feature Importance: This could rank the year built variable against other features in the dataset, showing its relative influence on the model's predictions.
- Model Diagnostics: This might include tests for normality of errors, multicollinearity (correlated features), or homoscedasticity (constant variance of errors).

Strategic Recommendations

Here are some potential recommendations:

- Focus on properties with features that have a strong positive correlation with sale price. This might include newer construction year (depending on market trends), larger square footage, or a specific number of bedrooms/bathrooms that is in high demand.
- Consider targeting specific neighbourhoods. Analyse areas with high average sale prices and good appreciation rates.
- Identify properties with features that have shown a stronger increase in value over time compared to others. This could be larger lot sizes, proximity to desirable amenities, or energy-efficient features.