**CW2 Summary – Linear Regression**

The Linear Regression method was implemented against the Ames Housing dataset and its results comprise this document summary. The study focused on creating a model which evaluates housing prices with multiple features.

**Dataset & Preprocessing**.

The research adopts AmesHousing.csv as its source data file. These preprocessing processes were implemented successively:

- Removed variables with more than 80% missing data.

A mode value serves as the replacement for missing categories in the feature set.

The analysis established median value as the replacement method for numeric features with missing data.

All categorical values received Label encoding as part of the preprocessing step.

The utility created a correlation heatmap to display the interconnections between different dataset variables.

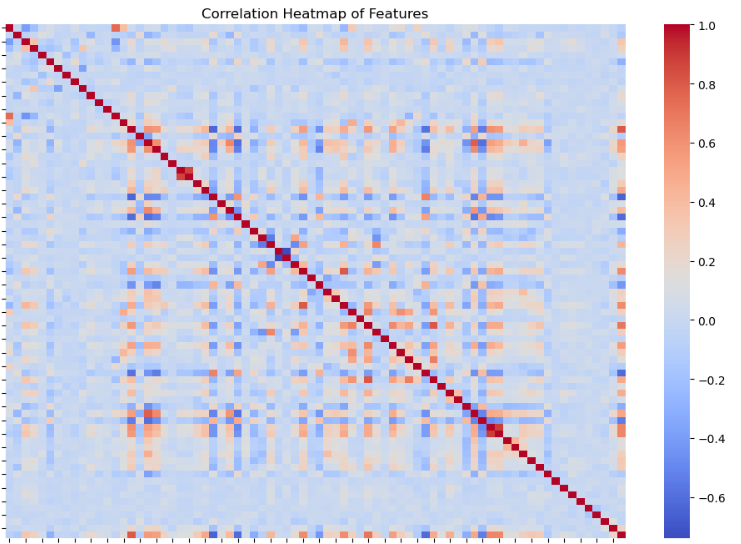
# Dataset & Preprocessing

The research uses the AmesHousing.csv file as its dataset. The following preprocessing steps were applied:  
- Removed variables with more than 80% missing data.  
- Mode value for missing categorical features.  
- Median value for missing numerical features.  
- Label encoding for categorical features.  
- Generated a correlation heatmap to visualize relationships between features.

# Model Evaluation

A correlation heatmap presents the relationships between features in the dataset design. Significant predictors for the model emerge from high correlation values between variables.

Graph 1- This image displays Feature Correlation relationships among data variables.



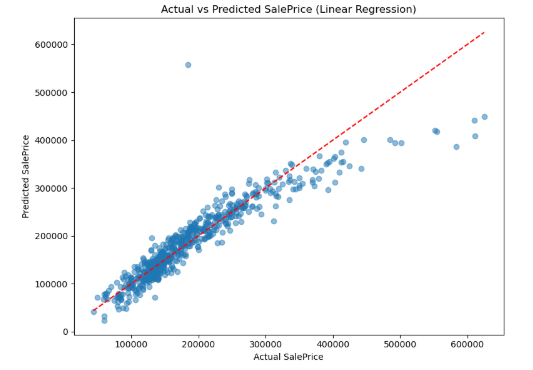
These features known as 'Overall Qual', 'Gr Liv Area' and 'Total Bsmt SF' demonstrate high positive relationships with the target outcome 'Sale Price'.

**Visualizations**

**Actual vs Predicted Sale Price**

The Linear Regression model predicted values of actual sale prices are presented in a scatter plot. The model performance quality can be determined by how close the scatter points are to the red reference line.

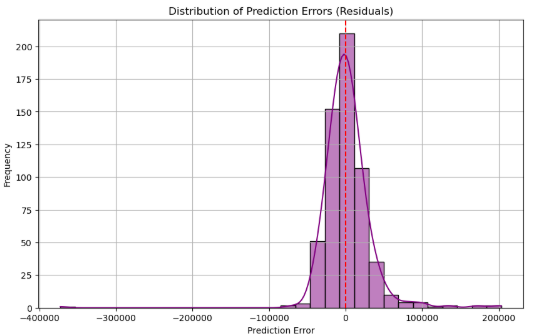
**The scatter plot for Linear Regression model displays Actual vs Predicted Sale Price relationships in Graph 2.**



**Residual Distribution**

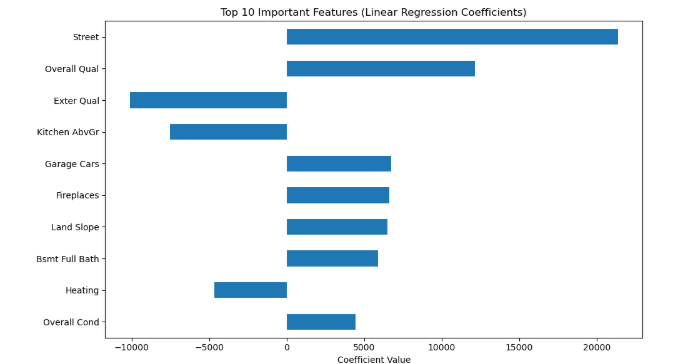
Analysis of errors from the model takes place through examining the residual distribution plot. The normal position with symmetric distribution indicates that the model shows accurate results.

The prediction error distribution of the model appears in Graph 3: Distribution of Prediction Errors (Residuals).



**These ten features represent the most influential elements according to linear regression coefficients (Linear Regression Coefficients)**

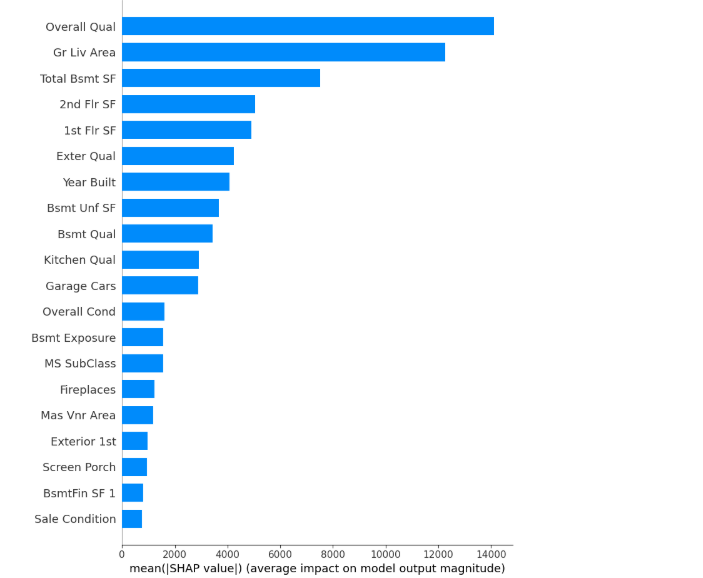
Below you can examine the bar plot that shows how the model has identified which features are its most influential components. The model's predictions receive their greatest impact from features named 'Overall Qual', 'Gr Liv Area', and 'Total Bsmt SF'.

**Graph 4 demonstrates important characteristics by evaluating linear regression model coefficients of the top ten features**.

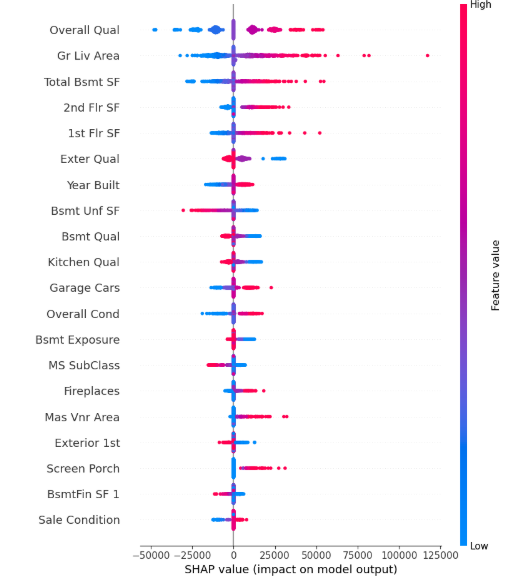
**SHAP Summary Plot**

Through the SHAP summary plot users obtain predictions analysis abilities by viewing feature contributions. The SHAP summary plot places features which significantly impact the model prediction at its top section.

**Model output features receive their influence from analysis through the SHAP Summary Plot presented in Graph 5.**



**The SHAP Summary Plot details how SHAP values distribute evenly across features in Graph 6.**



**Conclusion**

The Linear Regression model generated an R² score at 0.8611 which demonstrates its ability to accurately predict results. The main predictors of the target variable included 'Overall Qual' and 'Gr Liv Area' together with 'Total Bsmt SF'.

**Future Work**

The Linear Regression model showed strong performance yet future work opportunities exist to enhance multiple aspects of the solution.

1.The analysis should investigate new predictive models including Support Vector Machines (SVM) and Neural Networks because they would reveal intricate patterns within the dataset. By achieving proper tuning these advanced models become superior to basic linear regression models since they successfully detect non-linearities and feature correlations which earlier models struggle to process.

2.Feature selection followed by engineering processes has potential to enhance model performance when they create new meaningful features through variable interaction analysis and domain-specific transformations.

3.K-fold cross-validation with its usage allows us to overcome data overfitting problems while obtaining trustworthy model performance evaluations across distinct split systems.

4.The generalization capability of the model can be enhanced through hyperparameter tuning because it allows better regularization strengths to be found which decreases variance while improving predictive accuracy.

**References**

1. **Kaggle. (n.d.)**. Ames Housing Dataset. Retrieved from: <https://www.kaggle.com/datasets/prevek18/ames-housing-dataset>
2. **Scikit-learn documentation. (n.d.)**. LinearRegression. <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>
3. **SHAP Documentation. (n.d.)**. SHAP. <https://shap.readthedocs.io/en/latest/>