📌 Project Documentation: Feature Engineering & Model Preparation

1- Initial Data Inspection

The first step was to perform an **initial exploration** of the dataset. During this phase:

- **Checked the dataset structure**, identifying numerical and categorical features.
- **Explored missing values** and assessed their impact on model building.
- **Analyzed the target variable (SalePrice)**, observing its distribution.

2- Model Building Attempt & Challenge with Categorical Features

- Initially, a **Multiple Linear Regression** model was attempted.
- However, categorical variables prevented model training since **linear regression requires numerical features**.
- Given the presence of many categorical variables, **feature selection & transformation** were necessary.

3- Handling Categorical Features

To address the categorical feature issue:

1. **Separated categorical features** into a separate DataFrame (categorical_columns).
2. **Checked missing values** in categorical columns and imputed them as necessary.
3. **Visualized the target variable (SalePrice)** and found it to be **right-skewed**.
4. **Applied log transformation** to normalize the target variable.

4- Feature Selection Using ANOVA Test

To determine which categorical features significantly impact SalePrice, an **ANOVA test** was performed:

1. **Checked the ANOVA assumption**:
    - Each category in a categorical feature should have at least **20 observations**.
    - Identified features with **fewer than 20 observations** and **merged or removed categories** to improve reliability.
    - Used **boxplots** to explore category distributions before making adjustments.
2. **Performed ANOVA Test**:

- Calculated **F-statistic & p-value** for each categorical feature.
- Selected features that had **statistically significant differences across categories**.

5- Selected Features Based on ANOVA

From the ANOVA test, the following features were found to be most predictive:

✓ ExterQual
✓ KitchenQual
✓ BsmtQual
✓ GarageFinish
✓ GarageCond
✓ Foundation
✓ CentralAir
✓ GarageQual
✓ MasVnrType
✓ Electrical
✓ HeatingQC
✓ FireplaceQu

6- Refining Feature Selection via Correlation Analysis

To further **reduce multicollinearity**, **correlation analysis (corr function)** was used:

- Found **highly correlated features**: 'GarageFinish', 'GarageType', 'GarageCond', 'GarageQual'.
- Created a **boxplot analysis** to visualize relationships between these variables.
- Selected **GarageCond and GarageFinish** as the most relevant based on their **F-statistics**.

Final Feature Selection After Boxplot & F-Statistics Analysis: ✓ ExterQual
✓ KitchenQual
✓ BsmtQual
✓ GarageFinish
✓ GarageCond
✓ Foundation
✓ MasVnrType
✓ HeatingQC
✓ FireplaceQu

7- Handling Missing Values & Feature Engineering

- **Filled missing values for numerical features** using appropriate strategies.
- **Performed feature engineering**, including:
    - Filling GarageYrBlt using YearBuilt since they had **85% correlation**.
    - Created additional useful features to improve model performance.

8- Moving Data to SQL Server

After cleaning, transforming, and feature engineering, the dataset was **exported to SQL Server** for further analysis and modeling.

✅ **Summary of Work Done**

- **Explored dataset**, identified missing values, and prepared features.
- **Applied log transformation** to normalize the target variable.
- **Performed ANOVA & Correlation Analysis** to select meaningful categorical features.
- **Converted categorical variables to dummy variables** while reducing redundancy.
- **Engineered new features** and handled missing values.
- **Exported the final dataset to SQL Server** for modeling.