📌 **Project Documentation: Numerical Feature Selection & Final Data Preparation**

**1- Separating Numerical Features**

- In **File 1**, categorical features were selected using **ANOVA testing**.
- **File 2** focuses on selecting important **numerical features** from the dataset.
- The dataset contained both **numerical and categorical** features, so the **numerical features were separated into a new DataFrame**.

---

**2- Correlation Analysis for Feature Selection**

To determine which numerical features are important:

1. **Calculated the correlation matrix** using corr() function.
2. **Took the absolute values (abs())** to measure the strength of relationships.
3. **Kept features with correlation > 0.3** as they had a significant impact on SalePrice.

**Selected Numerical Features (Based on Correlation)**

✓ **OverallQual**
✓ **GrLivArea**
✓ **GarageCars**
✓ **GarageArea**

---

**3- Handling Highly Correlated Features**

- Found that **GarageCars and GarageArea** had a **correlation > 0.88**.
- Since they were highly collinear, one of them needed to be dropped.
- ✅ **Decided to keep GarageArea** and **remove GarageCars** to avoid multicollinearity.

---

**4- Keeping Selected Categorical Features from File 1**

- In **File 1**, categorical features were selected based on the **ANOVA test**.
- The following categorical features were retained:

✓ **ExterQual**
✓ **KitchenQual**
✓ **BsmtQual**
✓ **GarageFinish**

**5- Checking Correlation Among Categorical Features**

- Created a **correlation heatmap** for categorical features (dummy variables).
- Initially considered removing correlated categorical variables.
- However, after researching **multicollinearity in dummy variables**, realized that:
  - **Dummy variables within the same categorical feature group** are **mutually exclusive**.
  - This means that even if they are correlated, **they do not need to be removed**.

✅ **Final Decision: Kept all selected categorical features** despite correlation.

---

**6-Merging Selected Numerical & Categorical Features**

- After feature selection:
  - **Numerical features (OverallQual, GrLivArea, GarageArea)** were merged.
  - **Categorical features (ExterQual, KitchenQual, BsmtQual, GarageFinish)** were included.
- The final dataset is now **clean and ready** for **model training**.

---

✅ **Summary of Work Done in File 2**

✓ **Separated numerical features into a new DataFrame.**
✓ **Performed correlation analysis** and selected features with abs(correlation) > 0.3.
✓ **Dropped GarageCars due to high correlation with GarageArea.**
✓ **Retained categorical features selected in File 1.**
✓ **Checked correlation among categorical features (dummy variables) and confirmed they can be kept.**
✓ **Merged the final selected numerical & categorical features into one dataset.**

---

🚀 **Next Steps**

- The dataset is now **ready for model training**.
- Proceed with **building the regression model** using the selected features.