

Project Documentation: Handling Missing Values in LotFrontage

1- Understanding the Feature (LotFrontage)

- The goal was to **fill missing values** for the feature **LotFrontage**.
 - Before filling, an **exploratory data analysis (EDA)** was conducted:
 - Checked **basic statistics** of LotFrontage.
-

2- Finding Relationships Using Correlation Analysis

To determine how to fill missing values **most effectively**, correlation analysis was performed:

1. **Computed correlation (.corr())** between LotFrontage and other features.
 2. **Created a bar chart** to visualize correlation values.
 3. **Identified the two most correlated features:**
 - 1stFlrSF
 - LotArea
 4. **Checked the relationship between LotFrontage & these features.**
-

3- Removing Outliers Using Interquartile Range (IQR)

- **Why remove outliers?**
 - Outliers **distort correlation analysis**.
 - Removing them helps in **better understanding the pattern** in scatter plots.
- **Steps Taken:**
 1. **Applied IQR method** to detect and remove extreme values.
 2. **Plotted scatter plots** after removing outliers.
 3. **Checked correlation again** after removing outliers.

4- Choosing Between Spearman and Pearson Correlation

- Why compare Spearman & Pearson correlation?
 - Pearson correlation measures linear relationships.
 - Spearman correlation measures monotonic relationships (where variables move together but not necessarily linearly).
 - If the relationship is linear, Pearson is better.
 - If the relationship is monotonic, Spearman is better.
- Findings Based on the Image Provided:
 - LotFrontage vs LotArea: Pearson correlation = 0.426 → Linear relationship.
 - LotFrontage vs 1stFlrSF: Spearman correlation = 0.427 → Monotonic relationship.
 - Since the correlation values were similar, I analyzed the scatter plot and outlier impact.

✅ Final Decision:

- The relationship with 1stFlrSF was more structured and monotonic, making it a better predictor.
- Therefore, I chose 1stFlrSF to fill missing values in LotFrontage.

```
df[['LotFrontage', '1stFlrSF']].corr(method='spearman')
```

Executed at 2024.12.19 16:51:14 in 7ms

| | LotFrontage | 1stFlrSF |
|-------------|-------------|----------|
| LotFrontage | 1.000000 | 0.427678 |
| 1stFlrSF | 0.427678 | 1.000000 |


```
df[['LotFrontage', 'LotArea']].corr()
```

Executed at 2024.12.19 16:51:17 in 7ms

| | LotFrontage | LotArea |
|-------------|-------------|----------|
| LotFrontage | 1.000000 | 0.426095 |
| LotArea | 0.426095 | 1.000000 |

5- Filling Missing Values in LotFrontage

- After analyzing relationships, the best way to **fill missing values** was based on **1stFlrSF**.
- **Approach Used:**

1. **Calculated the ratio:** $\text{LotFrontage} / \text{1stFlrSF}$
2. **Computed the average ratio** from non-missing data:

```
no_outliers['Ratio'] = no_outliers['LotFrontage'] / no_outliers['1stFlrSF']
```

```
average_ratio = no_outliers['Ratio'].mean()
```

3. **Filled missing values using the ratio approach:**

```
df.loc[df['LotFrontage'].isna(), 'LotFrontage'] = df['1stFlrSF'] * average_ratio
```

- **Why this method?**
 - It maintains the **proportional relationship** between LotFrontage and 1stFlrSF.
 - It provides **more accurate estimations** based on the dataset's structure.

Summary of Work Done

- ✓ Explored LotFrontage to understand missing values.
- ✓ Analyzed correlations & identified 1stFlrSF & LotArea as key related features.
- ✓ Removed outliers using the IQR method to refine correlation analysis.
- ✓ Compared Spearman vs Pearson correlations to choose the best relationship type.
- ✓ Used the ratio approach to fill missing values based on 1stFlrSF.