Internship Weekly Report – Week 3

◈ Title Page

Name: Sandeep Ravaji Patel
Domain: Data Science
Week Number: Week 3

◈ Task Description

Objective:
To develop skills in data visualization and exploratory data analysis (EDA) using real-world datasets, focusing on generating meaningful visual insights and handling missing values effectively.

Tasks Completed:

Data Visualization:

- Plotted bar charts, histograms, and scatter plots using **Matplotlib** and **Seaborn**.
- Used plt.bar, plt.hist, sns.scatterplot, and sns.heatmap to generate clean and informative graphs.
- Compared visual outputs from different datasets (bengaluru_house_prices.csv and Pokemon.csv).

Exploratory Data Analysis (EDA):

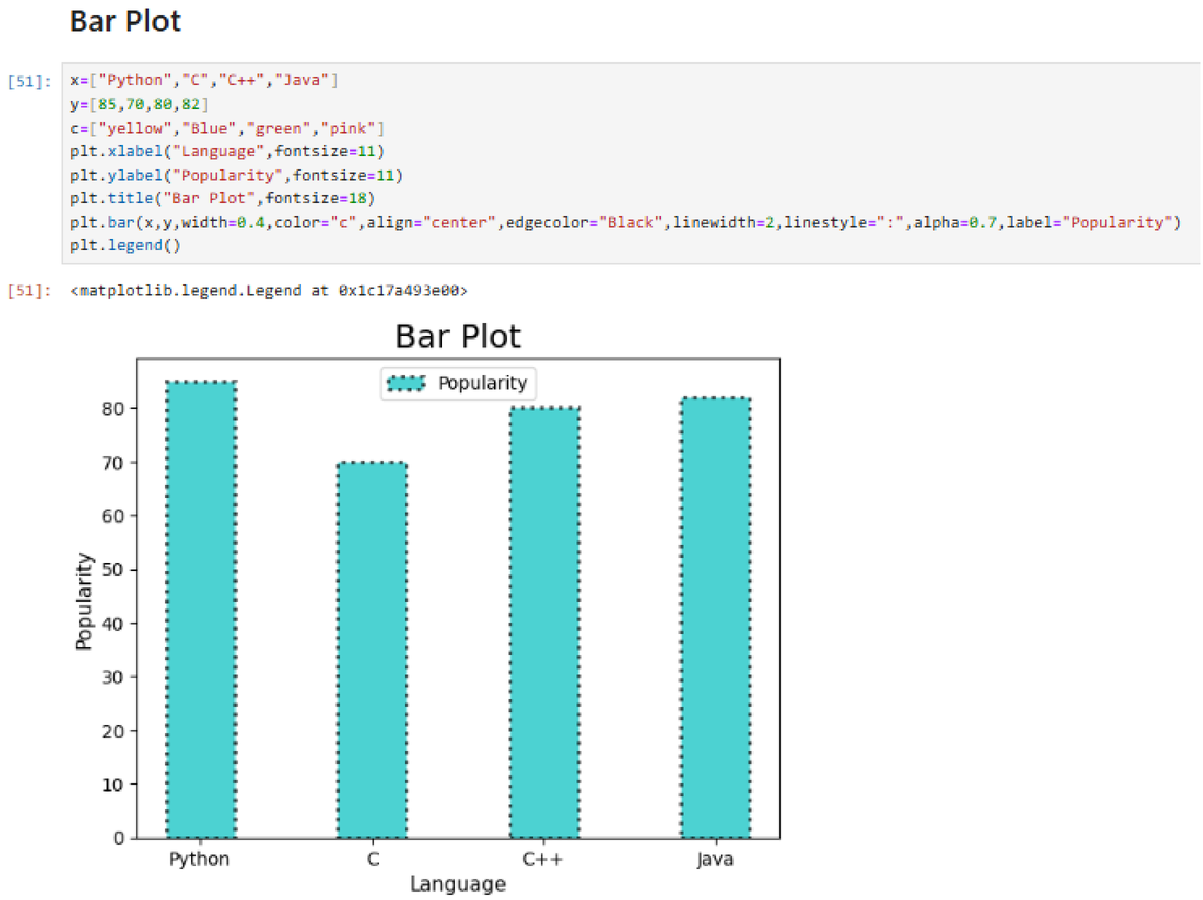
- Identified and handled missing data using isnull(), dropna(), and fillna().
- Conducted correlation analysis using corr() and visualized it with heatmaps.
- Analyzed distribution and relationship of features such as price, area, and Pokémon statistics.

Tools Used:

- **Matplotlib**
- **Seaborn**
- **Pandas**
- **Jupyter Notebook**

◈ Code Snippets / Design Screenshots

Example 1: Bar Chart of Pokémon Types

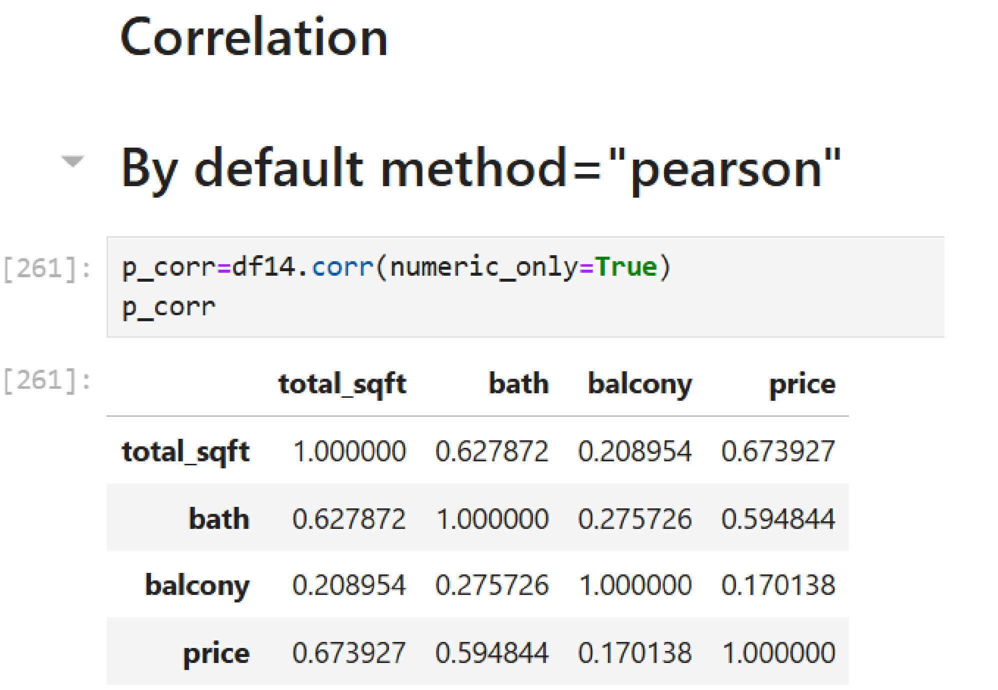


Bar Plot

```
[51]: x=["Python","C","C++","Java"]
      y=[85,70,80,82]
      c=["yellow","Blue","green","pink"]
      plt.xlabel("Language",fontsize=11)
      plt.ylabel("Popularity",fontsize=11)
      plt.title("Bar Plot",fontsize=18)
      plt.bar(x,y,width=0.4,color="c",align="center",edgecolor="Black",linewidth=2,linestyle=":",alpha=0.7,label="Popularity")
      plt.legend()

[51]: <matplotlib.legend.Legend at 0x1c17a493e00>
```

Example 2: Heatmap of Feature Correlations (Bengaluru Housing Dataset)



Correlation

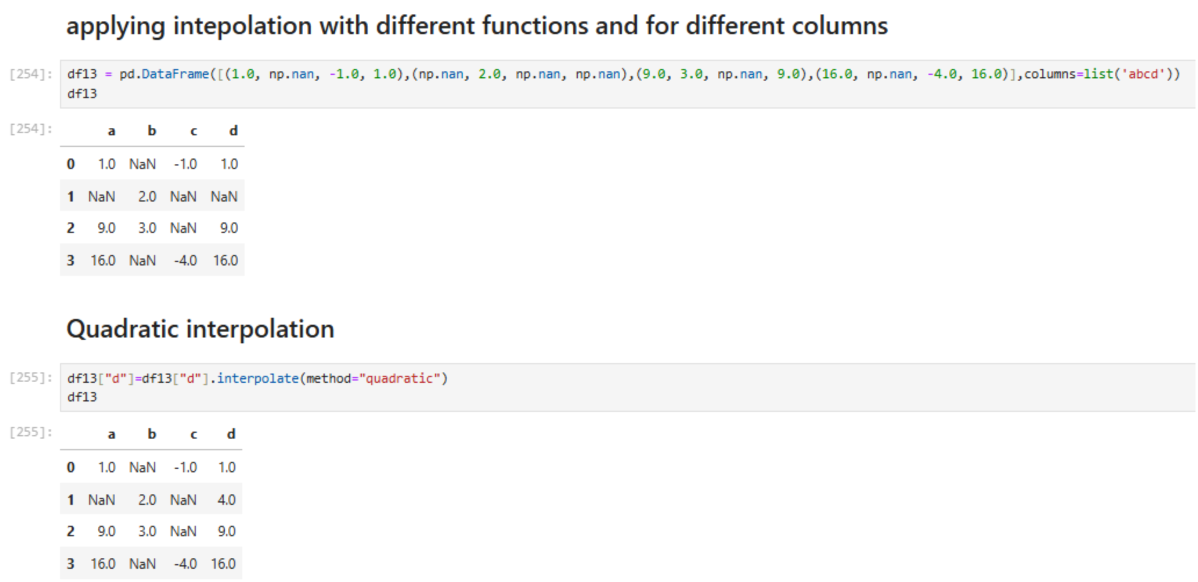
By default method="pearson"

```
[261]: p_corr=df14.corr(numeric_only=True)
       p_corr
```

[261]:

	total_sqft	bath	balcony	price
total_sqft	1.000000	0.627872	0.208954	0.673927
bath	0.627872	1.000000	0.275726	0.594844
balcony	0.208954	0.275726	1.000000	0.170138
price	0.673927	0.594844	0.170138	1.000000

Example 3: Handling Missing Data By Interpolation



applying intepolation with different functions and for different columns

[254]:

```
df13 = pd.DataFrame([(1.0, np.nan, -1.0, 1.0),(np.nan, 2.0, np.nan, np.nan),(9.0, 3.0, np.nan, 9.0),(16.0, np.nan, -4.0, 16.0)],columns=list('abcd'))
df13
```

[254]:

	a	b	c	d
0	1.0	NaN	-1.0	1.0
1	NaN	2.0	NaN	NaN
2	9.0	3.0	NaN	9.0
3	16.0	NaN	-4.0	16.0

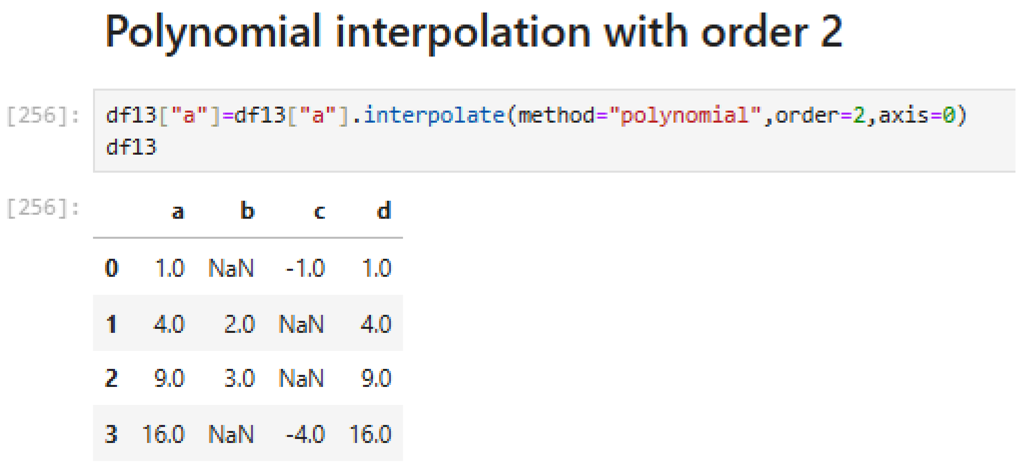
Quadratic interpolation

[255]:

```
df13["d"]=df13["d"].interpolate(method="quadratic")
df13
```

[255]:

	a	b	c	d
0	1.0	NaN	-1.0	1.0
1	NaN	2.0	NaN	4.0
2	9.0	3.0	NaN	9.0
3	16.0	NaN	-4.0	16.0



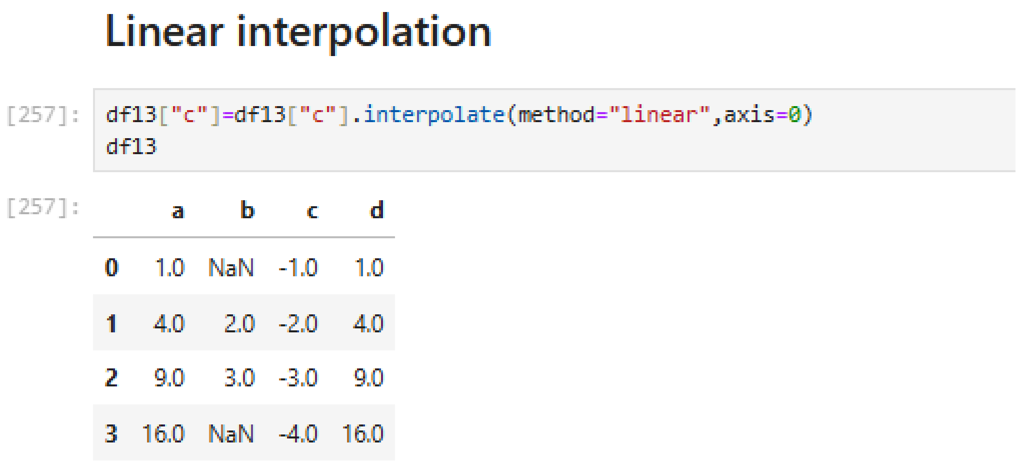
Polynomial interpolation with order 2

[256]:

```
df13["a"]=df13["a"].interpolate(method="polynomial",order=2,axis=0)
df13
```

[256]:

	a	b	c	d
0	1.0	NaN	-1.0	1.0
1	4.0	2.0	NaN	4.0
2	9.0	3.0	NaN	9.0
3	16.0	NaN	-4.0	16.0



Linear interpolation

[257]:

```
df13["c"]=df13["c"].interpolate(method="linear",axis=0)
df13
```

[257]:

	a	b	c	d
0	1.0	NaN	-1.0	1.0
1	4.0	2.0	-2.0	4.0
2	9.0	3.0	-3.0	9.0
3	16.0	NaN	-4.0	16.0

◈ Challenges Faced

Missing Data Issues:

- Some columns had extensive missing values.
- **Resolution:** Used median/mode imputation or dropped rows depending on data quality and quantity.

Inconsistent Data Types:

- Some numeric columns were read as strings due to formatting issues.
- **Resolution:** Cleaned and typecasted columns using astype(float) after cleaning symbols.

Plot Formatting and Readability:

- Legends and axis labels sometimes overlapped or were unreadable.
- **Resolution:** Added tight_layout(), rotated ticks, and adjusted figure size for clarity.

◈ Learning Outcome

Data Visualization Mastery:

- Gained confidence using Matplotlib and Seaborn to build clear and impactful visualizations.

EDA Techniques:

- Performed real-world data exploration with correlation analysis, value counts, and distribution plots.

Data Cleaning:

- Improved handling of null values and inconsistent formats across datasets.

◈ Next Steps

For **Week 4**, the focus will be on:

- **Machine Learning Basics:** Introduction to supervised models.
- **Modeling Practice:** Implementing Linear Regression.
- **Data Preparation:** Splitting data into training and testing sets.

❖ Resources

- **Matplotlib:** Matplotlib Guide
- **Seaborn:** Seaborn Documentation
- **Dataset 1:** Bengaluru House Prices – CSV
- **Dataset 2:** Pokémon Dataset – CSV