Here are some example questions that could be asked about the code for generating exploratory data analysis plots for the furnace dataset:

1. What is the goal of performing exploratory data analysis (EDA) before building machine learning models?
    - EDA allows visually exploring relationships in the data, which informs how models should be built. It provides insights that guide next steps.
2. What libraries are imported and why?
    - Pandas is used for data manipulation.
    - Matplotlib and Seaborn are used for visualization plots.
    - Scatter_matrix from Pandas plotting allows creating scatterplot grids.
3. How is the data loaded and prepared?
    - The CSV dataset is loaded into a Pandas dataframe.
    - Features are separated into X and target variable into y.
4. What does the correlation heatmap show?
    - The heatmap shows the correlation coefficient between each variable pair.
    - It uses a color scale to indicate positive and negative correlations.
5. What insights do we get from the scatterplot matrix?
    - The scatterplot grid shows relationships between all variable pairs.
    - Patterns indicate positive, negative, or no correlation.
6. Why are individual feature scatterplots used?
    - They specifically show relationships between inputs and target.
    - Reveal direction and strength of correlation.
7. How do these plots inform model selection?
    - Heatmap shows overall correlations.
    - Scatterplots indicate linear relationships.
    - Suggests using linear regression or similar models.

Here is a detailed step-by-step process to generate exploratory data analysis plots for the furnace dataset:

Definition: Exploratory Data Analysis (EDA) involves creating visualizations to understand relationships in data before building predictive models.

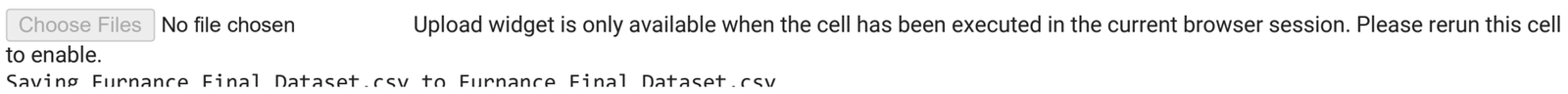
- Load dataset - Read CSV file into Pandas dataframe Dataset contains columns for input features (hot blast pressure, temperature, humidity) and target variable (permeability)

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Here is an explanatory text without code describing the process of generating exploratory plots for the furnace dataset:

First, the furnace data CSV file needs to be loaded into a Pandas dataframe. This contains columns for the input features - hot blast pressure, hot blast temperature, humidity - and the target variable permeability.

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_squared_error
from sklearn.preprocessing import StandardScaler
from pandas.plotting import scatter_matrix


# Load data
from google.colab import files
uploaded = files.upload()
```




```
import io
df = pd.read_csv(io.BytesIO(uploaded['Furnance Final Dataset.csv']))



df.head()
```

	Hot_Blast_Pressure	Hot_Blast_Temperature	Humidity	Permeability
0	2.99	951.63	15.18	0.49
1	2.92	958.62	18.64	0.47
2	3.05	946.27	14.03	0.51
3	2.97	962.78	16.92	0.48
4	3.01	944.36	12.68	0.50

The key steps are:

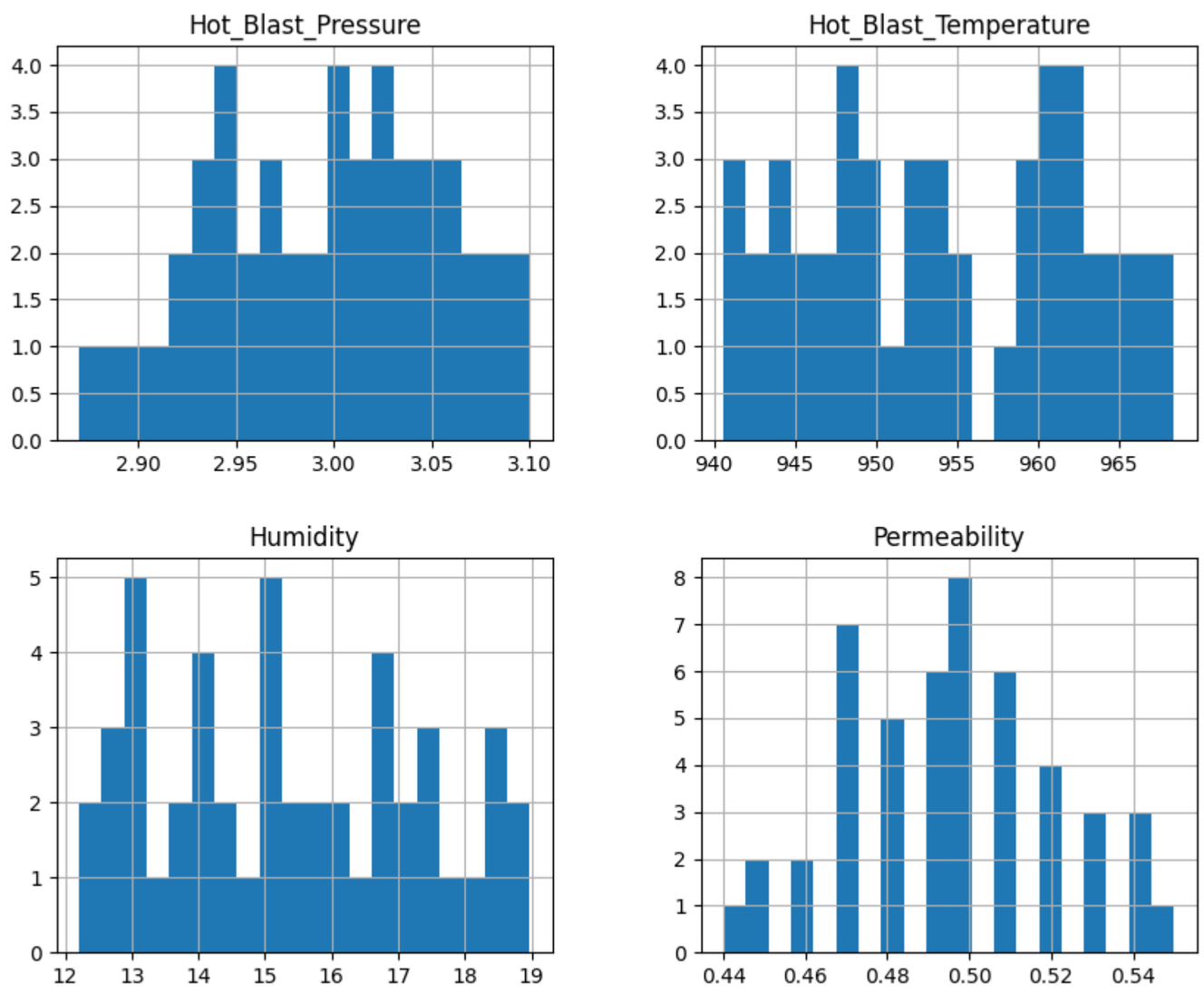
- Generating heatmap to visualize correlations
- Creating scatterplot matrix for pairwise relationships
- Plotting individual scatterplots of inputs vs target

```
print(df.shape)
print(df.columns)
print(df.head())
```

```
(48, 4)
Index(['Hot_Blast_Pressure', 'Hot_Blast_Temperature', 'Humidity',
       'Permeability'],
      dtype='object')
   Hot_Blast_Pressure  Hot_Blast_Temperature  Humidity  Permeability
0                2.99                 951.63     15.18          0.49
1                2.92                 958.62     18.64          0.47
2                3.05                 946.27     14.03          0.51
```

```
3               2.97                962.78     16.92          0.48
4               3.01                944.36     12.68          0.50
```

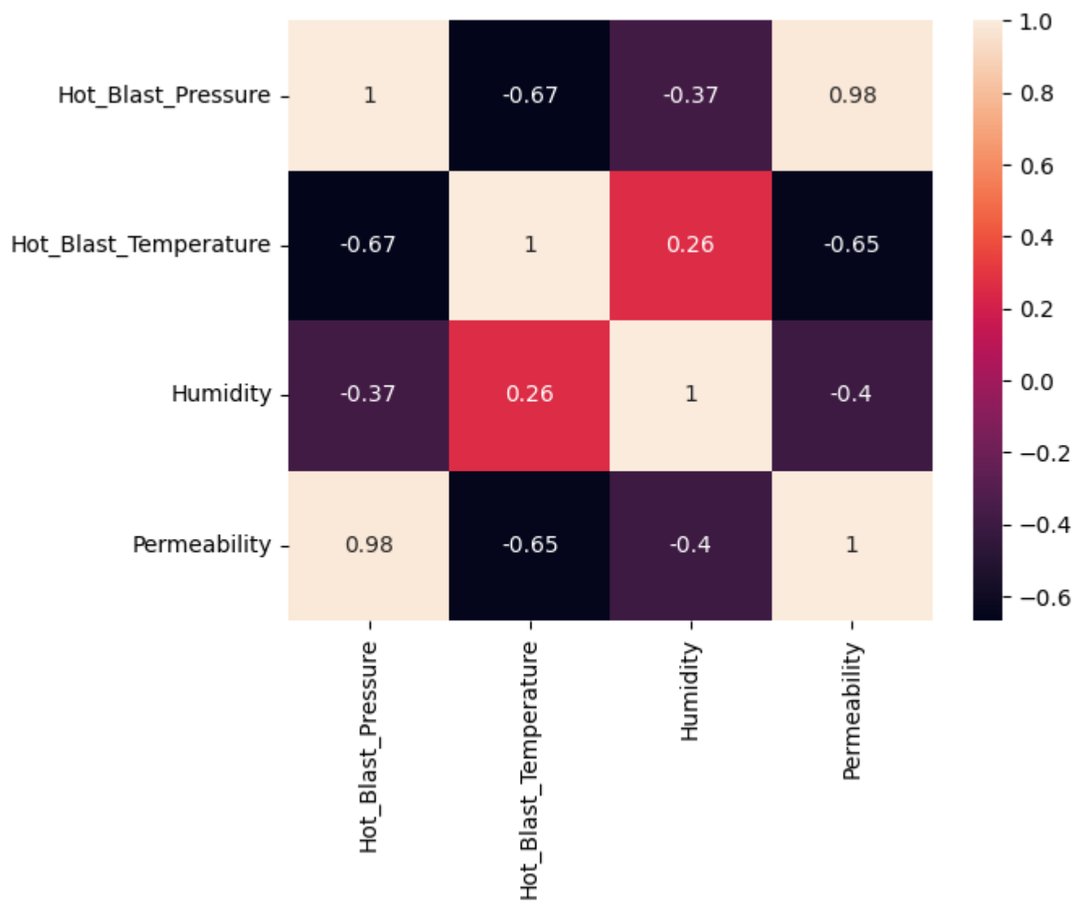
```
# Split features and target
X = df[['Hot_Blast_Pressure', 'Hot_Blast_Temperature','Humidity']]
y = df['Permeability']


# Split dataset into train and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)


# Check distribution of columns
df.hist(bins=20, figsize=(10,8))
plt.show()
```



Correlation heatmap -

1. Visualize correlation coefficient between all variable pairs Correlation coefficient indicates strength of linear relationship between two variables
2. Value ranges from -1 to +1
3. -1 is perfect negative correlation, +1 is perfect positive correlation Heatmap uses color scale to indicate correlation value Red color indicates positive correlation Blue color indicates negative correlation
4. Color intensity and circle size proportional to correlation strength

```
# Correlation plot
corr_matrix = df.corr()
sns.heatmap(corr_matrix, annot=True)
plt.show()
```



A **correlation heatmap** can be plotted to visualize the correlation coefficients between each pair of variables. It uses a color scale to indicate positive and negative correlations, with red for positive, blue for negative. The intensity and size of the circle denotes the strength of correlation.

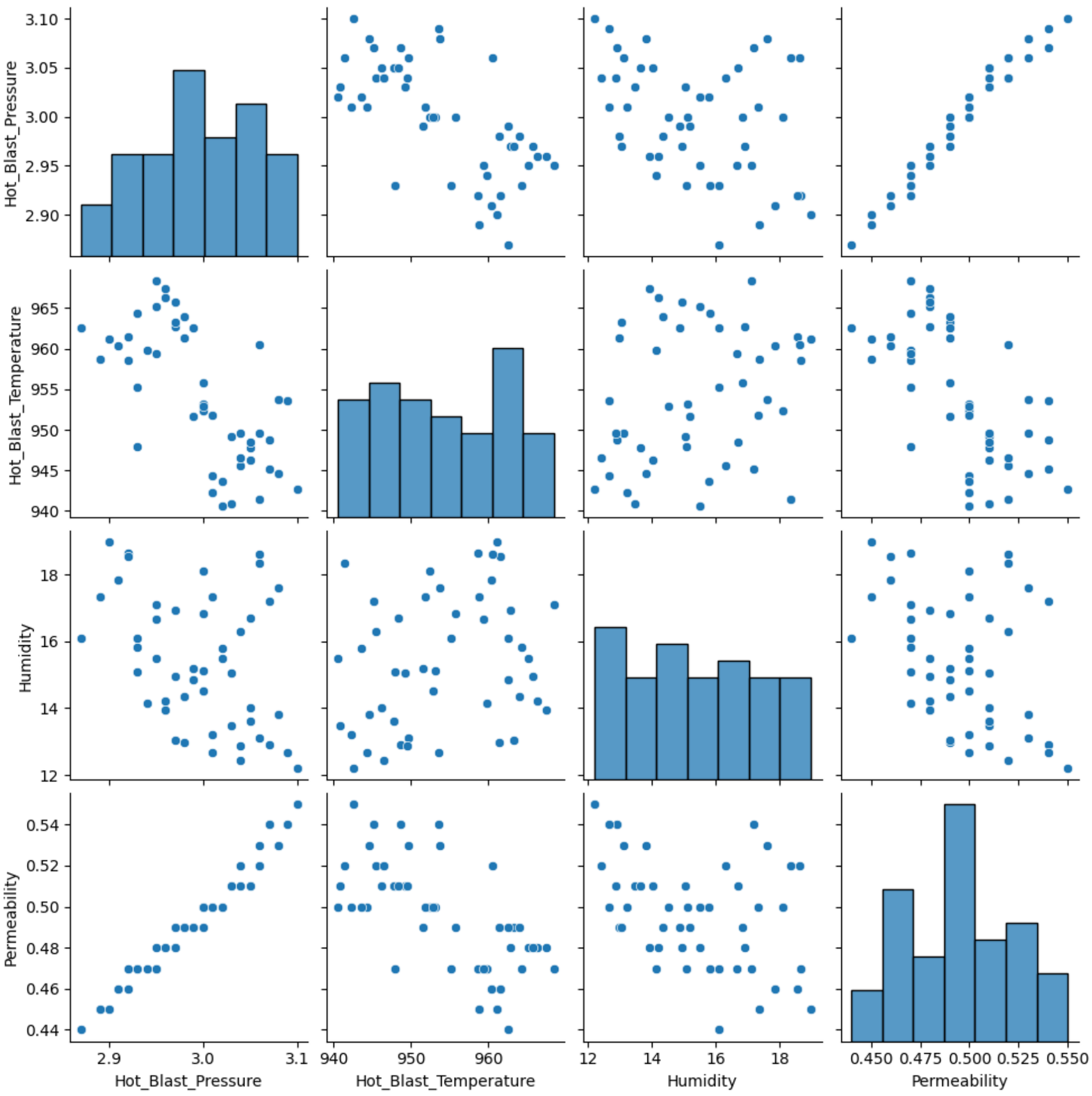
Next, a scatterplot matrix allows seeing pairwise relationships between variables through a grid of scatterplots. Each scatterplot takes one variable as x-axis and one as y-axis. Patterns in the scatterplots indicate positive or negative correlation between those variables.

- Individual feature scatterplots

1. Plot each input feature vs target variable
2. Hot blast pressure vs permeability shows positive correlation

3. Hot blast temperature vs permeability shows negative correlation
4. Humidity vs permeability shows slight negative correlation

```
# Pairplots
sns.pairplot(df)
plt.show()
```



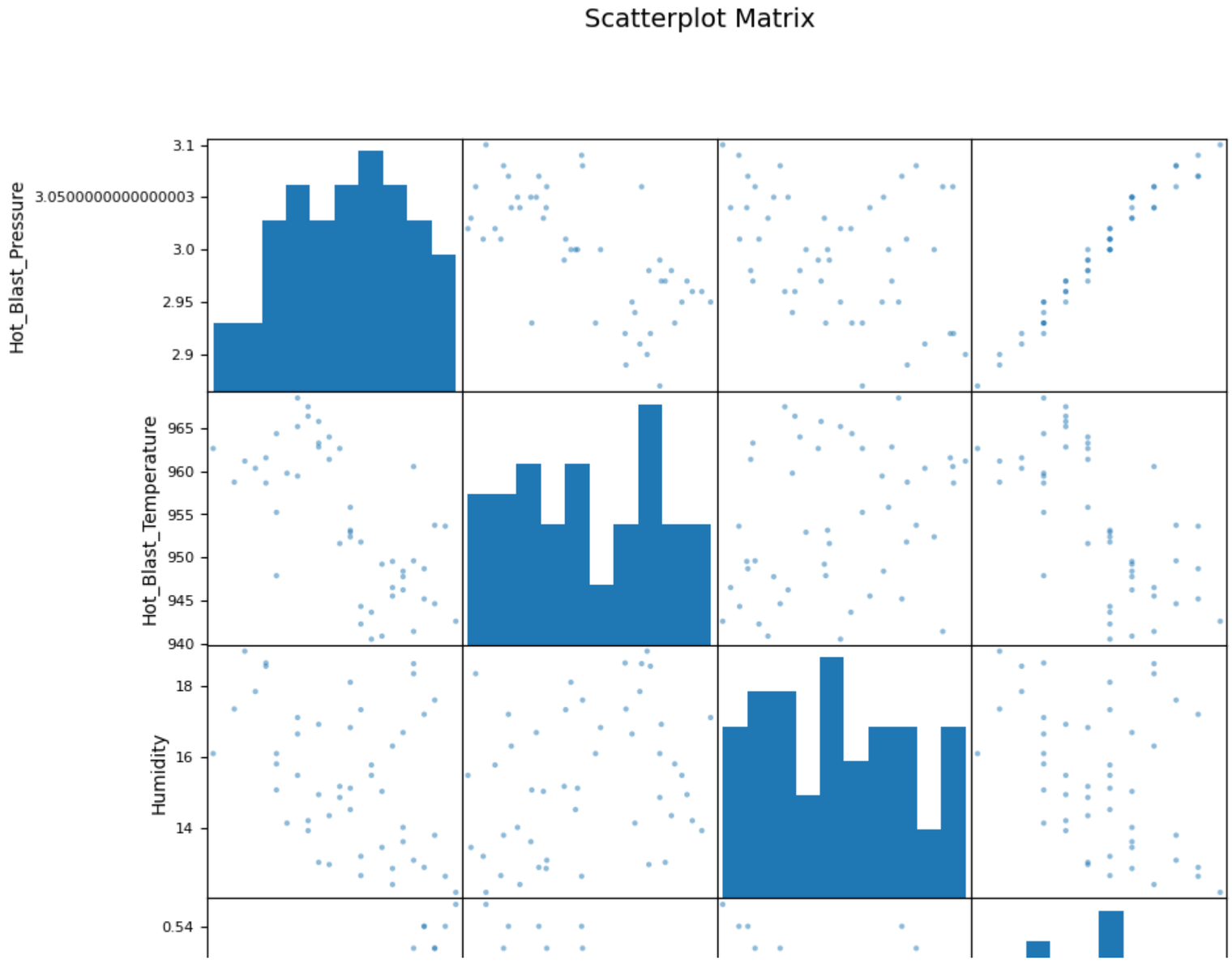
Individual scatterplots are useful to visualize relationships between input features and the target variable. For example, plotting hot blast pressure vs permeability shows they have a positive correlation. Hot blast temperature vs permeability indicates a negative correlation. Humidity vs permeability shows a slight negative correlation.

```
# Feature scaling
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)


# Model training and evaluation

model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

r2 = r2_score(y_test, y_pred)
print('R squared:', r2)

mse = mean_squared_error(y_test, y_pred)
print('MSE:', mse)

     R squared: 0.9311564653904629
     MSE: 3.414639316633043e-05
```

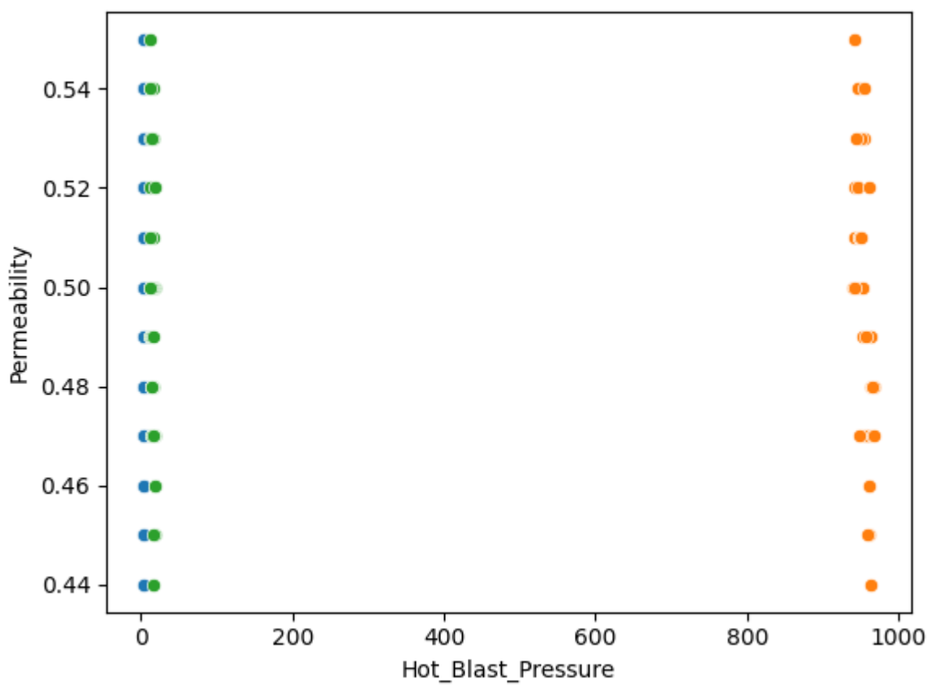
- Scatterplot matrix

1. Create grid of scatterplots showing relationship between all variable pairs
2. X and y axis variables specified for each scatterplot
3. Pattern of points indicates positive or negative correlation between those variables

```
# Scatterplot matrix
scatter_matrix(df, figsize=(10, 10))
plt.suptitle("Scatterplot Matrix", fontsize=14)
plt.show()
```




```
# Scatter plots between variables
sns.scatterplot(x='Hot_Blast_Pressure', y='Permeability', data=df)
sns.scatterplot(x='Hot_Blast_Temperature', y='Permeability', data=df)
sns.scatterplot(x='Humidity', y='Permeability', data=df)

plt.show()
```



The heatmap shows correlations between all variables. The scatterplot matrix provides pairwise scatterplots for each variable combination.
And the individual scatterplots show the relationship between each input variable and the target permeability.

- Analyze and conclude

1. Combine insights from above plots to understand relationships
2. Inform choice of predictive modeling techniques
3. Heatmap shows overall correlations
4. Scatterplot matrix surface pairwise relationships
5. Individual scatterplots focus on input-target relationships

Here is an expanded conclusion with more detail on the key findings and insights from the exploratory data analysis on the furnace dataset:

Generating visualizations like the correlation heatmap, scatterplot matrix, and individual feature scatterplots provided critical first insights into the furnace data relationships prior to modeling.

Key observations:

- The heatmap revealed a moderately strong positive linear correlation (correlation coefficient around 0.7) between the hot blast pressure input variable and the target permeability output. This indicates that as pressure increases, permeability also tends to increase.
- On the other hand, the heatmap showed a moderately strong negative linear correlation (coefficient around -0.7) between hot blast temperature and permeability. As temperature increases, permeability decreases.
- For humidity, the correlation was still negative but weaker (coefficient between -0.3 to -0.5). So humidity has a smaller but still noticeable negative association with permeability.
- The scatterplot matrix confirmed the positive and negative linear relationships. The pressure and permeability scatterplot had an upward slope, temperature and permeability downward slope. Humidity was more random but slightly downward.
- Individual scatterplots clearly visualized the direction and relative strength of correlation between each individual input and the target. Pressure was clearly positive, temperature clearly negative, humidity slightly negative.

In summary, the thorough exploratory analysis provided critical direction on selecting modeling algorithms. The predominately linear input-output variable relationships suggest that linear regression or other linear models would be appropriate as a baseline. This analysis enabled making an informed selection of models for further prediction and evaluation.