








# Data Analysis with Python

## Cheat Sheet: Exploratory Data Analysis

| Package/Method                 | Description   | Code Example  |
|--------------------------------|---|---|
| Complete dataframe correlation | Correlation matrix created using all the attributes of the dataset.   | <pre>1 df.corr()</pre>   |
| Specific Attribute correlation | Correlation matrix created using specific attributes of the dataset.  | <pre>1 df[['attribute1', 'attribute2', ...]].corr()</pre>    |
| Scatter Plot                   | Create a scatter plot using the data points of the dependent variable along the x-axis and the independent variable along the y-axis.   | <pre>1 from matplotlib import pyplot as plt 2 plt.scatter(df[['attribute_1']], df[['attribute_2']])</pre>    |
| Regression Plot                | Uses the dependent and independent variables in a Pandas data frame to create a scatter plot with a generated linear regression line for the data.  | <pre>1 import seaborn as sns 2 sns.regplot(x='attribute_1', y='attribute_2', data=df)</pre>    |
| Box plot                       | Create a box-and-whisker plot that uses the pandas dataframe, the dependent, and the independent variables.   | <pre>1 import seaborn as sns 2 sns.boxplot(x='attribute_1', y='attribute_2', data=df)</pre>    |
| Grouping by attributes         | Create a group of different attributes of a dataset to create a subset of the data.   | <pre>1 df_group = df[['attribute_1', 'attribute_2', ...]]</pre>    |
| GroupBy statements             | a. Group the data by different categories of an attribute, displaying the average value of numerical attributes with the same category.<br>b. Group the data by different categories of multiple attributes, displaying the average value of numerical attributes with the same category. | <pre>1 a) df_group = df_group.groupby(['attribute_1'], as_index=False).mean() 2 b) df_group = df_group.groupby(['attribute_1', 3 'attribute_2'], as_index=False).mean()</pre>  |

|                                 |  |  |
|---------------------------------|--|--|
| Pivot Tables                    | Create Pivot tables for better representation of data based on parameters                  | <div><div>1</div><div>grouped_pivot = df_group.pivot(index='attribute_1',columns='attribute_2')</div><div></div></div>   |
| Pseudocolor plot                | Create a heatmap image using a PsuedoColor plot (or pcolor) using the pivot table as data. | <div><div>1</div><div>from matplotlib import pyplot as plt</div><div>2</div><div>plt.pcolor(grouped_pivot, cmap='RdBu')</div><div></div></div>   |
| Pearson Coefficient and p-value | Calculate the Pearson Coefficient and p-value of a pair of attributes                      | <div><div>1</div><div>From scipy import stats</div><div>2</div><div>pearson_coef,p_value=stats.pearsonr(df['attribute_1'],</div><div>3</div><div>df['attribute_2'])</div><div></div></div> |