# 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. | `df.corr()` |
| Specific Attribute correlation | Correlation matrix created using specific attributes of the dataset. | `df[['attribute1','attribute2',...]].corr()` |
| 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. | `from matlplotlib import pyplot as`<br>`plt plt.scatter(df[['attribute_1']],df[['attribute_2']])` |
| 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. | `import seaborn as sns`<br>`sns.regplot(x='attribute_1',y='attribute_2', data=df)` |
| Box plot | Create a box-and-whisker plot that uses the pandas dataframe, the dependent, and the independent variables. | `import seaborn as sns`<br>`sns.boxplot(x='attribute_1',y='attribute_2', data=df)` |
| Grouping by attributes | Create a group of different attributes of a dataset to create a subset of the data. | `df_group = df[['attribute_1','attribute_2',...]]` |
| 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. | `a.`<br>`df_group =`<br>`df_group.groupby(['attribute_1'],as_index=False).mean()`<br>`b.`<br>`df_group = df_group.groupby(['attribute_1',`<br>`'attribute_2'],as_index=False).mean()` |
| Pivot Tables | Create Pivot tables for better representation of data based on parameters | `grouped_pivot =`<br>`df_group.pivot(index='attribute_1',columns='attribute_2')` |

| | | |
|---|---|---|
| Pseudocolor plot | Create a heatmap image using a PsuedoColor plot (or pcolor) using the pivot table as data. | 1. 1<br>2. 2<br><br>1. from matlplotlib import pyplot as plt<br>2. plt.pcolor(grouped_pivot, cmap='RdBu')<br><br>Copied! |
| Pearson Coefficient and p-value | Calculate the Pearson Coefficient and p-value of a pair of attributes | 1. 1<br>2. 2<br>3. 3<br><br>1. From scipy import stats<br>2. pearson_coef,p_value=stats.pearsonr(df['attribute_1'],<br>3. df['attribute_2'])<br><br>Copied! |