

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>df.corr()</pre>
Specific Attribute correlation	Correlation matrix created using specific attributes of the dataset.	<pre>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>from matplotlib import pyplot as plt 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>import seaborn as sns 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>import seaborn as sns 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>df_group = df[['attribute_1','attribute_2',...]]</pre>

GroupBy statements	<p>a. Group the data by different categories of an attribute, displaying the average value of numerical attributes with the same category.</p> <p>b. Group the data by different categories of multiple attributes, displaying the average value of numerical attributes with the same category.</p>	<pre>a) df_group = df.groupby(['attribute_1'],as_index=False).mean() b) df_group = df.groupby(['attribute_1','attribute_2'],as_index=False).mean()</pre>
Pivot Tables	Create Pivot tables for better representation of data based on parameters	<pre>grouped_pivot = df_group.pivot(index='attribute_1',columns='attribute_2')</pre>
Pseudocolor plot	Create a heatmap image using a PsuedoColor plot (or pcolor) using the pivot table as data.	<pre>from matplotlib import pyplot as plt plt.pcolor(grouped_pivot, cmap='RdBu')</pre>
Pearson Coefficient and p-value	Calculate the Pearson Coefficient and p-value of a pair of attributes	<pre>From scipy import stats pearson_coef,p_value=stats.pearsonr(df['attribute_1'],df['attribute_2'])</pre>



Skills Network