

Data Analysis with Python

Cheat Sheet: Data Wrangling

Package/Method	Description	Code Example
Replace missing data with frequency	Replace the missing values of the data set attribute with the mode common occurring entry in the column.	<div><div>1. 1</div><div>2. 2</div><div>1. MostFrequentEntry = df['attribute_name'].value_counts().idxmax()</div><div>2. df['attribute_name'].replace(np.nan,MostFrequentEntry,>df['attribute_name'].replace(np.nan,MostFrequentEntry, inplace=True)</div></div> <div>Copied!</div>
Replace missing data with mean	Replace the missing values of the data set attribute with the mean of all the entries in the column.	<div><div>1. 1</div><div>2. 2</div><div>1. AverageValue=df['attribute_name'].astype(<data_type>).mean(axis=0)</div><div>2. df['attribute_name'].replace(np.nan, AverageValue, inplace=True)</div></div> <div>Copied!</div>
Fix the data types	Fix the data types of the columns in the dataframe.	<div><div>1. 1</div><div>2. 2</div><div>3. 3</div><div>1. df[['attribute1_name', 'attribute2_name', ...]] =</div><div>2. df[['attribute1_name', 'attribute2_name', ...]].astype('data_type')</div><div>3. #data_type is int, float, char, etc.</div></div> <div>Copied!</div>
Data Normalization	Normalize the data in a column such that the values are restricted between 0 and 1.	<div><div>1. 1</div><div>1. df['attribute_name'] =</div><div>df['attribute_name']/df['attribute_name'].max()</div></div> <div>Copied!</div>
Binning	Create bins of data for better analysis and visualization.	<div><div>1. 1</div><div>2. 2</div><div>3. 3</div><div>4. 4</div><div>5. 5</div><div>6. 6</div><div>1. bins = np.linspace(min(df['attribute_name']),</div><div>2. max(df['attribute_name']),n)</div><div>3. # n is the number of bins needed</div><div>4. GroupNames = ['Group1','Group2','Group3,...]</div><div>5. df['binned_attribute_name'] =</div><div>6. pd.cut(df['attribute_name'], bins, labels=GroupNames, include_lowest=True)</div></div> <div>Copied!</div>
Change column name	Change the label name of a dataframe column.	<div><div>1. 1</div><div>1. df.rename(columns={'old_name':'new_name'}, inplace=True)</div></div> <div>Copied!</div>
Indicator Variables	Create indicator variables for categorical data.	<div><div>1. 1</div><div>2. 2</div><div>1. dummy_variable = pd.get_dummies(df['attribute_name'])</div><div>2. df = pd.concat([df, dummy_variable],axis = 1)</div></div> <div>Copied!</div>

