8/12/24, 7:43 PM about:blank

Data Analysis with Python

Cheat Sheet: Data Wrangling

```
Package/Method Description
                                                                                                                  Code Example
                     Replace the
                     missing
                                         1. 1
2. 2
                     values of the
                     data set
Replace missing
                     attribute with

    MostFrequentEntry = df['attribute_name'].value_counts().idxmax()
    df['attribute_name'].replace(np.nan,MostFrequentEntry,>df['attribute_name'].replace(np.nan,MostFrequentEntry, inpla

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



about:blank 1/1