Data Analysis with Python

Cheat Sheet: Data Wrangling

```
Package/Method Description
                                                                                                  Code Example
                  Replace the
                  missing
                  values of the
                                    1. 1
                   data set
Replace missing
                  attribute with
                                    1. MostFrequentEntry = df['attribute name'].value counts().idxmax()
data with
                  the mode
                                    2. df['attribute_name'].replace(np.nan,MostFrequentEntry,>df['attribute_name'].replace(np.nan,MostFrequentEntry, inplace=T
frequency
                  common
                  occurring
                                  Copied!
                  entry in the
                  column.
                  Replace the
                  missing
                                    1. 1
2. 2
                   values of the
                   data set
Replace missing
                  attribute with

    AverageValue=df['attribute_name'].astype(<data_type>).mean(axis=0)

data with mean
                                    df['attribute_name'].replace(np.nan, AverageValue, inplace=True)
                  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'] =
   df['attribute_name']/df['attribute_name'].max()
Data
                  that the
Normalization
                  values are
                  restricted
                                  Copied!
                  between 0
                  and 1.
                                    4.4
                  Create bins
                  of data for

    bins = np.linspace(min(df['attribute_name']),

Binning
                  better
                                    2. max(df['attribute_name'],n)
                  analysis and
                                    3. # n is the number of bins needed4. GroupNames = ['Group1','Group2','Group3,...]
                  visualization.
                                    5. df['binned_attribute_name']
                                    pd.cut(df['attribute_name'], bins, labels=GroupNames, include_lowest=True)
                                  Copied!
                  Change the
                                    1. 1
                  label name
Change column
                   of a

    df.rename(columns={'old_name':\'new_name'}, inplace=True)

name
                  dataframe
                                  Copied!
                  column.
                   Create
                  indicator
Indicator
                                    1. dummy_variable = pd.get_dummies(df['attribute_name'])
                  variables for
Variables
                                    2. df = pd.concat([df, dummy_variable],axis = 1)
                  categorical
                  data.
                                  Copied!
```

