The Engineering World #DataScience 20 & 21

May 31, 2018

AKKAL BAHADUR BIST
DATA SCIENTIST AT
KATHMANDU INSTITUTE OF APPLIED SCIENCES (KIAS)
Center for Conservation Biology (CCB)

1 OUTLIER ANALYSIS DETECTION WITH UNIVARIATE METHOD USING TUKEY BOXPLOTS

1.0.1 Extreme value analysis using Univariate Methods

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        from pylab import rcParams

In [2]: %matplotlib inline
        rcParams['figure.figsize'] = 5, 4

In [3]: df = pd.read_csv('iris_data_nepal.csv', header = None, sep = ',')
        df.columns = ['Special Length', 'Special Width', 'Petal Length', 'Petal Width', 'Species
        X = df.iloc[:,0:4].values
        y = df.iloc[:,4].values
        df[:5]


        ---------------------------------------------------------------------------

        UnicodeDecodeError                        Traceback (most recent call last)

        pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader._convert_tokens()


        pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader._convert_with_dtype()


        pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader._string_convert()
```

```
    pandas/_libs/parsers.pyx in pandas._libs.parsers._string_box_utf8()

    UnicodeDecodeError: 'utf-8' codec can't decode byte 0xff in position 0: invalid start by

During handling of the above exception, another exception occurred:

    UnicodeDecodeError                        Traceback (most recent call last)

    <ipython-input-3-815e685eb497> in <module>()
----> 1 df = pd.read_csv('iris_data_nepal.csv', header = None, sep = ',')
      2 df.columns = ['Special Length', 'Special Width', 'Petal Length', 'Petal Width', 'Spe
      3 X = df.iloc[:,0:4].values
      4 y = df.iloc[:,4].values
      5 df[:5]

    ~/anaconda3/lib/python3.6/site-packages/pandas/io/parsers.py in parser_f(filepath_or_buf
    707                     skip_blank_lines=skip_blank_lines)
    708
--> 709         return _read(filepath_or_buffer, kwds)
    710
    711     parser_f.__name__ = name

    ~/anaconda3/lib/python3.6/site-packages/pandas/io/parsers.py in _read(filepath_or_buffer
    453
    454     try:
--> 455         data = parser.read(nrows)
    456     finally:
    457         parser.close()

    ~/anaconda3/lib/python3.6/site-packages/pandas/io/parsers.py in read(self, nrows)
   1067                 raise ValueError('skipfooter not supported for iteration')
   1068
-> 1069         ret = self._engine.read(nrows)
   1070
   1071         if self.options.get('as_recarray'):

    ~/anaconda3/lib/python3.6/site-packages/pandas/io/parsers.py in read(self, nrows)
   1837     def read(self, nrows=None):
   1838         try:
-> 1839             data = self._reader.read(nrows)
```

```
1840         except StopIteration:
1841             if self._first_chunk:


 pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader.read()


 pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader._read_low_memory()


 pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader._read_rows()


 pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader._convert_column_data()


 pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader._convert_tokens()


 pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader._convert_with_dtype()


 pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader._string_convert()


 pandas/_libs/parsers.pyx in pandas._libs.parsers._string_box_utf8()


 UnicodeDecodeError: 'utf-8' codec can't decode byte 0xff in position 0: invalid start by
```

1.0.2 Identifying Outlier from Tukey boxplots

```
In [ ]: df.boxplot(return_type = 'dict')
        plt.plot()
```

```
In [ ]: Sepal_Width = X[:,1]
        iris_outliers = (Sepal_Width > 4)
        df(iris_outliers)
```

```
In [ ]: Sepal_Width = X[:,1]
        iris_outliers = (Sepal_Width < -.25)
        df(iris_outliers)
```

1.0.3 Applying Tukey outlier labeling

```
In [ ]: pd.options.display.float_format = '{:.1f}'.format
        X_df = pd.DataFrame(X)
        print x_df.describe()
```

2 MULTIVARIATE OUTLIER ANALYSIS DETECTION

2.0.1 Visually inspecting boxplots

```
In [ ]: df = pd.read_csv('iris_data_nepal.csv', header = None, sep = ',')
        df.columns = ['Special Length', 'Special Width', 'Petal Length', 'Petal Width', 'Species
        X = df.iloc[:,0:4].values
        y = df.iloc[:,4].values
        df[:5]
```

2.0.2 Looking at the scatterplot matrix

```
In [ ]: sb.boxplot(x ='Species', y = 'Sepal Length', data = df, palatte = 'hls')
```

```
In [ ]: sb.pairplot(df, hue = 'Species', platte = 'hls')
```