Cycle Data clean part 02-DBConnercct

April 27, 2018

1 Cycle Data Cleaning part 2

Declaration : The coding is abstract from Kevin mark ham youtube video seriese, Introduction to machine learning with scikit-learn video series. You can find link under resources section.

What are the **features**? - trip_id: A unique number to identify each trip

- From station Number: From station number where the trip Start
- Day: Day of the trip for example Monday, Tuesday etc.
- Month: Which month trip took place
- Duration: Total trip duration in minutes
- birthyear: Birth year of user
- Sex: Gender identification of user
- age: Current age of user

What is the **response**? - Station Number: To Station Number where the trip ends

2 Filling Bike Data with Weather Values

Weather data is consist on 600 observations where bikes data is consisting over a hundred thousand of tuples which mean it is not easy to combine. To achieve this we will, break both weather and bikes data Date columns and applying programming technique to achieve weather value for each bicycle trip.

3 Data Cleaning

```
In [1]: # load libraries and set styles, options
        import os,csv
        import numpy as np
        import pandas as pd
        import seaborn as sns
        import warnings; warnings.simplefilter('ignore')
        #from IPython.display import HTML
        #HTML('<iframe src=http://www.seattle.gov/documents/departments/sdot/newmobilityprogra
In [2]: %matplotlib inline
```

2. Read and verify data

```
In [3]: # read in a CSV
        df1 = pd.read_csv('C:/Users/mrferozi/Desktop/GitHub/Bike/dataset/cycle/weather_clean.cs
        df2 = pd.read_csv('C:/Users/mrferozi/Desktop/GitHub/Bike/dataset/cycle/trip_clean.csv'
```

```
In [4]: df1.dtypes
```

```
Out[4]: Date                           object
        Max_Temperature_F               int64
        Mean_Temperature_F              int64
        Min_TemperatureF                int64
        Max_Dew_Point_F                 int64
        MeanDew_Point_F                 int64
        Min_Dewpoint_F                  int64
        Max_Humidity                    int64
        Mean_Humidity                   int64
        Min_Humidity                    int64
        Max_Sea_Level_Pressure_In     float64
        Mean_Sea_Level_Pressure_In    float64
        Min_Sea_Level_Pressure_In     float64
        Max_Visibility_Miles            int64
        Mean_Visibility_Miles           int64
        Min_Visibility_Miles            int64
        Max_Wind_Speed_MPH              int64
        Mean_Wind_Speed_MPH             int64
        Max_Gust_Speed_MPH            float64
        Precipitation_In              float64
        Events                         object
        Mean_Temperature_C              int64
        Events_num                      int64
        month                           int64
        year                            int64
        dtype: object
```

```
In [5]: df2.dtypes
```

```
Out[5]: trip_id                   int64
        starttime                object
        stoptime                 object
        bikeid                   object
        tripduration            float64
        from_station_name        object
        to_station_name          object
        from_station_id          object
        to_station_id            object
        usertype                 object
```

```
gender                   object
birthyear                 int64
Sex_num                 float64
from_station_id_cat      object
from_station_id_num       int64
to_station_id_cat        object
to_station_id_num         int64
Day                      object
Day_cat                  object
Day_num                   int64
sthours                   int64
stphours                  int64
tripduration_minutes    float64
age                       int64
bmonth                    int64
Date                     object
year                      int64
dtype: object
```

```
In [6]:  # convert 'Time' to datetime format
        df1['Date'] = pd.to_datetime(df1.Date)
        df2['Date'] = pd.to_datetime(df2.Date)
```

```
In [7]: df1.dtypes
```

```
Out[7]: Date                          datetime64[ns]
        Max_Temperature_F                      int64
        Mean_Temperature_F                     int64
        Min_TemperatureF                       int64
        Max_Dew_Point_F                        int64
        MeanDew_Point_F                        int64
        Min_Dewpoint_F                         int64
        Max_Humidity                           int64
        Mean_Humidity                          int64
        Min_Humidity                           int64
        Max_Sea_Level_Pressure_In            float64
        Mean_Sea_Level_Pressure_In           float64
        Min_Sea_Level_Pressure_In            float64
        Max_Visibility_Miles                   int64
        Mean_Visibility_Miles                  int64
        Min_Visibility_Miles                   int64
        Max_Wind_Speed_MPH                     int64
        Mean_Wind_Speed_MPH                    int64
        Max_Gust_Speed_MPH                   float64
        Precipitation_In                     float64
        Events                                object
        Mean_Temperature_C                     int64
        Events_num                             int64
```