CS 487/519 Applied Machine Learning I

Project 2: Open ML project

October 05, 2018

Jithin Jacob Benjamin Jacob – 800681973
Joanna Augustine – 800656114
Group 9

MOTIVATION:

The key motivation behind this project is that carbon emissions contribute to climate change, which can have serious consequences for humans and their environment. According to the U.S. Environmental Protection Agency, carbon emissions, in the form of carbon dioxide, make up more than 80 percent of the greenhouse gases emitted in the United States. The burning of fossil fuels releases carbon dioxide and other greenhouse gases. These carbon emissions raise global temperatures by trapping solar energy in the atmosphere. So in this project we analyze a data on carbon footprint of individual's and using this data and machine learning techniques develop an algorithm that minimizes the carbon footprint of each individual while maintaining their quality of life.

PROBLEM DEFINITION:

The problem of high carbon footprint in the environment is a major issue and has a great influence in the climatic changes in the world. So this problem is kept forth by Wells Fargo whose high priority is to promote environmental sustainability. As an initiative they have produced a data containing all the daily activities of individual customers. We are to analysis on the daily activities of individuals customers in a way to accelerate the transition to a low-carbon economy. This has to be achieved without compromising on their daily priorities and needs. They believe that taking individual actions can encourage the collective responsibility to achieve this goal. So using Machine Learning we are to develop a data product that would help in analyzing the data and help individuals to optimize the balance between their carbon footprint and the quality of life.

The ultimate goal would be to recommend an environment friendly change to the everyday actions without lessening the individuals' quality of life. The data gives a peak into the lives of 1,000 individuals who rated several everyday activities (taking a long shower, driving a car, etc.) on a scale of 1-100 based on how important those activities are to their daily lives. So at the end the data product should produce a computer data program to find quality substitutes for activities that are high carbon emitters without reducing the happiness and utility that the individuals in the data obtain from these activities.

PROPOSED SOLUTION:

The solution that we are expecting from this analytical process is to refine the dataset in such a way that we can perform the basic three operations of loading the data, cleaning the data and thereby using the machine learning techniques to join the data which makes more relevance to the problem under discussion. So by the end of the problem we will get a more refined information on the dataset and thereby help in deciding on the alternatives that can be considered in order to arrive at our solution of low carbon footprint by the individuals.

LINK TO DATASET:

https://www.mindsumo.com/contests/campus-analytics-challenge-2018?&utm_campaign=send_drip_email&utm_source=mindsumo&utm_medium=email

(Data set is also placed in our Github repository inside the stage3 folder)

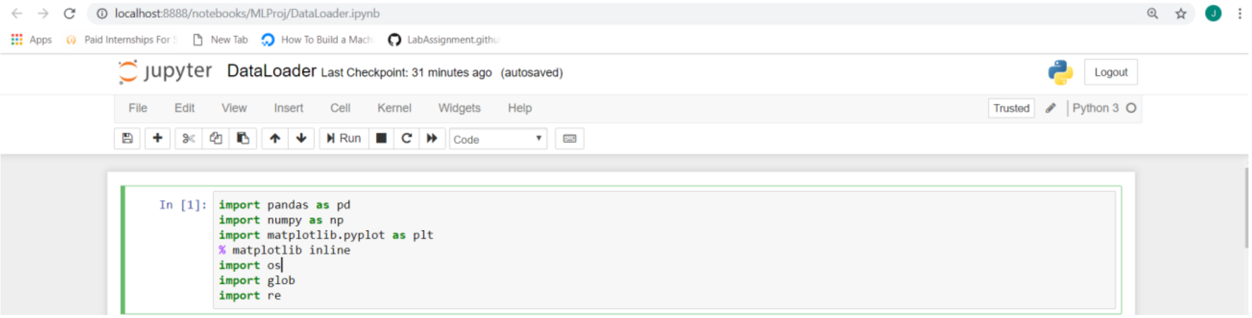
PROGRAM AND OUTPUTS:

(LOADING DATA)

DataLoader.py

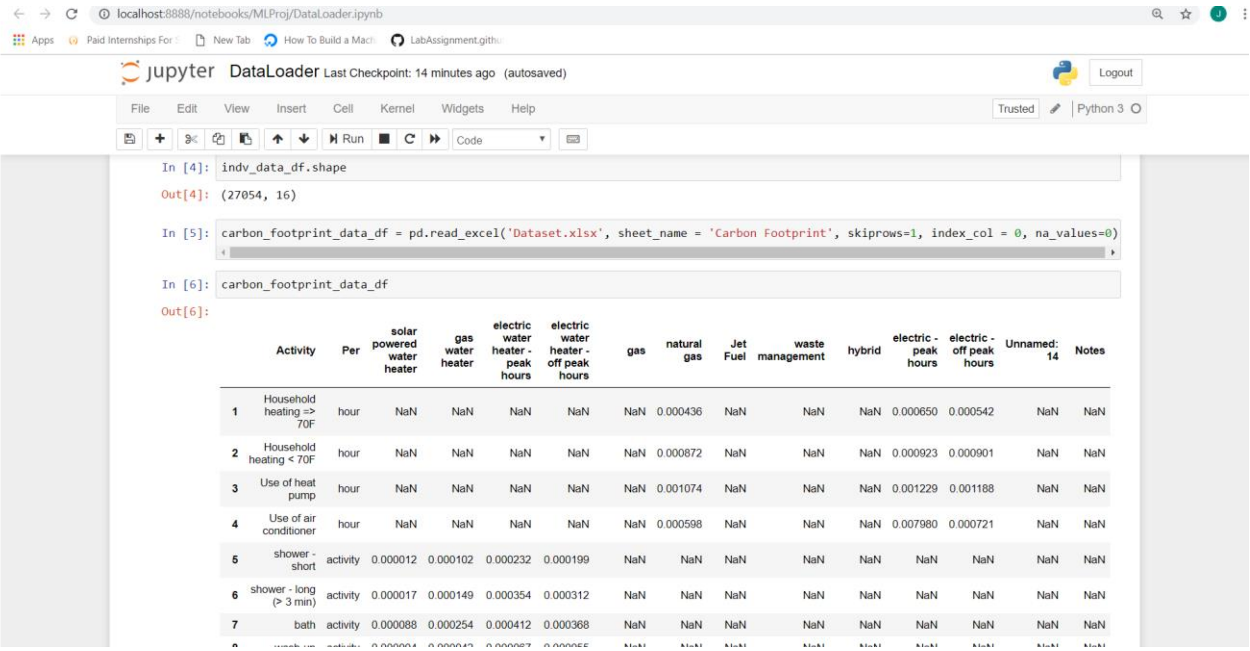
This jupyter notebook program is responsible to load the data from the excel file from both the sheets and just clean it up a little bit and save the data into csv files which is easier for loading in further analysis.

FIGURE 1: Importing the required libraries



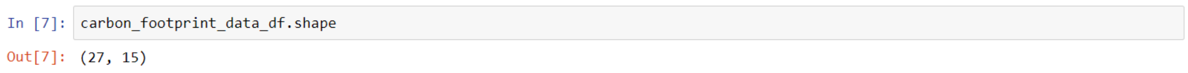
```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        % matplotlib inline
        import os
        import glob
        import re
```

FIGURE 2: Loading the Data from the Excel File

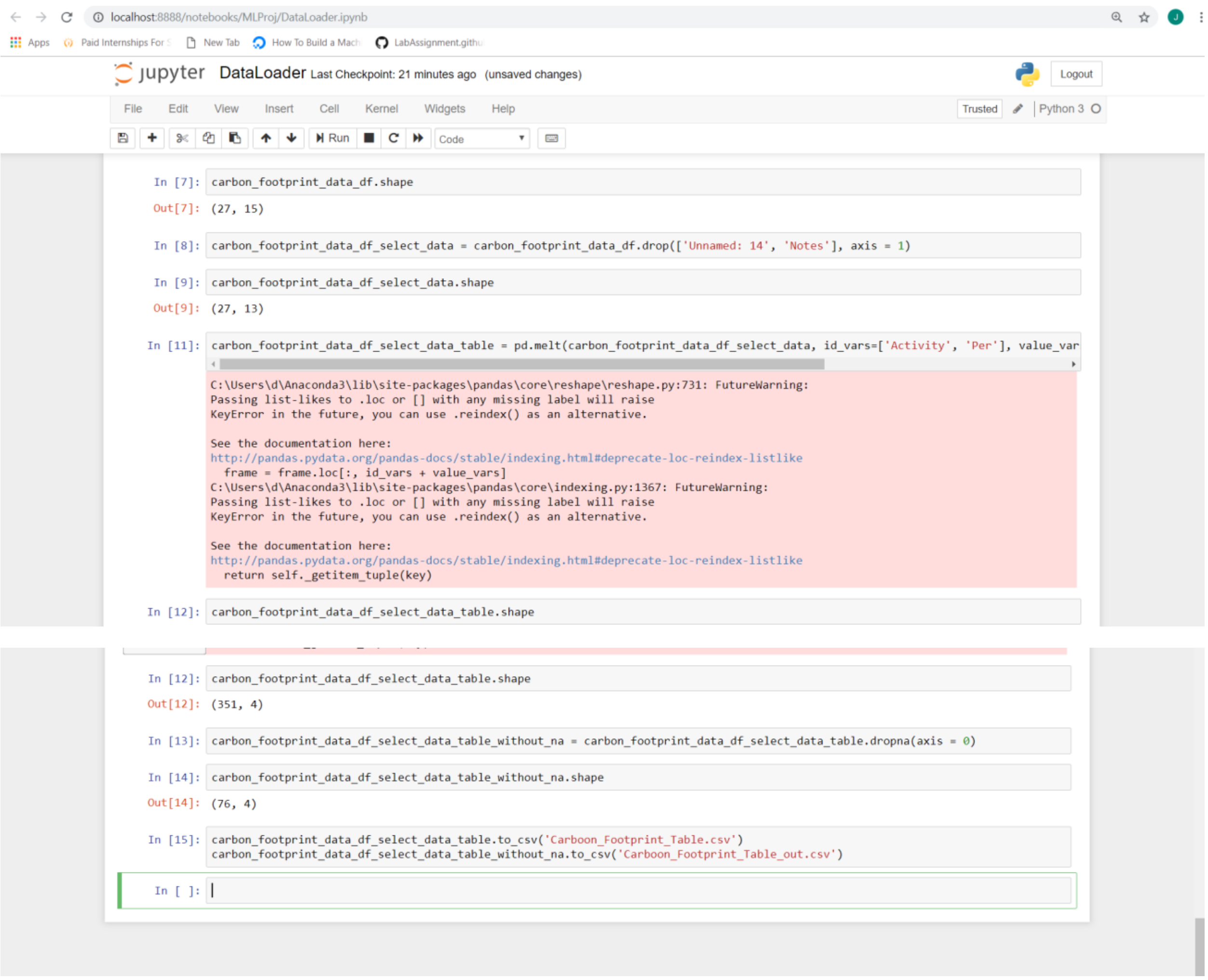


```
In [4]: indv_data_df.shape

Out[4]: (27054, 16)

In [5]: carbon_footprint_data_df = pd.read_excel('Dataset.xlsx', sheet_name = 'Carbon Footprint', skiprows=1, index_col = 0, na_values=0)

In [6]: carbon_footprint_data_df

Out[6]:
```

	Activity	Per	solar powered water heater	gas water heater	electric water heater - peak hours	electric water heater - off peak hours	gas	natural gas	Jet Fuel	waste management	hybrid	electric - peak hours	electric - off peak hours	Unnamed: 14	Notes
1	Household heating => 70F	hour	NaN	NaN	NaN	NaN	NaN	0.000436	NaN	NaN	NaN	0.000650	0.000542	NaN	NaN
2	Household heating < 70F	hour	NaN	NaN	NaN	NaN	NaN	0.000872	NaN	NaN	NaN	0.000923	0.000901	NaN	NaN
3	Use of heat pump	hour	NaN	NaN	NaN	NaN	NaN	0.001074	NaN	NaN	NaN	0.001229	0.001188	NaN	NaN
4	Use of air conditioner	hour	NaN	NaN	NaN	NaN	NaN	0.000598	NaN	NaN	NaN	0.007980	0.000721	NaN	NaN
5	shower - short	activity	0.000012	0.000102	0.000232	0.000199	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	shower - long (> 3 min)	activity	0.000017	0.000149	0.000354	0.000312	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	bath	activity	0.000088	0.000254	0.000412	0.000368	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	wash up	activity	0.000004	0.000042	0.000067	0.000055	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

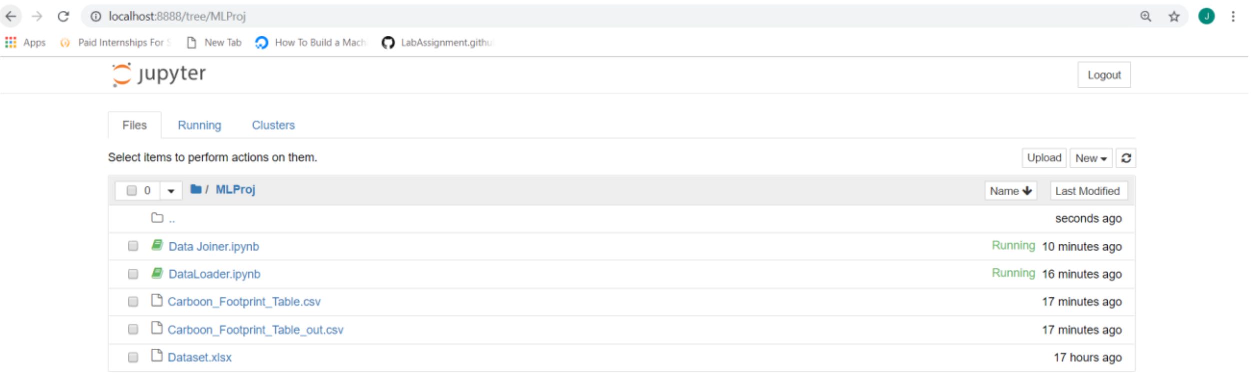


```
In [7]: carbon_footprint_data_df.shape
Out[7]: (27, 15)
```



```
In [7]: carbon_footprint_data_df.shape
Out[7]: (27, 15)

In [8]: carbon_footprint_data_df_select_data = carbon_footprint_data_df.drop(['Unnamed: 14', 'Notes'], axis = 1)

In [9]: carbon_footprint_data_df_select_data.shape
Out[9]: (27, 13)

In [11]: carbon_footprint_data_df_select_data_table = pd.melt(carbon_footprint_data_df_select_data, id_vars=['Activity', 'Per'], value_var

C:\Users\d\Anaconda3\lib\site-packages\pandas\core\reshape\reshape.py:731: FutureWarning:
Passing list-likes to .loc or [] with any missing label will raise
KeyError in the future, you can use .reindex() as an alternative.

See the documentation here:
http://pandas.pydata.org/pandas-docs/stable/indexing.html#deprecate-loc-reindex-listlike
  frame = frame.loc[:, id_vars + value_vars]
C:\Users\d\Anaconda3\lib\site-packages\pandas\core\indexing.py:1367: FutureWarning:
Passing list-likes to .loc or [] with any missing label will raise
KeyError in the future, you can use .reindex() as an alternative.

See the documentation here:
http://pandas.pydata.org/pandas-docs/stable/indexing.html#deprecate-loc-reindex-listlike
  return self._getitem_tuple(key)

In [12]: carbon_footprint_data_df_select_data_table.shape

In [12]: carbon_footprint_data_df_select_data_table.shape
Out[12]: (351, 4)

In [13]: carbon_footprint_data_df_select_data_table_without_na = carbon_footprint_data_df_select_data_table.dropna(axis = 0)

In [14]: carbon_footprint_data_df_select_data_table_without_na.shape
Out[14]: (76, 4)

In [15]: carbon_footprint_data_df_select_data_table.to_csv('Carboon_Footprint_Table.csv')
         carbon_footprint_data_df_select_data_table_without_na.to_csv('Carboon_Footprint_Table_out.csv')

In [ ]: 
```

FIGURE 3: Notebook After Performing Data Loader



Select items to perform actions on them.

□ 0 / MLProj

□ ..		seconds ago
□ Data Joiner.ipynb	Running	10 minutes ago
□ DataLoader.ipynb	Running	16 minutes ago
□ Carboon_Footprint_Table.csv		17 minutes ago
□ Carboon_Footprint_Table_out.csv		17 minutes ago
□ Dataset.xlsx		17 hours ago