CS 487/519 Applied Machine Learning I

Project 2: Open ML project – Stage 4

October 16, 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 to low carbon footprint of 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

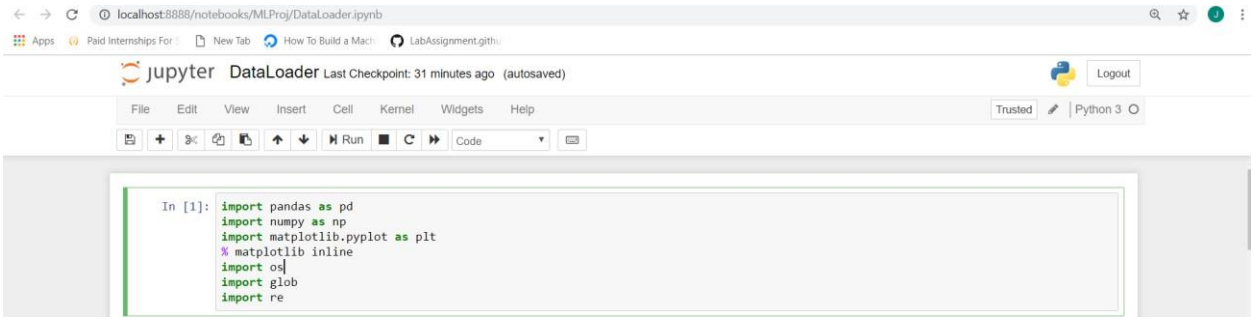
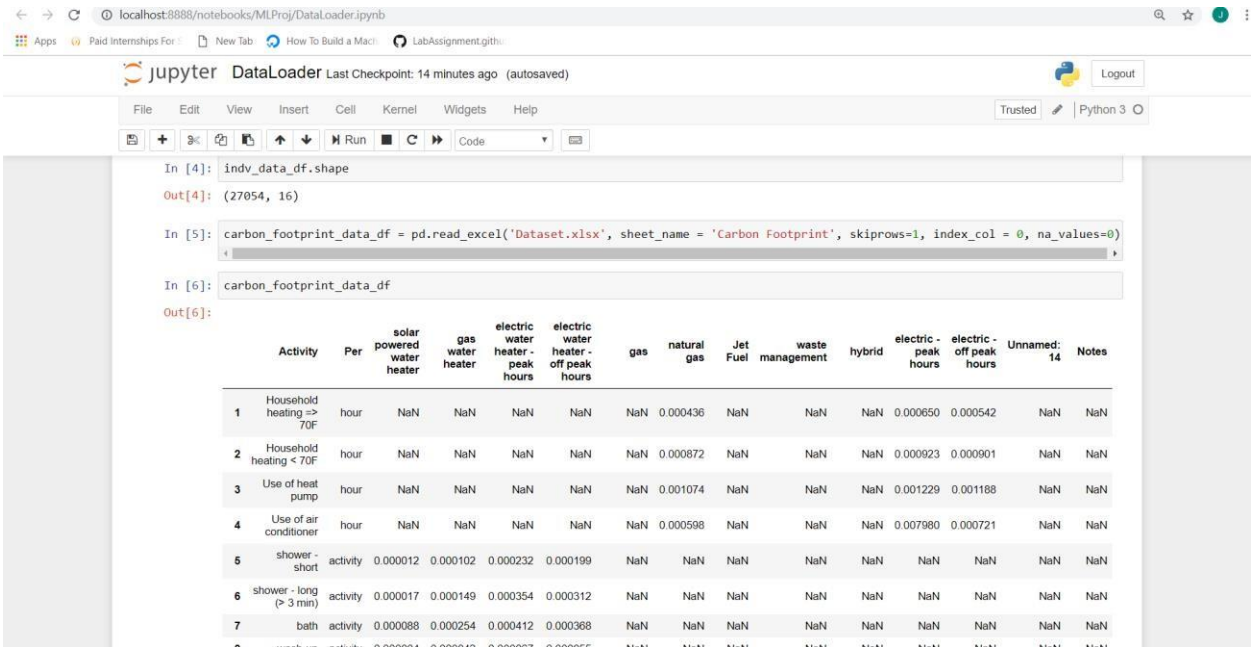


Figure 2: Loading the Data from the Excel File



	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

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

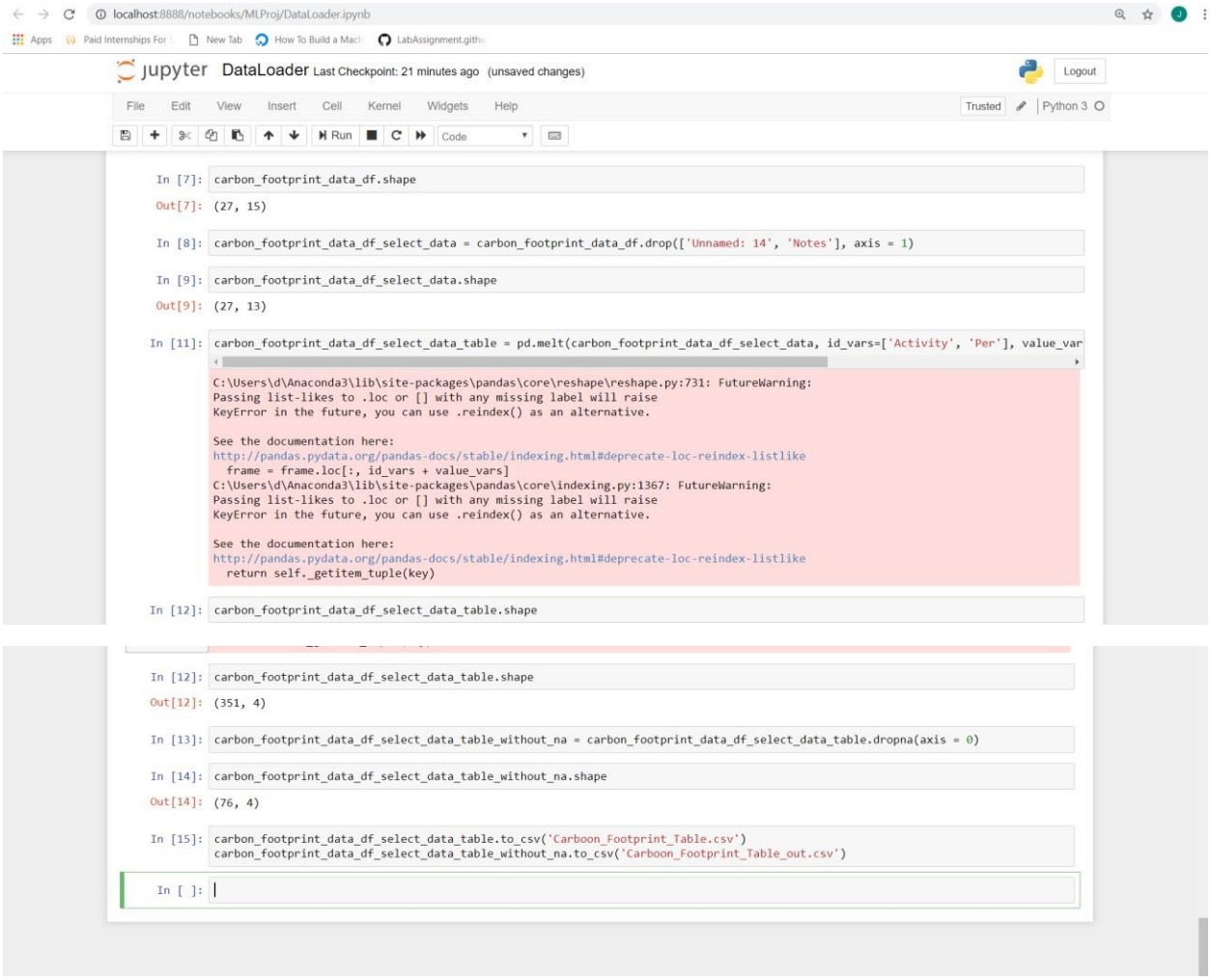
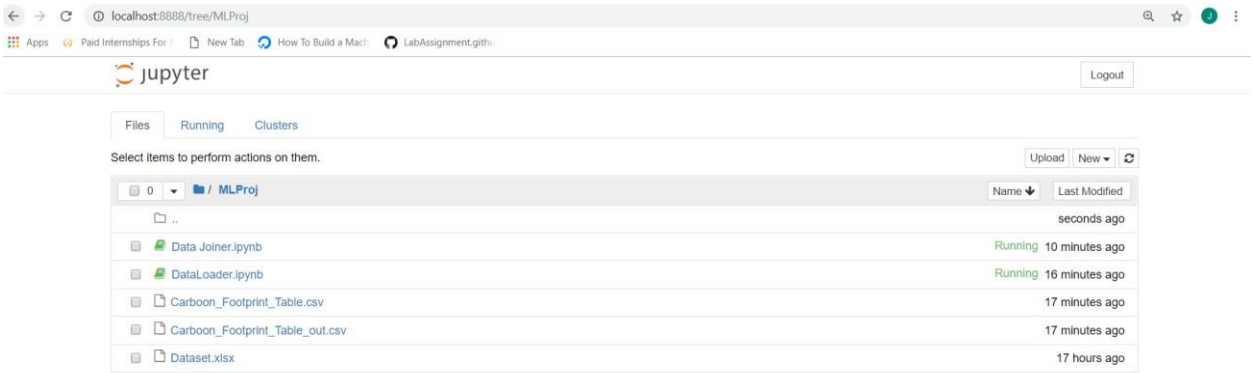


Figure 3: Notebook After Performing Data Loader



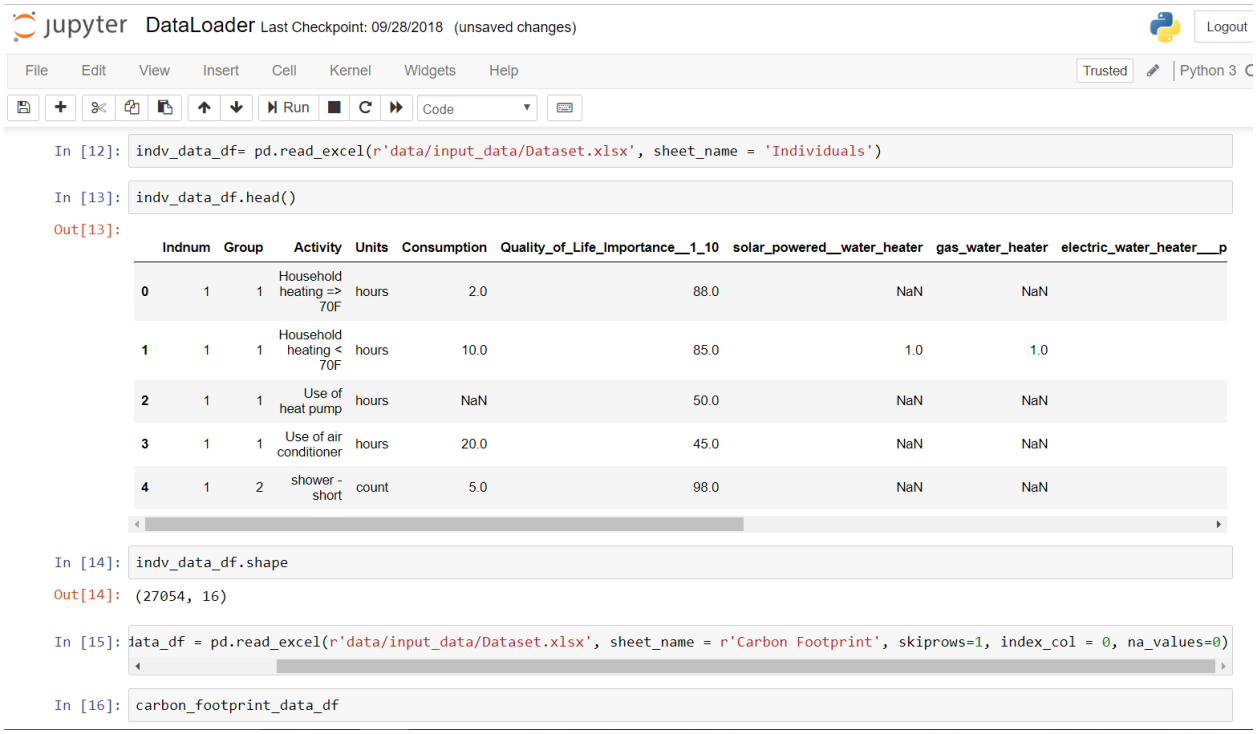
3

Stage 4 - Short Description:

In the stage 3 of the open Ml project we had performed the Data Loading process where we had two csv files namely Carbon Footprint with na and Carbon Footprint without na which were used for the project for finding the low carbon footprint index and in this stage we put them into the input_data and output_data files respectively to categorize them for further analysis**. (Note: na is the missing data values denoted in pandas.It may also be abbreviated as Not Available)**

Secondly in this stage 4 we perform the Data Cleaning of the Individual Person's Data with respect to his daily activities which yield to higher carbon footprint in the environment.

Figure 4: Loading the data from the original Dataset into the input_data folder



	Indnum	Group	Activity	Units	Consumption	Quality_of_Life_Importance__1_10	solar_powered__water_heater	gas_water_heater	electric_water_heater___p
0	1	1	Household heating => 70F	hours	2.0	88.0	NaN	NaN	
1	1	1	Household heating < 70F	hours	10.0	85.0	1.0	1.0	
2	1	1	Use of heat pump	hours	NaN	50.0	NaN	NaN	
3	1	1	Use of air conditioner	hours	20.0	45.0	NaN	NaN	
4	1	2	shower - short	count	5.0	98.0	NaN	NaN	

Figure 5: Extracting data and giving the sheet a name such as Carbon Footprint and displaying the data.

Out[16]:

	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
9	use of dishwasher	activity	0.000025	0.000165	0.000398	0.000311	NaN	NaN	NaN	NaN	NaN	0.000084	0.000078	NaN	NaN
10	use of clothes washer	activity	0.000033	0.000199	0.000433	0.000382	NaN	0.000154	NaN	NaN	NaN	0.000102	0.000093	NaN	NaN
11	use of clothes dryer	activity	NaN	NaN	NaN	NaN	NaN	0.000187	NaN	NaN	NaN	0.000132	0.000122	NaN	NaN
	use of														

Figure 6: Dropping the columns which contains the data which are less or not relevant to carbon footprint impact and defining carbon_footprint_data_df_select_data_ as the refined value, here the data with na is considered.

This method is more of a refining process to eliminate the unwanted columns.
(Note: The warning here is to use .reindex() as an alternative.)

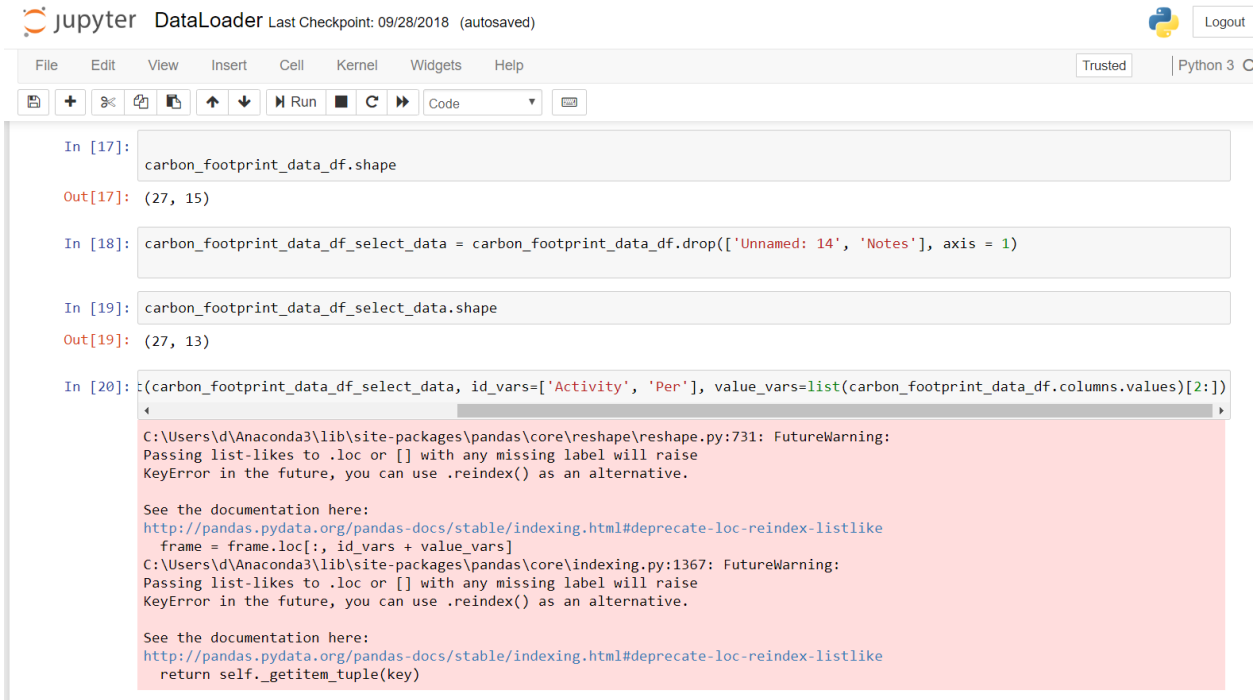


Figure 7: Dropping the columns which are not required and defining carbon_footprint_data_df_select_data_table_without_na as columns where axis=0 and where na is not present

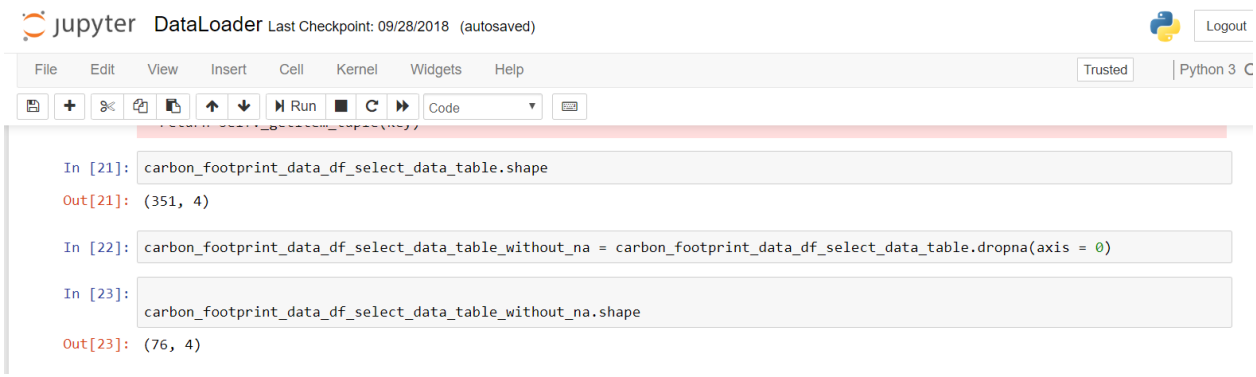
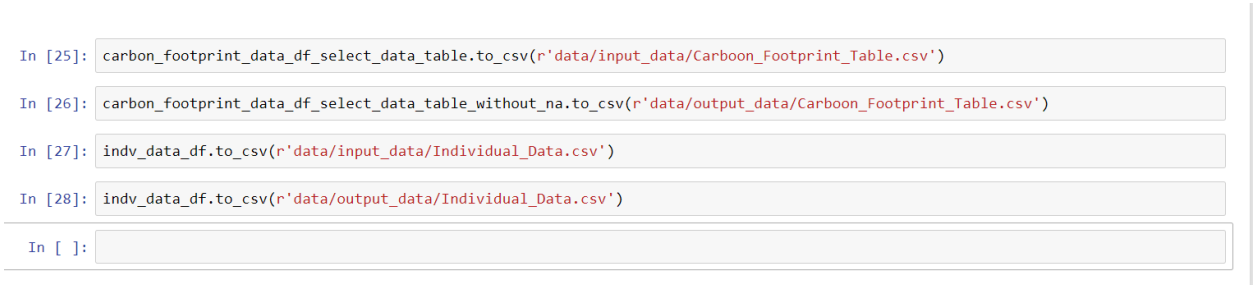
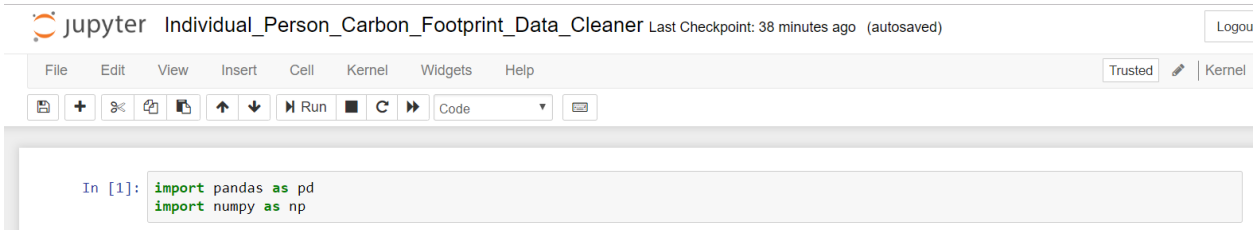


Figure 8:Exporting the data into csv files and saving them in the input_data and output_data folders



Individual Person Carbon Footprint Data Cleaner

Figure 9: First we import the required libraries



Secondly we load the data from our first and main Dataset which is “Dataset.xlsx” and we output the values as shown in the figure below.

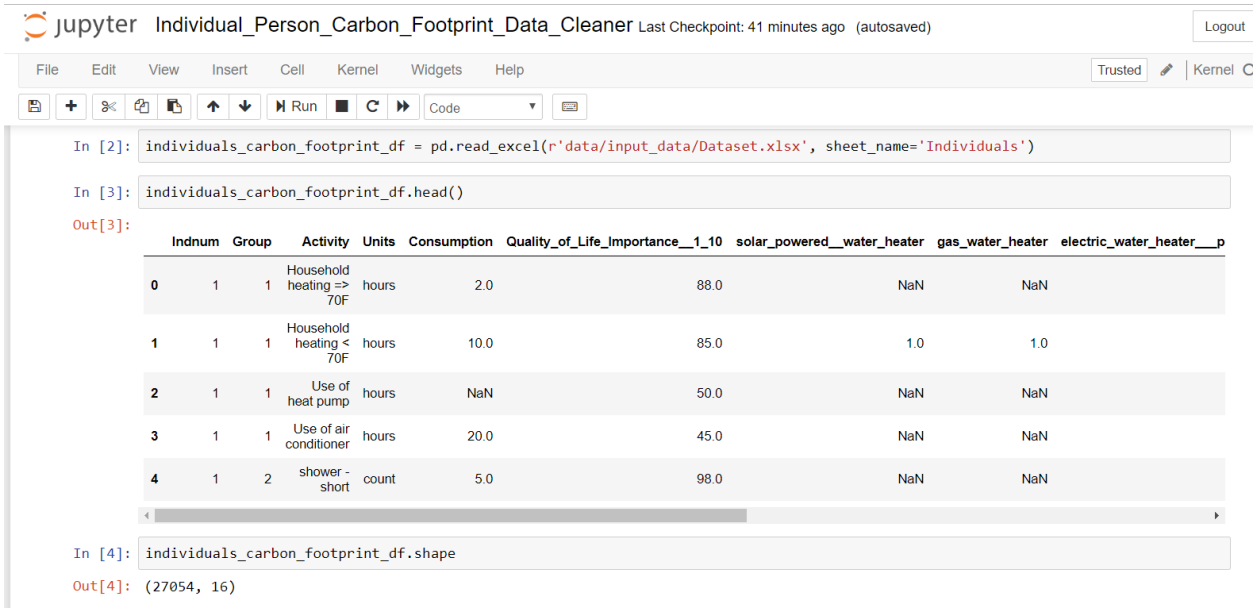


Figure 10: Augmenting the data

Thirdly we need to Augment the data which is basically converting the data from a 2 Dimensional Table to a 1 Dimensional Table by comparing the data of every individual vs the type of resource that he/she is using.

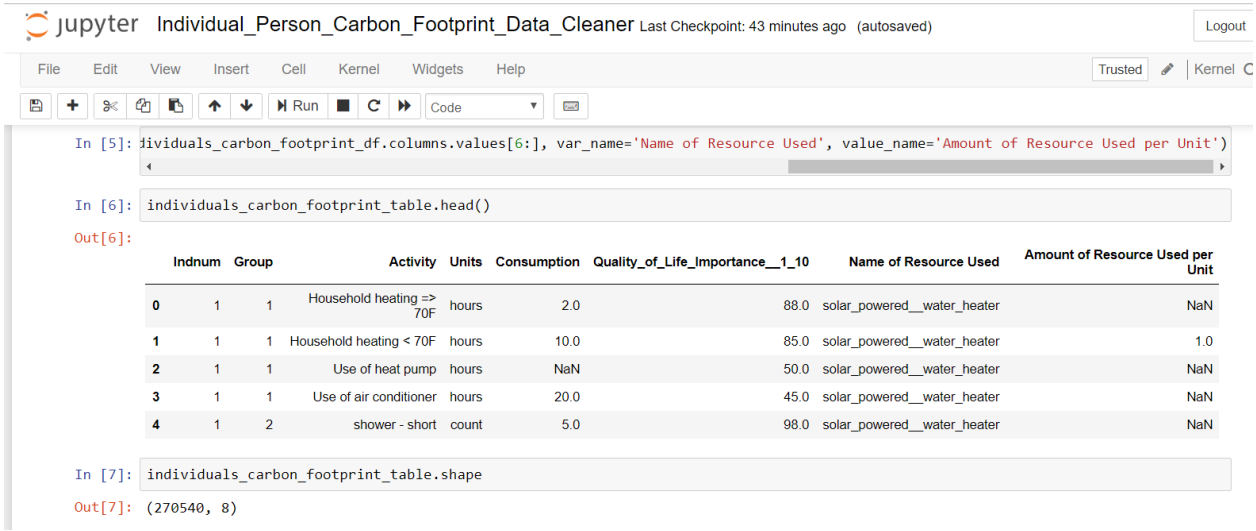
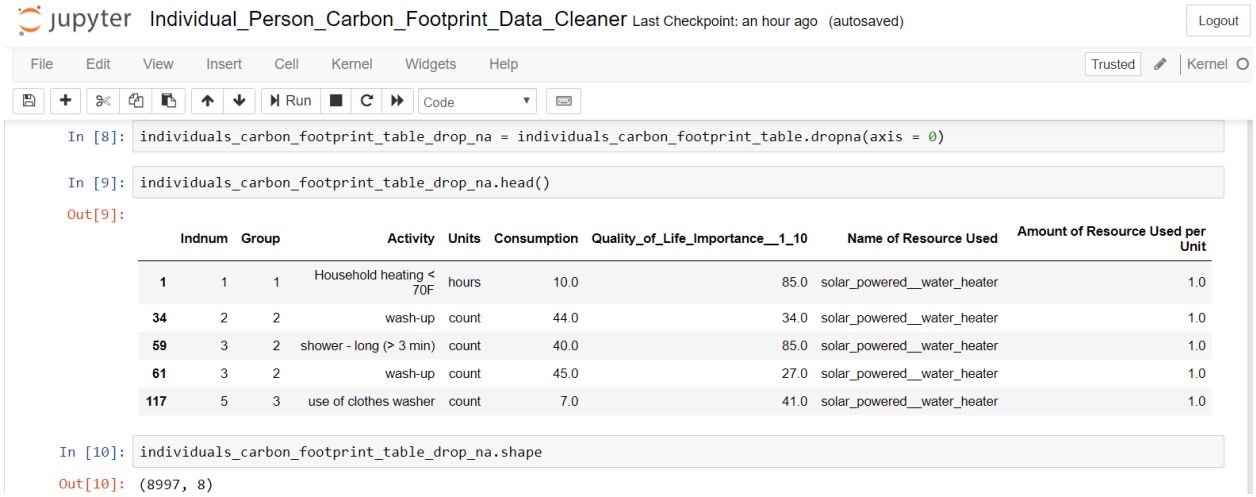
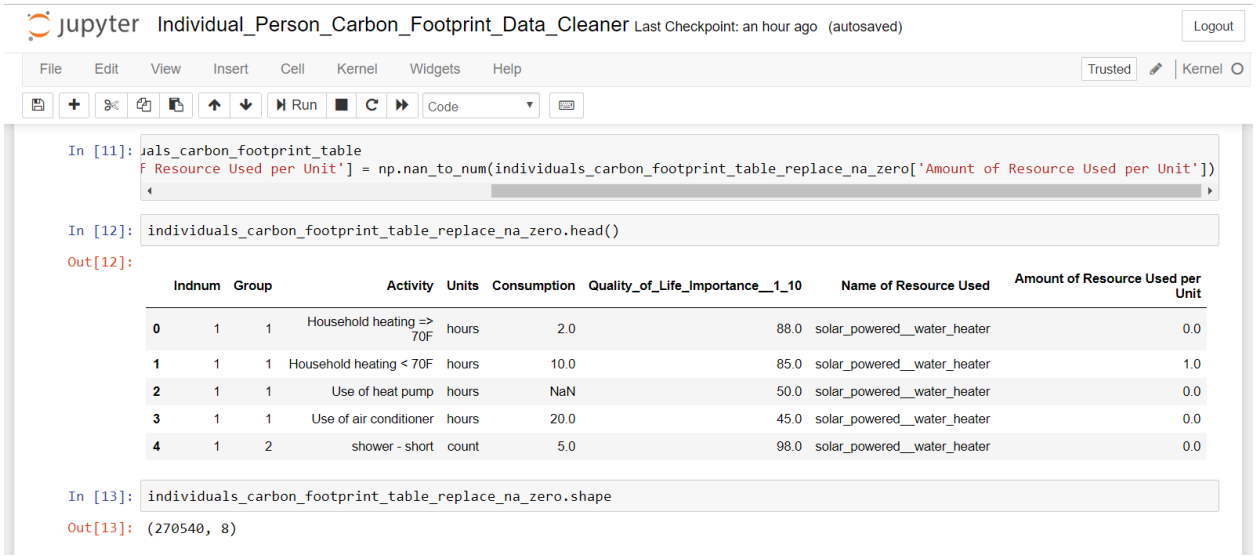


Figure 11:Removing the NaN values and replacing it with 0.0

The fourth part deals with all the NaN values which were outputted after the augmenting of the data has taken place.

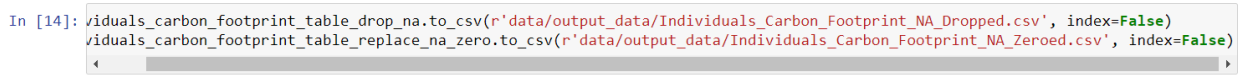


Next we replace the value of NaN with 0.0 wherever NaN is found

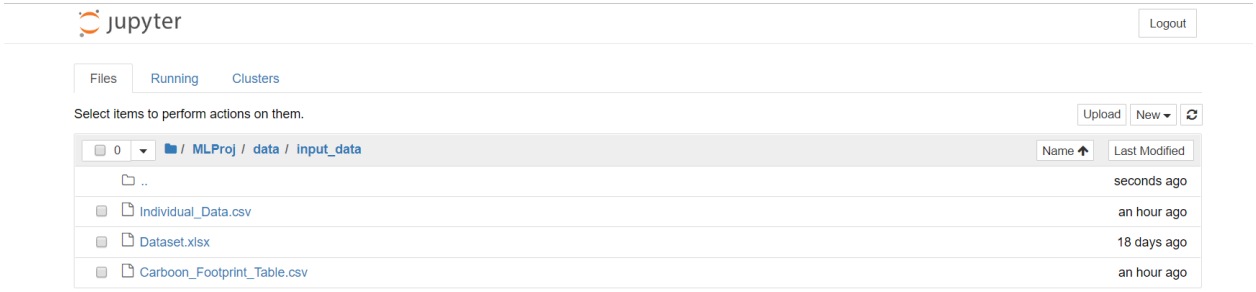


Finally we save the tables to the input_data and output_data locations

Figure 12:Saving the data to input_data and output_data locations after dropping na and replacing NaN by 0.0



Our notebook finally looks like this:



RESULT:

The above analysis have been made and the results have been shown in the figures above.These are to be updated in the future.