Project 1

What is the problem to investigate?

We are all interested in energy data. With analysis of energy data, we hope to obtain a detailed view of how energy consumption occurs in all fields of our daily life, such as residential, commercial and industrial fields. Based on historical data, we can also derive the possible consumptions in the future. As for these consumptions, we could optionally give some policies of saving energies.

Background information

Increasing environmental concerns and limited resources alter people that the significance of paying attention to saving energies. Energy consumption, monitoring, and management are crucial components of sustainable, eco-friendly infrastructures now and into the future.

In the past, a building energy consumption analysis system is available which focuses on accurately determining building energy expenditures, performance and costs. Besides, there is also energy consumption and cost-benefit analysis of hybrid and electric city buses. In 2017, real-time predictive systems for intelligent energy monitoring and management of electrical power networks are available. However, these systems are good tools for administrator to monitor and manage the electricity system to avoid any failure and interact with users. Instead of focusing on a specific energy consumption, such as bus fuel, air-condition electricity consumptions or management tools for owners, a need exists to analyze energy consumptions in a larger scale, including most fields of our daily life and show the detailed analysis results or predicted energy consumptions to the public. The object of this project is to provide the intuitive senses about energy consumptions.

References cited:

Hedley, Jay, Boris Tsypin, and Deepak Raghu. "Building energy consumption analysis system." U.S. Patent No. 8,756,024. 17 Jun. 2014.

Lajunen, Antti. "Energy consumption and cost-benefit analysis of hybrid and electric city buses." *Transportation Research Part C: Emerging Technologies* 38 (2014): 1-15.

Nasle, Adib. "Real-time predictive systems for intelligent energy monitoring and management of electrical power networks." U.S. Patent No. 9,557,723. 31 Jan. 2017.

What can be conducted using collected data?

For electricity and natural gas energy data, there are three classifications: residential, commercial, and industrial. These data are all necessary for this project. Because the goal of this project is to obtain a detailed view of how energy consumption occurs in most fields of our daily life. There are many attributes of these data, including housing_units, total_pop, num_establishments, elec_1kdollar, elec_mwh, gas_1kdollars, gas_mcf, elec_lb_ghg, gas_lb_ghg. And each row represents each city. Therefore, we can get many statistic data, such as ranking of electricity expenditure per housing unit, electricity expenditure per capita, and greenhouse gas emission. We can locate some cities where energy consumption is higher than others. There might be ought to strengthen measurements of saving energy. Besides, we can also predict the potential cities with high energy consumptions and take measures to save energy. To predict, there should be more historic data to collect in different years to analyze the change tendency.

Data Issues

1. "NOT FOUND" error when requesting json data using a zip code. Solution: detect the error code returned by the request, and save the data information in the csv files only when the errors are empty.
2. Cannot request json file. For some zip codes, it blocks or fails when I request the relevant json file. I set a conditional statement to ignore these special zip codes.
3. Missing values. Replace the missing value with its neighbors.
4. Noise (Incorrect values). Fortunately, there is no incorrect values, such as negative numeric data.

Collecting New data

The collected original data are stored in residential.csv, commercial.csv and industrial.csv. To merge them, I drop columns which include the maximum and minimum values. The merged file is electricity_and_natural_gas.csv.There are 22 columns.

Data Cleanliness

The cleanliness results are stored in the output.txt.
The fractions of missing values are:
merged data:

zip	0.0000%
city	0.0000%
housing_units	0.0148%
total_pop	0.0162%
commercial_num_establishments	1.8204%
industrial_num_establishments	4.9342%
residential_elec_1kdollars	0.0148%
commercial_elec_1kdollars	1.8204%
industrial_elec_1kdollars	4.9342%
residential_elec_mwh	0.0148%
commercial_elec_mwh	1.8204%
industrial_elec_mwh	4.9342%
residential_gas_1kdollars	0.0148%
commercial_gas_1kdollars	1.8204%
industrial_gas_1kdollars	5.0154%
residential_gas_mcf	0.0148%
commercial_gas_mcf	1.8204%
industrial_gas_mcf	4.9342%
residential_elec_lb_ghg	0.0148%
commercial_elec_lb_ghg	1.8204%
industrial_elec_lb_ghg	4.9342%
residential_gas_lb_ghg	0.0148%

```
commercial_gas_lb_ghg            1.8204%
industrial_gas_lb_ghg          4.9342%
```

Because the types of all attributes (except the zip and city) are all numeric type. Thus I regard negative values as incorrect values. And there are no incorrect values of all attributes.

Based on this information, the quality score could be the average fractions of data that are not noise or missing.

$$\frac{\sum_{col=1}^{n}(1 - missing\_rate\_column_i) \ * (1 - noise\_rate\_column_i)}{n}$$

According to this quality formula, the quality score is 97.84%. When the score is close to 1, this represents the cleanliness of the data is higher. The data is pretty clean.

Data Cleaning

A need exists to solve missing values. An important rule of the collected data is that data is indexed by zip codes. Sometimes there are more than one city of a zip code. If there is a missing value among these cities with same zip code, it is reasonable to deduce the approximate value of that city with data of other cities. Because their locations are near, so the population density and the energy expenditures are similar. If there is only one city for a zip code or all cities related with that zips have missing values, we can also use the median or average of near cities to replace the missing values. Luckily, the data are organized by zip codes, it is convenient to get data of neighbor cities whose zip codes are also neighbors. By default, I close the average of 4 close cities to replace the missing values. Simply imagine there are four directions to get sample.

The cleaned data are stored in cleaned_residential.csv, cleaned_commercial.csv, cleaned_industrial.csv and cleaned_electricity_and_natural_gas.csv.

Now, the cleanliness result is:

cleaned merged data:

```
  zip                                  0.0
city                                 0.0
housing_units                      0.0
total_pop                            0.0
commercial_num_establishments      0.0
industrial_num_establishments      0.0
residential_elec_1kdollars         0.0
commercial_elec_1kdollars          0.0
industrial_elec_1kdollars          0.0
residential_elec_mwh               0.0
commercial_elec_mwh                  0.0
industrial_elec_mwh                0.0
residential_gas_1kdollars          0.0
commercial_gas_1kdollars           0.0
```

```
industrial_gas_1kdollars     0.0
residential_gas_mcf          0.0
commercial_gas_mcf           0.0
industrial_gas_mcf           0.0
residential_elec_lb_ghg      0.0
commercial_elec_lb_ghg       0.0
industrial_elec_lb_ghg       0.0
residential_gas_lb_ghg       0.0
commercial_gas_lb_ghg        0.0
industrial_gas_lb_ghg        0.0
```

The data is cleaner.