DV1597 Exercise 1

Tobias Gustafsson

March 26, 2024

1 Datasets

I have chosen 3 different datasets from `http://www.data.gov`.

1.1 Dataset 1

Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System. (link).

This dataset contains data for a number of people with information on their physical activity and weight status. The dataset has 33 columns and 93249 rows. Some of the data in the columns is location (where they live), data source, weight status, physical activity, age, education, gender, income and race/ethnicity.

Questions: How is factors such as education and income related to weight status? What does the relation between physical activity and weight status look like?

1.2 Dataset 2

Water Quality Data. (link).

This dataset consist of water quality data that has been collected at 5 places for 30 years. The dataset has 17 columns and 2371 rows. The columns include read date, salinity, dissolved oxygen, pH, water temp, and a few more measurements.

Questions: How has the water quality changed over time? Do features like dissolved oxygen and pH change depending on the time of year?

1.3 Dataset 3

Electric Vehicle Population Size History By County. (link).

The dataset shows monthly data on how many electric vehicles were registred in a United States county. The dataset has 10 columns and 20819 rows. The columns include data on date, conty, state, vehicle (truck/passenger). Also the columns show the number of battery electric vehicles, plug-in hybrids, total electric vehicles, non-electric vehicles, and percentage of electric vehicles.

Questions: What states has the largest percentage of electric vehicles? Does the proportion of electrical vehicles increase with the total amount of cars in the county?

2 Data cleaning

Data cleaning plays a vital role in data analysis in several ways. Firstly, data cleaning can be used to deal with errors in the data and missing values. Another aspect is data compatibility, and the use of different units. For example if salary data is reported in US-dollars for some instances, and in Euro for some others, you can't compare these two directly. In this case data cleaning could be used for converting the values of one currency to the other.

3 Missing data

There are a few ways to deal with missing values. But the main idea is to keep the original data, and create a copy of it where you replace the missing data with suitable values. One method is to use replace the missing data with the mean value of the existing data.

Another method is the *nearest-neigbour*, where you try to find the instance(s) closest to the one with the missing value, looking at the values that are not missing. For example, if a survey responder hasn't entered their age, you compare the responses of that person to the others, and give them an age-value based on the respondent(s) that has answered similarly.

4 Tweet function

```
import pandas as pd

def get_hashtags(tweet):
    i = 0
    hashtags = []
    while i < len(tweet):
        if tweet[i] == '#':
            i += 1
            hashtag = ""
            while i < len(tweet):
                if tweet[i].isalpha() or tweet[i].isnumeric():
                    hashtag += tweet[i]
                    i+=1
                else:
                    break
            if len(hashtag) > 0:
                hashtags.append(hashtag)
        i+=1
    return hashtags


def extract_hashtags(dataframe):
    """Extracts the set of hashtags from a dataframe of tweets."""
    hashtags_list = dataframe["tweet"].apply(get_hashtags).tolist()
    return set([word for tweet in hashtags_list for word in tweet])

def main():
    data_url = "test_data.csv"
    dataframe = pd.read_csv(data_url, sep=',', encoding='utf-8')
    hashtags = extract_hashtags(dataframe)
    print(hashtags)

if __name__ == '__main__':
    main()
```