COMS4060A/7056A: Assignment # 1

University of the Witwatersrand
Muhammad Nasir
muhammad.nasir@wits.ac.za

August 2024

Introduction

This assignment is based on content from Lectures 1-4. You will be required to perform some basic data cleaning and exploration techniques on a prescribed dataset. The aim is to explore the dataset and make observations. You are required to provide a Jupyter notebook and a PDF for all your reasoning. You must discuss all your answers on the PDF with the plots attached. You must provide one Jupyter notebooks that will have a clear mention of the question number you are solving. You are free to use any library/framework in Python.

Problem Definition

As mentioned in the lectures, data exploration and visualisation is as much an art as it is a science. There is back and forth between the different sections, and the flow need not be linear. There are some points highlighted below that you will need to address for this hand-in, but the order is not important. The most important aspect of this assignment is that you are critical of the data, question your findings, and investigate your data in-depth. Ensure that you understand what the variables represent in your dataset and their datatypes. If there is specific domain knowledge that you find, highlight it and what the consequences are (this isn't required for this assignment though).

This assignment involves cleaning a dataset and providing insight into the contents of the data using plots and descriptive stats - it might appear long, but once you've got the hang of how to produce plots it shouldn't take too long. Make use of code from the Jupyter notebooks that have been provided throughout the lectures if it is useful.

Dataset Overview

You are provided with a dataset of fuel logs, `logbook_assignment1.csv`. The data comes from a company that provides a service for users to keep a logbook of their fuel usage - when and where they fill up, how much the fill up etc. The logbook also allows users to capture car services and repairs. The dataset has been scraped from their website and it is messy. Nominally, the dataset has the following features:

- `date_fueled`: The date that the user refueled their vehicle (usually in the format DD mmm YYYY).
- `date_captured`: The date that the user entered the information (usually in the format DD mmm YYYY).
- `user_url`: The URL for the user that captured this information.
- `odometer`: The odometer reading at the time of refueling.
- `gallons`: How many gallons they refueled with.
- `cost_per_gallon`: The cost per gallon in the local currency.
- `total_spent`: The total amount spent in the local currency.
- `mpg`: The computed fuel efficiency in miles per gallon (MPG).
- `miles`: The number of miles driven on this tank of fuel.

There is a lot of information available in this dataset that we would like to extract: we can use the currency as a proxy for country, we can look at fuel efficiency, we can analyse the cost of fuel throughout the world, for example.

Something to note on memory usage if you are using Pandas. In Pandas, everything is stored in memory (vs dask or modin where computation is done out of memory). Reading in a dataset of fairly modest size can use up a significant amount of your RAM. When reading in a dataset, Pandas will attempt to guess the datatype, but if it is unable to, it will revert to object (ie string). Getting the datatype correct for your columns can reduce memory use substantially: an integer will use far less memory than storing an integer as a string. Similarly for floats, dates, and bools. For strings, you might convert them to categoricals to reduce memory usage too. In this dataset, not only do you need to fix the datatypes to properly analyse the dataset, you will see a drop in memory usage and a big boost in performance after you have done this.

1 Data Cleaning

This dataset to a large extent relies on user input, and these are users from around the world. Looking at the data, you will find things like the following:

- `date_fueled` sometimes has a description of what the user did, instead of a date.

- Numerical fields like `gallons`, `miles`, `odometer` will have commas in them as a thousands delimeter (this depends on the location of the user, generally). So, instead of writing 1523.50, they might write 1,523.50. Converting this to a float in pandas will require editing the string.

- The fields relating to costs (`cost_per_gallon` and `total_spent`) have the currency symbol in the value (eg. R500 or $500). There are many different currencies used.

1.1 Date Fields

1. Identify what percentage of `date_fueled` entries that are not proper dates. [1]

2. If `date_fueled` is not entered correctly (or is not a date), and the date captured is a valid date, then fill in this value as a proxy. [1]

3. Convert the column to a date format, setting any invalid date fueled entries to NaT. [2]

4. Remove dates that are in the future, or dates that are earlier than 2005. [1]

5. Plot the distribution of fueling dates and comment on the results. [2]

1.2 Numeric Fields

1. Identify what percentage of `gallons, miles,` and `odometer` entries are missing. [3]

2. The `miles, gallons` and `mpg` columns are interdependent. If one is missing, the other two can be used to calculate it. [3]

3. The values will be read in as objects (or strings) by Pandas. Convert these values to float (note the point above about commas in the value). [5]

4. Plot the distributions and comment on the distributions. [3]

5. Compute the statistical description of the columns: mean, standard deviation, max, min, most frequent, and quartiles. Do these results make sense? [3]

2 Feature Engineering

We can use the existing features to create new features with more useable information. Add the following features:

1. Create a new column with the currency. (Something to keep in mind is that the Swiss Franc has a period in the abbreviation). [2]

2. Create a new column containing the float value of the total spend and the cost per gallon. (Swiss Franc comment as above). [2]

3. Car make, model, year, User ID: use the url (the last value in the URL is the user ID) [4]

The data is given in imperial units, and in SA, we use proper measurement standards.

1. litres filled: use the gallons - consider whether to use UK or US gallons. [2]

2. km driven: use the miles driven to compute this [1]

3. litres per 100km: use the two new features to calculate this. [1]

3 Vehicle Exploration

We will see in the next few questions (and you should be aware of it by now) that the data captured by users is not always accurate. In particular, the transaction level data. There is probably more accuracy in the user profile: their vehicle make and model, the year of the vehicle, and, hopefully, the currency they use. We'll look at vehicle and user profile information for the global population here, before we consider removing outliers and bad transaction data.

1. Plot the number of unique users per country (remember, we proxy this by currency). [2]

2. Look at the popularity of the app: plot the number of unique users per day. [2]

3. Look at the distribution of age of the vehicles per country - look at the year of the vehicle. Remember to look at the date it was refuelled, not the current date. [3]

4. Which makes and models of vehicles are the most popular? [2]

4 Fuel Usage

It is particularly difficult to identify outliers in this dataset, due, for example, to the multiple currencies. As an example, refilling a vehicle in South Africa would be maybe R1000, but in the US it would be $70. One would have to either perform outlier detection on each currency separately. (We could convert everything into a single currency, based on the time of the transaction, and use that, however, that is not required in this assigment.) We will focus on the top five currencies only (Rands will be one of them) to simplify things.

4.1 Outlier Removal

1. Identify the top 5 currencies by number of transactions. [2]

2. For each of the top 5 currencies separately, remove outliers by considering the total spend, litres, cost per litre, gallons, etc. Choose values you believe are reasonable and provide your reasoning. As an example of something you would want to look out for, there are some SA users that have their currency set to dollars. This will show a user refuelling with several hundred dollars, but only putting in tens of litres, which is clearly wrong. [10]

3. How many values have been removed after accounting for outliers? [1]

4.2 Fuel Efficiency

Now that you have a much cleaner dataset, we can start to look at some of the data more closely for insights. In particular, we want to look at the fuel efficiency in litres per 100km. In general, there are many confounding factors and unknown variables that can make an analysis of fuel efficiency difficult: engine size, vehicle type, fuel type (diesel vs petrol will show a massive difference), aircon usage, vehicle load, weather, transmission type. With this in mind, we need to be aware that the results found here are unlikely to be completely representative and accurate, but hopefully indicative. When you start this section, make sure you have removed outliers as indicated in the previous question.

1. Look at the difference in cost per litre per country for January 2022 - use the average currency conversion rate to Rands (quote your values and source). Are there any notable differences? Discuss reasons why this may/may not be the case. [5]

2. Looking at the odometer readings, find examples of where users have missed logging a fill-up. Give a basic rule for identifying this, and estimate how many there are in the dataset. [5]

3. Plot the average distance (in km) per tank per country. Which country has the largest average distance? Provide some explanations for why this might be the case. [5]

4. Do newer vehicles drive further distances between fill-ups? Provide a plot to show this. [4]

5. Take the top 5 most popular vehicles in SA (ie, those with currency set to R). Compute their fuel efficiency and discuss whether these values are realistic. [3]

6. Which vehicles are the most fuel efficient in each country? (Make sure the values are reasonable!!! You can look up values of fuel efficiency online to do a sanity check, but a value of 1l per 100km, or 100l per 100km are clearly wrong). [5]

7. Plot the difference in fuel efficiency for the top 5 Canadian vehicles between seasons. Would you expect to see big differences, and do you see them? [3]

8. Show the correlations between fuel efficiency and other features. You should find that there is a relative strongly correlation with distance travelled, the age of the vehicle, and the model of vehicle. [5]

9. Use a random forest to get a list of the most important variables. How different are they from each other, and how do these relate to the variables from the correlations above? [5]

4.3 Fuel Usage in SA

In South Africa, fuel prices are always adjusted at midnight on the first Tuesday of the month. If the price is going up, we expect there to be more people refuelling on a Tuesday than usual. If the price is going down, we might expect people to postpone refuelling until the Wednesday.

1. Filter the above dataset to focus on SA drivers. [1]

2. Plot the fuel prices over time for SA. [2]

3. Using a suitable plot, show if the difference in the number of people refueling on a Tuesday vs other days. [3]

4. Now reduce your dataset to only the entries on the 1st Tuesday and 1st Wednesday in SA every month. [2]

5. For each Tuesday and Wednesday, add an indicator for whether the price goes up or the price goes down that month. [4]

6. Do more people refuel on the first Wednesday of the month when the prices goes down? [2]

7. Do more people refuel on the first Tuesday of the month when the prices goes up? [2]

Submission

Work by yourself or in groups of up to four people. Submit your work to Moodle. Your submission should be a PDF and a Jupyter notebook. Total Marks Available: 105 Marks. For full score: 100 (5 bonus marks)

Deadline: 23 August 2024