Data Transformation for Question Answering

02/20/2020

Overview

- Attribute Generation
- Excluding/Including Attributes
- Lambda functions
- Filtering
- Grouping

Before we start...

Let’s join the slack station for this course:

https://join.slack.com/t/nyu-va-20spring/shared_invite/enQtOTYzNzg4ODM5MTU4LWFkOTgyNzg3MTBmMDk3YzA5NDYwZDQxMjU4MmU4MTU1NThlZjg5MTIyOTk2Mjc5MjJlODRiNDI3ZWU2NGJjNzc

Data

We are going to use a sampled nyc taxi trip data (Green taxi, 2016 June). Please download the data set here:

https://github.com/nyuvis/visual_analytics_course/blob/master/labs/lab2_Preprocessing/sampled_taxi.csv

Explanation of attributes:

https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf

Data Questions

- How do the payment types compare in terms of number of trips
- How many trips are recorded on different days? How about in different hours overall?
- What are average trip distances recorded on different days?
- What are the total trip duration of different vendors over time (days)

We’d better check what we have in the dataset first...

```
1  df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 18 columns):
 #   Column                 Non-Null Count   Dtype  
---  ------                 --------------   -----  
 0   VendorID               100000 non-null  int64  
 1   pickup_datetime        100000 non-null  object 
 2   dropoff_datetime       100000 non-null  object 
 3   Pickup_longitude       100000 non-null  float64
 4   Pickup_latitude        100000 non-null  float64
 5   Dropoff_longitude      100000 non-null  float64
 6   Dropoff_latitude       100000 non-null  float64
 7   Passenger_count        100000 non-null  int64  
 8   Trip_distance          100000 non-null  float64
 9   Fare_amount            100000 non-null  float64
 10  Extra                  100000 non-null  float64
 11  MTA_tax                100000 non-null  float64
 12  Tip_amount             100000 non-null  float64
 13  Tolls_amount           100000 non-null  float64
 14  improvement_surcharge  100000 non-null  float64
 15  Total_amount           100000 non-null  float64
 16  Payment_type           100000 non-null  int64  
 17  Trip_type              100000 non-null  float64
dtypes: float64(13), int64(3), object(2)
memory usage: 13.7+ MB
```

Question 1:

How do the payment types compare in terms of number of trips

What would you do?

Count the occurance of different values?

Method 1: Calculate it

```
df['Payment_type'].value_counts()
```

- Try to count the occurance of different days

Method 2: Plot it directly (some plot library does the aggregation internally)

* We will talk more about visualization in the next lab.

Try to answer the following question...

What are the most and least common number of passengers for a trip?/ What is the distribution of number of passengers in a trip?

* Hint: what column are you going to use to extract the information of number of passengers?

Question 2:

How many trips are recorded on different days? How about in different hours overall?

What would you do?

Attribute Transformation and Generation

1) Change the attribute `pickup_time` into the type of `datetime`
   - `df["pickup_datetime"] = pd.to_datetime(df.pickup_datetime)`

2) Extract Month, Week, Day, Weekday, Hour, etc. as new attributes
   - `df['month'] = df.pickup_datetime.dt.month`
   - Try to generate others by yourself

How about...

- How does the number of trips change over time on June 16 (by hour)?
  - Check what you can get from `df[df['day'] == 16]`
- What are the average numbers of trips in different hours in the dataset?

Question 3:

What is the average trip distance for each day?

What would you do?

Statistics over a pair of attributes

- Group By

```
df.groupby('day')
```

- Mean of a column

```
df[‘Trip_distance’].mean()
```

- Mean of a trip distances over days

```
df.groupby('day')[‘Trip_distance’].mean()
```

Try to answer the following question...

What is the daily average trip distance of Vendor 1?

Question 4:

What is the total trip duration of each vendor over time (days)?

What would you do?

Extract Trip Duration

- Get the trip duration

```
df['dropoff_datetime'] - df['pickup_datetime']
```

- Transform the duration into minutes

```
transformation = lambda x: x.components.hours*60 + x.components.minutes + x.components.seconds/60.0
df['duration'] = df['duration'].apply(transformation)
```

* You can put the definition of lambda function inside apply() directly...

** A lambda function is a small anonymous function.
*** A lambda function can take any number of arguments, but can only have one expression.

Question 4 - Show your answer

What is the total trip duration of different vendors/taxi drivers over time (days)

Practice

What is the total Fare_amount of Vendor 1 on different days?

Practice

What are the total number of trips from vender 1 on different days?

Practice

What is the total number of trips from Vender 2 for which the trip distances are larger than 10 miles over days?

* Hint:
  - Link conditions with `&`
  - Use () for conditions because `&` has a high priority