Analysis-Cleaning-Transformation

September 12, 2022

[2]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

[3]:

```
import os
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import seaborn as sns
```

Load data and Show information

[4]:

```
DATA_PATH = "/content/drive/MyDrive/Undergrad/Semester-7/
 ↪DWDM_Preprocessing-Assignment/Data"
df = pd.read_csv(os.path.join(DATA_PATH, "taxi_trip_data.csv"))
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000000 entries, 0 to 9999999
Data columns (total 17 columns):
 #   Column                Dtype
---  ------                -----
 0   vendor_id             int64
 1   pickup_datetime       object
 2   dropoff_datetime      object
 3   passenger_count       int64
 4   trip_distance         float64
 5   rate_code             int64
 6   store_and_fwd_flag    object
 7   payment_type          int64
 8   fare_amount           float64
 9   extra                 float64
 10  mta_tax               float64
 11  tip_amount            float64
 12  tolls_amount          float64
 13  imp_surcharge         float64
 14  total_amount          float64
 15  pickup_location_id   int64
 16  dropoff_location_id  int64
dtypes: float64(8), int64(6), object(3)
memory usage: 1.3+ GB
None
```

[5]:

```
# Show first 5 rows
print(df.head())
```

```
   vendor_id      pickup_datetime     dropoff_datetime  passenger_count  \
0          2  2018-03-29 13:37:13  2018-03-29 14:17:01                1
1          2  2018-03-29 13:37:18  2018-03-29 14:15:33                1
2          2  2018-03-29 13:26:57  2018-03-29 13:28:03                1
3          2  2018-03-29 13:07:48  2018-03-29 14:03:05                2
4          2  2018-03-29 14:19:11  2018-03-29 15:19:59                5

   trip_distance  rate_code store_and_fwd_flag  payment_type  fare_amount  \
0          18.15          3                  N             1         70.0
1           4.59          1                  N             1         25.0
2           0.30          1                  N             1          3.0
3          16.97          1                  N             1         49.5
4          14.45          1                  N             1         45.5

   extra  mta_tax  tip_amount  tolls_amount  imp_surcharge  total_amount  \
0    0.0      0.0       16.16         10.50            0.3         96.96
1    0.0      0.5        5.16          0.00            0.3         30.96
2    0.0      0.5        0.76          0.00            0.3          4.56
3    0.0      0.5        5.61          5.76            0.3         61.67
4    0.0      0.5       10.41          5.76            0.3         62.47

   pickup_location_id  dropoff_location_id
0                 161                    1
1                  13                  230
2                 231                  231
3                 231                  138
4                  87                  138
```

1 Introduction

The goal of this notebook is to clean and transform the data available for the purpose of later utilizing it in ML algorithms, or for data warehousing purposed.

The data cleaning process can begin to clear out outliers, missing values and other noise which might affect the results of the algorithm.

1.1 Background

Rides following similar paths in the past will likely take similar routes, and rides during the same hours of the day will also likely take roughly the same amount of time. This gives us a sort of rolling average for distance and time to make the calculation easier, what it doesn't give us, is how much of that distance is sitting in traffic, below 12mph, or driving at normal speeds above 12mph, nor does it account for sitting at red lights.

These values are hard to account for. While patterns can be detected when analysing the data through graphs and other visuals, it doesn't make for a very mathematical or repeatable prediction. We'll need the model to detect these patterns quickly and repeatably to get the most accurate predictions possible, which means some data, such as start and end times should be broken down into chunks that are easier for a machine to read such as the number of minutes per trip, the month, day, day of the week, and year (separately) .

1.2 Plan for Features

Here are the features of the current dataset that will be kept, as well as a few that will need to be created based on other features:
- pickup_timestamp
- dropoff_timestamp
- trip_distance
- fare_amount
- extra
- mta_tax
- imp_surcharge
- total_amount
- pickup_location_id
- dropoff_location_id

2 Data Analysis

The **correlation matrix** calculates how the change in one value effects a change in the other value, and assigns a value between -1 and 1 to that correlation.
Let's review what those correlation values mean before we move on:

- **-1:** A very strong negative correlation, when value A moves in one direction, value B moves in the opposite direction.

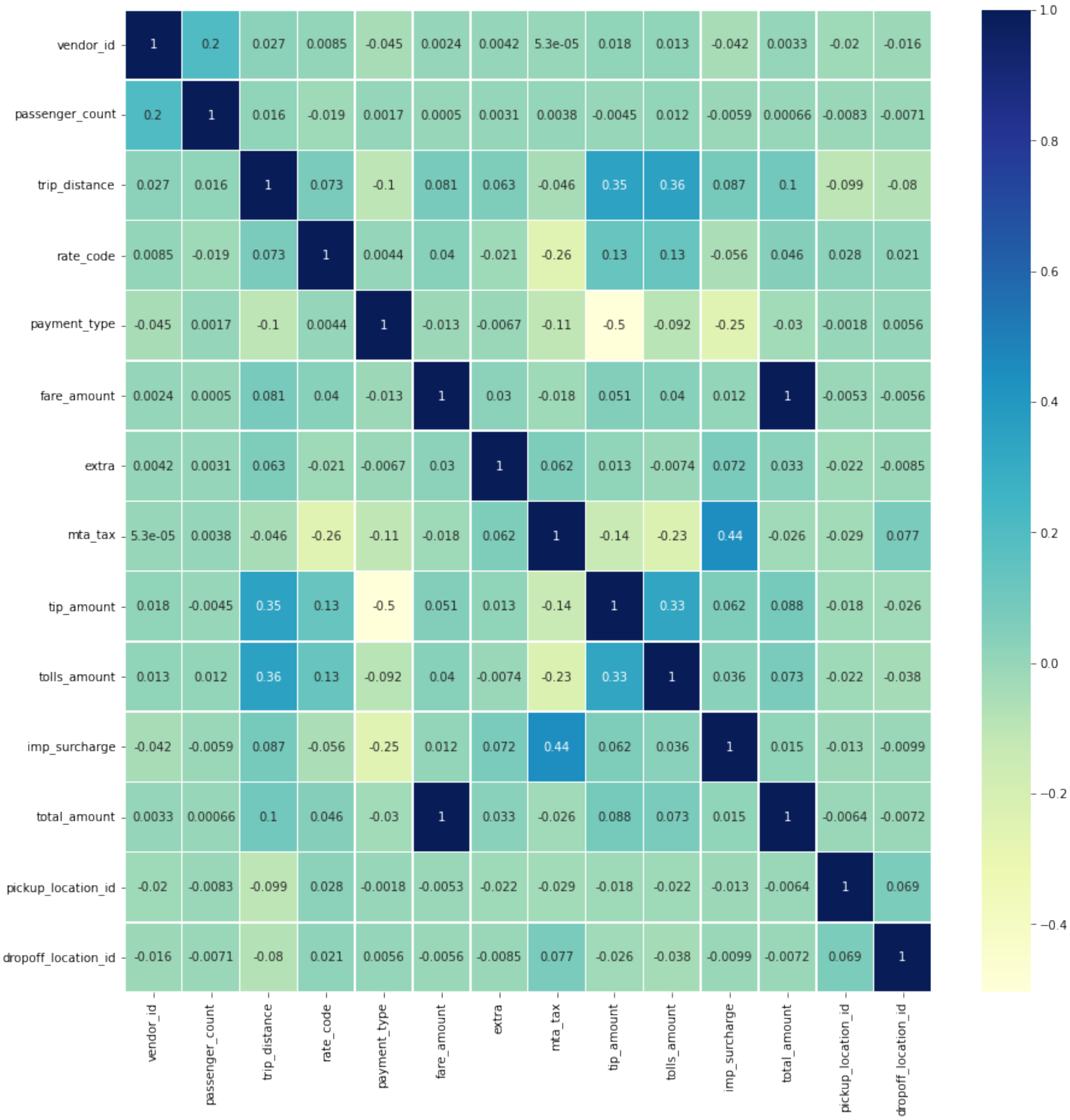
- **0:** No correlation between values A and B, when one moves, the other is not effected.

- **1:** A very strong positive correlation, as you can guess, this is the opposite of the negative correlation above. When value A moves in one direction, value B follows in the same direction.

This value isn't related to the rate of change, only the direction of change. Value A moves up, and value B either stays, moves up, or moves down.

```
[6]: # Generating the correlation matrix
     corr = df.corr()
```

```
[7]: # Drawing the heatmap
     fig, ax = plt.subplots(figsize=(15,15))
     ax = sns.heatmap(corr, cmap='YlGnBu', annot=True, linewidths=0.5);
```



There is already a list of known values that we should keeep. The only remaining values are:
- **vendor_id** - Vendor of data provider. This definitely won't be used for anything for our model here
- **rate_code** - The rate code at the end of the trip. Used likely to track certain charges. Has a correlation with tolls and tips but not much with anything else.

- **sotre__and__fwd__flag** - This is simply a fag that indicates whether a value was stored in vehicle memory before being recorded due to a lack of internet connection. This is useless to us, however, it's currently stored as a string and converting it to a value that can appear in a correlation matrix later might serve useful. While not likely, it could be possible that values are different from those not stored in memory, such as having a higher amount of errors, or some upload process might be altering values in an unexpected way.

In the end, only one column is being dropped right off the start and that's **Vendor ID**.

[8]:
```
df = df.drop('vendor_id', axis=1)
df.head()
```

[8]:
```
      pickup_datetime     dropoff_datetime  passenger_count  trip_distance  \
0 2018-03-29 13:37:13  2018-03-29 14:17:01                1          18.15
1 2018-03-29 13:37:18  2018-03-29 14:15:33                1           4.59
2 2018-03-29 13:26:57  2018-03-29 13:28:03                1           0.30
3 2018-03-29 13:07:48  2018-03-29 14:03:05                2          16.97
4 2018-03-29 14:19:11  2018-03-29 15:19:59                5          14.45

  rate_code store_and_fwd_flag  payment_type  fare_amount  extra  mta_tax  \
0         3                  N             1         70.0    0.0      0.0
1         1                  N             1         25.0    0.0      0.5
2         1                  N             1          3.0    0.0      0.5
3         1                  N             1         49.5    0.0      0.5
4         1                  N             1         45.5    0.0      0.5

  tip_amount  tolls_amount  imp_surcharge  total_amount  pickup_location_id  \
0      16.16         10.50            0.3         96.96                 161
1       5.16          0.00            0.3         30.96                  13
2       0.76          0.00            0.3          4.56                 231
3       5.61          5.76            0.3         61.67                 231
4      10.41          5.76            0.3         62.47                  87

  dropoff_location_id
0                   1
1                 230
2                 231
3                 138
4                 138
```

3 Data Cleaning

The following data cleaning steps are assessed and applied: 1. **Remove duplicate rows** - Carefully, as we only want to remove duplicate trips, not duplicates within the values themselves. These values are not required to be unique. 2. Check for **missing values** 3. Check for **zeros and empty strings**. These values won't be ”missing” but still aren't valid. Very few columns in this data have valid zeros

4.**Validate formatting** of data, especially dates
5. **Strip and normalize strings** - our data doesn't contain any strings, so we can skip this.

3.1 Remove Duplicates

[9]:

```
# Remove duplicates -
# Rename the dataframe from df to td for temporary data, thus not altering the␣
 ↪original dataframe until much later.
td = df.drop_duplicates()
# less than 1% dropped
print(f"{df.shape[0] - td.shape[0]} duplicate rows dropped. Thats {(df.shape[0]␣
 ↪- td.shape[0]) / df.shape[0] * 100}%")
print(f"{td.shape[0]} rows remain.")
```

```
607571 duplicate rows dropped. Thats 6.07571%
9392429 rows remain.
```

3.2 Remove Missing Values

[10]:

```
# Checking for missing values
for col in td.columns:
    missing = td[col].isna().sum()
    print(f"Missing values in {col}: {missing}")
```

```
Missing values in pickup_datetime: 0
Missing values in dropoff_datetime: 0
Missing values in passenger_count: 0
Missing values in trip_distance: 0
Missing values in rate_code: 0
Missing values in store_and_fwd_flag: 0
Missing values in payment_type: 0
Missing values in fare_amount: 0
Missing values in extra: 0
Missing values in mta_tax: 0
Missing values in tip_amount: 0
Missing values in tolls_amount: 0
Missing values in imp_surcharge: 0
Missing values in total_amount: 0
Missing values in pickup_location_id: 0
Missing values in dropoff_location_id: 0
```

3.3 Remove Zeros and Empty Strings

[11]:

```
# Checking for zeros in numeric columns
def check_for_zeros(td):
    for col in td.columns:
        zeros = td[td[col] == 0].shape[0]
        print(f"Zeros in {col}:{zeros}")

check_for_zeros(td)
```

```
Zeros in pickup_datetime:0
Zeros in dropoff_datetime:0
Zeros in passenger_count:85779
Zeros in trip_distance:264896
Zeros in rate_code:0
Zeros in store_and_fwd_flag:0
Zeros in payment_type:0
Zeros in fare_amount:12176
Zeros in extra:5048008
Zeros in mta_tax:285657
Zeros in tip_amount:2062555
Zeros in tolls_amount:6253748
Zeros in imp_surcharge:12433
Zeros in total_amount:5725
Zeros in pickup_location_id:0
Zeros in dropoff_location_id:0
```

Changes applied so far ...

- *passenger_count*, *trip_distance*, *fare_amount* and *total_amount* --- all contain zeros.
- It doesn't appear to be a large amount of the overall data.
- Without distance, we can't determine fare amount, even with distance, it's impossible to know which miles were driven above the 12mph threshold, and which were below.
- There isn't much of a choice but to drop these. However, **total_amount** can be corrected by simply adding all of the charge column values together, so I'll keep and fix these rows.

Dropping rows with 0 values in columns where 0 is not allowed

[12]:

```
# Dropping rows with 0 values in columns where 0 is not allowed
td = td.drop(['passenger_count'], axis=1)
td = td[td['trip_distance'] > 0]
td = td[td['fare_amount'] > 0]

check_for_zeros(td)
```

```
Zeros in pickup_datetime:0
Zeros in dropoff_datetime:0
Zeros in trip_distance:0
Zeros in rate_code:0
Zeros in store_and_fwd_flag:0
Zeros in payment_type:0
Zeros in fare_amount:0
Zeros in extra:4836000
Zeros in mta_tax:216469
Zeros in tip_amount:1886004
Zeros in tolls_amount:5980138
Zeros in imp_surcharge:1418
Zeros in total_amount:0
Zeros in pickup_location_id:0
Zeros in dropoff_location_id:0
```

After dropping rows with zero values in other columns, there remains no zeros in total_amount, so no corrections are necessary here

[13]:

```
# Checking how much of the original data ramains
remaining = td.shape[0] / df.shape[0] * 100
print(f"Remaining amount of original dataset: {remaining}%")
```

```
Remaining amount of original dataset: 90.99450999999999%
```

3.4 Validating Data Formats

Ensure that the dates are all readable date formats and exist in the same format such as mm/dd/yyyy, for example.

[14]:

```
# Converting to an actual Python/Pandas datetime object ensures that the data␣
 ↪is a valid datetime.
# Then, we  move on to exploring the datetimes available.
td['pickup_datetime'] = pd.to_datetime(td['pickup_datetime'])
td['dropoff_datetime'] = pd.to_datetime(td['dropoff_datetime'])

print('Done.')
```

```
Done.
```

Inference: All datetime stamps in the dataset are correctly formatted

The datetime columns are now split up into meaninful columns. The only dropoff information we really need to keep is the hour, and even then, only to calculate the length of the trip.

[15]:

```
td['year'] = pd.to_datetime(td['pickup_datetime']).dt.year
td['month'] = pd.to_datetime(td['pickup_datetime']).dt.month
td['day'] = pd.to_datetime(td['pickup_datetime']).dt.day
td['day_of_week'] = pd.to_datetime(td['pickup_datetime']).dt.dayofweek
td['hour_of_day'] = pd.to_datetime(td['pickup_datetime']).dt.hour

print('Done.')
```

```
Done.
```

3.5 Cleaning Date and Time Data

3.5.1 Validate Timestamps

[16]:

```
# Converting the datetime columns to a numpy array for vectorization
pickup_array = td['pickup_datetime'].values
dropoff_array = td['dropoff_datetime'].values
```

3.5.2 Validate Trip Durations

[17]:

```
# Getting the new timedelta, this takes less than a second to complete compared␣
 ↪to 15+ minutes with apply()
trip_duration = np.subtract(dropoff_array, pickup_array)

# Adding the resulting array to the dataframe in the trip_duration column
td['trip_duration'] = pd.Series(trip_duration)

# Converting the timedelta to number of seconds
td['trip_duration'] = td['trip_duration'].dt.total_seconds()

# Preview the results
td.head()
```

[17]:

```
      pickup_datetime    dropoff_datetime  trip_distance  rate_code  \
0 2018-03-29 13:37:13 2018-03-29 14:17:01          18.15          3
1 2018-03-29 13:37:18 2018-03-29 14:15:33           4.59          1
2 2018-03-29 13:26:57 2018-03-29 13:28:03           0.30          1
3 2018-03-29 13:07:48 2018-03-29 14:03:05          16.97          1
4 2018-03-29 14:19:11 2018-03-29 15:19:59          14.45          1

  store_and_fwd_flag  payment_type  fare_amount  extra  mta_tax  tip_amount  \
0                  N             1         70.0    0.0      0.0       16.16
1                  N             1         25.0    0.0      0.5        5.16
2                  N             1          3.0    0.0      0.5        0.76
3                  N             1         49.5    0.0      0.5        5.61
4                  N             1         45.5    0.0      0.5       10.41

   …  imp_surcharge  total_amount  pickup_location_id  dropoff_location_id  \
0  …            0.3         96.96                 161                    1
1  …            0.3         30.96                  13                  230
2  …            0.3          4.56                 231                  231
3  …            0.3         61.67                 231                  138
4  …            0.3         62.47                  87                  138
```

```
   year  month  day  day_of_week  hour_of_day  trip_duration
0  2018      3   29            3           13         2388.0
1  2018      3   29            3           13         2295.0
2  2018      3   29            3           13           66.0
3  2018      3   29            3           13         3317.0
4  2018      3   29            3           14         3648.0

[5 rows x 21 columns]
```

Now, the datetime columns can be dropped entirely.

```
[18]: td.drop(['pickup_datetime', 'dropoff_datetime'], axis=1, inplace=True)
```

Displaying the dataset's current state ...

```
[19]: td.head()
```

```
[19]:    trip_distance  rate_code store_and_fwd_flag  payment_type  fare_amount  \
0           18.15          3                  N             1         70.0
1            4.59          1                  N             1         25.0
2            0.30          1                  N             1          3.0
3           16.97          1                  N             1         49.5
4           14.45          1                  N             1         45.5

   extra  mta_tax  tip_amount  tolls_amount  imp_surcharge  total_amount  \
0    0.0      0.0       16.16         10.50            0.3         96.96
1    0.0      0.5        5.16          0.00            0.3         30.96
2    0.0      0.5        0.76          0.00            0.3          4.56
3    0.0      0.5        5.61          5.76            0.3         61.67
4    0.0      0.5       10.41          5.76            0.3         62.47

   pickup_location_id  dropoff_location_id  year  month  day  day_of_week  \
0                 161                    1  2018      3   29            3
1                  13                  230  2018      3   29            3
2                 231                  231  2018      3   29            3
3                 231                  138  2018      3   29            3
4                  87                  138  2018      3   29            3

   hour_of_day  trip_duration
0           13         2388.0
1           13         2295.0
2           13           66.0
3           13         3317.0
4           14         3648.0
```

Now that the dates have been broken down properly, a higher level of data clean-up can be performed.

- Any trips with a duration of 0 need to be dropped. These trips won't be useful, and are certainly due to a data entry error.
- Investigate what years are available in this dataset, how much of the dataset each year makes up, and begin investigating whether we should keep all years, or only specific years by visualizing trends in fare amounts when compared to trip duration and distance.

[20]:
```
td = td[td['trip_duration'] > 0]
```

[21]:
```
list_of_years = td.year.unique()
print(list_of_years)
```

```
[2018 2009 2017 2019 2008 2020 2003 2002 2001 2029 2032]
```

[22]:
```
for year in list_of_years:
    year_amount = td[td['year'] == year].shape[0]
    total_amount = td.shape[0]

    print(f"{year} makes up {(year_amount / total_amount) * 100}% of the␣
 ↪dataset")
```

```
2018 makes up 99.99841696039846% of the dataset
2009 makes up 0.0006356143854646263% of the dataset
2017 makes up 0.00020387631231884242% of the dataset
2019 makes up 0.00014391269104859462% of the dataset
2008 makes up 0.0004917016944160317% of the dataset
2020 makes up 1.1992724254049552e-05% of the dataset
2003 makes up 2.3985448508099104e-05% of the dataset
2002 makes up 2.3985448508099104e-05% of the dataset
2001 makes up 2.3985448508099104e-05% of the dataset
2029 makes up 1.1992724254049552e-05% of the dataset
2032 makes up 1.1992724254049552e-05% of the dataset
```

3.6 Eliminate Off-Trend Data

It's clear that this dataset is HEAVILY weighted towards 2018. For that reason, dropping anything from before 2018 can help avoid skewing the data towards old trends, while keeping anything newer than 2018 might reveal new trends.

If a dataset of such massive size consists of 99% of the same year, it's likely that the trips from newer years are either invalid data upon collection, and incomplete enough to actually show any trends.

All rows but 2018 are, therefore, dropped.

[23]:
```
td = td[td['year'] == 2018]
# Evaluate data stats after dropping
td.describe()
```

```
[23]:        trip_distance     rate_code  payment_type   fare_amount         extra  \
      count   8.338257e+06  8.338257e+06  8.338257e+06  8.338257e+06  8.338257e+06
      mean    9.120187e+00  1.154223e+00  1.180907e+00  3.178215e+01  3.469645e-01
      std     5.879868e+00  6.330880e-01  4.073165e-01  7.560952e+01  5.659283e-01
      min     1.000000e-02  1.000000e+00  1.000000e+00  1.000000e-02 -8.000000e+01
      25%     6.030000e+00  1.000000e+00  1.000000e+00  2.350000e+01  0.000000e+00
      50%     8.600000e+00  1.000000e+00  1.000000e+00  2.900000e+01  0.000000e+00
      75%     1.121000e+01  1.000000e+00  1.000000e+00  3.700000e+01  5.000000e-01
      max     7.655760e+03  9.900000e+01  4.000000e+00  1.874365e+05  2.020000e+01

                  mta_tax    tip_amount  tolls_amount  imp_surcharge  total_amount  \
      count  8.338257e+06  8.338257e+06  8.338257e+06   8.338257e+06  8.338257e+06
      mean   4.882261e-01  5.526390e+00  2.174295e+00   2.999538e-01  4.062672e+01
      std    8.265593e-02  4.568232e+00  3.748963e+00   3.744167e-03  7.668925e+01
      min    0.000000e+00  0.000000e+00 -5.760000e+00   0.000000e+00  3.100000e-01
      25%    5.000000e-01  2.000000e+00  0.000000e+00   3.000000e-01  2.915000e+01
      50%    5.000000e-01  5.550000e+00  0.000000e+00   3.000000e-01  3.755000e+01
      75%    5.000000e-01  7.910000e+00  5.760000e+00   3.000000e-01  4.901000e+01
      max    8.080000e+01  4.220000e+02  9.182500e+02   6.000000e-01  1.874378e+05

             pickup_location_id  dropoff_location_id       year         month  \
      count        8.338257e+06         8.338257e+06  8338257.0  8.338257e+06
      mean         1.528662e+02         1.476428e+02     2018.0  6.459984e+00
      std          6.017347e+01         7.560037e+01        0.0  3.423810e+00
      min          1.000000e+00         1.000000e+00     2018.0  1.000000e+00
      25%          1.320000e+02         8.800000e+01     2018.0  3.000000e+00
      50%          1.380000e+02         1.420000e+02     2018.0  6.000000e+00
      75%          1.860000e+02         2.290000e+02     2018.0  1.000000e+01
      max          2.650000e+02         2.650000e+02     2018.0  1.200000e+01

                      day   day_of_week   hour_of_day  trip_duration
      count  8.338257e+06  8.338257e+06  8.338257e+06   8.338257e+06
      mean   1.576347e+01  2.950375e+00  1.380998e+01   2.210049e+03
      std    8.640502e+00  1.930177e+00  6.231820e+00   4.865978e+03
      min    1.000000e+00  0.000000e+00  0.000000e+00   1.000000e+00
      25%    9.000000e+00  1.000000e+00  1.000000e+01   1.403000e+03
      50%    1.600000e+01  3.000000e+00  1.400000e+01   1.835000e+03
      75%    2.300000e+01  5.000000e+00  1.900000e+01   2.348000e+03
      max    3.100000e+01  6.000000e+00  2.300000e+01   3.200310e+05
```

3.7 Trip Fare - Sanity Checks

The value of *total_amount* should be equal to the sum of the *fare_amount*, *mta_tax*, *tip_amount*, *tolls_amount*, *imp_surcharge* and the *extra*.

Calculating total amounts and dropping rows whose values don't ”add up”...

3.7.1 Drop Fare Columns with Negative Values

```
[25]: init_count = len(td)
      td = td[td['fare_amount'] >= 0]
      td = td[td['extra'] >= 0]
      td = td[td['mta_tax'] >= 0]
      td = td[td['tip_amount'] >= 0]
      td = td[td['imp_surcharge'] >= 0]
      td = td[td['tolls_amount'] >= 0]
      final_count = len(td)

      print(f"Fraction of dataframe retained: {final_count / init_count * 100}%")
```

```
Fraction of dataframe retained: 99.99989206377305%
```

3.7.2 Verify that the Fare Values Add Up

```
[26]: # Calculating total amounts and dropping rows whose values don't "add up"...
      fare = td['fare_amount'].values
      extra = np.add(fare, td['extra'].values)
      mta_tax = np.add(extra, td['mta_tax'].values)
      tip_amount = np.add(mta_tax, td['tip_amount'].values)
      imp_surcharge = np.add(tip_amount, td['imp_surcharge'].values)
      calculated_total_amount = np.add(imp_surcharge, td['tolls_amount'].values)

      td['calculated_total_amount'] = pd.Series(calculated_total_amount)

      # validate calculated total by manually adding all relevant columns and␣
       ↪comparing to the calculated column
      td.head(10)
```

```
[26]:    trip_distance  rate_code store_and_fwd_flag  payment_type  fare_amount  \
      0          18.15          3                  N             1         70.0
      1           4.59          1                  N             1         25.0
      2           0.30          1                  N             1          3.0
      3          16.97          1                  N             1         49.5
      4          14.45          1                  N             1         45.5
      5          11.60          1                  N             1         42.0
      6           5.80          1                  N             1         24.0
      7           3.38          1                  N             1         25.0
      8          16.98          3                  N             1         85.0
      9           4.99          1                  N             1         22.0

         extra  mta_tax  tip_amount  tolls_amount  imp_surcharge  total_amount  \
      0    0.0      0.0       16.16         10.50            0.3         96.96
      1    0.0      0.5        5.16          0.00            0.3         30.96
```

```
2    0.0      0.5        0.76          0.00           0.3          4.56
3    0.0      0.5        5.61          5.76           0.3         61.67
4    0.0      0.5       10.41          5.76           0.3         62.47
5    0.0      0.5       14.57          5.76           0.3         63.13
6    0.0      0.5        4.95          0.00           0.3         29.75
7    0.0      0.5        5.16          0.00           0.3         30.96
8    0.0      0.0       15.00         12.50           0.3        112.80
9    1.0      0.5        4.76          0.00           0.3         28.56

   pickup_location_id  dropoff_location_id  year  month  day  day_of_week  \
0                 161                    1  2018      3   29            3
1                  13                  230  2018      3   29            3
2                 231                  231  2018      3   29            3
3                 231                  138  2018      3   29            3
4                  87                  138  2018      3   29            3
5                  68                  138  2018      3   29            3
6                 100                   87  2018      3   29            3
7                 144                  161  2018      3   29            3
8                  87                    1  2018      3   29            3
9                  13                  161  2018      3   29            3

   hour_of_day  trip_duration  calculated_total_amount
0           13         2388.0                    96.96
1           13         2295.0                    30.96
2           13           66.0                     4.56
3           13         3317.0                    61.67
4           14         3648.0                    62.47
5           14         3540.0                    63.13
6           14         1608.0                    29.75
7           15         2554.0                    30.96
8           15         5267.0                   112.80
9           16         1810.0                    28.56
```

Dropping incorrect `total_amount` values

[27]:
```
# Dropping incorrect `total_amount` values
init_count = len(td)
td = td[td['total_amount'] != td['calculated_total_amount']]
final_count = len(td)

print(f"Fraction of dataframe retained: {final_count / init_count * 100}%")

td.describe()
```

```
Fraction of dataframe retained: 99.7802655905653%
```

```
[27]:        trip_distance     rate_code  payment_type   fare_amount         extra  \
count     8.319926e+06  8.319926e+06  8.319926e+06  8.319926e+06  8.319926e+06
mean      9.126148e+00  1.154471e+00  1.180647e+00  3.179688e+01  3.470340e-01
std       5.882454e+00  6.336688e-01  4.070884e-01  7.558217e+01  5.652676e-01
min       1.000000e-02  1.000000e+00  1.000000e+00  1.000000e-02  0.000000e+00
25%       6.040000e+00  1.000000e+00  1.000000e+00  2.350000e+01  0.000000e+00
50%       8.600000e+00  1.000000e+00  1.000000e+00  2.900000e+01  0.000000e+00
75%       1.122000e+01  1.000000e+00  1.000000e+00  3.700000e+01  5.000000e-01
max       7.655760e+03  9.900000e+01  4.000000e+00  1.874365e+05  2.020000e+01

            mta_tax    tip_amount  tolls_amount  imp_surcharge  total_amount  \
count  8.319926e+06  8.319926e+06  8.319926e+06   8.319926e+06  8.319926e+06
mean   4.881855e-01  5.530809e+00  2.178277e+00   2.999539e-01  4.064989e+01
std    7.630298e-02  4.570137e+00  3.751520e+00   3.741069e-03  7.666306e+01
min    0.000000e+00  0.000000e+00  0.000000e+00   0.000000e+00  3.100000e-01
25%    5.000000e-01  2.000000e+00  0.000000e+00   3.000000e-01  2.915000e+01
50%    5.000000e-01  5.550000e+00  0.000000e+00   3.000000e-01  3.755000e+01
75%    5.000000e-01  7.910000e+00  5.760000e+00   3.000000e-01  4.906000e+01
max    2.150000e+01  4.220000e+02  9.182500e+02   6.000000e-01  1.874378e+05

       pickup_location_id  dropoff_location_id       year         month  \
count        8.319926e+06         8.319926e+06  8319926.0  8.319926e+06
mean         1.528594e+02         1.476394e+02     2018.0  6.460077e+00
std          6.015449e+01         7.559550e+01        0.0  3.423744e+00
min          1.000000e+00         1.000000e+00     2018.0  1.000000e+00
25%          1.320000e+02         8.800000e+01     2018.0  3.000000e+00
50%          1.380000e+02         1.420000e+02     2018.0  6.000000e+00
75%          1.860000e+02         2.290000e+02     2018.0  1.000000e+01
max          2.650000e+02         2.650000e+02     2018.0  1.200000e+01

                day   day_of_week   hour_of_day  trip_duration  \
count  8.319926e+06  8.319926e+06  8.319926e+06   8.319926e+06
mean   1.576325e+01  2.950070e+00  1.381030e+01   2.209896e+03
std    8.640600e+00  1.930190e+00  6.231147e+00   4.865080e+03
min    1.000000e+00  0.000000e+00  0.000000e+00   1.000000e+00
25%    9.000000e+00  1.000000e+00  1.000000e+01   1.403000e+03
50%    1.600000e+01  3.000000e+00  1.400000e+01   1.835000e+03
75%    2.300000e+01  5.000000e+00  1.900000e+01   2.348000e+03
max    3.100000e+01  6.000000e+00  2.300000e+01   3.200310e+05

       calculated_total_amount
count             7.656260e+06
mean              4.064563e+01
std               7.969093e+01
min               3.100000e-01
25%               2.915000e+01
50%               3.755000e+01
```

```
75%                 4.906000e+01
max                 1.874378e+05
```

4 Data Transformation

The following data transformation steps are applied to the cleansed data, to facilitate further data analysis and mining.

4.1 Compute and Add Driving Speed

A `driving_speed` attribute in addition to the existing data attributes can be useful to analyze traffic data in different geographic locations of the city. Average speed can be directly correlated with the traffic density.

[30]:

```
trip_distance_array = td['trip_distance']
trip_duration_array = td['trip_duration']

# driving_speed = trip_dist (in miles) / trip_duration (in sec) * 3600 sec
driving_speed = np.divide(trip_distance_array, trip_duration_array)*3600

td['driving_speed'] = pd.Series(trip_duration)
```

4.2 Compute and Add Tipping Rate

`tipping_rate` can help to analyze the general proportion of tipping that cab riders usually pay for their rides.

[33]:

```
tip_amount_array = td['tip_amount']
total_amount_array = td['total_amount']

# tipping rate = tip_amount / total_amount
tipping_rate = np.divide(tip_amount_array, total_amount_array)

td['tipping_rate'] = tipping_rate
```

5 Finishing Up

The *total_amount* column did a lot more than just clean totals, but it actually checked all of the other total effecting columns at the same time. If any errors occurred in any column, the calculated total would have differed from the calculated total.

Missing *mta_tax*, and incorrect *toll_amount* values are dropped.

```
[ ]: # this is a quick, easy way to de-allocate the memory assigned to df, which␣
      ↪holds the original dataframe
     # this was necessary else the write to csv function of Pandas (to_csv) would␣
      ↪max out the allowed memory in the notebook environment on Kaggle.
     df=[]
```

Save the cleaned dataframe to a CSV.

```
[ ]: td.to_csv(os.path.join(DATA_PATH, 'taxi-trip-data_2018_cleaned.csv'))
     print('Done!')
```

6 Save the notebook

```
[34]: !apt-get install texlive texlive-xetex texlive-latex-extra pandoc
      !pip install pypandoc
```

```
Reading package lists… Done
Building dependency tree
Reading state information… Done
pandoc is already the newest version (1.19.2.4~dfsg-1build4).
pandoc set to manually installed.
The following package was automatically installed and is no longer required:
  libnvidia-common-460
Use 'apt autoremove' to remove it.
The following additional packages will be installed:
  fonts-droid-fallback fonts-lato fonts-lmodern fonts-noto-mono fonts-texgyre
  javascript-common libcupsfilters1 libcupsimage2 libgs9 libgs9-common
  libijs-0.35 libjbig2dec0 libjs-jquery libkpathsea6 libpotrace0 libptexenc1
  libruby2.5 libsynctex1 libtexlua52 libtexluajit2 libzzip-0-13 lmodern
  poppler-data preview-latex-style rake ruby ruby-did-you-mean ruby-minitest
  ruby-net-telnet ruby-power-assert ruby-test-unit ruby2.5
  rubygems-integration t1utils tex-common tex-gyre texlive-base
  texlive-binaries texlive-fonts-recommended texlive-latex-base
  texlive-latex-recommended texlive-pictures texlive-plain-generic tipa
Suggested packages:
  fonts-noto apache2 | lighttpd | httpd poppler-utils ghostscript
  fonts-japanese-mincho | fonts-ipafont-mincho fonts-japanese-gothic
  | fonts-ipafont-gothic fonts-arphic-ukai fonts-arphic-uming fonts-nanum ri
  ruby-dev bundler debhelper gv | postscript-viewer perl-tk xpdf-reader
  | pdf-viewer texlive-fonts-recommended-doc texlive-latex-base-doc
  python-pygments icc-profiles libfile-which-perl
  libspreadsheet-parseexcel-perl texlive-latex-extra-doc
  texlive-latex-recommended-doc texlive-pstricks dot2tex prerex ruby-tcltk
  | libtcltk-ruby texlive-pictures-doc vprerex
The following NEW packages will be installed:
  fonts-droid-fallback fonts-lato fonts-lmodern fonts-noto-mono fonts-texgyre
  javascript-common libcupsfilters1 libcupsimage2 libgs9 libgs9-common
  libijs-0.35 libjbig2dec0 libjs-jquery libkpathsea6 libpotrace0 libptexenc1
  libruby2.5 libsynctex1 libtexlua52 libtexluajit2 libzzip-0-13 lmodern
  poppler-data preview-latex-style rake ruby ruby-did-you-mean ruby-minitest
  ruby-net-telnet ruby-power-assert ruby-test-unit ruby2.5
  rubygems-integration t1utils tex-common tex-gyre texlive texlive-base
  texlive-binaries texlive-fonts-recommended texlive-latex-base
  texlive-latex-extra texlive-latex-recommended texlive-pictures
  texlive-plain-generic texlive-xetex tipa
0 upgraded, 47 newly installed, 0 to remove and 32 not upgraded.
Need to get 146 MB of archives.
After this operation, 460 MB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu bionic/main amd64 fonts-droid-fallback
all 1:6.0.1r16-1.1 [1,805 kB]
Get:2 http://archive.ubuntu.com/ubuntu bionic/main amd64 fonts-lato all 2.0-2
[2,698 kB]
Get:3 http://archive.ubuntu.com/ubuntu bionic/main amd64 poppler-data all
0.4.8-2 [1,479 kB]
Get:4 http://archive.ubuntu.com/ubuntu bionic/main amd64 tex-common all 6.09
[33.0 kB]
Get:5 http://archive.ubuntu.com/ubuntu bionic/main amd64 fonts-lmodern all
2.004.5-3 [4,551 kB]
Get:6 http://archive.ubuntu.com/ubuntu bionic/main amd64 fonts-noto-mono all
20171026-2 [75.5 kB]
Get:7 http://archive.ubuntu.com/ubuntu bionic/universe amd64 fonts-texgyre all
20160520-1 [8,761 kB]
Get:8 http://archive.ubuntu.com/ubuntu bionic/main amd64 javascript-common all
11 [6,066 B]
Get:9 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libcupsfilters1
amd64 1.20.2-0ubuntu3.1 [108 kB]
Get:10 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libcupsimage2
amd64 2.2.7-1ubuntu2.9 [18.6 kB]
Get:11 http://archive.ubuntu.com/ubuntu bionic/main amd64 libijs-0.35 amd64
0.35-13 [15.5 kB]
Get:12 http://archive.ubuntu.com/ubuntu bionic/main amd64 libjbig2dec0 amd64
0.13-6 [55.9 kB]
Get:13 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libgs9-common
all 9.26~dfsg+0-0ubuntu0.18.04.16 [5,093 kB]
Get:14 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libgs9 amd64
9.26~dfsg+0-0ubuntu0.18.04.16 [2,265 kB]
Get:15 http://archive.ubuntu.com/ubuntu bionic/main amd64 libjs-jquery all
3.2.1-1 [152 kB]
Get:16 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libkpathsea6
amd64 2017.20170613.44572-8ubuntu0.1 [54.9 kB]
Get:17 http://archive.ubuntu.com/ubuntu bionic/main amd64 libpotrace0 amd64
1.14-2 [17.4 kB]
Get:18 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libptexenc1
amd64 2017.20170613.44572-8ubuntu0.1 [34.5 kB]
```

```
Get:19 http://archive.ubuntu.com/ubuntu bionic/main amd64 rubygems-integration
all 1.11 [4,994 B]
Get:20 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 ruby2.5 amd64
2.5.1-1ubuntu1.12 [48.6 kB]
Get:21 http://archive.ubuntu.com/ubuntu bionic/main amd64 ruby amd64 1:2.5.1
[5,712 B]
Get:22 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 rake all
12.3.1-1ubuntu0.1 [44.9 kB]
Get:23 http://archive.ubuntu.com/ubuntu bionic/main amd64 ruby-did-you-mean all
1.2.0-2 [9,700 B]
Get:24 http://archive.ubuntu.com/ubuntu bionic/main amd64 ruby-minitest all
5.10.3-1 [38.6 kB]
Get:25 http://archive.ubuntu.com/ubuntu bionic/main amd64 ruby-net-telnet all
0.1.1-2 [12.6 kB]
Get:26 http://archive.ubuntu.com/ubuntu bionic/main amd64 ruby-power-assert all
0.3.0-1 [7,952 B]
Get:27 http://archive.ubuntu.com/ubuntu bionic/main amd64 ruby-test-unit all
3.2.5-1 [61.1 kB]
Get:28 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libruby2.5
amd64 2.5.1-1ubuntu1.12 [3,073 kB]
Get:29 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libsynctex1
amd64 2017.20170613.44572-8ubuntu0.1 [41.4 kB]
Get:30 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libtexlua52
amd64 2017.20170613.44572-8ubuntu0.1 [91.2 kB]
Get:31 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libtexluajit2
amd64 2017.20170613.44572-8ubuntu0.1 [230 kB]
Get:32 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libzzip-0-13
amd64 0.13.62-3.1ubuntu0.18.04.1 [26.0 kB]
Get:33 http://archive.ubuntu.com/ubuntu bionic/main amd64 lmodern all 2.004.5-3
[9,631 kB]
Get:34 http://archive.ubuntu.com/ubuntu bionic/main amd64 preview-latex-style
all 11.91-1ubuntu1 [185 kB]
Get:35 http://archive.ubuntu.com/ubuntu bionic/main amd64 t1utils amd64 1.41-2
[56.0 kB]
Get:36 http://archive.ubuntu.com/ubuntu bionic/universe amd64 tex-gyre all
20160520-1 [4,998 kB]
Get:37 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 texlive-
binaries amd64 2017.20170613.44572-8ubuntu0.1 [8,179 kB]
Get:38 http://archive.ubuntu.com/ubuntu bionic/main amd64 texlive-base all
2017.20180305-1 [18.7 MB]
Get:39 http://archive.ubuntu.com/ubuntu bionic/universe amd64 texlive-fonts-
recommended all 2017.20180305-1 [5,262 kB]
Get:40 http://archive.ubuntu.com/ubuntu bionic/main amd64 texlive-latex-base all
2017.20180305-1 [951 kB]
Get:41 http://archive.ubuntu.com/ubuntu bionic/main amd64 texlive-latex-
recommended all 2017.20180305-1 [14.9 MB]
Get:42 http://archive.ubuntu.com/ubuntu bionic/universe amd64 texlive all
2017.20180305-1 [14.4 kB]
```

```
Get:43 http://archive.ubuntu.com/ubuntu bionic/universe amd64 texlive-pictures
all 2017.20180305-1 [4,026 kB]
Get:44 http://archive.ubuntu.com/ubuntu bionic/universe amd64 texlive-latex-
extra all 2017.20180305-2 [10.6 MB]
Get:45 http://archive.ubuntu.com/ubuntu bionic/universe amd64 texlive-plain-
generic all 2017.20180305-2 [23.6 MB]
Get:46 http://archive.ubuntu.com/ubuntu bionic/universe amd64 tipa all 2:1.3-20
[2,978 kB]
Get:47 http://archive.ubuntu.com/ubuntu bionic/universe amd64 texlive-xetex all
2017.20180305-1 [10.7 MB]
Fetched 146 MB in 5s (28.2 MB/s)
Extracting templates from packages: 100%
Preconfiguring packages …
Selecting previously unselected package fonts-droid-fallback.
(Reading database … 155685 files and directories currently installed.)
Preparing to unpack …/00-fonts-droid-fallback_1%3a6.0.1r16-1.1_all.deb …
Unpacking fonts-droid-fallback (1:6.0.1r16-1.1) …
Selecting previously unselected package fonts-lato.
Preparing to unpack …/01-fonts-lato_2.0-2_all.deb …
Unpacking fonts-lato (2.0-2) …
Selecting previously unselected package poppler-data.
Preparing to unpack …/02-poppler-data_0.4.8-2_all.deb …
Unpacking poppler-data (0.4.8-2) …
Selecting previously unselected package tex-common.
Preparing to unpack …/03-tex-common_6.09_all.deb …
Unpacking tex-common (6.09) …
Selecting previously unselected package fonts-lmodern.
Preparing to unpack …/04-fonts-lmodern_2.004.5-3_all.deb …
Unpacking fonts-lmodern (2.004.5-3) …
Selecting previously unselected package fonts-noto-mono.
Preparing to unpack …/05-fonts-noto-mono_20171026-2_all.deb …
Unpacking fonts-noto-mono (20171026-2) …
Selecting previously unselected package fonts-texgyre.
Preparing to unpack …/06-fonts-texgyre_20160520-1_all.deb …
Unpacking fonts-texgyre (20160520-1) …
Selecting previously unselected package javascript-common.
Preparing to unpack …/07-javascript-common_11_all.deb …
Unpacking javascript-common (11) …
Selecting previously unselected package libcupsfilters1:amd64.
Preparing to unpack …/08-libcupsfilters1_1.20.2-0ubuntu3.1_amd64.deb …
Unpacking libcupsfilters1:amd64 (1.20.2-0ubuntu3.1) …
Selecting previously unselected package libcupsimage2:amd64.
Preparing to unpack …/09-libcupsimage2_2.2.7-1ubuntu2.9_amd64.deb …
Unpacking libcupsimage2:amd64 (2.2.7-1ubuntu2.9) …
Selecting previously unselected package libijs-0.35:amd64.
Preparing to unpack …/10-libijs-0.35_0.35-13_amd64.deb …
Unpacking libijs-0.35:amd64 (0.35-13) …
Selecting previously unselected package libjbig2dec0:amd64.
```

```
Preparing to unpack …/11-libjbig2dec0_0.13-6_amd64.deb …
Unpacking libjbig2dec0:amd64 (0.13-6) …
Selecting previously unselected package libgs9-common.
Preparing to unpack …/12-libgs9-common_9.26~dfsg+0-0ubuntu0.18.04.16_all.deb
…
Unpacking libgs9-common (9.26~dfsg+0-0ubuntu0.18.04.16) …
Selecting previously unselected package libgs9:amd64.
Preparing to unpack …/13-libgs9_9.26~dfsg+0-0ubuntu0.18.04.16_amd64.deb …
Unpacking libgs9:amd64 (9.26~dfsg+0-0ubuntu0.18.04.16) …
Selecting previously unselected package libjs-jquery.
Preparing to unpack …/14-libjs-jquery_3.2.1-1_all.deb …
Unpacking libjs-jquery (3.2.1-1) …
Selecting previously unselected package libkpathsea6:amd64.
Preparing to unpack …/15-libkpathsea6_2017.20170613.44572-8ubuntu0.1_amd64.deb
…
Unpacking libkpathsea6:amd64 (2017.20170613.44572-8ubuntu0.1) …
Selecting previously unselected package libpotrace0.
Preparing to unpack …/16-libpotrace0_1.14-2_amd64.deb …
Unpacking libpotrace0 (1.14-2) …
Selecting previously unselected package libptexenc1:amd64.
Preparing to unpack …/17-libptexenc1_2017.20170613.44572-8ubuntu0.1_amd64.deb
…
Unpacking libptexenc1:amd64 (2017.20170613.44572-8ubuntu0.1) …
Selecting previously unselected package rubygems-integration.
Preparing to unpack …/18-rubygems-integration_1.11_all.deb …
Unpacking rubygems-integration (1.11) …
Selecting previously unselected package ruby2.5.
Preparing to unpack …/19-ruby2.5_2.5.1-1ubuntu1.12_amd64.deb …
Unpacking ruby2.5 (2.5.1-1ubuntu1.12) …
Selecting previously unselected package ruby.
Preparing to unpack …/20-ruby_1%3a2.5.1_amd64.deb …
Unpacking ruby (1:2.5.1) …
Selecting previously unselected package rake.
Preparing to unpack …/21-rake_12.3.1-1ubuntu0.1_all.deb …
Unpacking rake (12.3.1-1ubuntu0.1) …
Selecting previously unselected package ruby-did-you-mean.
Preparing to unpack …/22-ruby-did-you-mean_1.2.0-2_all.deb …
Unpacking ruby-did-you-mean (1.2.0-2) …
Selecting previously unselected package ruby-minitest.
Preparing to unpack …/23-ruby-minitest_5.10.3-1_all.deb …
Unpacking ruby-minitest (5.10.3-1) …
Selecting previously unselected package ruby-net-telnet.
Preparing to unpack …/24-ruby-net-telnet_0.1.1-2_all.deb …
Unpacking ruby-net-telnet (0.1.1-2) …
Selecting previously unselected package ruby-power-assert.
Preparing to unpack …/25-ruby-power-assert_0.3.0-1_all.deb …
Unpacking ruby-power-assert (0.3.0-1) …
Selecting previously unselected package ruby-test-unit.
```

```
Preparing to unpack …/26-ruby-test-unit_3.2.5-1_all.deb …
Unpacking ruby-test-unit (3.2.5-1) …
Selecting previously unselected package libruby2.5:amd64.
Preparing to unpack …/27-libruby2.5_2.5.1-1ubuntu1.12_amd64.deb …
Unpacking libruby2.5:amd64 (2.5.1-1ubuntu1.12) …
Selecting previously unselected package libsynctex1:amd64.
Preparing to unpack …/28-libsynctex1_2017.20170613.44572-8ubuntu0.1_amd64.deb
…
Unpacking libsynctex1:amd64 (2017.20170613.44572-8ubuntu0.1) …
Selecting previously unselected package libtexlua52:amd64.
Preparing to unpack …/29-libtexlua52_2017.20170613.44572-8ubuntu0.1_amd64.deb
…
Unpacking libtexlua52:amd64 (2017.20170613.44572-8ubuntu0.1) …
Selecting previously unselected package libtexluajit2:amd64.
Preparing to unpack
…/30-libtexluajit2_2017.20170613.44572-8ubuntu0.1_amd64.deb …
Unpacking libtexluajit2:amd64 (2017.20170613.44572-8ubuntu0.1) …
Selecting previously unselected package libzzip-0-13:amd64.
Preparing to unpack …/31-libzzip-0-13_0.13.62-3.1ubuntu0.18.04.1_amd64.deb …
Unpacking libzzip-0-13:amd64 (0.13.62-3.1ubuntu0.18.04.1) …
Selecting previously unselected package lmodern.
Preparing to unpack …/32-lmodern_2.004.5-3_all.deb …
Unpacking lmodern (2.004.5-3) …
Selecting previously unselected package preview-latex-style.
Preparing to unpack …/33-preview-latex-style_11.91-1ubuntu1_all.deb …
Unpacking preview-latex-style (11.91-1ubuntu1) …
Selecting previously unselected package t1utils.
Preparing to unpack …/34-t1utils_1.41-2_amd64.deb …
Unpacking t1utils (1.41-2) …
Selecting previously unselected package tex-gyre.
Preparing to unpack …/35-tex-gyre_20160520-1_all.deb …
Unpacking tex-gyre (20160520-1) …
Selecting previously unselected package texlive-binaries.
Preparing to unpack …/36-texlive-
binaries_2017.20170613.44572-8ubuntu0.1_amd64.deb …
Unpacking texlive-binaries (2017.20170613.44572-8ubuntu0.1) …
Selecting previously unselected package texlive-base.
Preparing to unpack …/37-texlive-base_2017.20180305-1_all.deb …
Unpacking texlive-base (2017.20180305-1) …
Selecting previously unselected package texlive-fonts-recommended.
Preparing to unpack …/38-texlive-fonts-recommended_2017.20180305-1_all.deb …
Unpacking texlive-fonts-recommended (2017.20180305-1) …
Selecting previously unselected package texlive-latex-base.
Preparing to unpack …/39-texlive-latex-base_2017.20180305-1_all.deb …
Unpacking texlive-latex-base (2017.20180305-1) …
Selecting previously unselected package texlive-latex-recommended.
Preparing to unpack …/40-texlive-latex-recommended_2017.20180305-1_all.deb …
Unpacking texlive-latex-recommended (2017.20180305-1) …
```

```
Selecting previously unselected package texlive.
Preparing to unpack …/41-texlive_2017.20180305-1_all.deb …
Unpacking texlive (2017.20180305-1) …
Selecting previously unselected package texlive-pictures.
Preparing to unpack …/42-texlive-pictures_2017.20180305-1_all.deb …
Unpacking texlive-pictures (2017.20180305-1) …
Selecting previously unselected package texlive-latex-extra.
Preparing to unpack …/43-texlive-latex-extra_2017.20180305-2_all.deb …
Unpacking texlive-latex-extra (2017.20180305-2) …
Selecting previously unselected package texlive-plain-generic.
Preparing to unpack …/44-texlive-plain-generic_2017.20180305-2_all.deb …
Unpacking texlive-plain-generic (2017.20180305-2) …
Selecting previously unselected package tipa.
Preparing to unpack …/45-tipa_2%3a1.3-20_all.deb …
Unpacking tipa (2:1.3-20) …
Selecting previously unselected package texlive-xetex.
Preparing to unpack …/46-texlive-xetex_2017.20180305-1_all.deb …
Unpacking texlive-xetex (2017.20180305-1) …
Setting up libgs9-common (9.26~dfsg+0-0ubuntu0.18.04.16) …
Setting up libkpathsea6:amd64 (2017.20170613.44572-8ubuntu0.1) …
Setting up libjs-jquery (3.2.1-1) …
Setting up libtexlua52:amd64 (2017.20170613.44572-8ubuntu0.1) …
Setting up fonts-droid-fallback (1:6.0.1r16-1.1) …
Setting up libsynctex1:amd64 (2017.20170613.44572-8ubuntu0.1) …
Setting up libptexenc1:amd64 (2017.20170613.44572-8ubuntu0.1) …
Setting up tex-common (6.09) …
update-language: texlive-base not installed and configured, doing nothing!
Setting up poppler-data (0.4.8-2) …
Setting up tex-gyre (20160520-1) …
Setting up preview-latex-style (11.91-1ubuntu1) …
Setting up fonts-texgyre (20160520-1) …
Setting up fonts-noto-mono (20171026-2) …
Setting up fonts-lato (2.0-2) …
Setting up libcupsfilters1:amd64 (1.20.2-0ubuntu3.1) …
Setting up libcupsimage2:amd64 (2.2.7-1ubuntu2.9) …
Setting up libjbig2dec0:amd64 (0.13-6) …
Setting up ruby-did-you-mean (1.2.0-2) …
Setting up t1utils (1.41-2) …
Setting up ruby-net-telnet (0.1.1-2) …
Setting up libijs-0.35:amd64 (0.35-13) …
Setting up rubygems-integration (1.11) …
Setting up libpotrace0 (1.14-2) …
Setting up javascript-common (11) …
Setting up ruby-minitest (5.10.3-1) …
Setting up libzzip-0-13:amd64 (0.13.62-3.1ubuntu0.18.04.1) …
Setting up libgs9:amd64 (9.26~dfsg+0-0ubuntu0.18.04.16) …
Setting up libtexluajit2:amd64 (2017.20170613.44572-8ubuntu0.1) …
Setting up fonts-lmodern (2.004.5-3) …
Setting up ruby-power-assert (0.3.0-1) …
Setting up texlive-binaries (2017.20170613.44572-8ubuntu0.1) …
update-alternatives: using /usr/bin/xdvi-xaw to provide /usr/bin/xdvi.bin
(xdvi.bin) in auto mode
update-alternatives: using /usr/bin/bibtex.original to provide /usr/bin/bibtex
(bibtex) in auto mode
Setting up texlive-base (2017.20180305-1) …
mktexlsr: Updating /var/lib/texmf/ls-R-TEXLIVEDIST…
mktexlsr: Updating /var/lib/texmf/ls-R-TEXMFMAIN…
mktexlsr: Updating /var/lib/texmf/ls-R…
mktexlsr: Done.
tl-paper: setting paper size for dvips to a4:
/var/lib/texmf/dvips/config/config-paper.ps
tl-paper: setting paper size for dvipdfmx to a4:
/var/lib/texmf/dvipdfmx/dvipdfmx-paper.cfg
tl-paper: setting paper size for xdvi to a4: /var/lib/texmf/xdvi/XDvi-paper
tl-paper: setting paper size for pdftex to a4:
/var/lib/texmf/tex/generic/config/pdftexconfig.tex
Setting up texlive-fonts-recommended (2017.20180305-1) …
Setting up texlive-plain-generic (2017.20180305-2) …
Setting up texlive-latex-base (2017.20180305-1) …
Setting up lmodern (2.004.5-3) …
Setting up texlive-latex-recommended (2017.20180305-1) …
Setting up texlive-pictures (2017.20180305-1) …
Setting up tipa (2:1.3-20) …
Regenerating '/var/lib/texmf/fmtutil.cnf-DEBIAN'… done.
Regenerating '/var/lib/texmf/fmtutil.cnf-TEXLIVEDIST'… done.
update-fmtutil has updated the following file(s):
        /var/lib/texmf/fmtutil.cnf-DEBIAN
        /var/lib/texmf/fmtutil.cnf-TEXLIVEDIST
If you want to activate the changes in the above file(s),
you should run fmtutil-sys or fmtutil.
Setting up texlive (2017.20180305-1) …
Setting up texlive-latex-extra (2017.20180305-2) …
Setting up texlive-xetex (2017.20180305-1) …
Setting up ruby2.5 (2.5.1-1ubuntu1.12) …
Setting up ruby (1:2.5.1) …
Setting up ruby-test-unit (3.2.5-1) …
Setting up rake (12.3.1-1ubuntu0.1) …
Setting up libruby2.5:amd64 (2.5.1-1ubuntu1.12) …
Processing triggers for mime-support (3.60ubuntu1) …
Processing triggers for libc-bin (2.27-3ubuntu1.5) …
Processing triggers for man-db (2.8.3-2ubuntu0.1) …
Processing triggers for fontconfig (2.12.6-0ubuntu2) …
Processing triggers for tex-common (6.09) …
Running updmap-sys. This may take some time… done.
Running mktexlsr /var/lib/texmf … done.
Building format(s) --all.
```

```
        This may take some time… done.
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting pypandoc
  Downloading pypandoc-1.8.1-py3-none-any.whl (20 kB)
Installing collected packages: pypandoc
Successfully installed pypandoc-1.8.1
```

[35]:

```
!cd /content/drive/MyDrive/Undergrad/Semester-7/DWDM_Preprocessing-Assignment
!ls
!jupyter nbconvert --to PDF "Analysis-Cleaning-Transformation.ipynb"

# Save processed CSV
td.to_csv(os.path.join(DATA_PATH, 'taxi-trip-data_2018_cleaned.csv'))
print('Done!')
```

```
[NbConvertApp] WARNING | pattern 'Analysis-Cleaning-Transformation.ipynb'
matched no files
This application is used to convert notebook files (*.ipynb)
        to various other formats.

        WARNING: THE COMMANDLINE INTERFACE MAY CHANGE IN FUTURE RELEASES.

Options
=======
The options below are convenience aliases to configurable class-options,
as listed in the "Equivalent to" description-line of the aliases.
To see all configurable class-options for some <cmd>, use:
    <cmd> --help-all

--debug
    set log level to logging.DEBUG (maximize logging output)
    Equivalent to: [--Application.log_level=10]
--show-config
    Show the application's configuration (human-readable format)
    Equivalent to: [--Application.show_config=True]
--show-config-json
    Show the application's configuration (json format)
    Equivalent to: [--Application.show_config_json=True]
--generate-config
    generate default config file
    Equivalent to: [--JupyterApp.generate_config=True]
-y
    Answer yes to any questions instead of prompting.
    Equivalent to: [--JupyterApp.answer_yes=True]
--execute
    Execute the notebook prior to export.
    Equivalent to: [--ExecutePreprocessor.enabled=True]
```

```
--allow-errors
    Continue notebook execution even if one of the cells throws an error and
include the error message in the cell output (the default behaviour is to abort
conversion). This flag is only relevant if '--execute' was specified, too.
    Equivalent to: [--ExecutePreprocessor.allow_errors=True]
--stdin
    read a single notebook file from stdin. Write the resulting notebook with
default basename 'notebook.*'
    Equivalent to: [--NbConvertApp.from_stdin=True]
--stdout
    Write notebook output to stdout instead of files.
    Equivalent to: [--NbConvertApp.writer_class=StdoutWriter]
--inplace
    Run nbconvert in place, overwriting the existing notebook (only
            relevant when converting to notebook format)
    Equivalent to: [--NbConvertApp.use_output_suffix=False
--NbConvertApp.export_format=notebook --FilesWriter.build_directory=]
--clear-output
    Clear output of current file and save in place,
            overwriting the existing notebook.
    Equivalent to: [--NbConvertApp.use_output_suffix=False
--NbConvertApp.export_format=notebook --FilesWriter.build_directory=
--ClearOutputPreprocessor.enabled=True]
--no-prompt
    Exclude input and output prompts from converted document.
    Equivalent to: [--TemplateExporter.exclude_input_prompt=True
--TemplateExporter.exclude_output_prompt=True]
--no-input
    Exclude input cells and output prompts from converted document.
            This mode is ideal for generating code-free reports.
    Equivalent to: [--TemplateExporter.exclude_output_prompt=True
--TemplateExporter.exclude_input=True]
--log-level=<Enum>
    Set the log level by value or name.
    Choices: any of [0, 10, 20, 30, 40, 50, 'DEBUG', 'INFO', 'WARN', 'ERROR',
'CRITICAL']
    Default: 30
    Equivalent to: [--Application.log_level]
--config=<Unicode>
    Full path of a config file.
    Default: ''
    Equivalent to: [--JupyterApp.config_file]
--to=<Unicode>
    The export format to be used, either one of the built-in formats
            ['asciidoc', 'custom', 'html', 'latex', 'markdown', 'notebook',
'pdf', 'python', 'rst', 'script', 'slides']
            or a dotted object name that represents the import path for an
            `Exporter` class
    Default: 'html'
    Equivalent to: [--NbConvertApp.export_format]
--template=<Unicode>
    Name of the template file to use
    Default: ''
    Equivalent to: [--TemplateExporter.template_file]
--writer=<DottedObjectName>
    Writer class used to write the
                                        results of the conversion
    Default: 'FilesWriter'
    Equivalent to: [--NbConvertApp.writer_class]
--post=<DottedOrNone>
    PostProcessor class used to write the
                                        results of the conversion
    Default: ''
    Equivalent to: [--NbConvertApp.postprocessor_class]
--output=<Unicode>
    overwrite base name use for output files.
                can only be used when converting one notebook at a time.
    Default: ''
    Equivalent to: [--NbConvertApp.output_base]
--output-dir=<Unicode>
    Directory to write output(s) to. Defaults
                                  to output to the directory of each notebook.
To recover
                                  previous default behaviour (outputting to the
current
                                  working directory) use . as the flag value.
    Default: ''
    Equivalent to: [--FilesWriter.build_directory]
--reveal-prefix=<Unicode>
    The URL prefix for reveal.js (version 3.x).
            This defaults to the reveal CDN, but can be any url pointing to a
copy
            of reveal.js.
            For speaker notes to work, this must be a relative path to a local
            copy of reveal.js: e.g., "reveal.js".
            If a relative path is given, it must be a subdirectory of the
            current directory (from which the server is run).
            See the usage documentation
            (https://nbconvert.readthedocs.io/en/latest/usage.html#reveal-js-
html-slideshow)
            for more details.
    Default: ''
    Equivalent to: [--SlidesExporter.reveal_url_prefix]
--nbformat=<Enum>
    The nbformat version to write.
            Use this to downgrade notebooks.
    Choices: any of [1, 2, 3, 4]
    Default: 4
    Equivalent to: [--NotebookExporter.nbformat_version]

Examples
--------

    The simplest way to use nbconvert is

            > jupyter nbconvert mynotebook.ipynb

            which will convert mynotebook.ipynb to the default format (probably
HTML).

            You can specify the export format with `--to`.
            Options include ['asciidoc', 'custom', 'html', 'latex', 'markdown',
'notebook', 'pdf', 'python', 'rst', 'script', 'slides'].

            > jupyter nbconvert --to latex mynotebook.ipynb

            Both HTML and LaTeX support multiple output templates. LaTeX
includes
            'base', 'article' and 'report'.  HTML includes 'basic' and 'full'.
You
            can specify the flavor of the format used.

            > jupyter nbconvert --to html --template basic mynotebook.ipynb

            You can also pipe the output to stdout, rather than a file

            > jupyter nbconvert mynotebook.ipynb --stdout

            PDF is generated via latex

            > jupyter nbconvert mynotebook.ipynb --to pdf

            You can get (and serve) a Reveal.js-powered slideshow

            > jupyter nbconvert myslides.ipynb --to slides --post serve

            Multiple notebooks can be given at the command line in a couple of
            different ways:

            > jupyter nbconvert notebook*.ipynb
            > jupyter nbconvert notebook1.ipynb notebook2.ipynb

            or you can specify the notebooks list in a config file, containing::
                c.NbConvertApp.notebooks = ["my_notebook.ipynb"]

            > jupyter nbconvert --config mycfg.py

To see all available configurables, use `--help-all`.
```

```
      ␣
↪---------------------------------------------------------------------------

       KeyboardInterrupt                         Traceback (most recent call␣
↪last)

       <ipython-input-35-3a656b49859b> in <module>
         3
         4 # Save processed CSV
   ----> 5 td.to_csv(os.path.join(DATA_PATH, 'taxi-trip-data_2018_cleaned.csv'))
         6 print('Done!')


       /usr/local/lib/python3.7/dist-packages/pandas/core/generic.py in␣
↪to_csv(self, path_or_buf, sep, na_rep, float_format, columns, header, index,␣
↪index_label, mode, encoding, compression, quoting, quotechar, line_terminator,␣
↪chunksize, date_format, doublequote, escapechar, decimal, errors,␣
↪storage_options)
      3480             doublequote=doublequote,
      3481             escapechar=escapechar,
   -> 3482             storage_options=storage_options,
      3483         )
      3484


       /usr/local/lib/python3.7/dist-packages/pandas/io/formats/format.py in␣
↪to_csv(self, path_or_buf, encoding, sep, columns, index_label, mode,␣
↪compression, quoting, quotechar, line_terminator, chunksize, date_format,␣
↪doublequote, escapechar, errors, storage_options)
      1103             formatter=self.fmt,
      1104         )
   -> 1105         csv_formatter.save()
      1106
      1107         if created_buffer:


       /usr/local/lib/python3.7/dist-packages/pandas/io/formats/csvs.py in␣
↪save(self)
       255             )
       256
   --> 257             self._save()
       258 
       259     def _save(self) -> None:


       /usr/local/lib/python3.7/dist-packages/pandas/io/formats/csvs.py in␣
↪_save(self)
       260         if self._need_to_save_header:
       261             self._save_header()
   --> 262         self._save_body()
       263 
       264     def _save_header(self) -> None:


       /usr/local/lib/python3.7/dist-packages/pandas/io/formats/csvs.py in␣
↪_save_body(self)
       298             if start_i >= end_i:
       299                 break
   --> 300             self._save_chunk(start_i, end_i)
       301 
       302     def _save_chunk(self, start_i: int, end_i: int) -> None:


       /usr/local/lib/python3.7/dist-packages/pandas/io/formats/csvs.py in␣
↪_save_chunk(self, start_i, end_i)
       305         df = self.obj.iloc[slicer]
       306 
   --> 307         res = df._mgr.to_native_types(**self._number_format)
       308         data = [res.iget_values(i) for i in range(len(res.items))]
       309 


       /usr/local/lib/python3.7/dist-packages/pandas/core/internals/managers.py␣
↪in to_native_types(self, **kwargs)
       464         in formatting (repr / csv).
       465         """
   --> 466         return self.apply("to_native_types", **kwargs)
       467 
       468     def is_consolidated(self) -> bool:


       /usr/local/lib/python3.7/dist-packages/pandas/core/internals/managers.py␣
↪in apply(self, f, align_keys, ignore_failures, **kwargs)
       325                     applied = b.apply(f, **kwargs)
       326                 else:
   --> 327                     applied = getattr(b, f)(**kwargs)
       328             except (TypeError, NotImplementedError):
```

```
    329                 if not ignore_failures:


    /usr/local/lib/python3.7/dist-packages/pandas/core/internals/blocks.py␣
↪in to_native_types(self, na_rep, quoting, **kwargs)
    639     def to_native_types(self, na_rep="nan", quoting=None, **kwargs):
    640         """convert to our native types format"""
--> 641         result = to_native_types(self.values, na_rep=na_rep,␣
↪quoting=quoting, **kwargs)
    642         return self.make_block(result)
    643


    /usr/local/lib/python3.7/dist-packages/pandas/core/internals/blocks.py␣
↪in to_native_types(values, na_rep, quoting, float_format, decimal, **kwargs)
   2053
   2054     if isinstance(values, (DatetimeArray, TimedeltaArray)):
-> 2055         result = values._format_native_types(na_rep=na_rep, **kwargs)
   2056         result = result.astype(object, copy=False)
   2057         return result


    /usr/local/lib/python3.7/dist-packages/pandas/core/arrays/_mixins.py in␣
↪method(self, *args, **kwargs)
     60         flags = self._ndarray.flags
     61         flat = self.ravel("K")
---> 62         result = meth(flat, *args, **kwargs)
     63         order = "F" if flags.f_contiguous else "C"
     64         return result.reshape(self.shape, order=order)


    /usr/local/lib/python3.7/dist-packages/pandas/core/arrays/timedeltas.py␣
↪in _format_native_types(self, na_rep, date_format, **kwargs)
    431
    432         formatter = get_format_timedelta64(self._ndarray, na_rep)
--> 433         return np.array([formatter(x) for x in self._ndarray])
    434
    435     #␣
↪----------------------------------------------------------------


    /usr/local/lib/python3.7/dist-packages/pandas/core/arrays/timedeltas.py␣
↪in <listcomp>(.0)
    431
    432         formatter = get_format_timedelta64(self._ndarray, na_rep)
--> 433         return np.array([formatter(x) for x in self._ndarray])
    434
```

```
       435     #␣
↪--------------------------------------------------------------


       /usr/local/lib/python3.7/dist-packages/pandas/io/formats/format.py in␣
↪_formatter(x)
      1798         if not isinstance(x, Timedelta):
      1799             x = Timedelta(x)
   -> 1800         result = x._repr_base(format=format)
      1801         if box:
      1802             result = f"'{result}'"


       /usr/local/lib/python3.7/dist-packages/pandas/_libs/tslibs/timedeltas.
↪pyx in pandas._libs.tslibs.timedeltas._Timedelta._repr_base()


       KeyboardInterrupt:
```