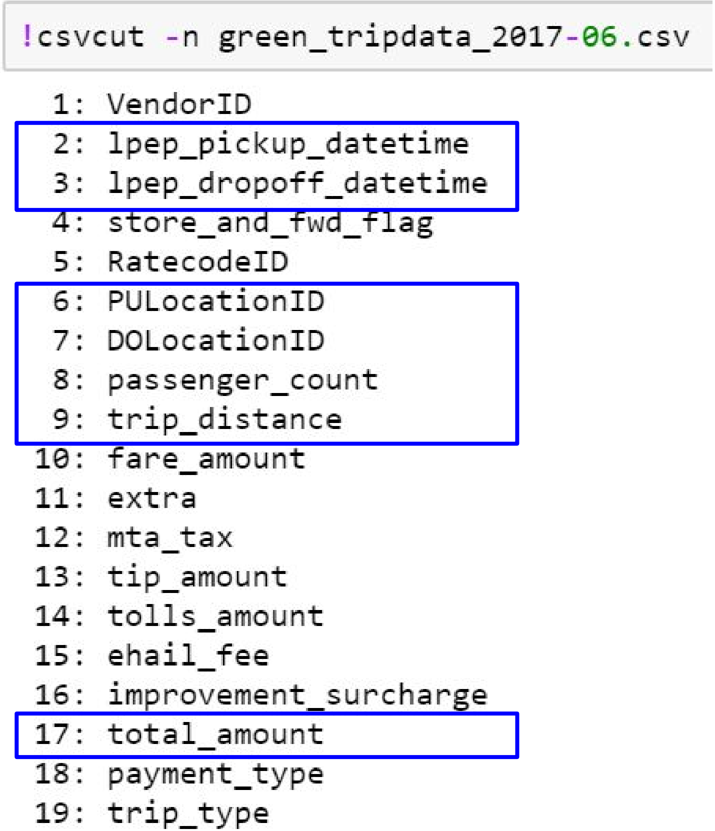
NYC Taxi & Limousine Commission

NYC Taxi

Guangyu Xing
Agnes Jiang
Pei-Hsuan Hsia
Jiwei Zeng

Data Introduction & Tools

- NYC Green Cab Transactions
- June 2017
  - About 1 million rows
- Interesting Information
- Tools
  - Spark
  - Spark-SQL
  - Matplotlib & seaborn



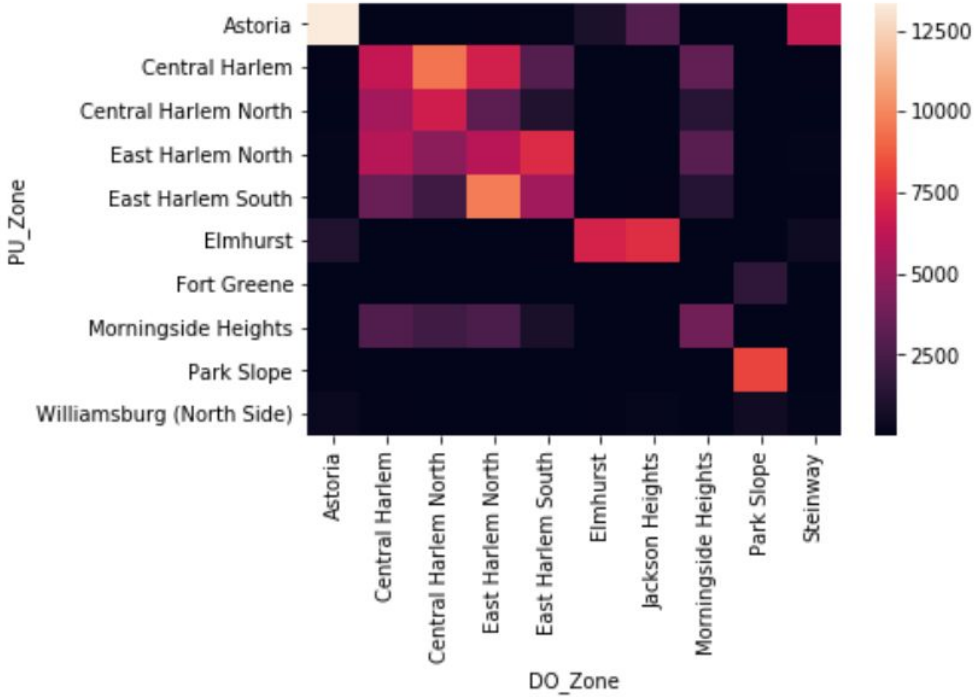
```
!csvcut -n green_tripdata_2017-06.csv
```

```
  1: VendorID
  2: lpep_pickup_datetime
  3: lpep_dropoff_datetime
  4: store_and_fwd_flag
  5: RatecodeID
  6: PULocationID
  7: DOLocationID
  8: passenger_count
  9: trip_distance
 10: fare_amount
 11: extra
 12: mta_tax
 13: tip_amount
 14: tolls_amount
 15: ehail_fee
 16: improvement_surcharge
 17: total_amount
 18: payment_type
 19: trip_type
```

Data Process -- Wrangling

- Drop unused columns
- Convert data types
- Create a new column named valid_data
- Define **VALID DATA**
    - Trip distance is not 0
    - Different pick-up and drop-off time
    - Transaction is not cancelled
- **98.41%** of transactions are valid

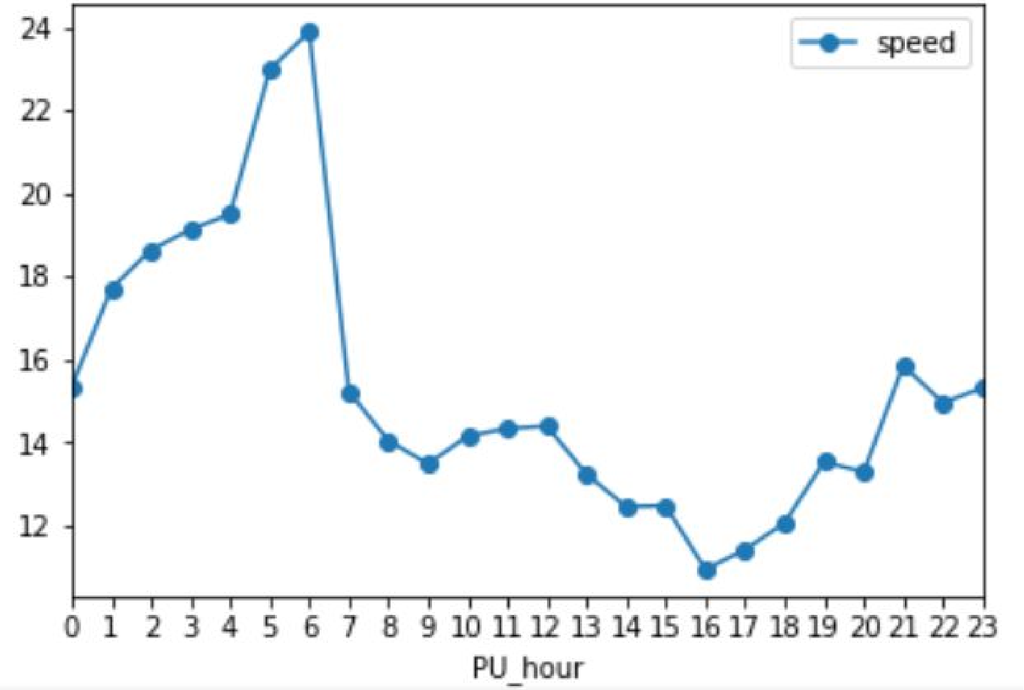
The popular trips are among the same zone.



- Based on top 10 most frequent pick-up and drop-off locations

- Heat map to show the relationship between these locations

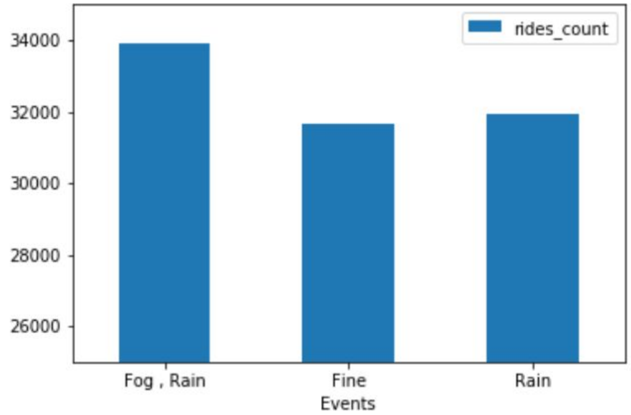
Rush Hours: 9 a.m. & 4 p.m.



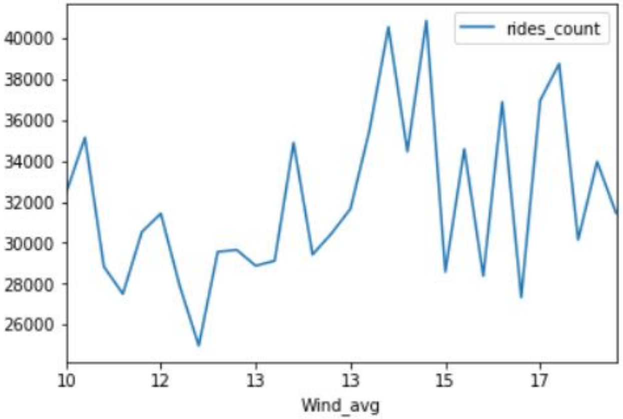
- The line chart shows the relationship between speed and pick-up hour

- In the afternoon rush hours, speed is the lowest

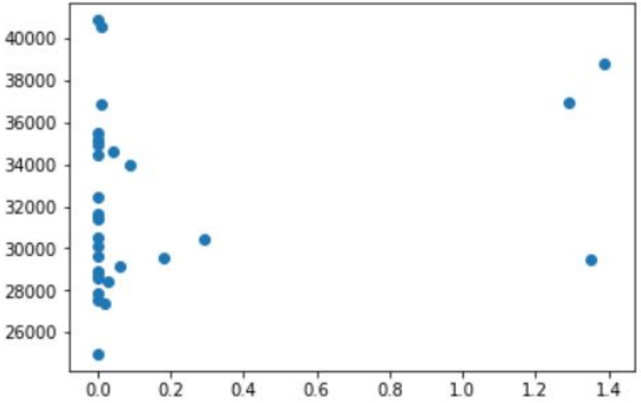
Fog: Taxi ride maker



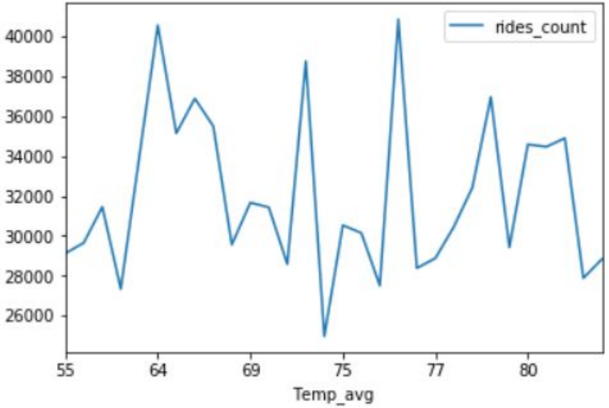
Foggy, Rainy, and Fine weather



Wind



Precipitation



Temperature

Q & A