**Assignment 1-Jinghui Yuan-4177204**

**Problem 1 (30 pt)** Download the taxi cab data of the New York City for the month of October 2012 from here [http://www.nyc.gov/html/tlc/html/about/trip\_record\_data.shtml.](http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml) Plot the data points of daily taxi trips for each day between October 24 to October 30 (7 days). You should use the datashader tool. We recommended a tutorial on datashader in lecture 4 (see at the end of Data\_exploration\_visualization.ipynb).

In addition to visualizing the taxi cab data, separately plot the following items:

* + A distribution of the number of daily trips within this period.
  + The pdf and cdf of trip distance, taxi fare and tips
  + Plot a scatter matrix of all the reasonable variables (exclude variables such as vendor\_id, datetime, and coordinates)

Briefly describe the most interesting findings. Hurricane Sandy made landfall on October 29, 2012; did you find any interesting trends in taxicab data due to Sandy?

1. Daily taxi trips

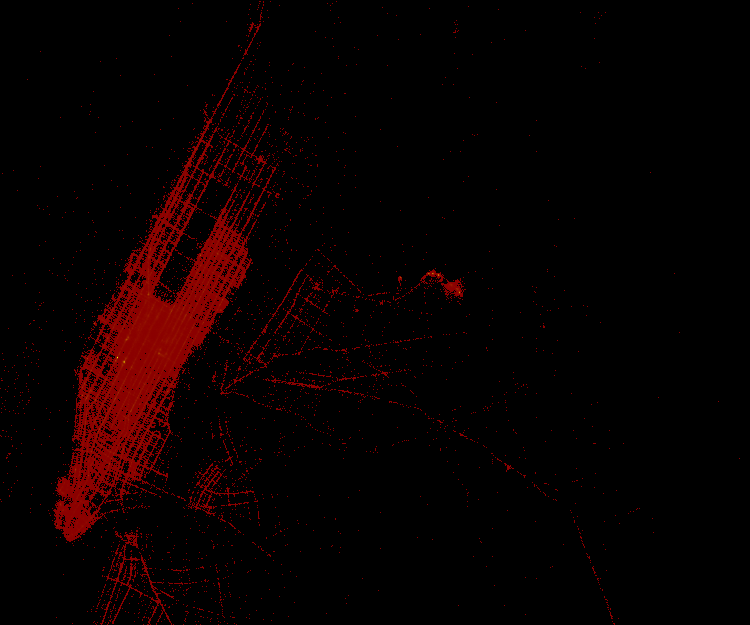


Figure 1. Daily taxi trips (October 24-pick up)

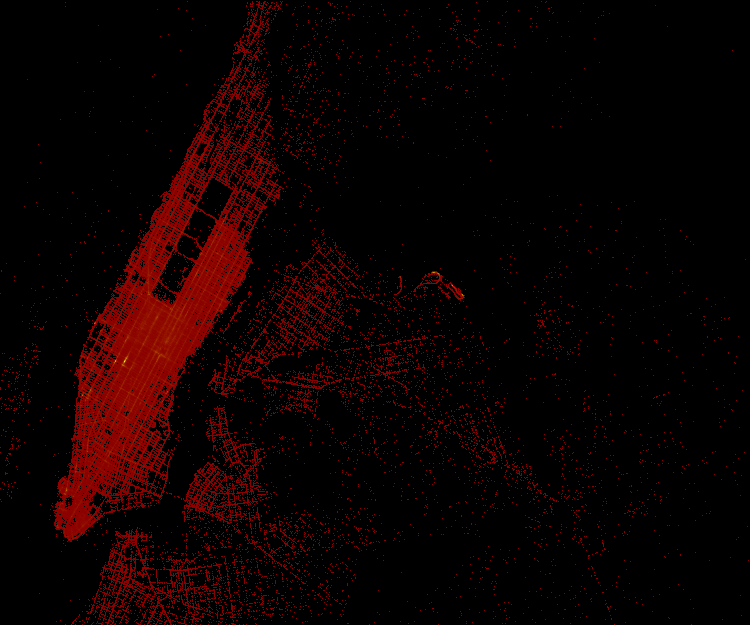


Figure 2. Daily taxi trips (October 24-drop off)

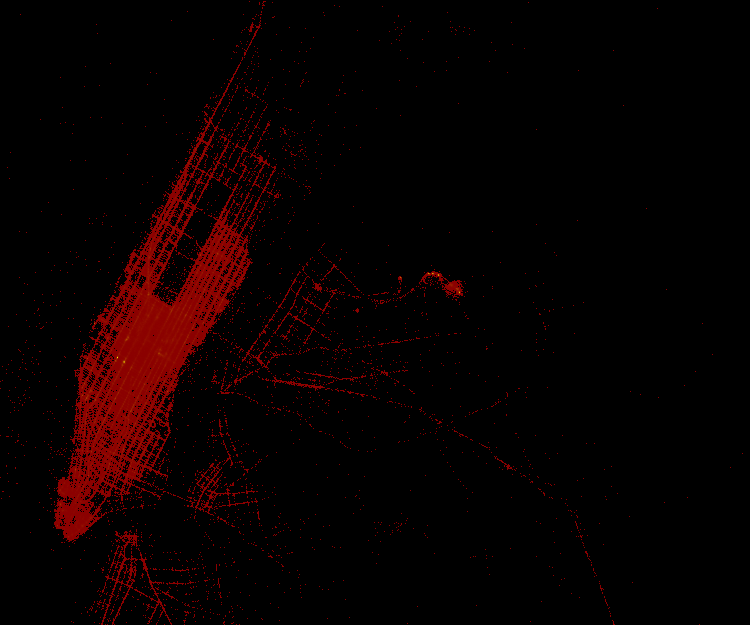


Figure 3. Daily taxi trips (October 25-pick up)

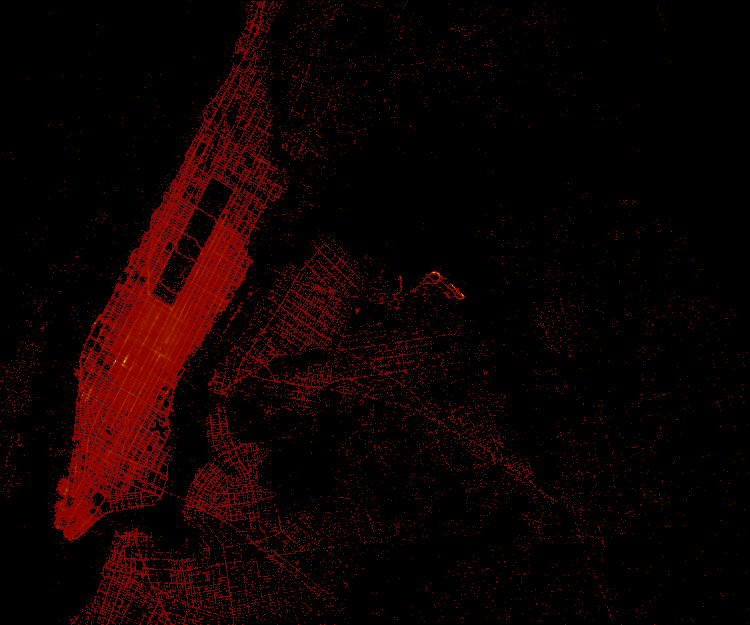


Figure 4. Daily taxi trips (October 25-drop off)

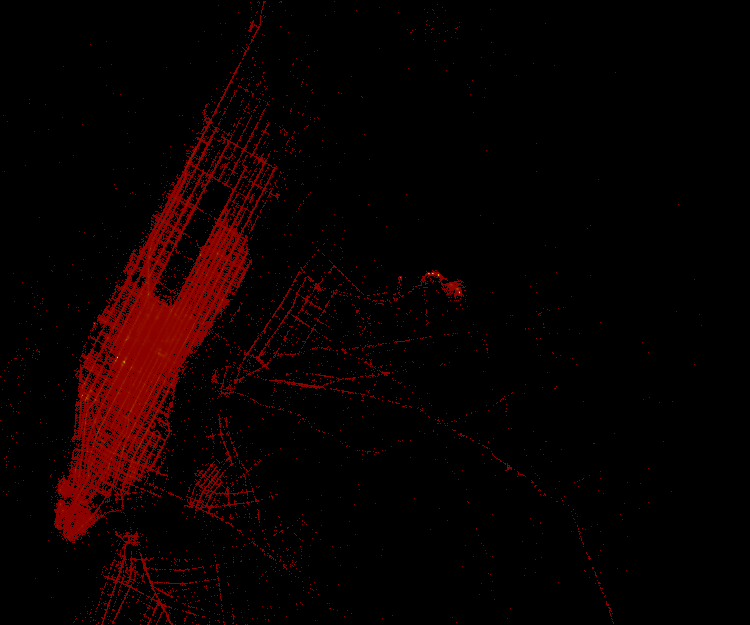


Figure 5. Daily taxi trips (October 26-pick up)

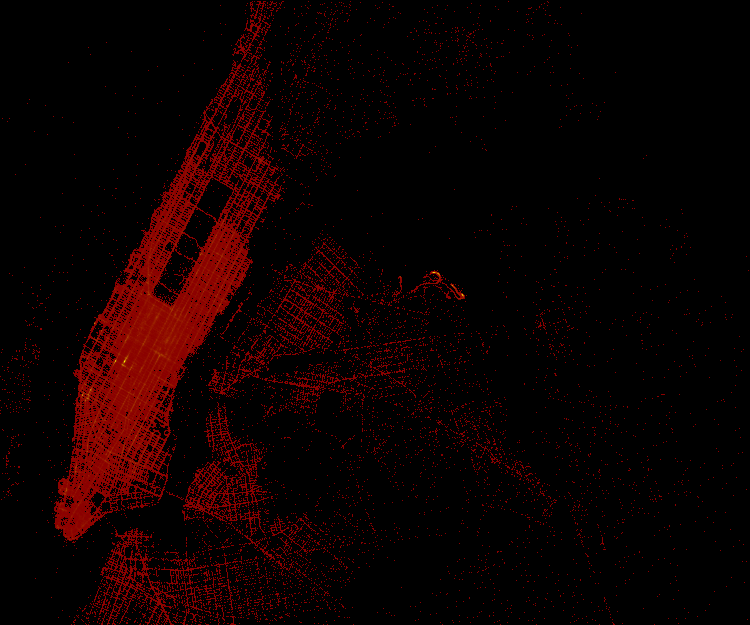


Figure 6. Daily taxi trips (October 26-drop off)

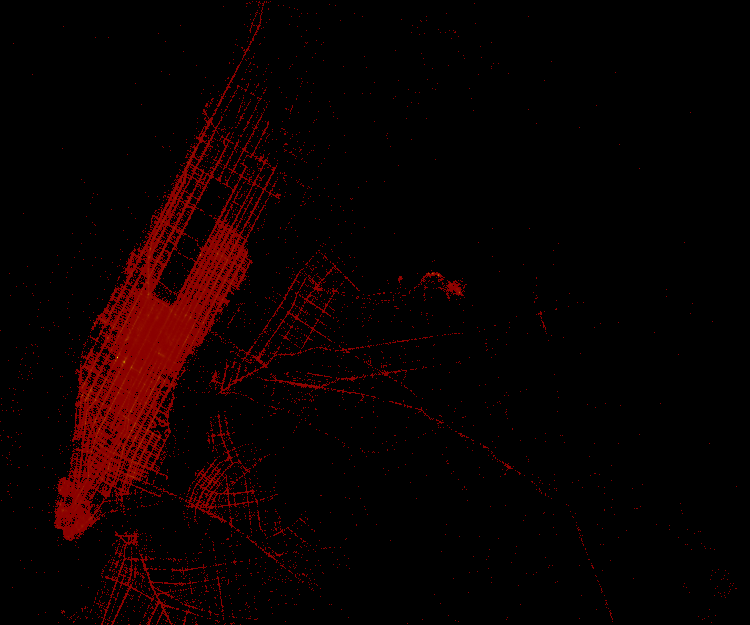


Figure 7. Daily taxi trips (October 27-pick up)

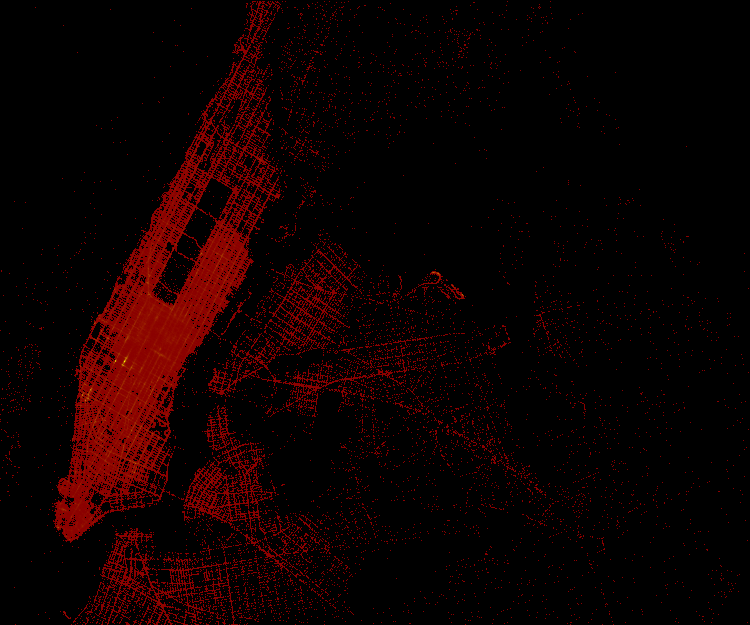


Figure 8. Daily taxi trips (October 27-drop off)

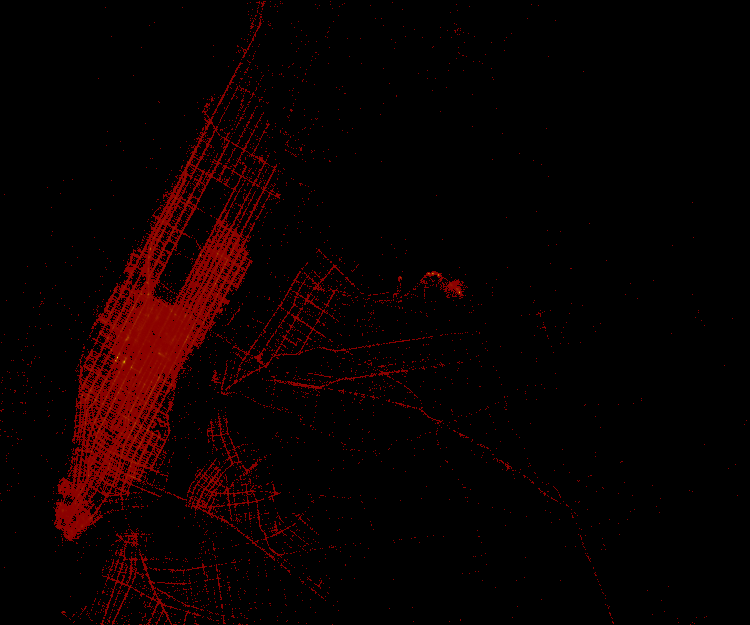


Figure 9. Daily taxi trips (October 28-pick up)

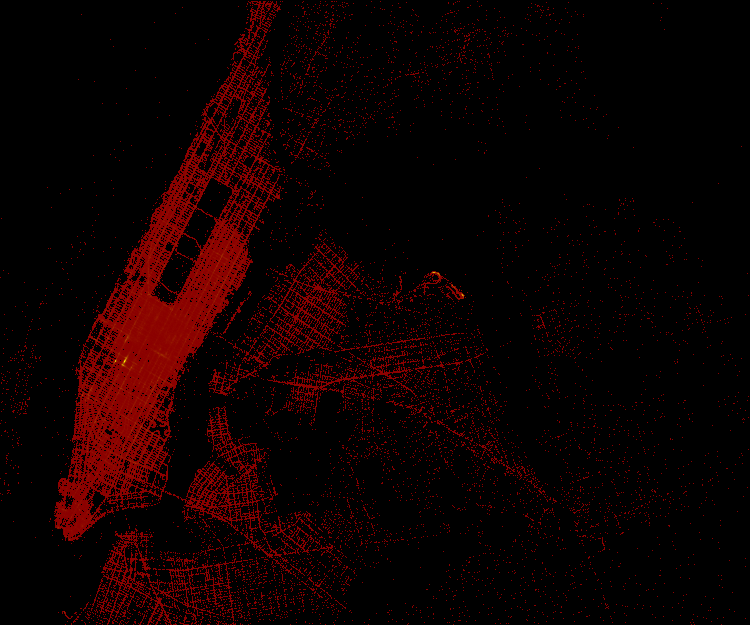


Figure 10. Daily taxi trips (October 28-drop off)

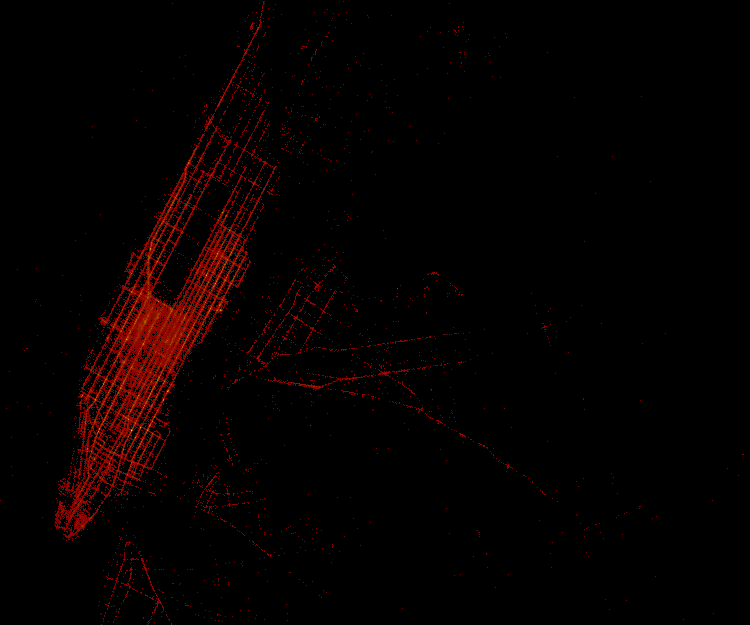


Figure 11. Daily taxi trips (October 29-pick up)

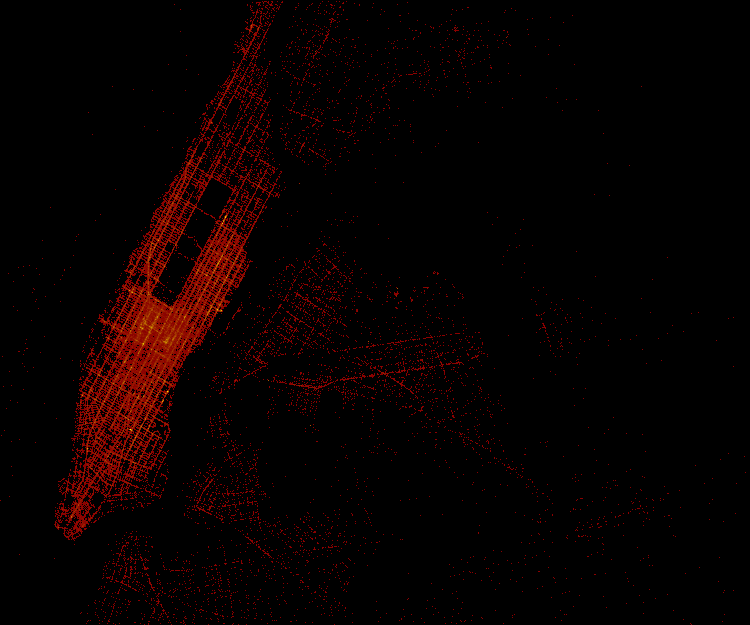


Figure 12. Daily taxi trips (October 29-drop off)

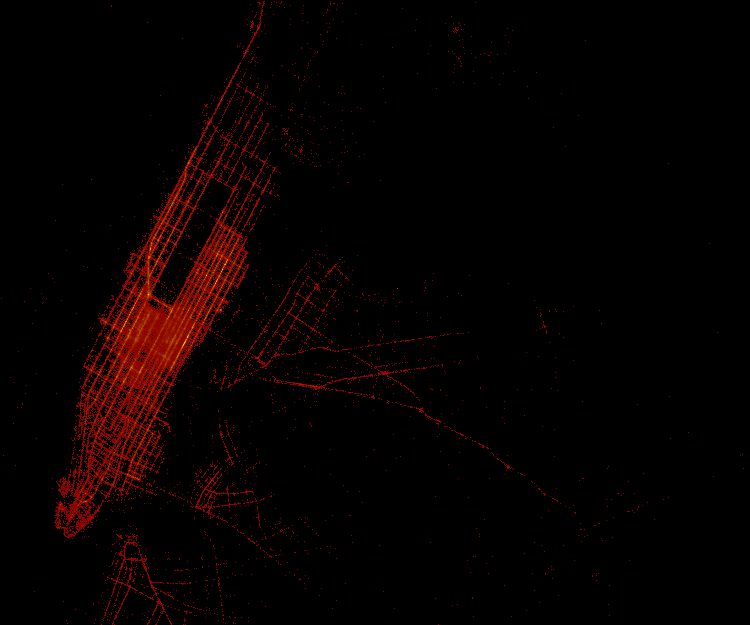


Figure 13. Daily taxi trips (October 30-pick up)

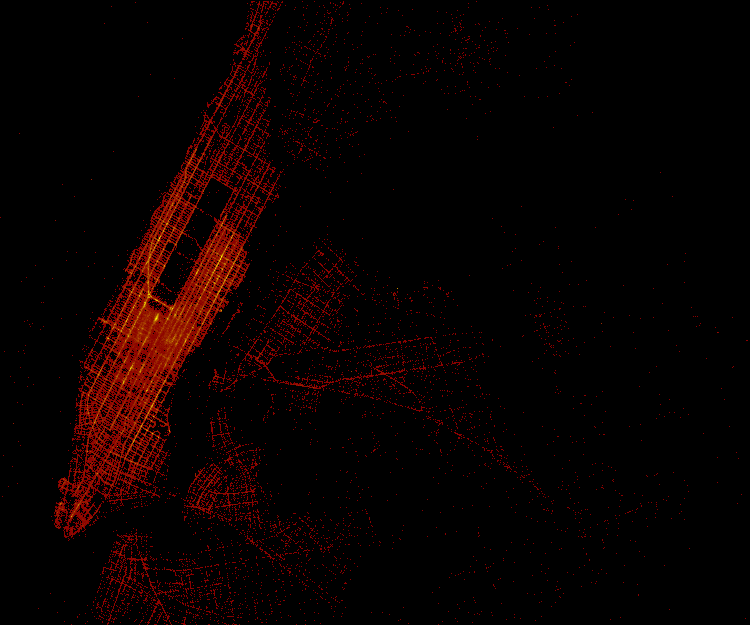


Figure 14. Daily taxi trips (October 30-drop off)

1. Distribution of the number of daily trips

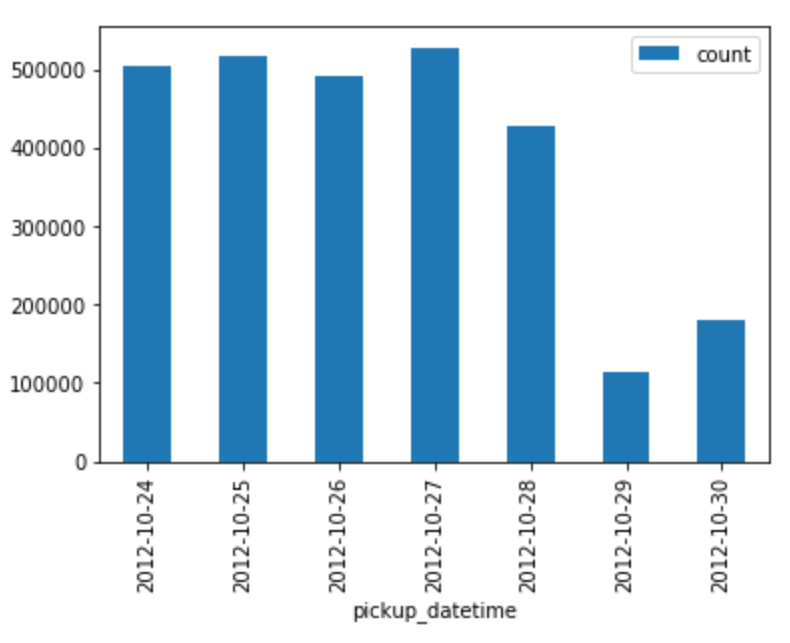


Figure 15. Bar chart of the distribution of the number of daily trips

1. The pdf and cdf of trip distance, taxi fare and tips
2. Trip distance

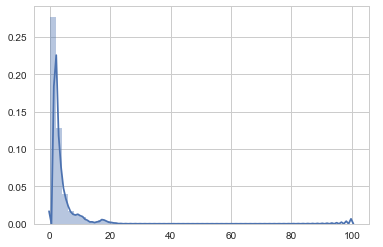


Figure 16. Pdf of trip distance

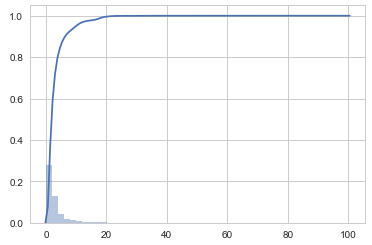


Figure 17

1. Taxi fare

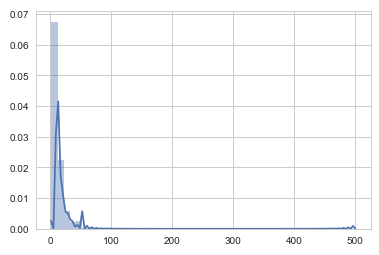


Figure 18

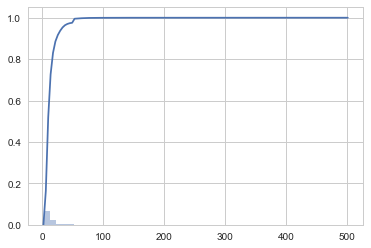


Figure 19

1. Tips

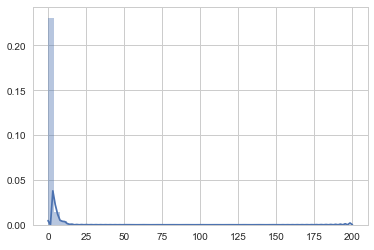


Figure 20

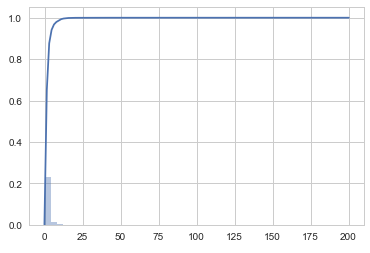


Figure 21

1. Scatter matrix of all the reasonable variables

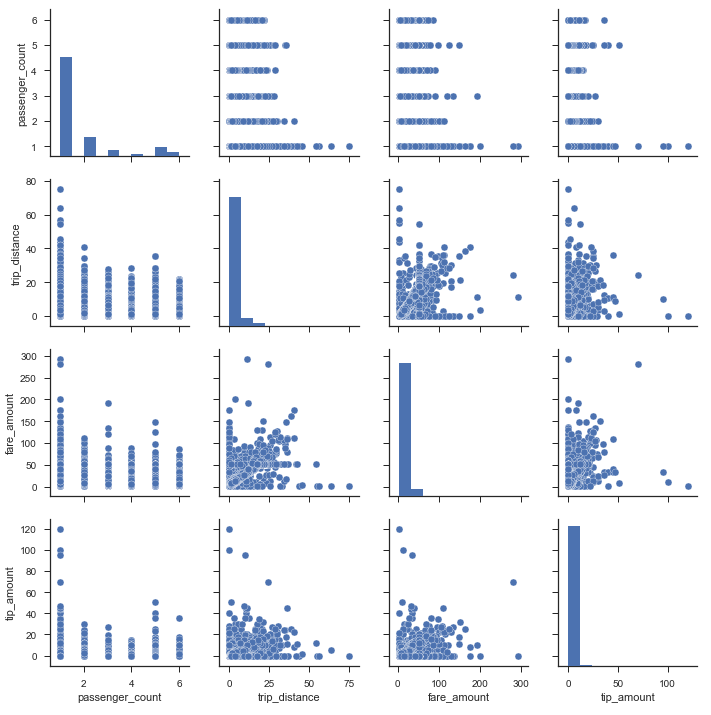


Figure .

**Problem 2 (10 pt)** Let 𝑋 be a random variable which denotes if a patient has AIDS and let 𝑇1

and 𝑇2 be the outcomes of two clinical tests with the following error profiles:

|  |  |  |
| --- | --- | --- |
| 𝑝(𝑡1|𝑥) | 𝑥 = 𝐻𝐼𝑉 − | 𝑥 = 𝐻𝐼𝑉 + |
| 𝑡1= 𝐻𝐼𝑉 − | 0.91 | 0.07 |
| 𝑡1= 𝐻𝐼𝑉 + | 0.09 | 0.93 |

|  |  |  |
| --- | --- | --- |
| 𝑝(𝑡7|𝑥) | 𝑥 = 𝐻𝐼𝑉 − | 𝑥 = 𝐻𝐼𝑉 + |
| 𝑡2= 𝐻𝐼𝑉 − | 0.98 | 0.01 |
| 𝑡2= 𝐻𝐼𝑉 + | 0.02 | 0.99 |

Compute

1. 𝑝(𝑥 = +| 𝑡1 = +, 𝑡2 = +)

2. 𝑝(𝑥 = +| 𝑡1 = +, 𝑡2 = −)

3. 𝑝(𝑥 = +| 𝑡1 = −, 𝑡2 = +)

4. 𝑝(𝑥 = +| 𝑡1 = −, 𝑡2 = −)

You may assume that the prior probability 𝑝 𝑥 = + is 0.005

**Problem 3 (30 pt)** You can read about the ECML/PAKDD discovery challenge 2006 which dealt with email spam detection here: [http://www.ecmlpkdd2006.org/challenge.html.](http://www.ecmlpkdd2006.org/challenge.html) Your task is to download the dataset for task A from [http://www.ecmlpkdd2006.org/data\_task\_a.zip.](http://www.ecmlpkdd2006.org/data_task_a.zip)

**Implement** and train a Naive Bayes classifier using the data found in task\_a\_labeled\_train.tf file. You can test the performance of your classifier on the data found in task\_a\_u00\_tune.tf . In your report, briefly describe how you approached the problem, what results you obtained, what practical difficulties you faced, and how you overcame these difficulties.

**Problem 4 (30 pt)** You can read about the bikesharing dataset here https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset. This dataset records the number bikes rented in Capital bikeshare system in Washington D.C. and many other weather and environmental attributes.

Download the dataset. Using the daily data from day.csv file, create a label based on the attribute “cnt” into three categories {low, medium, high} demand based on some threshold values according to your judgment. Your task will be to create a nearest neighbor classifier for such a dataset using scikit-learn and evaluate your classifier’s performance. In your report, briefly describe on how you approached the problem, what interesting results you obtained, what practical difficulties you faced, and how you overcame these difficulties.