NYC Yellow Taxi Data Set Questions

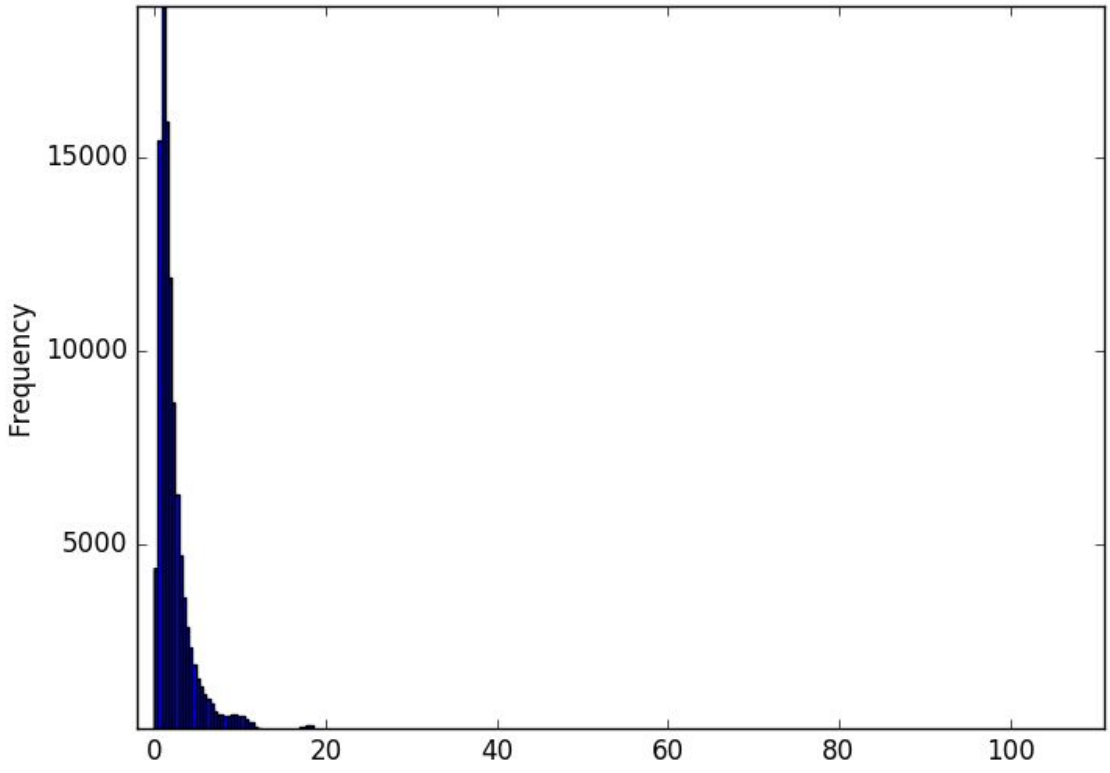
Team Data Science @ Sac State

Varun Ved
Brandon Sherman
Yash Farooqui

1. Which records appear to be outliers and why?

- We had to make some assumptions to answer this question

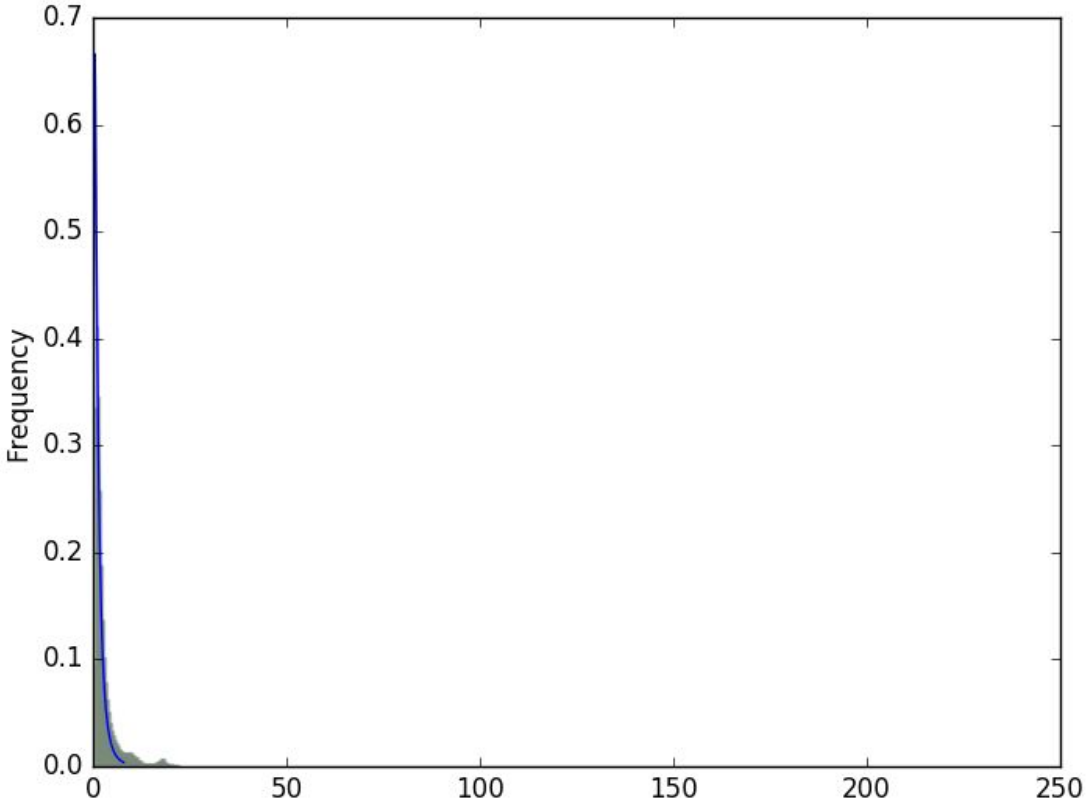
2. What is the distribution of traveled trip length? Where is the farthest traveled?



(Picture of the trip-length distribution)

It appears that the distribution is a lognormal and comparing to a fitted log-normal distribution, this looks correct. Of course, the fit is skewed by the presence of two less prominent categories of trip - long trips and medium length trips, both of which are roughly gaussian.

Hence, the distribution is really a log-normal distribution added to two gaussians.



The parameter for this distribution is 0.89, located with mean of -0.1 at a scale of 2.0. The location parameter is likely a misestimate. While one gaussian can be easily accounted for, the other is significantly more difficult.

3. Do the longitude and latitude coordinates make sense? Are there any suspicious coordinates?

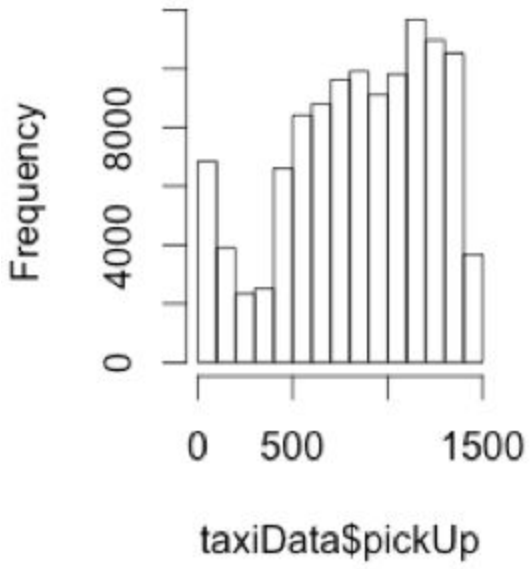
Most of them make sense. There are a few cases where the coordinates are way off - for example record 205.

4. Why are most improvement_surcharges at 0.3? (involves online research)

NYC added a 30 cent surcharge on all trips starting on January 1, 2015.

5. What do the rate code id's represent? What does "99" most likely represent?

6. What is the distribution of pickup times? Are there any periods of the day that are busiest? Least busiest?



7. What about frequency of taxi trips on a week day basis? Are there any week days that are busier than others?

Sun	Mon	Tues	Wed	Thurs	Fri	Sat
16073	13768	18249	18585	18650	15179	14181

Tues, Wed, and Thurs are busier than the other days. Monday is pretty slow.

8. On average, how long do trips last? How does trip time length vary according to pickup hour?

How about trip time length according to week day?

On average they last 16.67 minutes.

Average trip length by pickup hour

0	1	2	3
16.13618	16.67670	16.79930	16.36735

4	5	6	7

16.03538	18.00741	18.16612	16.09126
8	9	10	11
16.19945	16.50723	16.73754	17.71304
12	13	14	15
16.51437	15.68125	16.12189	15.94477
16	17	18	19
16.04023	16.02599	17.02972	16.26035
20	21	22	23
17.22819	16.38672	18.21351	17.23837

Sun	Mon	Tues	Wed
16.23051	15.77399	16.96488	17.27122
Thurs	Fri	Sun	
16.53349	17.89416	15.74191	

9. How does miles traveled per minute depend per hour? When does a taxi cab drive fastest?
Slowest?

0	1	2	3
0.3025271	0.2076562	0.2115402	0.2229261
4	5	6	7
0.2002511	0.2165827	0.2202546	0.1944824
8	9	10	11
0.2164488	0.1949695	0.1961721	0.2023677
12	13	14	15
0.2071444	0.1981695	0.1988888	0.2069761
16	17	18	19
0.1973579	0.2529747	0.2260038	0.1978620
20	21	22	23
0.2006586	0.2725675	0.2581652	0.2118209

Taxi cabs drive the fastest between midnight and 1AM.
Taxi cabs drive the slowest between 7AM and 8AM.

10. How does miles traveled per minute depend by week day? When does a taxi cab drive fastest?

Slowest?

Sun	Mon	Tues	Wed
0.2083335	0.2400412	0.2246437	0.2026020
Thurs	Fri	Sat	
0.2006964	0.2173545	0.2367364	

Taxi cabs drive the fastest on Monday.
Taxi cabs drive the slowest on Sundays.

11. How does tip rate and/or tip amount change on an hourly basis? Week day basis?

0	1	2	3
1.758904	1.726826	1.776954	1.758200
4	5	6	7
1.761415	1.754135	1.701004	1.729139
8	9	10	11
1.680233	1.744674	1.710492	1.753029
12	13	14	15
1.719866	1.730210	1.735935	1.746033
16	17	18	19
1.757487	1.697374	1.752764	1.753050
20	21	22	23
1.757296	1.718496	1.730635	1.823596

Sun	Mon	Tues	Wed
1.547208	1.682059	1.780666	1.857649
Thurs	Fri	Sat	
1.796655	1.810945	1.664840	

12. Which code id is most expensive? Least expensive?

"14.9043282725652" "1"
"63.5858912869692" "2"
"87.7424110671935" "3"
"86.5851282051282" "4"

"67.1398404255317" "5"
"3.3"   "6"
"18.1866666666667" "99"

SELECT avg(total_amount), RatecodeID from data
GROUP BY RatecodeID;

rate code id 3 is the most expensive on average, 6 is the cheapest

---
SUM

SELECT sum(total_amount) as sumtotal, RatecodeID from data
GROUP BY RatecodeID
ORDER BY sumtotal DESC;

"1661996.55000202" "1"
"159091.899999997" "2"
"25244.5799999999" "5"
"22198.83"   "3"
"3376.82"   "4"
"54.56""99"
"3.3"   "6"

by sum, 1 is the most expensive and 6 is the cheapest

13. Which vendor receives a higher total amount? Is this difference significant? Practically significant?

"1"   "16.1149895577359"
"2"   "16.5081844894292"

SELECT VendorID, avg(total_amount) from data
GROUP BY VendorID;

#2 receives more, but not by a significant amount
It's not particularly significant nor significant because delta(1,2)/sum(1,2)

14. Which vendor receives a higher tip amount? Is this difference significant? What about tip rate?

SUM

"1"	"92666.1299999988"
"2"	"106985.560000005"

SELECT VendorID, sum(tip_amount) from data
GROUP BY VendorID;

AVG

"1"	"1.71264586837191"
"2"	"1.76607943477839"

SELECT VendorID, avg(tip_amount) from data
GROUP BY VendorID;

Is this difference practically significant?
A t-test gives us a p-value of 0.00036 which is well within practical significance.

15. How many roundtrip records are there?
1807

SELECT count(*)
FROM data
WHERE pickup_longitude = dropoff_longitude AND pickup_latitude = dropoff_latitude;

Harder questions:

16. What variables improve the chances of tipping at the end of a trip? What variables seem to decrease the chances of tipping at the end of a trip? (e.g. Do larger parties increase the chance of tipping the taxi driver?)

SELECT passenger_count, payment_type, tip_amount, total_amount
FROM data
WHERE tip_amount > 0
GROUP BY passenger_count;
"0"	"1"	"8"	"95.3"
"1"	"2"	"0"	"13.3"

"2"	"1"	"0.95"	"5.75"
"3"	"2"	"0"	"13.8"
"4"	"1"	"1"	"6.3"
"5"	"1"	"3.46"	"20.76"
"6"	"1"	"2.94"	"12.74"

This tells us that 5 passengers got the most tips. However, the fact that the driver sometime reported 0 but this only happened 7 times so we're gonna say that it was probably a mistake from the driver.

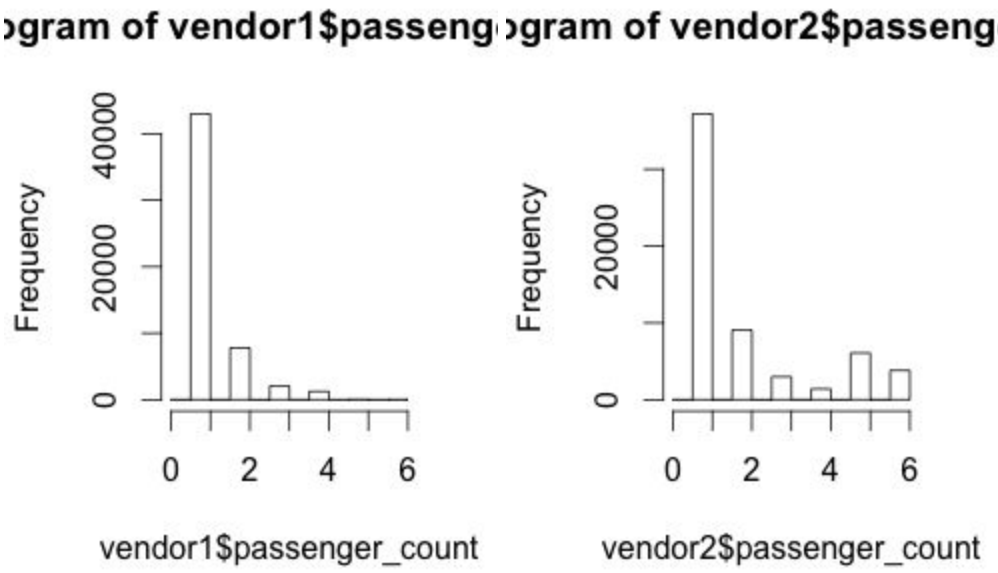
17. What are the pricing relationships that determine total amount paid for a trip? Is there a base cost and if so what is it?

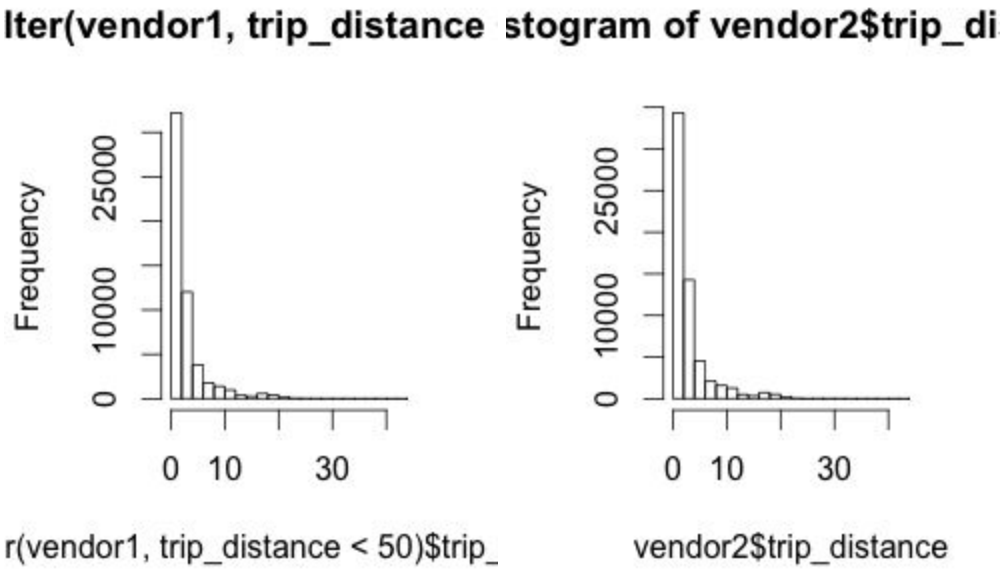
18. Where are pickup locations most densely concentrated? How about weekday? By hour?

19. Where are drop off locations most densely concentrated? How about by weekday? By hour?

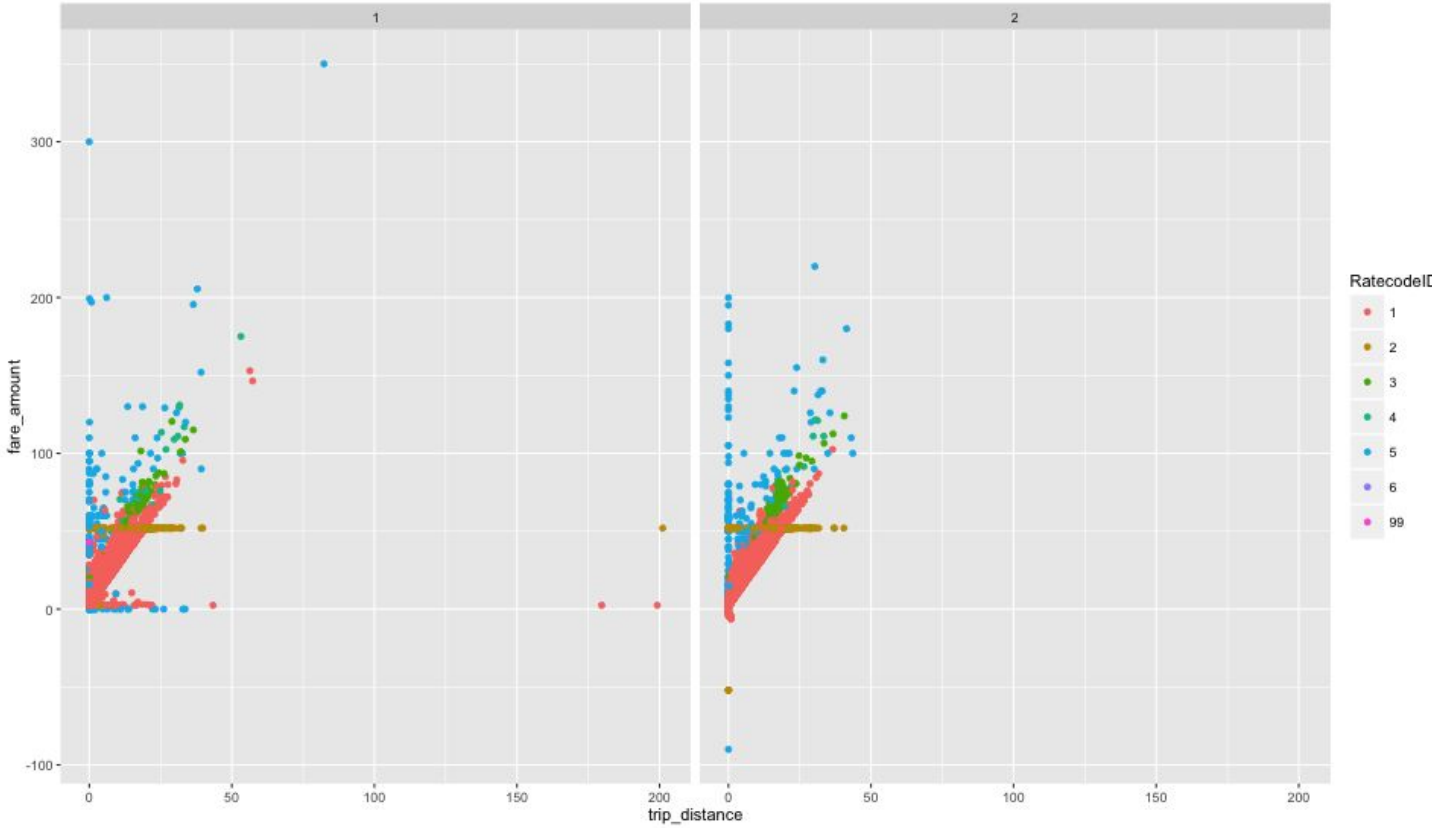
20. What do you think are the main differences between the two vendors? Use the provided data to support your conclusions.

There are much more passenger_counts for values 3 and higher in vendor2.





Plot of trip_distance VS fare_amount with RatecodeID being the color.Split by vendor



\insertWordsToMakeThisSoundIntelligent.
It seems like vendor1 occassionally gives very good deals to it's passengers charging a much smaller amount for the distance they traveled.

21. If you had a choice, which vendor's taxi would you take (1 or 2) and why?

We cut off at a trip distance of 8 miles as the vast majority take very short trips. Hence, assuming a "homo economicus"-like assumption, it is rational to pick vendor1 as there is a slightly lower average cost.

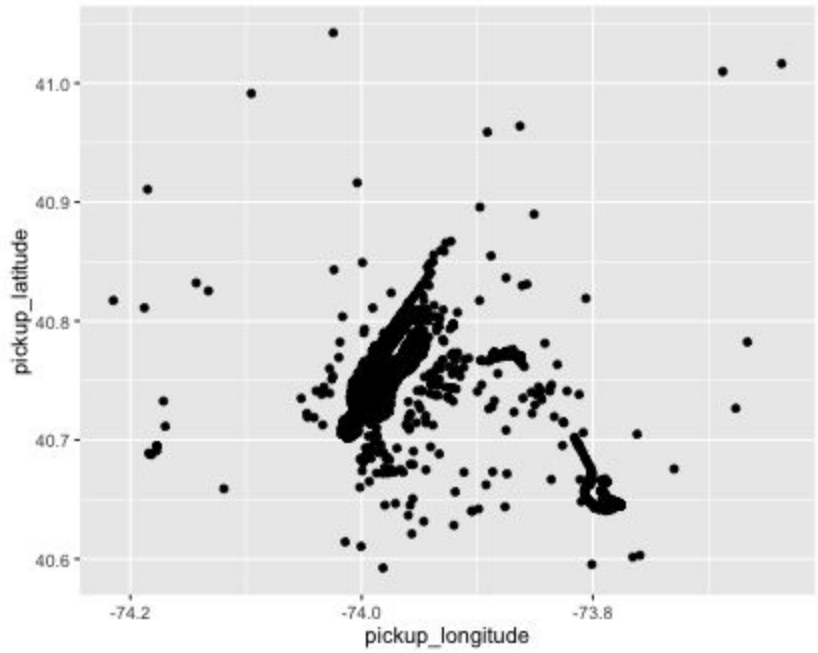
The mean of fares for trip_distances are $10.54 for vendor1 and $10.62 for vendor2.

22. How do the holidays impact ridership? Do people tip more or less than usual and is this difference significant?

<u>Very hard questions (and borderline privacy infringement…?)</u>

23. Identify an individual based on travel patterns.

24. Which local neighborhood pickup locations (besides airports) have the highest total amount per taxi trips ratio?



This preliminary graph shows the pickup locations in New York that led to a total cost of greater than $40. The fact to realize is that this graph approximates New York itself. High price pickups seem to be roughly uniformly distributed across New York City.

25. Do taxi drivers use the most optimal route (in terms of time and cost)?

26. What is the most number of cabs hailed in a 20 second time window, regardless of location and with regards to location?

27. Given a drop off location, what is the probability that a cab can pick up another person or group within the next 1 minute that is within a 200 meter radius? Within a T minute window and an N meter radius? Create a model to compute this probability based on any inputs of your choosing.