1. (5pts) Not including the header row, how many rows are there in the file?

55918

1. (10pts) One might posit that a single passenger is the most common taxi ride. Which was more common: a single passenger\_count or a passenger\_count with more than 1 person?

39670 For One Passenger

15752 For more than One Passenger

Most common is One Passenger! Interestingly, there are some observations with 0 passengers in this dataset, but those were filtered out.

1. (15pts) It is possible that people who paid a toll went further than people who didn’t pay a toll. What is the average trip\_distance for people who had tolls\_amount greater than 0? What is the average trip\_distance for people who paid nothing in tolls?

Fare Amount > 0 Avg Trip Distance = 2.9340971339574677

Fare Amount = 0 Avg Trip Distance = 1.731578947368421

Those who paid a toll did indeed go further on average.

1. (20pts) What are the average fare\_amounts for each month and which month has the highest average fare\_amount? You should use the pickup date/time as the month to which a row belongs. You should take the sum of the fare\_amounts and divide it by the total number of rows for that month. To ensure we have reliable data, you should filter out all rows where the **fare\_amount** is less than or equal to 0. You should filter out all rows where the (fare\_amount / trip\_distance) is greater than 10,000. An obvious side-effect of this is to filter out all rows with a trip\_distance of 0.

The average fare\_amounts for each month are:

+----------+------------------+------+

|Monthstamp| MonthFares|Ranked|

+----------+------------------+------+

| 10368000|13.595211546936978| 1|

| 23587200|13.511325791855203| 2|

| 18316800|13.353933418693982| 3|

| 15638400|13.326148683531233| 4|

| 28857600|13.234322371699053| 5|

| 26265600|13.231794294294295| 6|

| 7776000|12.936855154414964| 7|

| 13046400|12.845854004252303| 8|

| 20995200|12.652934236947791| 9|

| 5097600|12.528635446532867| 10|

| 0|12.381402298850574| 11|

| 2678400|12.148956584312785| 12|

+----------+------------------+------+

The Top Month is **May (05):**

+---------+----------+

|BestMonth|Monthstamp|

+---------+----------+

| 05| 10368000|

+---------+----------+

1. (15pts) Who tips better: people who pay with a credit card before noon or people who pay with credit card from noon until the end of the day (i.e. 12:00:00pm - 11:59:59pm)? We will use the dropoff time to determine when a person paid. The payment\_type has a numeric score of 1 if the person paid by a credit card. To figure out who tipped the best, take the sum of the tip\_amount and divide it by the sum of the fare\_amount. You should print out the (sum(tip\_amount) / sum(fare\_amount)) for each of the 2 requested groups.

It appears people who pay with credit cards after noon tip slightly better overall!

+------------------+

|BeforeNoon\_TipRate|

+------------------+

|0.1968107145335785|

+------------------+

taxi\_BeforeNoon: Unit = ()

+-------------------+

| AfterNoon\_TipRate|

+-------------------+

|0.20209688769666953|

+-------------------+

1. (15pts) What are the top ten worst rides with respect to time and distance? In other words, who sat in the taxi the longest and went the shortest distance? You should remove any ride whose trip\_distance is 0. We want to maximize the following calculation:

(number of seconds in the taxi / trip\_distance)

To ensure you are picking the correct row, your answer should include all information about the row (i.e. the entire row). Your answers must include a new column representing the above calculation. You only need to display 10 answers and do not need to worry about ranks.

NOTE: This table is concatenated, but the **FIRST COLUMN** Shows the WorstRides output we are after:

|**WorstRides**|VendorID|tpep\_pickup\_datetime|tpep\_dropoff\_datetime|passenger\_count|trip\_distance|RatecodeID|store\_and\_fwd\_flag|PULocationID|DOLocationID|payment\_type|fare\_amount|extra|mta\_tax|tip\_amount|tolls\_amount|improvement\_surcharge|total\_amount|PickUp\_Time|DropOff\_Time|Time\_In\_Car|PickUp\_TimeStamp|DropOff\_TimeStamp|

* | **157160.0**| 2|06/08/2018 12:23:...| 06/08/2018 02:34:...| 1| 0.05| 2| N| 132| 142| 1| 52.0| 0.0| 0.5| 15.84| 0.0| 0.3| 68.64|12:23:36 AM| 02:34:34 AM| 7858| 1416| 9274|
* | **36000.0**| 2|04/25/2018 08:21:...| 04/25/2018 08:27:...| 1| 0.01| 1| N| 142| 142| 2| 2.5| 0.5| 0.5| 0.0| 0.0|
* |**13380.441640378549**| 2|04/19/2018 12:14:...| 04/19/2018 11:48:...| 2| 6.34| 1| N| 164| 116| 2| 22.5| 0.5| 0.5| 0.0| 0.0|
* | **12275.0**| 2|03/27/2018 12:50:...| 03/27/2018 01:07:...| 1| 0.08| 1| N| 230| 230| 2| 10.5| 0.0| 0.5| 0.0| 0.0|
* | **12000.0**| 2|05/02/2018 03:43:...| 05/02/2018 03:45:...| 1| 0.01| 5| N| 142| 100| 2| 68.0| 0.0| 0.0| 0.0| 0.0|
* | **10200.0**| 2|08/31/2018 07:45:...| 08/31/2018 07:48:...| 1| 0.02| 1| N| 143| 143| 2| -4.0| -1.0| -0.5| 0.0| 0.0|
* | **8700.0**| 2|09/08/2018 03:30:...| 09/08/2018 03:31:...| 1| 0.01| 2| N| 132| 132| 1| 52.0| 0.0| 0.5| 11.71| 5.76|
* | **8200.0**| 2|04/28/2018 02:35:...| 04/28/2018 02:36:...| 1| 0.01| 1| N| 249| 249| 2| 3.0| 0.5| 0.5| 0.0| 0.0|
* |**7153.7037037037035**| 2|02/05/2018 07:00:...| 02/05/2018 09:09:...| 1| 1.08| 1| N| 141| 162| 2| 6.0| 0.0| 0.5| 0.0| 0.0|
* | **7050.0**| 2|07/26/2018 06:10:...| 07/26/2018 06:19:...| 6| 0.08| 1| N| 161| 161| 1| 7.0| 1.0| 0.5| 1.76| 0.0| 0.3| 10.56|06:10:11 PM| 06:19:35 PM| 564| 65411| 65975|

1. (30pts) Imagine that you work in NYC. You are considering working for a ride-sharing company to make a few extra dollars later in the evening. However, you don't have that much time to spend driving other people around. You realize you have 60 minutes that you can start your rides between 4:00:00pm and finish by 11:00:00pm1. You have determined that any total\_amount that is over $200 is not going to be possible so those ought to be filtered out. You want to maximize the amount of money you could earn so you want to find the best 60 minute period between 4:00:00pm and 11:00:00pm, inclusive, to start your rides. You do not care about days. In other words, a ride that starts at 9:45:13pm on June 12th is in the same 60 minute time slot as a ride that starts at 9:45:13pm on August 15th.

+-------------------------+------------------+

|Best\_AvgWageHour |Hour\_Average\_Total|

+-------------------------+------------------+

|04:30:23 PM - 05:30:23 PM|18.142735687350395|

+-------------------------+------------------+