Computational and Mathematical Modelling

2017 Competition

Team control number: #734

October 3rd, 2017

Table of Contents

Problem 1: Matching the Typing Patterns

Summary sheet:

Interpretation of the Problem

For the typing problem, eleven individuals have been submitted to two different typing methods, measuring the speed (word-per-minute) and the accuracy of the writer. The results are divided in three categories: one where the identity of the subject is known and two where it is unknown. Knowing the typing speed and accuracy of an individual may fluctuate accordingly to the task requested, we can estimate a general typing pattern, using data recorded while typing several independent samples. It is then possible to match the individuals with the samples obtained anonymously.

Overview of Methods and Statement of Conclusions

The problem has been divided into two main steps. The first objective was to group the relevant data together; since the participant #4 did not do the Letter typing exercise, we could only investigate on #4 with the Quote task. Therefore, the data have been divided into two separate sets: the Quote results and the Letter results. Both the Quote and Letter sets contained each result for the speed and accuracy of the known and unknown writers. The following method has been used twice to match independently the category two and the category three with the first one:

Using Lloyd's algorithm, the numbers of each sections (Given set and Unknown set) were them reduced to centroid points of both known and unknown writers. It was then possible to evaluate the distance between the centroids of the known and unknown clusters to match the officers and the anonymous result sets. Since it occurred that a few points were assigned to a same individual, Hall's Stable Marriage theorem was used to determine the correct matching of the results.

Through the application of the model, I was able to match the eleven officers with their anonymous Quote and Letter tests. Although there is a small uncertainty as some of the results were significantly close and Hall's stable marriage theorem was used to decided which result was the most appropriate, the results found are demonstrated in this paper.

Solution:

Assumptions/ Justifications

1. The typing pattern of the officer is the same in both the known and unknown data sets.

Justification: Although the typing speed and accuracy of an individual differs accordingly to the task he is doing, W.Chang [1] argues that the typing pattern of an individual is developed in a natural and unique way, similarly to handwriting. It is therefore very unlikely that the writing pattern of one of the eleven officers is different in the anonymous tests than the identified ones.

2. For the third category containing the Letter tests, the officer #4 do not have a match with the samples Q to Z.

Justification: It is given in the problem that the officer #4 did not do the Letter test. The results for ''Given Letter Speed'' and ''Given Letter Accuracy'' of Person 4 are indeed non-existent (N/A) in the excel sheet containing the values of each tests taken by all the officers.

3. There is only one possible match between the officers and the second and third categories.

Justification: Although this assumption opens the door to a possible mistake in the matching of the officers and the anonymous results, as the anonymous tests are not exactly the same as the identified ones, there is no exact match possible. Therefore, since there will not be any match percentage of 100%, we must assume that the closest values to a perfect match percentage is the correct one.

4. The results for the anonymous individual 'K' for the alternate Quote test #1 are considered as if they were results of the regular Quote test #1.

Justification: The individual K had an alternate Quote test #1 that has not been done by any other officer. However, according to the Difficulty Averages for each Quote test that are given to us shows that the alternative test and the test #1 both have the same Difficulty average of 60. Since the difficulty is the same, the results of alternative test are considered the same as if they were the results of test #1.

5. For each individual, known or unknown, the cluster of points of this individual is only represented by its own centroid and do not affect the centroid of another's cluster.

Justification: As the amount of data is significant, the distribution of the data on the graphs can become quite confusing. As there is also a large amount of overlapping points, some values of one individual could be considered in another person's cluster of points and therefore affecting its centroid value. Hence, the results to the tests of one individual is considered as its one and only cluster, with a centroid using only this same individual's parameters.

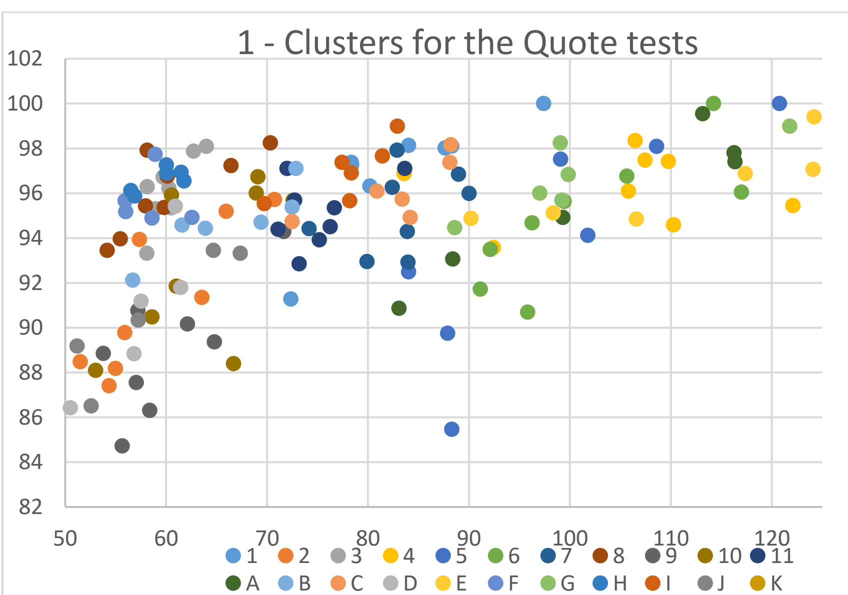
6. There is only one iteration necessary for the Cluster's centroids.

Justification: As the clusters are respectively pre-determined by the results of one individual only, there is no need to do multiple iterations to find multiple new positions of the centroids.

Usage of Method(s), Variables and Parameters

The first step was to gather the relevant information together. As mentioned previously, the results were divided in two categories: The given and anonymous results for the quote tests, and the given and anonymous results for the letter tests.

The values of speed (word-per-minute, or WPM) were treated as X values, and the accuracy as Y values. For both the quote and letter tests, the X and Y values were used to create a graph with the clusters of each known individual 1 to 10 A to K for the quote test, and Q to Z for the letter test.



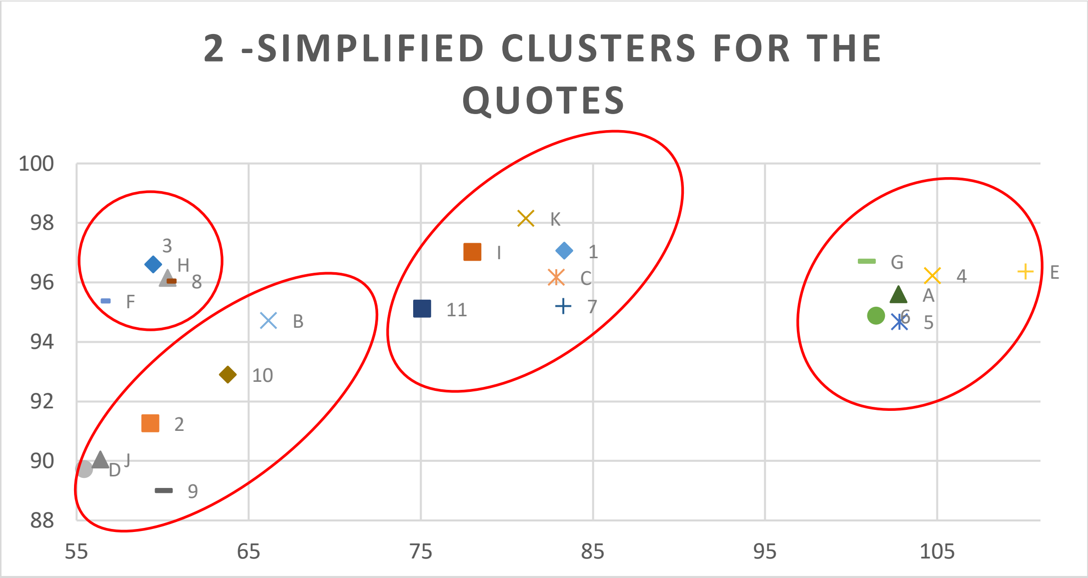
As it can be seen in the graph shown above, it is very difficult to identify the different individuals. That's when the creation of centroids becomes very useful. This is the formula to calculate de centroid of a cluster:

$$\arg min \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{y,x \in S_i} \|x - y\|^2$$

S_i here is the the number of sets (eleven in our case)

$\sum_{y,x \in S_i} \|x - y\|^2$ is the Eucledian distance.

Using the formula mentionned above, it is applied to the twenty-two clusters that were created in the graph 1. The following graph shows the centroids of the clusters of the quote tests for the Person 1 to the Person 11 and for the individuals A to K.



Thanks to the new graph, four main clusters can be observed. It is also much easier to see the similarities between the known typing pattern and the anonymous typing patterns. Using again the Eucledian distance, this time between the centroids, it is possible to find the matches between 1-10 and A-K. When the Eucledian distance is minimal, the two centroids can be related as a "match".

Below are the results of the Eucledian distance between the individuals 1-10 and the individuals A-K:

	A	B	C	D	E	F	G	H	I	J	K
Person 1	19.4799	17.3404	1.0007	28.8428	26.8141	26.9216	17.5803	23.8827	5.3268	27.8480	2.4793
Person 2	43.6879	7.6895	24.1007	4.1414	51.1072	4.9884	41.9780	5.3510	19.5877	3.1462	22.8862
Person 3	42.4673	6.0282	22.5871	8.0525	49.8442	3.9078	40.6192	0.9559	17.7337	7.2506	20.9105
Person 4	2.0727	38.6123	21.8530	49.7217	5.4055	48.2772	3.8544	45.2812	26.7398	48.7436	23.7049
Person 5	0.9249	36.6729	19.9986	47.6430	7.5055	46.3647	2.7874	43.4123	24.9284	46.6706	21.9929
Person 6	1.4923	35.3020	18.6166	46.3012	8.8114	44.9910	1.9062	42.0338	23.5446	45.3271	20.6075
Person 7	19.4854	17.1321	1.0460	28.3710	26.8864	26.8126	17.6969	23.8634	5.5764	27.3827	3.6654
Person 8	42.4661	6.0048	22.5874	7.9709	49.8448	3.8887	40.6208	1.0090	17.7391	7.1645	20.9206
Person 9	43.1892	8.3421	23.9048	4.6911	50.6006	7.3175	41.5537	7.6249	19.6418	3.8335	22.9371
Person 10	39.0552	2.9806	19.3637	8.9384	46.4710	7.7388	37.3081	5.7117	14.7974	7.9420	18.0935
Person 11	27.6768	8.9428	7.8668	20.3724	35.0748	18.6226	25.8730	15.7016	3.4888	19.3769	6.7468

As it can be noticed, Person 1 and Person 2 both have two possible match. Since it is assumed that only one match is possible for each individual, we use Hall's Stable Marriage problem (HSM).

The HSM problem is used to match two sets of same size, using an ordering of the preferences [2] for each elements. There is also the fact that once an element is matched with another one, it become unavailable and cannot be matched again. Following this theory, Person 1 has a better match with the individual C as the Eucledian distance is smaller than for the individual K. As person 1 become unavailable, the following match attempt for K is with Person 7, which is avaible. Hence, 7 and K become a pair. The same logic is used with Person 2, and D will get paired with Person 9.

Using the same model and logic, every steps used to paire Person 1-10 and A-K are now used to match Person 1-10 with Q-Z. The following table is obtained :

.	Q	R	S	T	U	V	W	X	Y	Z
Person 1	6.5440	13.2002	4.1412	1.4506	5.7901	5.8226	4.3991	5.5143	4.5438	17.8562
Person 2	5.9637	12.6222	4.3302	1.2776	6.3543	5.4583	4.1801	5.0547	4.9708	18.4367
Person 3	6.7352	11.1914	4.5100	7.5804	9.6986	1.7891	2.2609	3.1174	6.7601	21.4931
Person 5	4.0464	9.9440	4.5785	4.7082	8.9819	1.8118	1.9026	1.3118	6.5219	21.2344
Person 6	11.0715	4.8297	17.8763	17.0723	22.5923	12.9544	14.8828	12.3704	20.0276	34.8396
Person 7	0.9901	5.9506	9.4870	6.7692	13.1093	5.8564	6.9219	4.5467	11.1542	25.2509
Person 8	11.4831	18.0542	4.8227	6.0771	0.7901	9.3545	7.4144	9.6337	2.7516	12.9386
Person 9	11.6211	18.2030	5.0190	6.1648	0.7473	9.5385	7.6010	9.8066	2.9457	12.7880
Person 10	24.8673	31.4251	17.3447	19.2160	12.6785	22.2715	20.3149	22.8056	15.1500	1.1953
Person 11	6.7827	13.4119	3.7316	1.8843	5.4733	5.7252	4.1954	5.5167	4.1176	17.6051

Using the HSM problem again, W will get paired with Person 1 as it is the only Person available with a small score matching W.

Conclusion to the problem

In conclusion, Using Lloyd's Theorem and Hall's Stable Marriage problem, it became possible to match the known officers with the anonymous tests taken.

After dividing the data into two distinct sets and creating the corresponding clusters, the distance between the centroids of the said so clusters have been calculated to find the shortest distance between the unknown centroids and the identified ones. Using Hall's stable marriage problem, the pairs have been made according to their corresponding similarities.

The following results have been found:

1	2	3	4	5	6	7	8	9	10	11
C	J	H	E	A	G	K	F	D	B	I
W	T	V		X	R	Q	Y	U	Z	S

Analysis and Assessment

This model has the major flaw of not being exact. Hall’s stable marriage problem/theorem allows to get a definitive answer that is correct 99.9% of the time. However, there could be mistakes and no way in the current model to see it.

On the other hand, this model uses basic principles, allowing changes. Indeed, if the changes made in the model are incorrect, due to the simplicity of this model it would be flagrant, and therefore minimizing the possibility of having a major error without noticing it.

It does react positively to the changes of variables, thanks to the fixed ‘’framework’’ of the model. The equations are relatively simple, without complex relationships between the different parameters, hence forbidding crossed-linked error; There is always a maximum of two variables interacting at the same time.

Problem 2: Taxi Drivers in New York City

Summary Sheet:

Interpretation of the Problem

For this problem, we take into consideration that there are times in the day when taxis are vacant (or unoccupied). Some say that to get customers, they should drive to the city center where there is usually a lot of people. Knowing the exact latitude and longitude of all pickups and drop-offs of taxi drivers in New York City (Manhattan), we are asked to prove or dismiss the popular idea that more customers are in the center of the city. From there, two questions have been raised: First, what should a taxi driver do when it's car is vacant? Secondly, what should the head of a taxi company advise to his drivers?

Overview of Methods and Statements

The problem has been divided in multiples steps:

For the first question, the main concern was to find new customers for the taxi drivers. The main concern was to find where were the largest pickup areas, and whether or not these areas were in the center of the city. For that I: Associated each pickup location to one of the nineteen districts of Manhattan, Calculated the percentage of the pickups locations in the center area of the city, and Calculated which districts had the largest pickup percentages.

For the second question regarding the head of the taxi company, the focus was on the monetary aspect of the problem. As a head of a company, the most key factor is money. If the runs are further apart but gives a significant amount of money, the longer runs are favored. For that I: calculated the average total amount and the districts that are above that general average, and then the average of runs per district.

Overall, the main method used for this problem is Lloyd's algorithm (k-mean clustering). It is used to separate the runs into districts that are represented as centroids, using their longitude and latitude location.

Using this method, I was able to identify the nineteen districts of Manhattan as centroids, to create clusters with the latitude and longitude of the pickups location, and to identify the main clusters responsible for 1) the most customers and 2) the larger average total amount per run for each district.

Basic mathematics such as calculating averages and sums have also been used.

Solution:

Assumptions/ Justifications

1) The data given in the short sample of 500,000 taxi runs are representative of the general pool of values given for the entire year.

Justification: As I did not have enough time to write a functional program that would calculate the results for the large version of data and since I used Excel for my calculations, the amount of values for the large sample is simply too important for excel to run. Since it was offered to download a shortest version of the data, it will be considered as representative of the large sample.

2) The number of customers per districts remain the same through the years.

Justification: As the question asks for the head of a company to make decisions, it is assumed that throughout the years, the customer distribution will remain the same.

3) Only pickup locations are relevant for the first sub-problem (what should a taxi driver do if his taxi is vacant?).

Justification: Since the problem asks for us to determine where the taxi drivers should get their customers, only pickup locations are relevant. Also, it could happen that a customer wants a ride from inside toward outside of Manhattan, however these kinds of runs are rare and will not be considered in the calculations.

4) The center of the city is defined by eight (8) districts: Central park, Chelsea, Garment, Gramercy, Midtown East, Midtown West, Murray Hill, and Time Squarer.

Justification: [3] The New York Times made an article about the different districts of New York City and their location. Hence, the districts mentioned above are and will be considered as the center of the city for this paper.

Usage of Method(s), Variables and Parameters

The first step was to identify the relevant variables. In a document given to us contains a list of the first 500,000 taxi run. This list contains multiple variables for each trip: the time of the pickup and the drop-off, the location of the pickups and drop-off using their latitude and longitude, the trip distance, the fare amount, the extras, the MTA taxes, the improvement surcharge, the tip and tolls amount, and finally the total amount charged to the passenger.

Such amount of variable can be confusing and needs to be reduced to the relevant ones only. Using the assumptions mentioned above, only the pickup locations will be use when it comes to location. For the monetary aspect, the variables can be reduced to the total amount charged to the passenger as in includes all the other money related variables.

Moreover, since Lloyd's algorithm is used, the centroids of the clusters will be the location of the districts. There are nineteen (19) districts in Manhattan [4] (locations found with [5] Latlong.com in function of their longitude and latitude). The clusters are created using the latitude and longitude of the pickup location for all the taxi trips.

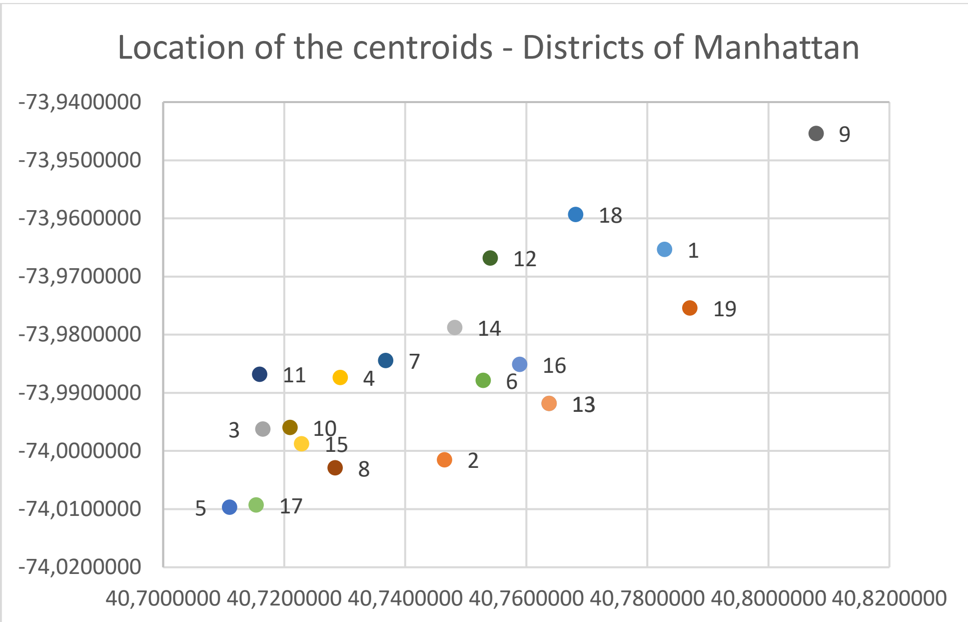
Finally, it has appeared while writing this paper that the distance of the trips was also an essential element. As it will be discussed further in this paper, the difference in the averages of the total amount charged to the clients per districts comes from the distance between the pickup location and the drop-off location. It seems as if the customers of different districts have different travel habits according to their location.

Hence, for this problem, the variables used are: The locations of each districts (found with [5] Latlong.com), the location of the pickups, the total amount charged to the clients, and the distance between the pickup location and the drop-off location.

Manhattan is divided into nineteen districts that have been assigned a number from 1 to 19:

1	Central park	11	Lower East Side
2	Chelsea	12	Midtown East
3	Chinatown	13	Midtown West
4	East village	14	Murray hill
5	Financial district	15	SoHo
6	Garment	16	Time Square
7	Gramercy	17	Tribeca
8	Greenwich	18	Upper East Side
9	Harlem	19	Upper west side
10	Little Italy		

We then create a first template containing location of the districts as centroids:



It is immediately possible to observe that the centroids shape the island of Manhattan. This confirms that the centroids and their locations must be exact.

Lloyd's algorithm is then applied to the 498,674 pickups location. As a reminder from the previous problem, This is the formula to calculate de centroid of a cluster:

$$\arg min \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{y,x \in S_i} \|x - y\|^2$$

S_i here is the the number of sets (eleven in our case)

$\sum_{y,x \in S_i} \|x - y\|^2$ is the Eucledian distance.

The clusters with the pickups location values are too close for us to be able to distinguish the different clusters. However, using Excel and the IF condition, the pickup locations are assigned to the districts they belong to.

Example of the IF condition:

=(IF(AF2=M2;$M$1;IF(AF2=N2;$N$1;IF(AF2=O2;$O$1;IF(AF2=P2;$P$1;IF(AF2=Q2;$Q$1;IF(AF2=R2;$R$1;IF(AF2=S2;$S$1;IF(AF2=T2;$T$1;IF(AF2=U2;$U$1;IF(AF2=V2;$V$1;IF(AF2=W2;$W$1;IF(AF2=X2;$X$1;IF(AF2=Y2;$Y$1;IF(AF2=Z2;$Z$1;IF(AF2=AA2;$AA$1;IF(AF2=AB2;$AB$1;IF(AF2=AC2;$AC$1;IF(AF2=AD2;$AD$1;IF(AF2=AE2;$AE$1;0)))))))))))))))))))

The results are displayed in the following table:

Column1	Districts	Number of runs	Percentage of runs
1	Central park	16091	3,2268%
2	Chelsea	6	0,0012%
3	Chinatown	0	0,0000%
4	East village	0	0,0000%
5	Financial district	17845	3,5785%
6	Garment	2	0,0004%
7	Gramercy	175	0,0351%
8	Greenwich	1	0,0002%
9	Harlem	153908	30,8634%
10	Little Italy	1	0,0002%
11	Lower East Side	169922	34,0748%
12	Midtown East	20177	4,0461%
13	Midtown West	3	0,0006%
14	Murray hill	1	0,0002%
15	SoHo	0	0,0000%
16	Time Square	0	0,0000%
17	Tribeca	23	0,0046%
18	Upper East Side	118322	23,7273%
19	Upper west side	2197	0,4406%
	total	498674	

Using the data for the nine districts mentioned earlier that represents the center of the city, it is calculated that only 7% of the runs starts in the Center of the city.

The locations are first divided into two categories : 1 if the location belongs to one of the nine districts, 0 if it belongs to another district:

=IF(AG2=$AL$2;1;IF(AG2=$AL$3;1;IF(AG2=$AL$17;1;IF(AG2=$AL$7;1;IF(AG2=$AL$15;1;IF(AG2=$AL$13;1;IF(AG2=$AL$14;1;IF(AG2=$AL$8;1;0))))))))

Then, the 1 locations are added and divided by the total runs:

Taxi trips starting in the Center of the City
36455
7%

It is then concluded that no, if your taxi is vacant, going to the center of the city will not get you more customers.

From there, the question was: If the center of the city is not the main way to get customers, where are the customers?

As it can be seen in the table above, Harlem represents 30.86% of the pickups, Lower East Side 34.07%, and Upper East Side 23.72%. None of these locations are in the Center of the city and yet they represent 88.65% of the pickups.

Now, as the first question has been answered, the following step was to work on the head of the company question. As mentioned in the Assumptions section, the key element for the head of a company is monetary. Hence, the total amount of money charged to the clients is the variable used for this section.

First of all, a general average of the total amount has been calculated. The general total amount is $14.34. Then, using the IF function again, the 498,674 taxi trips are divided into two categories: 1 if the total amount is above or equal to the average, 0 if it is not.

Example of function used: =IF(L2>=$AK$6;1;0)

Then, the number of trips having a total amount superior or equal to the average were computed per districts. The average per district was also calculated. The results are shown in the following table:

	Districts	Above Average		Average per District
1	Central park	4791	2.863%	13.629
2	Chelsea	4	0.002%	100.2
3	Chinatown	0	0.000%	0
4	East village	0	0.000%	0
5	Financial district	8114	4.849%	17.3849
6	Garment	2	0.001%	26.8
7	Gramercy	77	0.046%	15.0986
8	Greenwich	1	0.001%	68
9	Harlem	43440	25.958%	14.4481
10	Little Italy	1	0.001%	17.8
11	Lower East Side	68031	40.652%	15.5532
12	Midtown East	10125	6.050%	17.3095
13	Midtown West	3	0.002%	128.75
14	Murray hill	1	0.001%	18.3

15	SoHo	0	0%	0
16	Time Square	0	0%	0
17	Tribeca	16	0.010%	85.8314
18	Upper East Side	32148	19.210%	13.2672
19	Upper west side	596	0.356%	13.38620784
	total	167350	100.000%	

This table shows that:

1) The number of total amount above the average per districts is not very relevant. Since there are a larger population taking the taxi in Harlem, Lower East side, and Upper East side as it was proved earlier, the percentage of total amount above the average is necessarily important. Even if a person in Time square gave a thousand dollar, it would still represent a small percentage as it is only a single person versus the 68,031 individuals above the average who took the taxi in Lower East side

2) Unlike the previous point, the average total amount per district is more revealing of the customer's habits. The trips in the districts with the most customers seems to have a lower average of total amount of money charged to the client. Indeed, Midtown West and Chelsea both have averages above $100, while Harlem, Lower East Side, and Upper East side have averages around $14-$15.

As the head of a taxi company, the question following these results would be: why?

The answer to this question, seems to rely on the trip habit of the customers, and more specifically the customers of each districts. Hence, the following table shows the percentage of users who stays in the same district from the pickup to the drop-off:

	Districts	Same districts ?	Pickup	%same district
1	Central park	2874	16091	18%
2	Chelsea	4	6	67%
3	Chinatown	0	0	0%
4	East village	0	0	0%
5	Financial district	5439	17845	30%
6	Garment	2	2	100%
7	Gramercy	2	175	1%
8	Greenwich	1	1	100%
9	Harlem	107766	153908	70%
10	Little Italy	1	1	100%
11	Lower East Side	113883	169922	67%
12	Midtown East	5771	20177	29%
13	Midtown West	2	3	67%
14	Murray hill	1	1	100%
15	SoHo	0	0	0%
16	Time Square	0	0	0%
17	Tribeca	18	23	78%
18	Upper East Side	86564	118322	73%
19	Upper west side	522	2197	24%

The results show that Harlem, Lower East Side, and Upper East side's customers tend to stay in the same district around 70% of the time. Some results are excluded such as the percentages of Garment or Little Italy as the sample size is too small to get a proper representation (i.e 1 or 2 individual versus 85,564 or 113,883 individuals).

Hence, the smaller average of total amount charged to the client can be explained by the habits of the customers, which tends to differ depending on their pickup district.

Finally, the last step was to determine whether it is more profitable to do a larger amount of trip for a smaller amount of total charges, or a lesser number of trips but for a greater amount of total charges.

As Harlem, Lower East Side, and Upper East Side represent 88% of the total trips accomplished within the time frame of the given data, and with an total amount average of $14.45, $13.27, and $15.55 respectively, the total amount of money for those three districts is $6,436,311.

Out of a total of $7,350,273.25, those three districts represent also 87.6% of the total amount of money. Therefore, it can be concluded that it is more profitable to make multiple small trips than fewer longer trips.

Conclusion to the Problem

In conclusion, using Lloyd's algorithm and by modelling the data, it became possible to answer the two questions of the problem:

First, it is not true that when the taxi is vacant, the taxi driver should go to the center of the city. As demonstrated earlier, the larger group of customers are in three distinct districts: Harlem, Lower East Side, and Upper East Side. These three districts represent 88% of the total trips. So as a taxi driver, if my taxi is vacant, I would head toward one of these three districts, whichever is the nearest.

Second, as a head of a taxi company, my main interest is the money. Hence, using the calculations made earlier, it is more profitable for my company to run multiple small distances than fewer large distances. It is also more profitable to be in a district where the average of the total amount is lesser but with a larger pool of customers. As the head of a taxi company, I would recommend to my drivers to take clients that stays in the same district or within a small radius of their pickup position, to be able to do more trips.

Analysis and Assessment

The model is viable if there are no major changes in the city. As New York City is a city that is constantly changing, this model would need to be updated frequently to keep up with New York's life style. However, if it is considered that the first 498,674 trips given as data are representative of the larger sample of data, then this model is quite viable. There are not a substantial number of variable, hence preventing major calculation mistakes or too complicated formulas. Since this model is mostly accessible to most of the population, it is also a positive aspect of it.

If there had to be a major change in the population's habits, the results above would be false but one of the good points of this model is that it can be adapted quickly. With a little more time, there could have been a program written for this model, where it would only be necessary to change the variables that were modified and the ''skeleton'' of the model would remain unchanged.

References

[1] - Anon, (2017). *Privacy: Gone with typing.* [online] Available at: http://athena.ecs.csus.edu/~changw/Pubs/IRI05.pdf

[2] - Brilliant.org. (2017). *Applications of Hall's Marriage Theorem | Brilliant Math & Science Wiki*. [online] Available at: https://brilliant.org/wiki/applications-of-hall-marriage-theorem/

[3] - Nytimes.com. (2017). *The Heart of Manhattan*. [online] Available at: http://www.nytimes.com/1994/09/21/opinion/the-heart-of-manhattan.html

[4] – Frommers.com. (2017). *Neighborhoods in Brief in New York City | Frommer's*. [online] Available at: http://www.frommers.com/destinations/new-york-city/neighborhoods-in-brief

[5] - Latlong.net. (2017). *Times Square, Manhattan, NY, USA Map Lat Long Coordinates*. [online] Available at: https://www.latlong.net/place/times-square-manhattan-ny-usa-7560.html