sensors

MDPI

Advanced Network Analysis: Flow of taxi passengers in NYC

Rodrigo Sarroeira [1], Wendel Vilaça [2]

[1] ISCTE-IUL; rcdfs@iscte-iul.pt
[2] ISCTE-IUL; ------

Abstract: In this paper a study of the flow of taxy passengers in New York City is carried out. The database in use contains almost 1.5 million observations. Each observation corresponds to a taxi trip inside the New York State. Several transformations to the data are applied to enable possibility of analyzing the data using network analysis technics. An exploratory analysis is also carried out to allow a better understanding of the variables. The goal is to understand the areas that present a higher flow of people. The periods with the highest number of trips will also be analyzed, to understand what causes them. To accomplish these defined goals, we resort to the NetworkX python package, used for creating, modeling, and analyzing the graph. For visualization purposes, the Folium python package is imported, given that the data is geo-referenced. Our study revealed that the hours with more trips and flow of passengers correspond to the rush hours, caused by the movimentation of people to and from work. The afternoon rush hour, between 6 and 7 pm is the period of the day with a higher intensity of the flow of passengers. 5 am is the hour that presents less flow of passengers. The following network analysis metric are calculated: Mean degree (80.86), density (0.196), average assortativity (0.102), average cluster coefficient (0.645), average betweeness (0.003), and the average closeness (0.414).

Keywords: Network analysis; Data science; Mobility; Data mining; Visualization.

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Received: date
Accepted: date
Published: date

1. Introduction

This paper focuses on the analysis of a dataset containing information on taxi trips in the New York State during the year of 2016. This database was extracted from a Kaggle competition and contains 1 458 644 observations and 11 columns. The goal of this competition is to develop a model to predict the total duration of a taxi trip given the other variables. Instead of predicting the time taken by a certain trip, we will use network analysis to mine the data and extract important features, patterns, and conclusions. To enable the use of network analysis algorithms, first, it is needed to prepare the database for network analysis, by applying a preprocess phase to the data, for later analysis. The raw database does not allow network analysis, given that each trip would generate two new nodes, created by their longitude and latitude coordinates.

This dataset consists of nominal, discrete, continuous quantitative variables. In it we can find information related to each taxi trip, such as, the collection time, geographic coordinates of the pickup and dropout locations, number of passengers, time taken, and several other variables.

With this project we intend to answer some questions related to our data. Firstly, we want to understand which are the hotspot locations, in terms of the flow of passengers. Secondly, we want to study what are the periods with highest flow of people inside NYC. Also, a study of the density of the pickup and dropout locations is carried out, this study will show not only the flow of passengers, but also their direction. To answer these questions, a spatial and temporal analysis of the data must be carried out. To get insight on our data, several network analysis metrics are calculated and interpretated. Also, visualization technics are applied to allow the draw of conclusions.

2. Materials and Methods

For this project, we based ourselves on the concepts of the CRISP-DM methodology (Cross Industry Standard Process for Data Mining). It is a technique used from Data Mining and which is also used in data science, in view of its robustness. This methodology is responsible for bringing together the best practices in data mining, allowing the management and analysis of data through Business Intelligence projects and tools to be carried out more efficiently. Figure 1 shows that steps that constitute the CRISP-DM methodology.

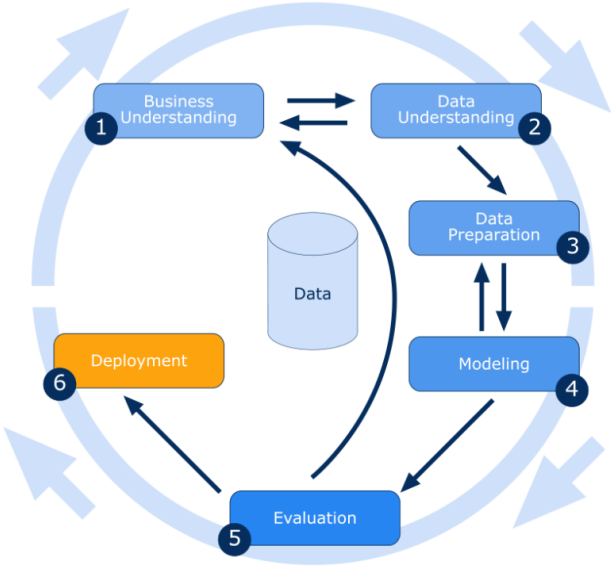


Figure 1. CRISP-DM diagram.

In this subsection, the flow and structure of the project is described. Figure 2 presents the flow of the project visually. The development of this study is done in python, given its flexibility, simplicity, and strong packages to deal with data stored in networks. The first phase of this project began with the Business and data understanding (Subsection 3.1). This phase allowed defining the goals of the project and understand the meaning of each variable in the dataset. Secondly, the Data collection (Subsection 3.2) phase is developed, here the importation of the dataset is carried out. The first focus of our project is developed in the Data preprocessing (Subsection 3.3) phase. In this phase, a set of transformations, using the pandas python package, is applied to the database, to clean it, add new variables, and most importantly, prepare it for network analysis. After the cleaning of the dataset, the data is ready for the Exploratory analysis (Subsection 3.4) phase, here charts and statistics are calculated to gather important and useful knowledge on the available variables. The next phase consists of the import of the cleaned dataset to a graph object, this phase is called Graph creation (Subsection 3.5). To model the graph the networkX python package, used for modelling networks, is used. This tool presents useful functions and metrics to help the extraction of information from a network. Once the data is modelled into a graph object, the Graph study – Metrics phase (Subsection 3.6) starts, here we apply network analysis algorithms to extract information on the data. The Graph study – Visualization (Subsection 3.7) phase is carried out to allow a visual understanding of the data and its interactions. In this phase, the folium python package is fundamental to allow the visualization of the network, given that the data is geo-referenced. Finally, the conclusions (Section 4) of our study are presented in the last section.

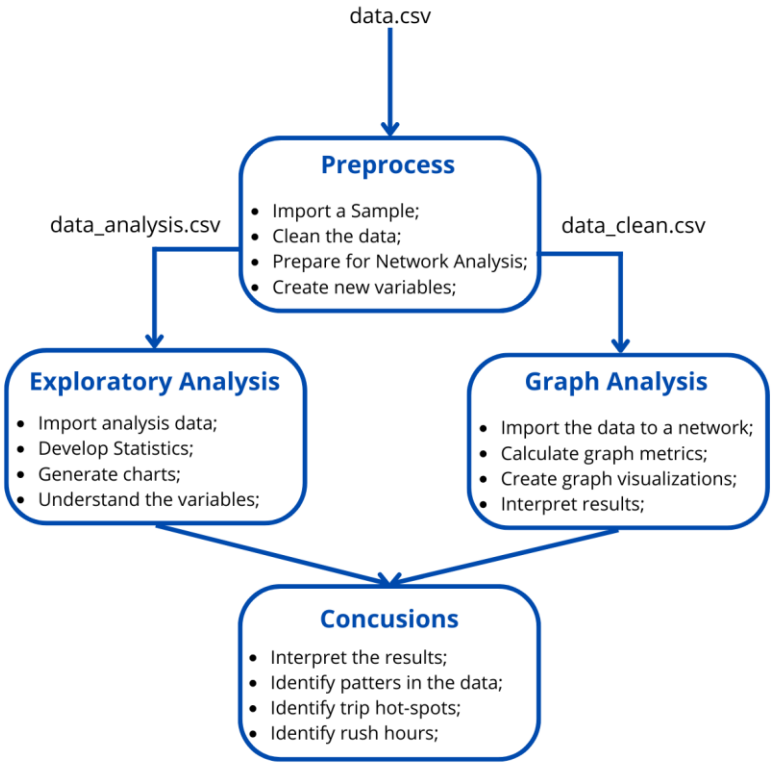


Figure 2. Methodology and development workflow

4. Data

This is the main section of this article, here all the interactions with the data are presented, from the data understanding, collection, preprocessing, modeling, to analysis. These steps will be divided into subsections as section 2 (Materials and Methods) shows.

3.1. Bussiness and data understanding

In this subsection the variables available within the database are contextualized. For each variable, its datatype, meaning, and relevance are discussed. The first variable is the id, this variable serves as a unique identifier of each trip and may be useful in detecting duplicated observations. It is a categorical variable, given that it contains characters and numbers. Secondly, the "vendor id" variable is a numeric variable that identifies the provider associated with the trip record. This variable is only interesting for analysis. Both "pickup datetime" and "drop-off datetime" are datetime variables that serve as a timestamp to indicate the beginning and ending of each trip, respectively. These two variables are very important to our study, given that we want to understand the periods with higher flow of people, for that our data needs to be contextualized in time. The "passenger count" variable is of type integer and represent the number of passengers in each trip. This variable is key in the development of our study, since it contains information relative to the number of people travelling. The following four continuous variables, "pickup longitude", "pickup latitude", "drop-off longitude", "drop-off latitude", save information relative to the pickup and the dropout locations, in other words, the location where the trip started and the location where the trip ended. The "store and fwd flag" variable indicates whether the trip record was held in vehicle memory or not, it is a Boolean variable. The last variable is the "trip duration" which is a continuous variable that represents the duration of the trip in seconds.

3.2. Data collection

The dataset in use was extracted from a kaggle competition (XXX), the data comes in a csv file. To interpert the data, it is imported to a pandas dataframe structure, enabling easy manipulation and visualization of the information. Given that our dataset contains almost 1.5 M observations, and some of the algorithms that are applied are have a high time complexity, a subset of the data is selected. The sample is set to 20% of the original data, to select this information a random method is implemented, a seed is set to allow verification. The sample contains 291 728 rows. To simplify some processes some of the variables are renamed. The following variables, "pickup latitude", "pickup longitude", "drop-off latitude", "drop-off longitude", "pickup datetime", "drop-off datetime", and "passenger count", are renamed to, "x1", "y1", "x2", "y2", "time1", "time2", "n pass". These are the initial variables, during the development of this project other variables will be created with the goal of enriching the data to enable a better analysis. Figure 2 shows the result of the importation of the database.

	id	vendor_id	time1	time2	n_pass	y1	x1	y2	x2	store_and_fwd_flag	trip_duration
1329268	id1356976	1	2016-06-27 21:53:40	2016-06-27 21:59:12	1	-74.013336	40.702694	-74.011681	40.713924	N	332
1345788	id3724158	2	2016-05-27 01:14:10	2016-05-27 01:43:49	2	-73.781799	40.644669	-73.970329	40.762581	N	1779
293496	id2080002	1	2016-03-21 13:23:09	2016-03-21 13:33:37	1	-73.983582	40.762287	-74.002769	40.760578	N	628
614139	id2142952	2	2016-03-23 16:25:35	2016-03-23 16:32:43	1	-73.970474	40.758896	-73.975914	40.744888	N	428
1331566	id0732479	1	2016-04-01 22:59:15	2016-04-01 23:07:37	1	-73.974625	40.762012	-73.969872	40.752270	N	502
...	...	...	...	...	...	...	...	...	...	...	...
397700	id2593421	1	2016-04-04 08:56:21	2016-04-04 09:21:08	1	-73.977188	40.787697	-73.978951	40.751228	N	1487
327345	id1023664	2	2016-02-12 14:34:24	2016-02-12 14:36:28	1	-73.950996	40.791630	-73.948624	40.797401	N	124
350877	id1799316	1	2016-05-28 08:32:52	2016-05-28 08:44:59	1	-73.946335	40.776424	-73.976402	40.759563	N	727
1198343	id3443806	2	2016-01-03 11:05:24	2016-01-03 11:08:23	1	-73.995796	40.732956	-73.999039	40.738346	N	179
28617	id2984025	2	2016-03-02 16:25:13	2016-03-02 16:29:49	1	-74.007355	40.743382	-73.998306	40.755585	N	276

291728 rows × 11 columns

Figure 3. Data table

3.3. Data Preprocessing

Firstly, we studied the database to find missing values, but none were found. Secondly, given that our study focuses on the NYC and its roundabouts, the observations are filtered by their longitude and latitude coordinates, to fit the area under analysis. We defined the location of study as all the points in which the latitude is between 40.683280 and 40.846102, and the longitude is between -74.030592 and -73.892724. Trips where the pickup or drop-off locations don't belong to our defined interval are removed from the database. With this filter, 10,41% of the observations were removed.

To enrich the data de distance variable is created. Given that for each trip the database contains the pickup and drop-off locations, it is easy to calculate the distance between these two points. To create this variable the Euclidean distance is taken into consideration, formula 1 presents the expression used to calculate the distance. In this case, n is equal to two, given that a point in earth is referred to as a set of two coordinates, latitude, and longitude. After calculating the distance for each trip, a new column, containing this information, is added to the dataset.

$$d(x, y) = \sqrt{\sum_{i}^{n} (y_i - x_i)^2}$$

Formula 1. Euclidean distance

The next step marks the beginning of the dataset transformation for network analysis. For each trip, the database contains the coordinates of the pickup and the dropout locations. This information is not very useful to import to a network, because the coordinates are continuous variables. This would be a problem, because almost every trip would generate two new nodes in the graph, created by the coordinates of the drop-off and pickup locations. The result of the importation of this dataset to a network object would be the

following: a network with density close to zero, n nodes, and n / 2 edges. To overcome this problem, a data transformation method is implemented.

The transformation consists in using the coordinates to assign each trip to a pickup area and a drop-off area. This will allow analyzing the data through a network analysis point of view. Each node will correspond to a certain area of NYC, and each connection represents a trip between these two areas.

Before assigning each trip to a specific influence area, we must create them. The process of creating the influence areas is simple, we define the interval of interest, equal to the one presented before, and define a step, that represents the distance gap between areas. This step is defined as 0.01º. The algorithm generates 270 influence areas, all of them contained within the defined interval. Figure 4 shows the influence areas along with 100 trips, where the red point corresponds to the pickup and the green one to the drop-off.

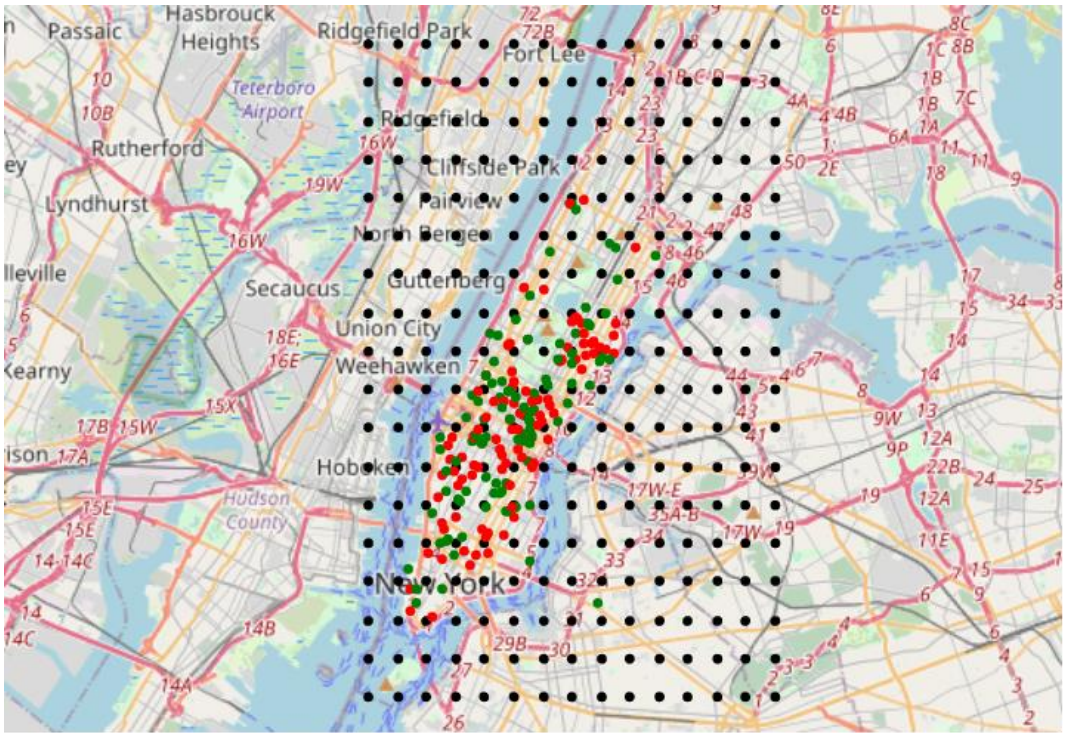


Figure 4. Influence areas and trips

Now that the influence areas are created, we need to assign each trip to two areas, a pickup area, and a drop-off area. To do it, the Euclidean distance (Formula 1.) will be used again. For each trip, the distance between its pickup location and all the influence areas must be computed, the trip is assigned to the area from which it is the closer. The same process must be repeated for the drop-off location. This algorithm presents a very high time complexity given that for each trip 540 distances must be computed, 270 distances for the pickup area and 270 distances for the drop-off area. The Euclidean distance itself, with an n equal to two, preforms 2 subtractions, one addition, two squares, and one square root. To prepare the full dataset, 997 709 760 (291 728 * 570 * 6) operations must be performed, this number almost reaches 1 billion. The time complexity of this algorithm is the

reason why we decided to work only with 20% of the data, otherwise the algorithm would take more than 24 hours to compute all the operations.

Finally, as the data in use is related to trips and taxis, we identified that an analysis from the perspective of time spent on trips, their daily, monthly, and annual occurrences could give us a holistic view of the data. Therefore, this specific data can expand the context to better understand its behavior, as well as identifying outliers. To enrich the data and make it easier to access, several transformations are applied to the time variables. These dataset transformations bring more value to the analysis and decision making. The "year", "month", and "hour" variables are extrapolated from the "time1" variable. The "trip pickup date", and "trip pickup time" variables are transformed from the "time1" variable. Finally, we create the "trip dur min" that transforms the "trip duration" variable, that is expressed in seconds, to minutes. These variables will be fundamental in the Exploratory analysis (Subsection 4.4) phase, and in the Graph study – Visualization (Subsection 4.7).

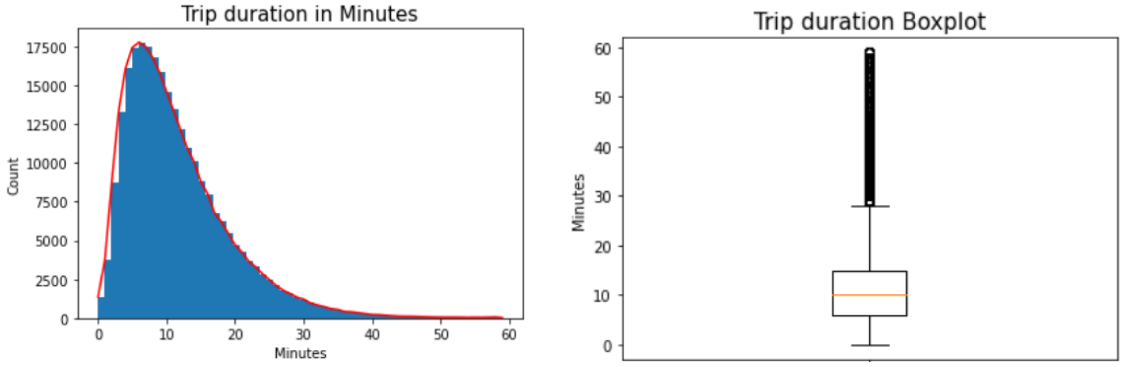
After this detailed preprocessing phase, two datasets are saved. The "data_clean.csv" that will be used to create the network object in the Graph creation (Subsection 4.5) phase, and the "data_analysis.csv" that is used in the Exploratory analysis phase (Subsection 4.4).

3.4. Exploratory Analysis

The present subsection is developed with the goal of extracting insight on the variables of the dataset. Charts are generated to understand how the taxi trips are distributed through the days and hours. The duration of the trip and the number of passengers is also studied. The maps will study the spatial distribution of the pickups and the drop-offs.

Through the engineering of resources for temporal variables, we created a variable "trip_dur_min" that represents in an agglutinated way the minutes spent on each trip, so we have as objective the distribution of data referring to the time spent, as we can see in figure 5, image (a). This histogram demonstrates the behavior of this variable as well as its representation defined by the Poisson distribution, considering that its application occurs when the number of possible occurrences is much higher than the average number of occurrences.

In figure 5 (b) the boxplot allows the visualization of the distribution and outliers of the time taken by the trips. The outliers of the data represent values that are very extreme, in this case there are only upper outliers, which is normal given that this variable has a Poisson distribution. Also, by interpreting the boxplot it is possible to identify the 25% and 75% quantiles, as well as the average. The 25% quantile is around 6, the average value is of 10, meaning that on average a trip is expected to take 10 minutes, finally, the 75% quantile is 27.

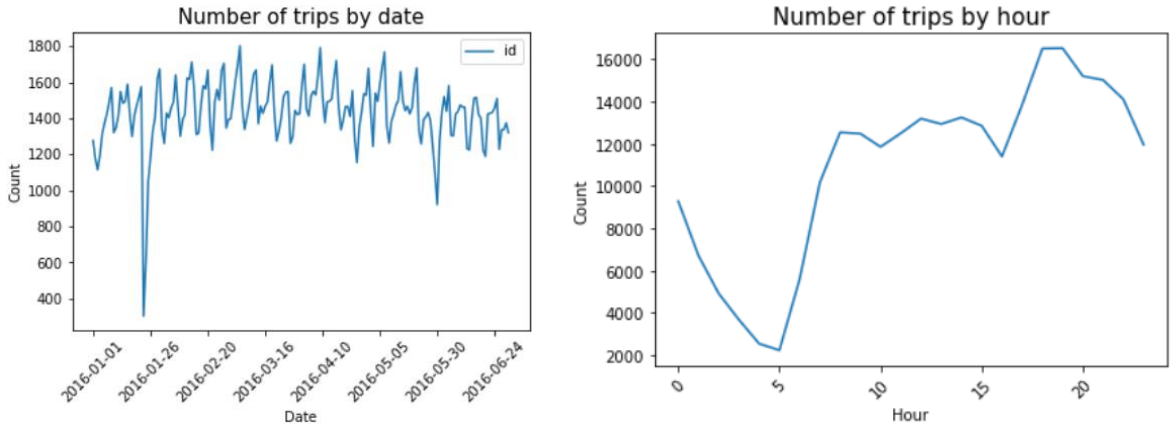


(**a**) (**b**)

Figure 5. (**a**) Distribution of the trip duration; (**b**) Boxplot of the trip duration.

In figure 6, image (a) we analyze the distribution of the numbers of trips appropriate to their frequencies by date, that is, over the period under analysis, in order to understand the distribution of the time series. In this time series, we identified an almost homogeneous frequency distribution, but which differed more sharply at the end of January, more specifically on 01/23/2016, where there were only 302 registered trips. This drastic decrease in the number of trips was due to the fact that the Jonas blizzard took over New York with more than 68 centimeters of snow, being classified as the 2nd largest accumulation of snow since 1869 with a storm.

To create the chart on figure 6 (b), the hour variable is used, first the observations are grouped by the hour of their pickup time, then a count is applied to each gorup, giving us the number of trips by hour. In it we can observe a distribution that accumulates between 8 am and 10 pm, with its peak at 7 pm, better known as rush hour, on the other hand, at 5 am we have the lowest flow of trips. There is a sharp decrease in the number of trips between 9 pm and 5 am.



(**a**) (**b**)

Figure 6. (**a**) Number of trips by day; (**b**) Number of trips by hour.

Another chart was developed, this time using the number of passengers. Figure 7 presents a bar chart with the number of trips segmented by the number of passengers. This way, we were able to identify the highest frequency of trips for each group, in this case, individual trips occupy about 75% of the total trips. The second most frequent number of passengers on a trip is two, followed by five. The category that presents the lowest frequency is the category of trips with four passengers.

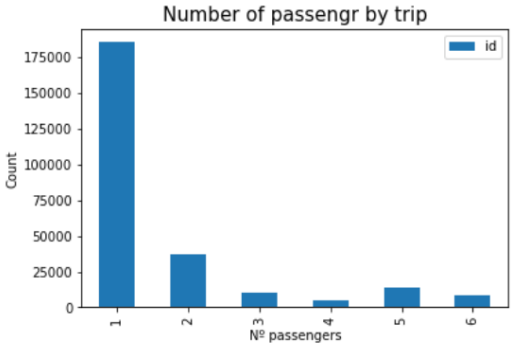
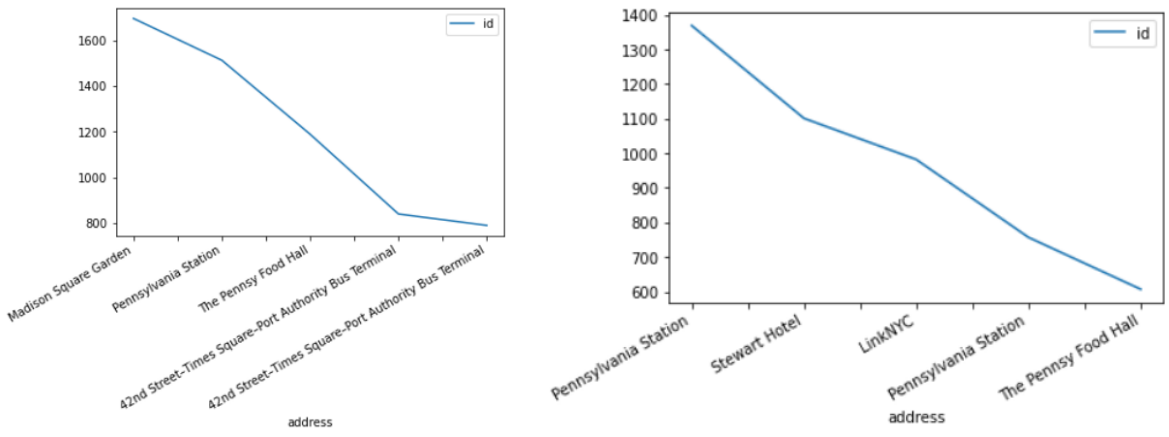


Figure 7. Number of passengers by trip

The following figure presents two line charts. Figure 8 (a) presents the five areas with higher number of pickups. It is possible to see that the Hadison Square Garden is the location that records more pickups, with 1696 occurencies. Secondly, the Pennsylvania Station presents 1513 ocurencies, aproximatly. The Pennnsy Food Hall is the third location with higher iontensity of pickups (1189). Finally, the last two locations actually correspond to the same area, which is the Bus Terminal of Times Square, both of these entries together present 1630 occurencies, meaning that, this location is actually the second biggest hotspot, in terms of pickups.

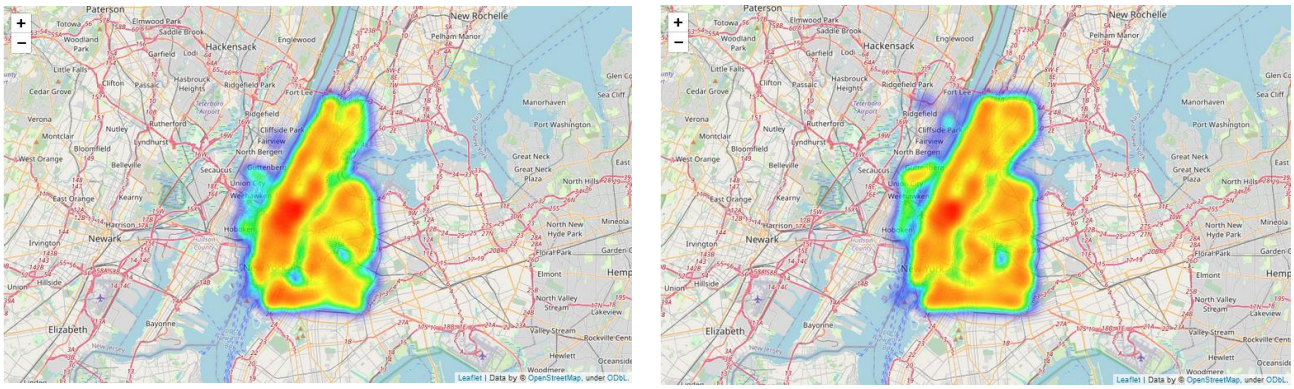
On the other hand, Figure 8 (b) presents the same analysis, this time, for the drop-off hotspots. In this chart, the Pennsylvania Station presents two distinct entries, together the number of drop-offs in this location is of 2126. Secondly, the Stewart Hotel presents 1101 registered drop-offs, this makes sence given that this location corresponds to an hotel. The LinkNYC and the Pennsy Food Hall present 982 and 607 drop-offs, respectively.



(**a**) (**b**)

Figure 8 (**a**) Pickup hotspots; (**b**) Drop-off hotspots.

Figure 9 continues the study of the pickup and drop-off locations. To understand the spread of these locations across the area under analysis, an Heatmap, created with the folium package, is used. Figure 9 (a) represents the distribution of pickup locations, while figure 9 (b) shows the same information for the drop-off locations. It is possible to understandf that the areas with higher intensity of both pickups and drop-off is in the central part of the NYC, given that this area presents a dark red. By comparing the shapes, it is possible to see that the distribution of the drop-off and pickups is very similar. There are only some locations where a considerable difference can be noticed, for example the upper left corner presents a higher density of drop-offs than pickups.



(**a**) (**b**)

Figure 9. (**a**) Pickups heatmap; (**b**) Drop-offs heatmap.

4.3. Graph creation

Given that the data is ready to be imported to a graph structure and the exploratory analysis phase allowed a better understanding of the variables, it is viable to start modelling the graph. To create the graph object containing the information relative to all the trips, the networkX python pakage is used. Before adding the nodes and the edges to the graph, one more processing is applied. Instead of adding one edge per trip to the graph, the information is merged with the goal of creating a lighter graph that contains the same information.

The idea is simple, all the trips that start in the same area and end in the same area will be grouped. Let's imagine the following example: there are three trips connecting area x and area y, these trips transported 3, 1, and 2 passengers. Instead of adding three edges to the graph, only one edge is added, but this edge has two different weights, the number of trips (3) and the number of passengers (3 + 1+ 2 = 6). Following this methodology, we ensure that no information is lost, furthermore, the graph will be more efficient in terms of time complexity, given that the number of edges is substantially reduced.

After this second preprocess to the data, the number of nodes is equal to 207, instead of the initial 270 areas. This means that 63 influence areas do not present any trips in the database, therefore these areas are not included as nodes. The number of edges presents a very big reduction, instead of the supposed 291 728, only 8 370 are inserted to the graph, this is a reduction of 97.13%. Again, this reduction in the number of edges does not represent a loss of information, because the information relative to the number of trips and

passengers between two areas is saved in the weights of each edge. This transformation will allow calculating graph metrics more efficiently. The graph is created as a directed graph, given that each edge represents the intensity of flow from area x to area y.

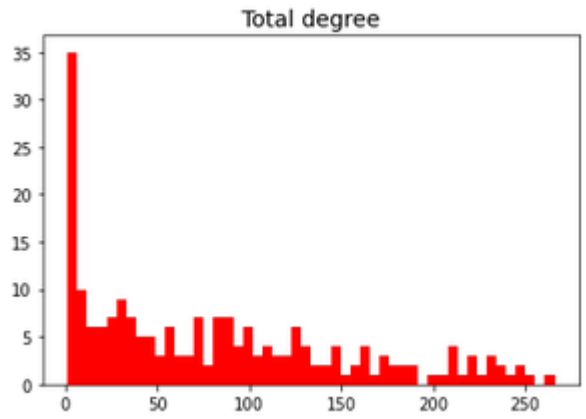
3.6. Graph study – Metrics

In this subsection, several network analysis metrics are calculated. These metrics allow the drawing of conclusions relative to the network. First, a study of the node degree is carried out. Given that our graph is directed, it is important to study the total degree, the in-degree, and the out-degree. In the context of our data, the total degree of a given area represents the number of areas that are connected to area x. The in-degree of area x represents the number of areas that contain trips that end in area x. The out-degree of area x represents the number of areas that are connected to area x, given that the trip started in area x.

Figure 10 (a) presents a table containing information on the node degrees of the network, while figure 10 (b) shows the distribution of the node degrees. The area that is more connected to other areas presents 266 edges linked to it. On average, each area is connected to 80 other areas. There are areas that don't present any trips that end on it, because the minimum in-degree is zero. The area that contains more areas pointing to it presents an in-degree of 99. Given that the minimum out-degree is 0, there is at least one area that does not contain any pickup location. The area from which more trips go to other areas presents a maximum out-degree of 168. Finally, the average out-degree and in-degree are the same, which was already expected, given that the sum of all the in-degrees must be equal to the sum of out-degrees, thus both values divided by the number of nodes, must also be equal. Figure 10 (b) shows that there are many areas that present a low degree, in other words, areas that are connected to a few areas. It is possible to notice that there are fewer nodes presenting higher degrees. Nodes with high degree correspond to areas where the flow of passengers, or at least, the intensity of trips is higher.

	Degree	In Degree	Out Degree
Min	1.000000	0.000000	0.000000
Mean	80.869565	40.434783	40.434783
Max	266.000000	99.000000	168.000000

(**a**)



(**b**)

Figure 10. (**a**) Node degree statistics; (**b**) Node degree distribution.

Another important network analysis metric is the density of the network. Basically, the density of the network represents the proportion between the existing edges and all the possible combinations of edges. In the context of our problem, the density of the network shows that only 19,6% of all the connections between two areas are present in the database. This metric is calculated resorting to a networkx function and confirmed using the mathematical formula. The mathematical formula is presented in Formula 2.

$$D = \frac{Nº\ edges}{Nº\ nodes \cdot (Nº\ nodes - 1)}$$

Formula 2. Network density

The clustering coefficient is a very famous measure of the degree to which nodes in a graph tend to cluster together. Basically, for each node the proportion of possible triangles with its neighbors is calculated. For exampe if the cluster coefficient of a node is 0.5, it means that half of all the possible triangles with its neighboors are present in the netwrok. Figure 11 shows the distribution of the cluster coefficient of the nodes. It is possible to observe that most of the nodes present cluster coefficients between 0.4 and 0.9. It is also possible to see that tehre are 13 areas that are fully connected to its neighbors, therefore presenting a cluster coeficient of 1. 20 areas present a cluster coefficient of 0, meaning that they do not tend to be clustered to its neighboors. The average cluster coefficient of the network is calculated resorting to networkx, its value is of 0.65, which is considered a high value for this metric. Therefore we can conclude that the nodes in this network tend to cluster together, creating triagle between neighboors.

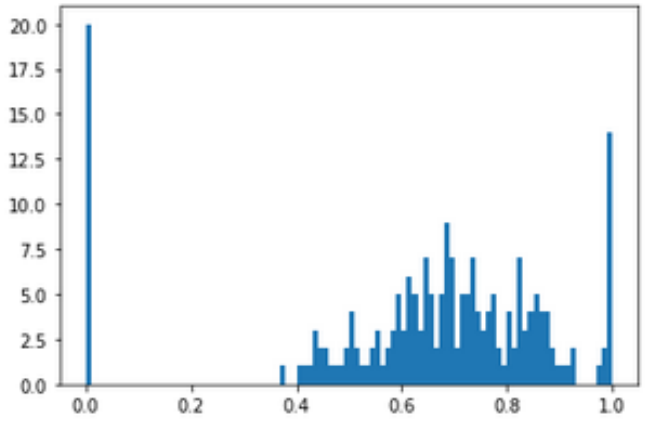
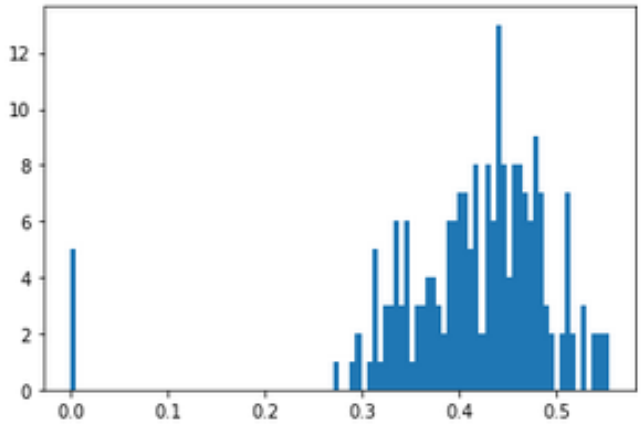


Figure 11. Distribution of the cluster coefficient of the nodes

The assortativity is a metric that tries to study weather nodes that are similar tend to connect between them or not. If this metric is 1, then simillar nodes only connect between them. If the value is –1 than it means that nodes only connect to nodes that are different from them. If the value is zero it means that there is no pattern regarding the similarity of nodes that constitude the edges. In this case, the assortativity levels is calculated using the degree of the nodes as the similarity criteria, the average assortativity value is –0.10. Since the value is close to zero it means that the relation is not very intense, still, given that this value is negative it means that nodes with higher degrees tend to connect to nodes with lower degree and vice-versa.

Figure 12 presents the distribution of the closseness of the nodes. The closeness is a metric that indicates how close a node is to the rest od the network, basically this is a centrality metric, that indicates how important a node is to the network. A node that is connected to all the other nodes would have a closeness value of 1. While a node with zero conection would get a closeness value of zero. It is possible to see that the majoraty of nodes presents a closeness value beteween 0.38 and 0.5. These values are not low. The average closeness value of this network is 0.41, meaning that on average, each node tens to be connected to 41% of the remaining nodes. This shows that our graph is connected.



3.7. Graph study – Visualization

To create the visualization present in figure 13, several steps are carried out. Firstly, the map object is created resorting to the folium python package. The second step is to add the nodes of the graph to the map, each node corresponding to an influence area. The third step is to add the edges, representing the flow of passengers between areas. To allow a better visualization transformations are applied to the edges. Firstly, the edges are sorted by their weight (number of passengers), this way the first edges to be added to the map are the ones that present a lower number of passengers. This will make the edges with higher weight be on top on of the others, enabling us to easily understand what are the areas that present stronger connections. Also, the weight of the connections influences their color, width, oppacity. This map was the first to be developed and it allows us to identify the stronger connections between areas, in this case, the areas in the middle of the NYC are the ones strongly connected. Since our goal is, not only to identify the areas with higher flow of people, but also when the flow is higher, this map will be repeated for each hour of the day. Figure 14 will present 24 maps, each refering to a determined hour of the day, this will allow us to identify the hour with higher number of trips.

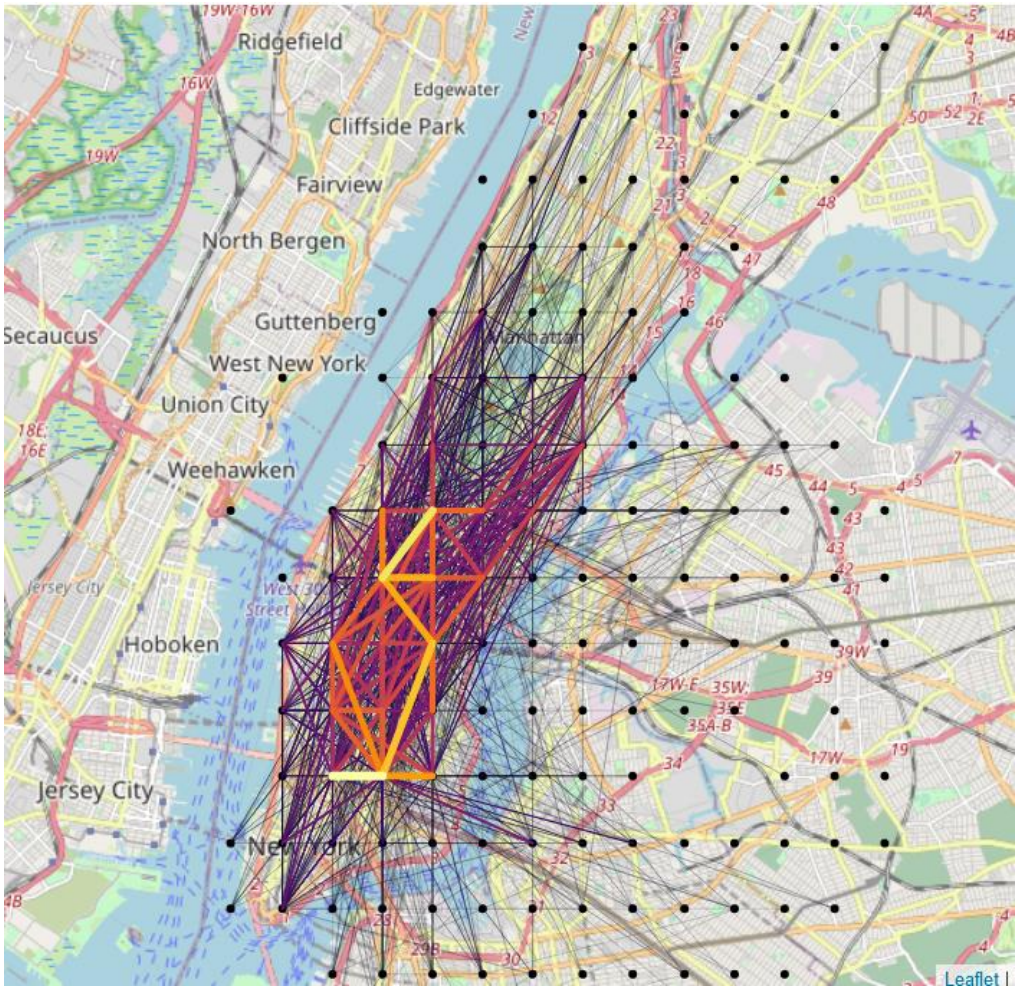


Figure 13. Network representation

Figure 14 presents 24 maps, one for each hour. It is important to point out that the color scale is the same for all the maps, allowing comparison between them. Starting at midnight, it is possible to see that the flow of passengers reduces until 5 am, when it reaches its lowest level. Between 6 and 7 am, the flow of passengers has its biggest increase, due to the beginning of the working day, these two hours correspond to the morning rush hour. As figure 6 (a) shows, between 7 am and 4 pm, the number of trips is the almost same, but we see that the maps of 6 and 7 am present connections with a very high flow of passengers. After these two hours, the flow of people starts to be more uniformly distributed through the area under analysis. From 4 pm to 7 pm the flow of people starts to rise again, reaching its prime at 6 and 7 pm. These two hours correspond to the afternoon rush hour, caused by the end of the working day. After these two hours, the flow of passenger starts to slowly decrease until midnight.

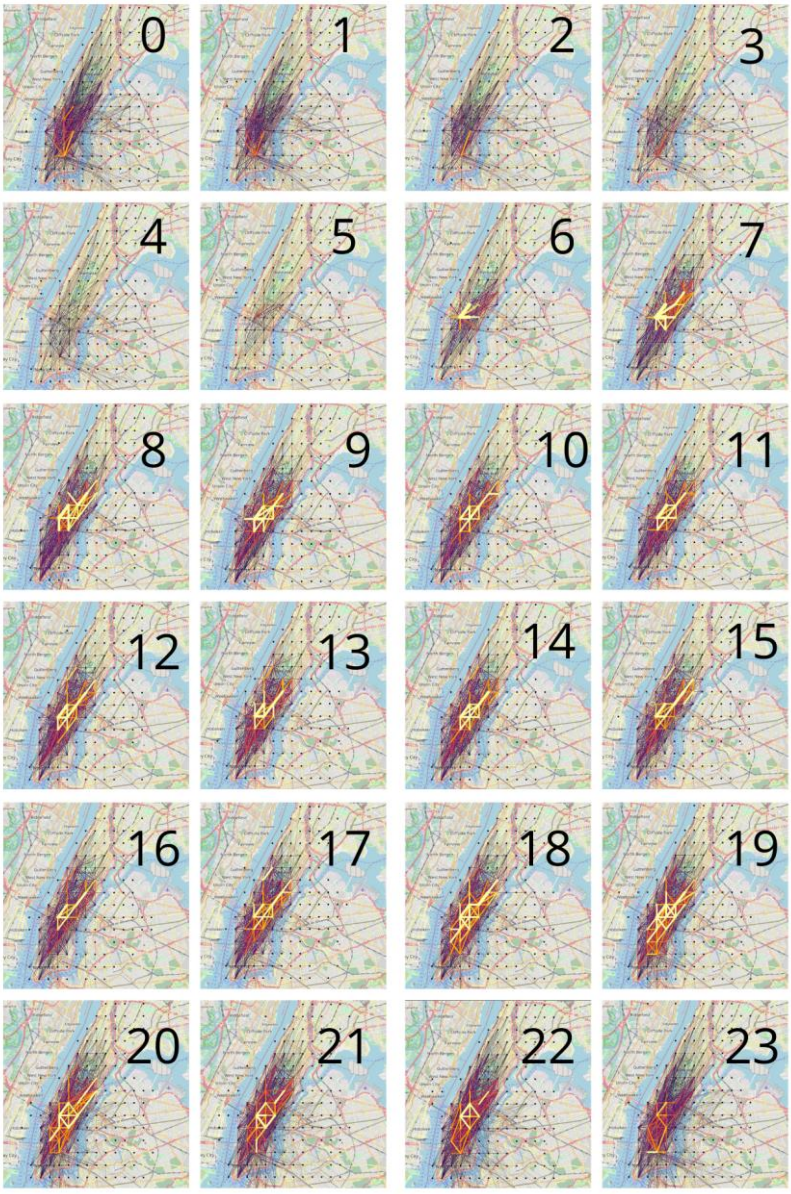


Figure 14. Hourly evolution of the flow of passengers in NYC

4. Conclusion

This project was able to answer the proposed questions. First, we were able to preprocess the dataset to properly apply netwrok analysis technics. Secondly, we were able to identify the locations that present higher flow of people. This area corresponds to to the center part of NYC, as the map shows. Also, we identified the hours when there is higher flow of passengers in NYC. It was possible to understand that the hours with higher flow of passengers correspond to the rush hours, caused by the movimentation of people from their house to their work and vice-versa. We applied netwrok analysis metrics to extract information about our network, It presents a density of 0.196. The network is non-assortative, meaning that the nodes with a degree tend to connect with nodes of a different degree, although this is true, the strength of this pattern is not very intense. The average cluster coeficient presented a high value (0.645), meaning that the nodes tend to cluster with each other. Also, the average closeness level of the nodes is considerable (0.414).