ICD PROYECT

*Note: Sub-titles are not captured in Xplore and should not be used

1$^{\text{nd}}$ Enmanuel Magallanes Pinargote
Manta, Ecuador
fmagalla@espol.edu.ec

2$^{\text{st}}$ Josue Cobos Salvador
Guayaquil, Ecuador
jcobos@espol.edu.ec

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

In this document we will answer business question about Shippify business model. The questions posed here will help identify elusive shipping cost using exploratory data analysis. Our final goal will to get a model as accurate as possible that is able to predict the arrival time of an delivery or if it will have a delay in shipping or not

II. Datasets

For this research we will define two datasets, the main dataset, which contains information about tasks of Shippify and the traffic dataset, which contains information about intensity of traffic on Santiago de Chile between the dates previously delimited in the main dataset.

A. Describing the main dataset

This dataset contains information about the tasks of Santiago of Chile. The main dataset comes from Shippify's backup databases, which have task records up to December 2020. For this research, the analysis was limited to the city of Santiago of Chile between July and August 2020, which translates into a total of 251823 deliveries. Each file is a task with its respective information. In the Table I could see more details about type and description of each columns on this dataset.

B. Size of deliveries

The dataset is mostly made up of Medium weighted tasks, with 65.8% of deliveries, and X Small weighted tasks, with 29.1% of deliveries. The Small, Large and X Long categories are the ones that appear the least in the dataset.

In the Figure II-F The task distance is very similar between the X Small, Medium and X Large weighted tasks. It has a median close to 20 and its distribution follows a similar shape, which is very slightly skewed to the right. On the other hand, the Small-weighted tasks have the smallest distance, with a median close to 10 and their distribution is strongly skewed to the right. Long-weighted tasks have a median close to 15 and are slightly skewed to the right.

C. Delivery type

Most of the tasks are Slot type tasks, 88.7%, which are tasks that have a collection date for the package(s), with a maximum time window of one hour from the specified time. On the other hand, Flex type packages, 11.3%, are those tasks that have a collection interval of more than one hour, with a maximum of 12 hours.

From the Figure **??** we can determine that flex type deliveries are the ones that present a higher correlation between the cost and distance variables, mainly between deliveries with size X Small and Small, are the ones that present a weak positive correlation between cost and distance, 0.61 and 0.67 respectively.

D. Tasks around

The feature taskAround means the number of tasks that a delivery had in a time window of 3 hours with 3 kilometers around, so the average is 172 tasks around, half of the records have had up to 124 tasks around and 3/4 of the deliveries had up to 269 tasks around.

From the figure II-F we can see how most of the deliveries have had between 50 and 250 tasks around them, although we can also notice a whisker at the end of the box plot which means that there are several scattered data but they do not have great relevance in this set.

E. Cost

The feature cost means the value in CLP that a driver collected from a delivery. The average is 844(CLP), half of the deliveries cost up to 1500(CLP) and 3/4 of the deliveries cost up to 2000(CLP).

In the figure II-F we can see how most of the deliveries accumulate to the left of the density plot with values ranging between 0 and 5000(CLP). To a lesser extent we find values that exceed 35000(CLP) which is equivalent to 45(USD).

F. Delta

Delta means the delivery time minus the planned delivery time in units of hours. For this feature the average is -3.70 hours, on the other hand half of the deliveries were made up to 1.6 hours in advance and 3/4 of the deliveries are delivered

Identify applicable funding agency here. If none, delete this.

TABLE I
COLUMNS OF MAIN DATASET

Column	Type	Description
id	string	Delivery ID
route_id	string	Route ID, only if the delivery is part of a route, if delivery is single delivery route_id is empty
creation_date	DateTime64	Date when the delivery was created
delivery_type	DateTime64	If categorical column that specify type of delivery
city	Int8	A ID that represents the city of delivery
cost	Float64	The price of the route when was created
total_size	Int16	Represent the size of delivery
distance	FLoat64	Km from pickup and drop off of delivery
company_id	Int8	ID of company own of delivery
pickup_dt	DateTime64	Date scheduled for pickup
pickup_effective	DateTime64	Date when the shipper arrival to pickup location
pickuplat	Float64	Latitude of pickup location
pickuplong	Float64	Longitude of pickup location
pickup_location	list of objects	Metadata of pickup location like name of street or place
delivery_dt	DateTime64	Date scheduled for drop off
delivery_effective	DateTime64	Date when the shipper arrival to drop off location and delivery to client
deliverylat	Float64	Latitude of drop off location
deliverylong	Float64	Longitude of drop off location
delivery_location	list of objects	Metadata of drop off location like name of street or place
items	list of objects	List with metadata of items of the task

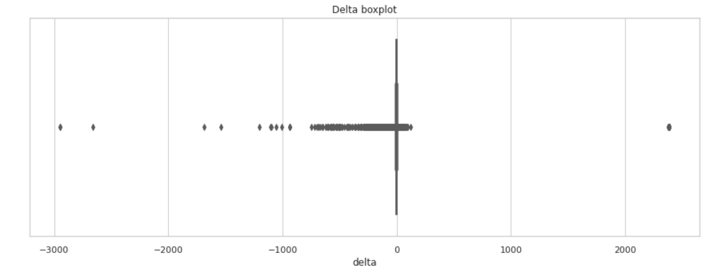


Fig. 1. Cost density plot

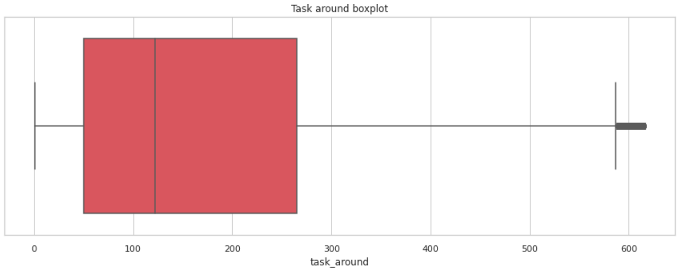


Fig. 3. Task around box plot

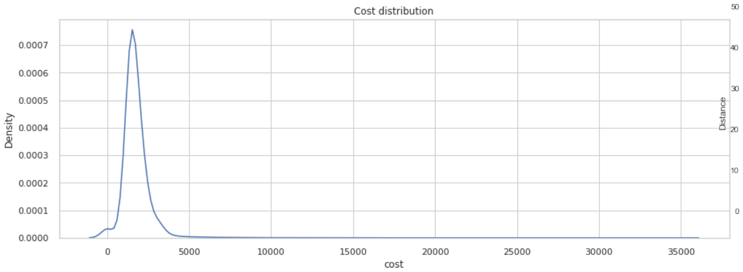


Fig. 2. Cost density plot

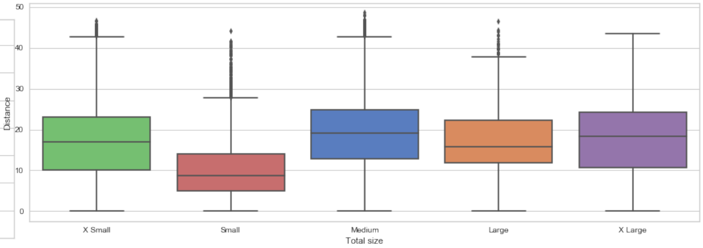


Fig. 4. Mean time difference

with a maximum delay of 25 minutes. This allows us to conclude that Shippify does a good job in meeting the planned delivery time.

In the following plot II-F we can see a rather scattered box plot, this is due to the fact that, without being a majority, there is a large amount of data outside the quartiles.

G. Describing the traffic dataset

This dataset contains information about the traffic intensity of Santiago de Chile. Each row has a identifier and information about shape, speed limit, streetName and traffic intensity of street on Santiago de Chile. In the table II could see more details about this dataset.

III. Data exploring

In this section we will describe some analyses that were done on the datasets to find out what factors or conditions affect the timeliness of a delivery.

A. Analysis of delivery sizes

In the main dataset we find fields to determine the total weight of a delivery. Therefore, an analysis was started to find out if this variable affects the timeliness of the delivery. The weights of the deliveries are divided into 5 categories:

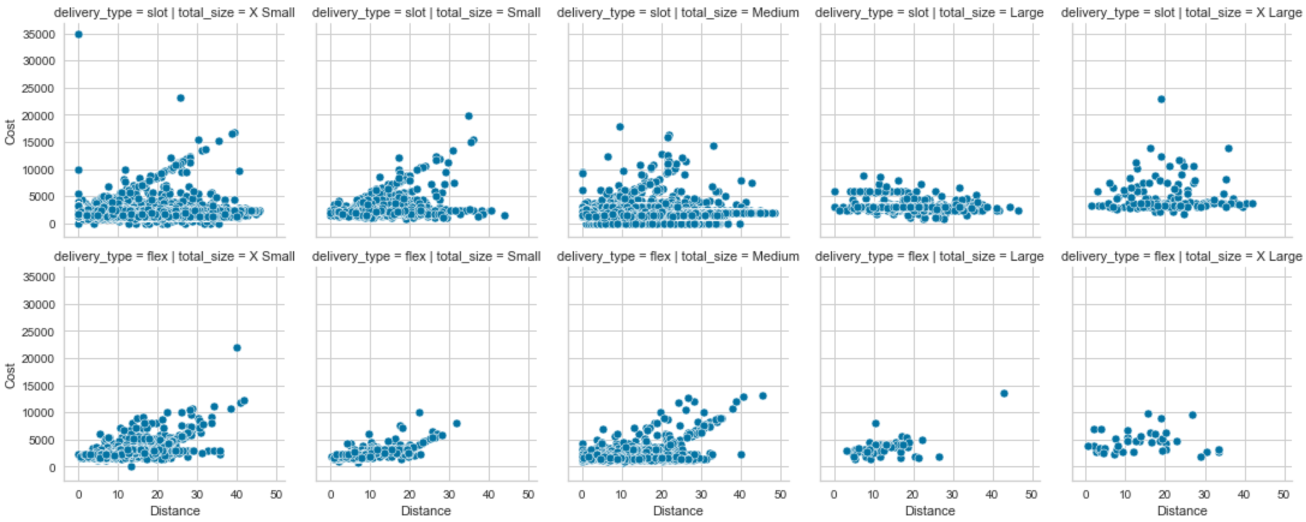


Fig. 5. Mean time difference

TABLE II
Columns of trafficT dataset

Column	Type	Description
segmentId	string	ID of street
speedLimit	Int16	Speed limit on street
streetNmae	string	Name of the street
distance	Float64	Distnace on kilometers of the street
shape	list of objects	Is a list of points (latitude and longitude) to represent a polygon on map of the street
segmentProbeCounts	list of objects	Is a list of object when each one has the field timeset, that represent the time window, and the field trafficIntenisty corresponding to that time window

X Small ($0.000\,216\,\mathrm{m}^3$), Small ($0.02\,\mathrm{m}^3$), Medium ($0.2\,\mathrm{m}^3$), Large ($0.6\,\mathrm{m}^3$), X Large ($1.2\,\mathrm{m}^3$).

For this analysis, the tasks were grouped by size and then categorized by timeliness of delivery, late or not late. We can see in the Figure III-A that there are more deliveries with size X Small and Medium than the others. However, what is important here is to analyze the percentage of tasks delivered on time and those that are not. There are no significant differences between Small and Small X-size tasks.

In contrast, Medium, Large and X Large tasks have a significant difference in the percentage of on-time deliveries. This indicates that the size of deliverables affects on-time delivery when their size is equal to or larger than Medium. The differences on percentages is showed in Figure III-A.

B. Analysis neighboring deliveries

To begin, we define neighboring deliveries as all those deliveries that are less than or equal to X kilometers away from the same and their delivery date is within T hours of the same.

To carry out this analysis we first have to determine what would be the time window (hours) and radius (kilometers) around a task to be used. For this we analyzed the distance (Figure III-B) and average time difference (Figure III-B) of a route in the delivery path, from the first package to the last one. The following results were obtained:

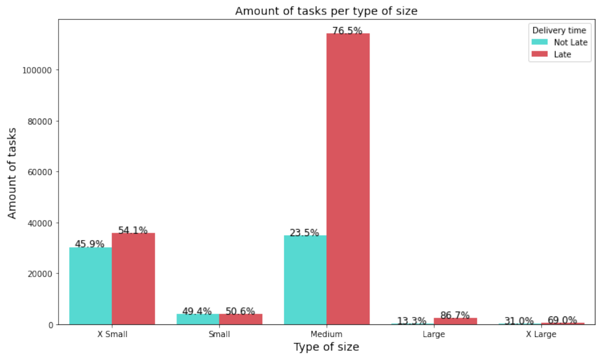


Fig. 6. Percentage of tasks overdue and not overdue by size type

- Mean time difference: 3.12 hours
- Mean distance: 3.20 Km.

After having calculated the number of neighboring deliveries for each one. They were grouped into classes of 100 deliveries each. Obtaining the Figure III-B, in that figure we can see that as the number of neighboring deliveries increases, the percentage of late deliveries also increases. This tells us that the higher the concentration or density of deliveries around you, the more likely it is that you will be late in reaching the final customer.

In the Figure III-B can see more clearly how, as the number

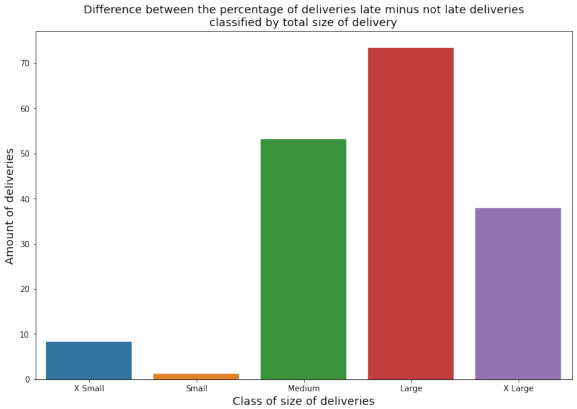


Fig. 7. Percentage differences of tasks overdue and not overdue by size type

of neighboring deliveries increases, the difference between the percentage of overdue tasks minus those not overdue increases.

In the figure III-B we can see how there is a high concentration of deliveries with a high number of neighboring deliveries at one point on the map. On the other hand, the further away from that point the number of neighboring deliveries decreases and deliveries larger than Medium are more frequent.

C. Distance to economic center and Delivery delay analysis

In the main data set each event like pick up and delivery has a date time register. We have a estimated and effective date time for each event. To calculate the delay time we subtract effective to estimated and we save the result as seconds. This variable was not categorized because we are interested in knowing if there is a correlation between the distance to the economic center of each delivery and the delivery delay. To achieve this goal we assume that Santiago (as comuna) is the economic center based on it economy and PIB. To calculate distance we use Haversine method witch equation goes like:

$$d = 2r\arcsin(\sqrt{\sin^2(\frac{x_2-x_1}{2}) + \cos(x_1)\cos(x_2)\sin^2(\frac{y_2-y_1}{2})}))$$

Where

$$d$$

is the distance between 2 points,

$$r$$

is the earth radius and

$$x_2, x_1, y_2, y_1$$

are components of

$$X, Y$$

points.One we Once we obtained the 2 metrics of our interest we proceeded to eliminate the few outliers we found in the metric of delivery distances to the economic center using z scores, and filtered out those entries where z score is less than 3. We can check this in the Figure III-B. Finally, we show our metrics in a scatter plot where at first glance we notice that there are no correlations of any kind. Being strict with the literature we proceeded to look for a correlation with the following results: III-B. Therefore due to the non-existent correlation between the delay of shipments (delay_delivery) and distance to economic center (dst_to_ec_center), we can conclude that no matter how far your order is, it does not necessarily mean that it has a longer delay time.

D. Traffic and delivery delay analysis

To develop this analysis, we obtained a third-party data set that quantified traffic in units based on coordinates and a time zone. Since each delivery has its location and date-time we combined the data, once the data was combined we proceeded to show in a dot plot the delivery delay time against the traffic unit that was detected in the delivery area within the established time. The following III-B. plot intuitively shows a low relationship between these two metrics. Once we performed a correlation obtaining the following values shown in III-B. Therefore we can conclude that the traffic that usually has the area where deliveries are made does not have a direct impact on the delay time.

IV. Finding common characteristics on tasks with delay through clustering

For apply a good clustering first we need to determinate the distribution of feature that will be used. Depending of the distribution type it will require a transformation or normalization.

- Power law distribution: apply log transform scale.
- A normal (Gaussian) distribution: apply normalization (z-score, min-max, divide by max).
- Any different distribution (like bi-modal or Poisson): apply quantiles (bins).

The variables to be used are distance cost delta, latitude on drop off, longitude on drop off, neighboring deliveries, delivery type and Total size. The delivery type and total size are categorical, so max-min normalization will be applied on these. In the other hand, the rest of variables doesn't have a clear distribution, so we will use quantiles and max-min normalization.

The Elbow method was used to determinate that the optimal numbers of clusters is four. Resulting in the next clusters characteristics:

- Cluster 1: this cluster contains deliveries that arrive on time. The distances present in this cluster are the shortest of all clusters. The payment received by the shipper is close to but below the global average. Finally, the concentration of neighboring deliveries is the lowest along with cluster 3.
- Cluster 2: this cluster presents tasks that register a high delay at the time of their deliveries. The distances are the longest of the clusters. However, the payment received by the shipper is the lowest of all clusters. Finally, they have a high concentration of neighboring deliveries.
- Cluster 3: this cluster has slightly backlogged tasks. The distances are the second highest of all clusters. In

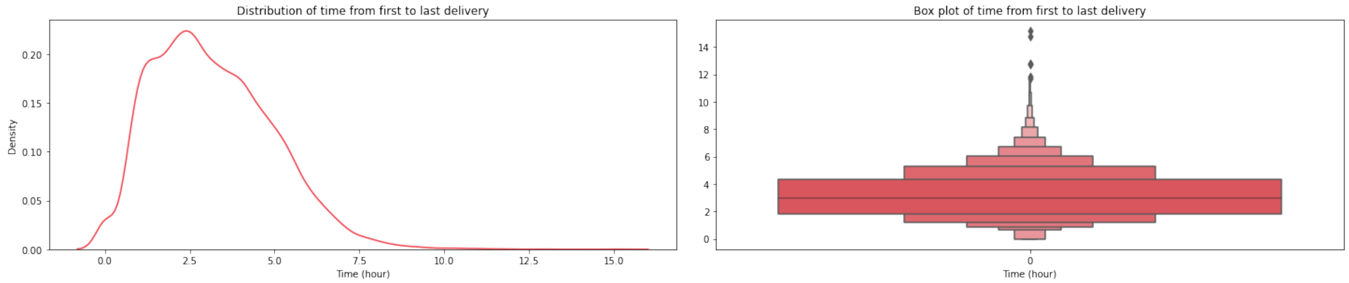


Fig. 8. Mean time difference

this occasion shippers are paid very close to the global average. Finally, the number of neighboring deliveries is the lowest of all clusters.

- Cluster 4: this cluster contains deliveries that arrive on time. The delivery distances are the second shortest, only behind Cluster 0, and the payment received by the shipper is also low. Finally, the concentration of neighboring deliveries is the lowest along with cluster 0 deliveries.

In summary, clusters 1 and 2 are the ones with late deliveries. Conditions such as a very low payment to shippers and a longer delivery distance and high number of neighbors are conditions that cause deliveries to suffer a high delay at the time of delivery.

V. Temporal analysis

A. *Cost vs date*

We limit the response to the end of August because since September there is not enough data to be analyzed. According V-A in our analysis we found seasonality once again with respect to daily cost of deliveries, it always starts high and has to drop at the end of the month, this cycle repeats itself every week.

B. *Delivery type amount vs date*

For this conclusion we must take into account that deliveries are divided into 2 types ”flex” and ”slot”, for the first type we see that from June to September there is seasonality in an interval of 1 week. For the same type ’slot’, according to V-B , we see that from September to December 2020 we do not have a downward trend of daily deliveries. For the slot type we notice that from mid-June there is a clear downward trend until September. Probably this downward trend is due to ”business rules” that are unknown to us.

VI. Social network analysis

For our analysis we construct a directed graph where the nodes are ’commune’ and the arcs translate to ”has made a shipment to”. Our graph has 32 nodes, i.e. each and every commune in the province of Santiago de Chile. It also has 716 nodes, which represents a priori a highly connected network, taking into account that the maximum possible connections would be 1024 due to the fact that

$$possibleConnections = |nodes| \cdot |nodes|$$

A. *Cohesion*

From an initial plot VI-A we can observe at a glance a highly cohesive graph in terms of path length. To corroborate these intuitions we calculated its average clustering coefficient, which gave a value of 0.86, a fairly high value. Finally we compared it with its equivalent random graph and once again calculated the average clustering coefficient, giving a value of 0.68, with which we can conclude that the graph is highly cohesive. This high cohesiveness translates to business information to the fact that the vast majority (if not all) of the communes are related to all the others.

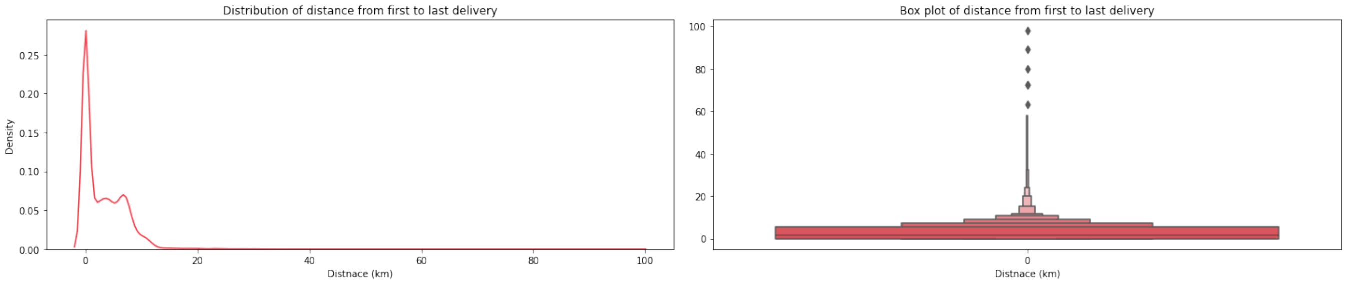
B. *Most important comunas in terms of sending deliveries*

According to VI-B The ’comuna’ Pudahuel has sending deliveries to all of 32 ’comunas’ of Santiago (Including it self). ”Santiago” and ”Recoleta” has sending deliveries to 31 ”comunas” (Including it self).

C. *Most important comunas in terms of receipt of shipments*

According to VI-C The ’comuna’ Las Condes has received deliveries of 30 ’comunas’ of Santiago (Including it self). ”Providencia” has received deliveries of 29 ’comunas’ and ”ÑUÑOA” 28 (Including it self).

References


Distribution of distance from first to last delivery

Box plot of distance from first to last delivery

Fig. 9. Mean distance

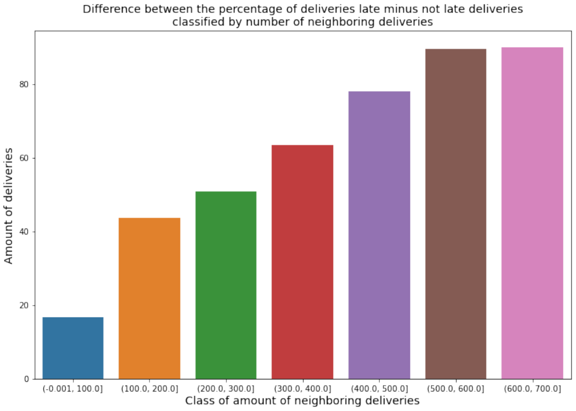


Fig. 10. Amount of deliveries per neighboring deliveries classes

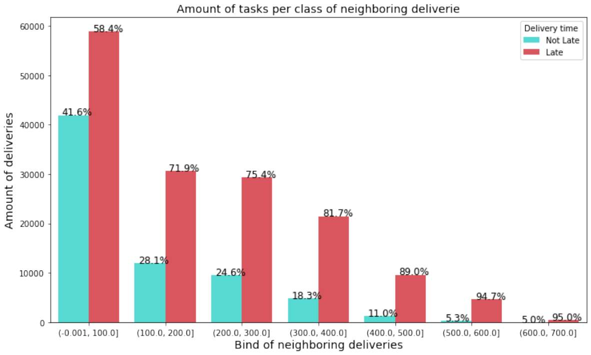


Fig. 11. Plot of percentage differences

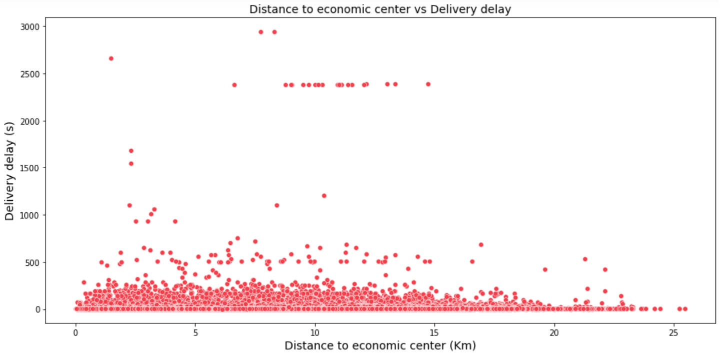
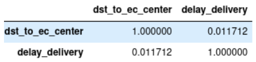


Fig. 12. Distance to economic center vs Delivery delay



	dst_to_ec_center	delay_delivery
dst_to_ec_center	1.000000	0.011712
delay_delivery	0.011712	1.000000

Fig. 13. Distance to economic center and Delivery delay correlation

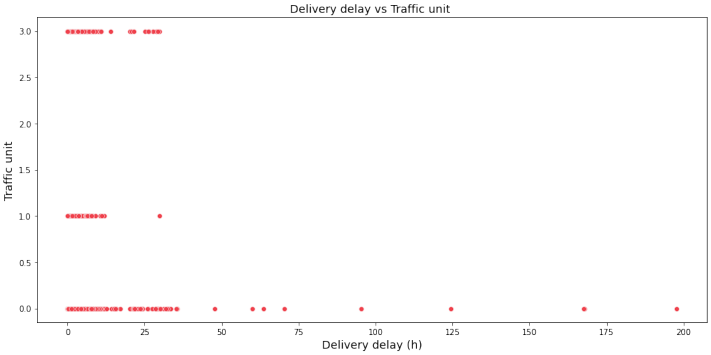
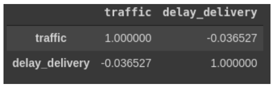


Fig. 14. Delivery delay vs Traffic unit plot



	traffic	delay_delivery
traffic	1.000000	-0.036527
delay_delivery	-0.036527	1.000000

Fig. 15. Delivery delay and Traffic unit correlation

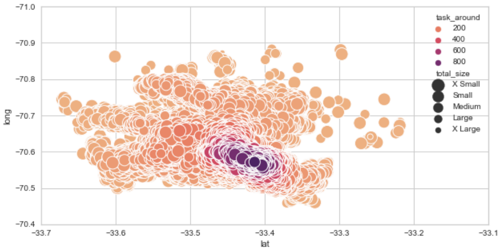


Fig. 16. Mean time difference

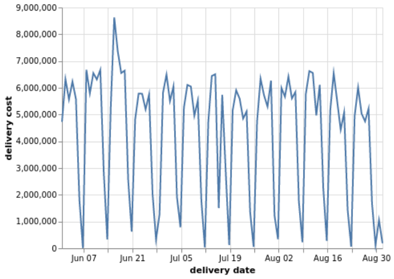


Fig. 17. Delivery cost vs date

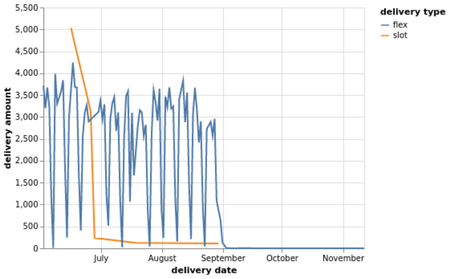


Fig. 18. Delivery type amount vs date

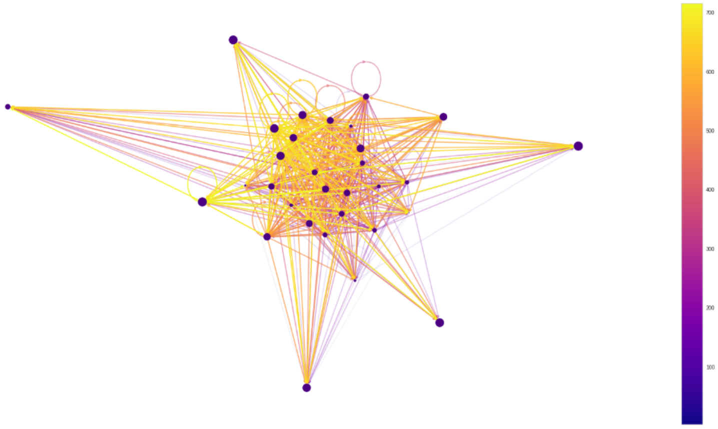


Fig. 19. Comunas graph plot

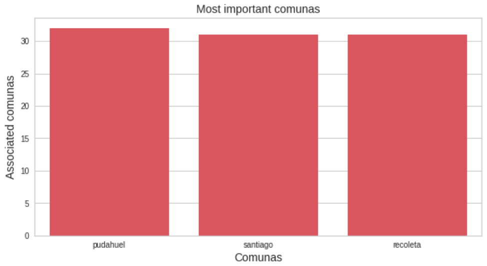


Fig. 20. Comunas with the highest number of shipments

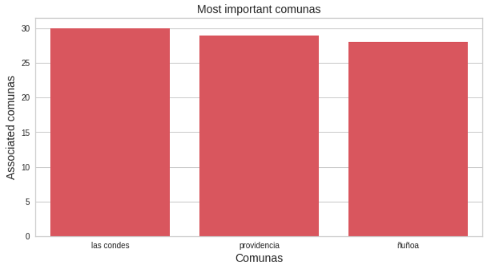


Fig. 21. Comunas that receive the most shipments