

*Springboard Introduction to Data Science*

---

# AN ANALYSIS OF THE PERFORMANCE: GLS-ITALY, FIELD OFFICE OF BRESCIA



*By Simone Zanetti*

---

# PROBLEM DEFINITION

---

- ❖ EXPANSION OF THE DELIVERY SECTOR
- ❖ EACH DRIVER HAS TO PERFORM THE PROCESS OF DELIVERY OF GOODS AS WELL AS THE RECALL
- ❖ THE PROCESS OF DELIVERIES HAS TO FACE A SERIES OF ENDOGENOUS VARIABLES ( n. of packages, n.driver, etc.) , AND EXOGENOUS (traffic, availability of clients, etc.) .

---

# PROJECT AIM

---

- ❖ NECESSITY TO ASSESS THE PERFORMANCE OF DELIVERY FOR EACH DRIVER
- ❖ UNDERSTAND THE FACTORS THAT MORE INFLUENCE THE PROCESS OF DELIVERY.
- ❖ SET THE BASIS AND THE NEEDS FOR A FUTURE ANALYSIS WHOSE AIM IS TO DEFINE “OPTIMISED WAYS OF ORGANISATION OF THE DELIVERY PROCESS”.

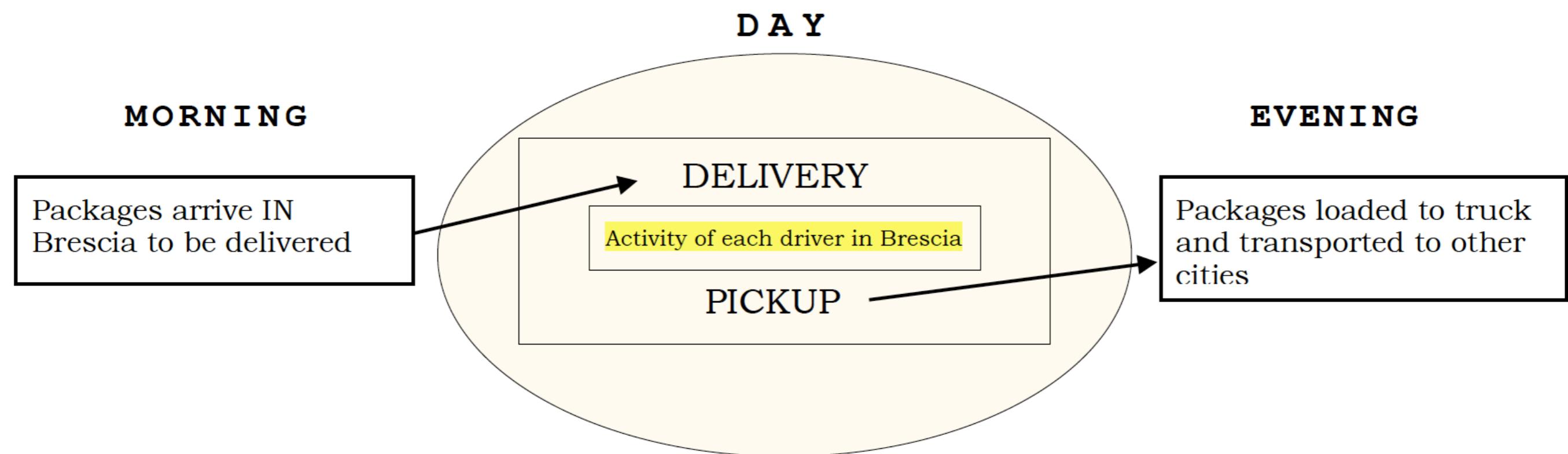
---

# DATA EXTRACTION

---

- ❖ DATA KINDLY PROVIDED BY THE FIELD OFFICE OF GLS-BRESCIA (ITALY)
- ❖ ALL DATA REGARD THE PROCESS OF DELIVERY IN THE MONTH OF MARCH 2017
- ❖ IMPORTANT TO UNDERSTAND THE ACTIVITY OF THE COMPANY for each driver (see following page)

# THE ACTIVITY OF EACH DRIVER



**Figure I:** Daily activity of the company with a focus on the

---

# DATA EXPLORATION: dataset I

---

- ❖ **DATASET I:** Record of all deliveries performed in the month by each driver. ( each row is a delivery )
- ❖ **FUNDAMENTAL VARIABLES**, such as the *day of delivery*, the *hours of delivery*, the *address and district of delivery*, the *kg for each package delivered* ( all referred to the driver )
- ❖ **VARIABLES NOT TAKEN IN CONSIDERATION**, such as the *addressee of the delivery*, the *day of departure* to the field office of Brescia, the *day of arrival* at the field office of Brescia, etc.

---

# DATA EXPLORATION: dataset II

---

- ❖ **DATASET II:** Daily summary of the services performed by each driver, including recall of goods. ( each row is the summary of a day of one driver )
- ❖ **FUNDAMENTAL VARIABLES**, such as the *tot of deliveries per day*, the *tot of picked up goods*, the *tot of pack loaded on the truck*, the *tot of pack arrived each day* for the specific driver.
- ❖ **VARIABLES NOT TAKEN IN CONSIDERATION**, such as the *name of the driver*, the *tot kg of picked up goods* and the *tot kg of delivered goods*, for each driver.

---

# DATA LIMITATIONS

---

- ❖ **LACK OF PRECISE DATA IN RELATIONSHIP WITH THE PICKED UP GOODS.** ( vs precise data for the delivered packages )
- ❖ **SEE PAG.5 TO OBSERVE WHY THIS IS A LIMIT**  
( Note: the process of recall of goods is about 10% of the daily activity of a driver)
- ❖ **LACK OF ENOUGH DATA** ( Necessity to have observations for at least one year )

---

# ANALYSIS LIMITATIONS

---

- ❖ **TURN ADDRESSES INTO COORDINATES** ( over 100,000 addresses and only 2,500 queries per day with google geocode)
- ❖ **ABSENCE OF ADDRESSES AND PICKEDUP\_TIME FOR THE RECALLED GOODS** ( as a consequence, no possibility to observe the itinerary in a truthful way )
- ❖ **PERSONAL COMPETENCES:** Necessity to remind myself this is my first project and, as such, necessity to simplify the analysis on such a raw dataset, setting the field for a future and more detailed analysis.

---

# DATA WRANGLING PHASE\*

---

- ❖ Loaded **Tidyverse package** ( in particular, utilisation of Dplyr and Tidyr, and in the following section Ggplot2 )
- ❖ **DEAL WITH MISSING VALUES** ( replaced them when needed, with *mutate()* and *replace()*, erased them when not needed, with *filter()* )
- ❖ **FIX THE DIFFERENCES OF SPELLING** ( with *sub()* and *gsub()* )
- ❖ **TURN LOCATIONS INTO COORDINATES TO ASSOCIATE THEM TO THEIR POSTAL CODE** ( with *geocode()* from ggmap package, and *left\_join()* )

\* For a detailed overview of the process of Data Wrangling check the document *Data Wrangling Project*)

---

# DATA WRANGLING PHASE, 2\*

---

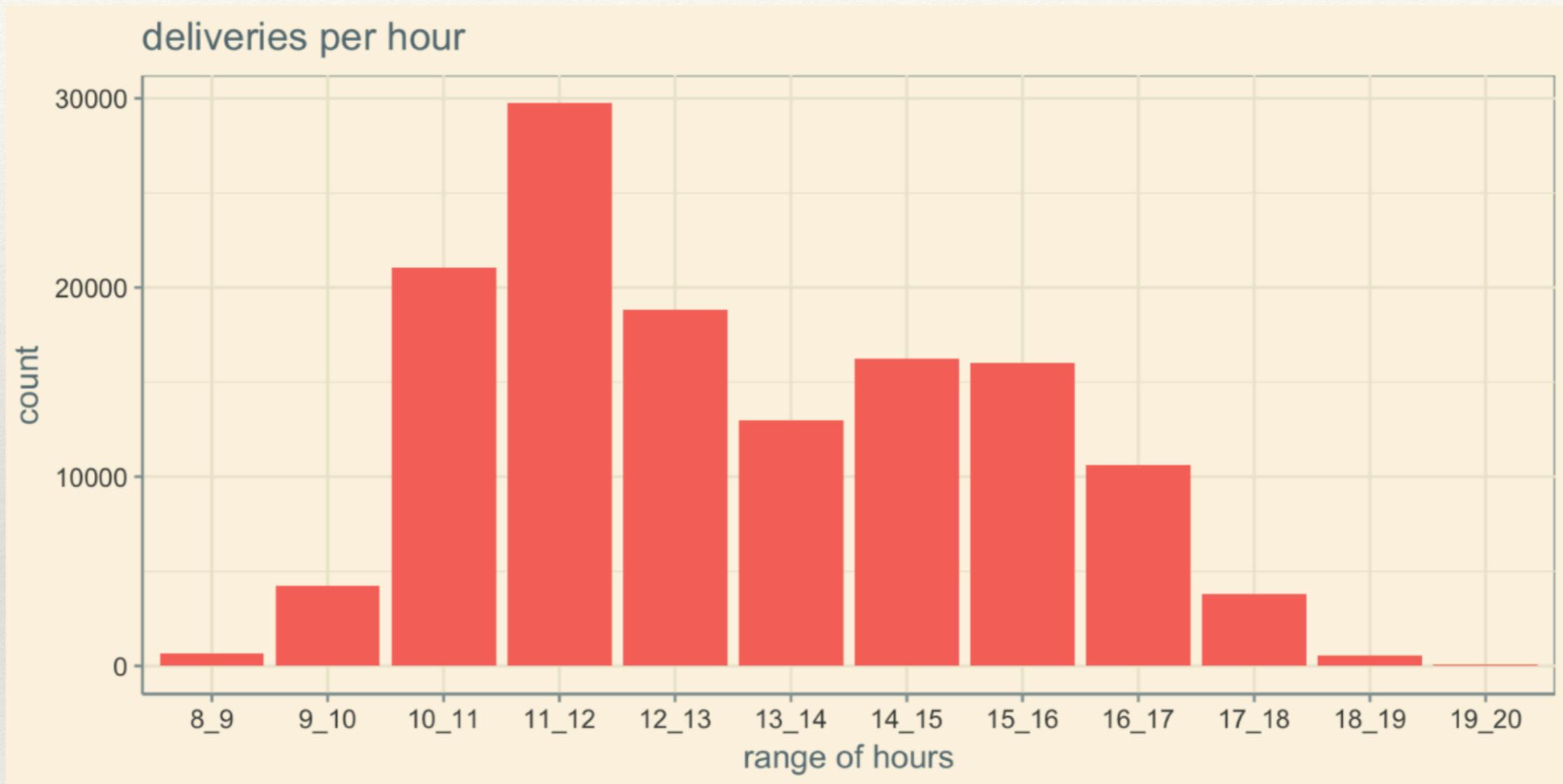
## ❖ WORK WITH TIME AND DATES

- Obtain the variable `day_delivery` ( without unnecessary month and year, with `as.Date()` and `separate()` )
- Obtain the weekday of delivery for each date ( with `format(., "%a")` )
- Turn Time from a string to `as.PosixCT` format

## ❖ CREATE DATASET II BY BINDING EACH DATASET\* TOGETHER (\*the daily summary for all the drivers)

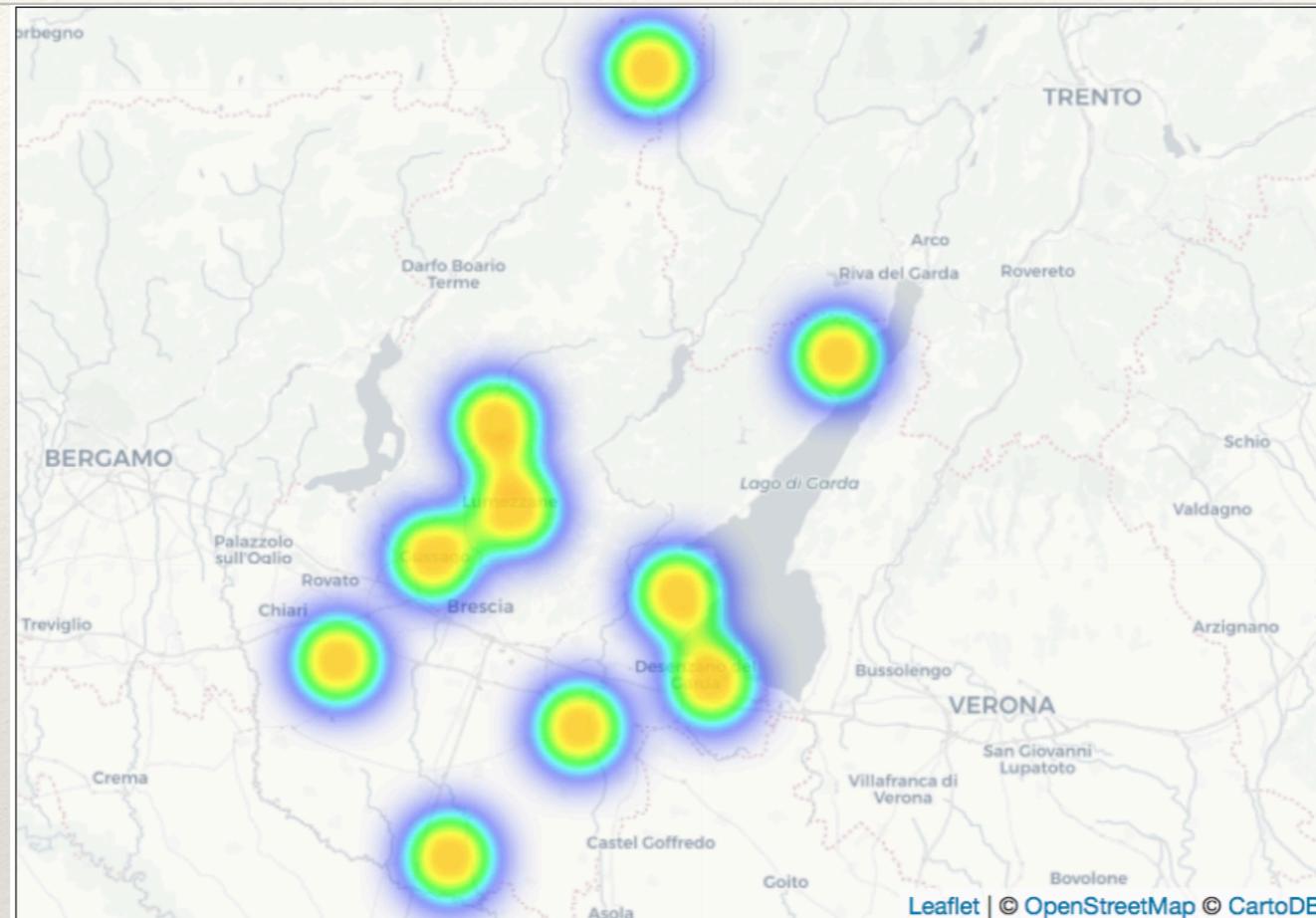
\* For a detailed overview of the process of Data Wrangling check the document *Data Wrangling Project*

# DATA VISUALISATION 1



The plot shows a **peak of deliveries on the range between 11 am and 12 pm**. After that pick, the curve regularly decreases until the end of the day with the exception of the range **between 13 and 14**, where the number of deliveries first decreases before to increase again on the following hour range. From this point of view, it is possible to observe that the **morning between 10 and 13 the majority of the deliveres are done**.

# DATA VISUALISATION 2



**Figure IV:** heat map containing the top five district of deliveries, after the city centre of Brescia

Due to the fact that for the **centre of Brescia** it was not possible to obtain one postal code, it has been identified with the observation “25121/25136”. This represents an important outlier of this distribution. For this observation, the number of monthly deliveries is **31846**, with the second postal code with maximum number of **deliveries which is 9111**. This postal code is **25020** which is an area south of Brescia which includes several important district of the city. **Third postal code is 25080** which is an area **close to the Lake Garda**, where commercial activities and tourism are really frequent. This postal code is right adjacent to the fourth most frequently delivered postal code, which is **25010** and it represents the **northern part of the Garda Lake**.

# DATA VISUALISATION 3

- ❖ CREATION OF THE DEPENDENT VARIABLE IN ORDER TO ANALYSE THE PERFORMANCE FOR EACH DRIVER:

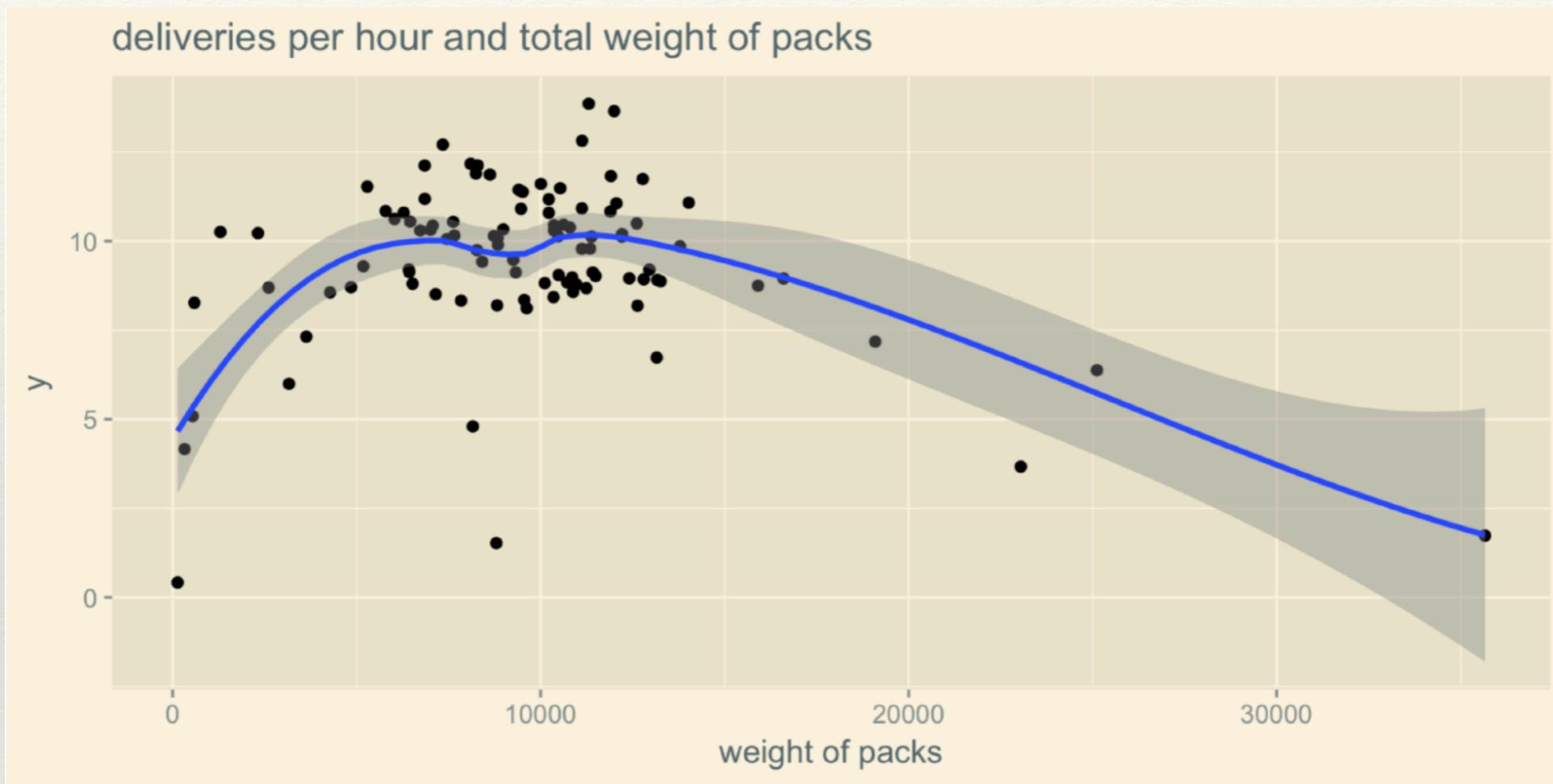
$$y = \frac{\text{tot. deliveries}}{\text{worked . hours}}$$

for each driver

- ❖ CREATION OF AGGREGATE DATA TO BE COMPARED TO THE DEPENDENT VARIABLE:

- total packs loaded on the van
- total packs arrived for each driver / zone
- total packs not delivered at the end of the days
- total packs picked up
- total weight of the packs picked up

# DATA VISUALISATION 4



The **weight of the packages delivered** seems to have a **negative influence** on the number of packages delivered per hour

# DATA VISUALISATION 5



The graph shows a **positive relationship** between the **y analysed** and the **total amount of services done**

---

# PREDICTIVE MODEL

---

- ❖ OBTAIN A MODEL THAT COULD PREDICT THE RATIO BETWEEN DELIVERIES AND HOUR FOR EACH DRIVER, BASED ON THE PREDICTOR OBTAINED
- ❖ MODELS INVOLVED:
  - LINEAR REGRESSION
    1. Traditional
    2. Lasso
    3. Ridge
  - REGRESSION TREE

# PREDICTIVE MODEL: CONCLUSIONS

- ❖ METHOD TO TEST MODELS:

$$\text{mean}((\text{predicted.y} - \text{effective.y})^2)$$

- ❖ COMPARISON BETWEEN MODELS

```
best_model_fit <- c(mse.lasso,mse.ridge,mse.tree)
names(best_model_fit) <- c("mse.lasso","mse.ridge","mse.tree")
best_model_fit
```

```
## mse.lasso mse.ridge mse.tree
## 0.5102480 0.4931289 0.9085693
```

The comparison between the models suggests the **Penalised linear model Ridge** to be the best to predict the number of deliveries per hour

---

# CONCLUSIONS

---

- ❖ PROJECT AS A STARTING POINT FOR FUTURE ANALYSIS
- ❖ IN ORDER TO DO SO, THERE ARE SOME NECESSITIES/SUGGESTS FOR THE COMPANY TO FOLLOW (See conclusion 2)

# CONCLUSIONS 2

---

- ❖ **MORE OBJECTIVE AND DETAILED RECORDS OF DATA** ( more automatisation in the data recorded, more objectivity when the data are entered in the database)
- ❖ **MORE DETAILS FOR THE ACTIVITY OF RECALL OF GOODS** ( allowing to perform precise analysis of the itinerary followed by each driver )
- ❖ **NECESSITY OF THE COMPANY TO INCREASE THE COLLABORATION WITH THEIR CLIENTS** ( who are necessary for the success of the process delivery / recall )
  - **Educating them** about how a process of delivery works
  - **Providing them more choices** about the time, place they prefer to have the good delivered, picked up.