# The Influence of Train Trips on BlaBlaCar Rides in France

**AL MASSRI Mostafa** **EL OSMAN Mohamad** **NAKAD Elie**

mostafa.almassri@telecom-paris.com, mohamad.elosman@telecom-paris.fr, elie.nakad@telecom-paris.fr

**LAITENBERGER Ulrich** **PAPE Louis**

## Abstract

*In recent years, the transportation landscape in France has undergone significant changes with the proliferation of BlaBlaCar rides, which are informal transportation services provided by individuals using their own vehicles. The objective of this research paper is to examine how trains, as a formal mode of transportation, affect the demand for BlaBlaCar rides in France.*

## Introduction

Ridesharing services like BlaBlaCar have grown in popularity in recent years as a convenient and cost-effective alternative to traditional forms of transportation. However, the success of these services is often dependent on the availability and reliability of public transportation options, such as trains. The impact of trains on BlaBlaCar rides is an important topic to consider because it can affect the demand for ridesharing services, as well as the overall efficiency and convenience of transportation systems. Understanding how trains impact BlaBlaCar rides can help cities and transportation providers to better plan and optimize transportation networks, ultimately improving the travel experience for all.

The research question we are answering: ***" To what extent does the pricing structure of traditional train transportation affect the pricing strategies of BlaBlaCar drivers? "*** is important for understanding the relationship between public transportation options and the adoption of ridesharing services. This information can aid in transportation planning and optimization, as well as provide insight for ridesharing companies on how to better market their services. Additionally, it is interesting to examine how the convenience and efficiency of transportation networks, impacted by train services, can affect the demand for BlaBlaCar rides in urban areas.

Furthermore, we collected data on the availability and reliability of train services in urban areas as well as the demand for BlaBlaCar rides, using BlaBlaCar’ s API but we got blocked by the cites (Captcha, Blocking Messages, etc.) so we went to alternative methods that took a lot of time to achieve. We then analyzed the data using descriptive statistics and regression analysis to identify patterns and relationships. The data was also visualized using the python libraries to easily identify trends and patterns. Overall, our research aimed to investigate the impact of train services on the demand for BlaBlaCar rides in urban areas by using a combination of data collection, statistical analysis, and visualization techniques.

The results of the regression analysis indicate that there is a strong positive and statistically significant relationship between the train price per km and the BlaBlaCar price per km. This suggests that BlaBlaCar drivers may be taking into consideration the prices of train trips when setting the prices for their own trips. This inference is supported by the finding that the impact of train prices on BlaBlaCar prices is even stronger when the departure city is Paris. Thus, it can be suggested that BlaBlaCar drivers may be monitoring the prices of train trips and adjusting their own prices accordingly.

## Background

In this project, we followed the outline we put in our first assignment to collect data then proceed in a data visualization and conclude our findings with a regression

Firstly, we reached out to BlaBlaCar to request an API through email in order to collect data for our project. After receiving the API, we were able to successfully implement it into our code and begin making requests. Initially, we were only able to make 1000 requests per day, but by optimizing our collecting code, we managed collect 10000 rows using only 10 requests. This allowed us to gather a larger amount of data, which was critical for the success of our project.

BlaBlaCar is a long-distance carpooling platform that connects drivers traveling from one city to another with passengers looking for a ride. The company was founded in France in 2006 and has since expanded to operate in 22 countries around the world. The platform allows users to share the cost of fuel and tolls and helps reduce congestion and pollution on the roads.

The market for long-distance carpooling is relatively new, but it has grown rapidly in recent years. According to the company, BlaBlaCar now has over 60 million users worldwide and has completed over 100 million rides. In terms of revenue, the company reported €140 million in 2017 and €185 million in 2018.

The growth of the market can be attributed to a number of factors. One is the increasing awareness of environmental issues and the desire to reduce carbon emissions. Carpooling is seen as a more sustainable alternative to traditional forms of transportation such as flying or taking the train. Additionally, the high cost of fuel and tolls has made carpooling a more attractive option for cost-conscious travelers.

In recent years, BlaBlaCar has also expanded beyond Europe and has established a presence in Latin America, Asia, and North America. The company's goal is to become the global leader in long-distance carpooling and to continue to grow its user base and revenue.

The BlaBlaCar French site is the platform's main website for users in France. It allows users to search for and book rides between cities in France and other countries in Europe. The site is available in French and it allows users to search for rides by entering their desired destination and departure location, as well as the date of travel. Users can also filter their search results by the number of seats available, departure time, and price. The website also allows riders to view the driver's profile and reviews from previous passengers.

In addition, the French site also provides users with information on the company's policies and safety features, such as the ability to rate drivers and share feedback. Users can also find information on BlaBlaCar's various services such as their bus service and car-sharing service.

Users can also find on the site the company's blog and news, where they can learn about new features and updates to the service, as well as tips and advice for using the platform. There is also a section of the website dedicated to the company's commitment to sustainability and the environmental impact of carpooling.

## Data collection strategy

We mainly utilized an API from the BlaBlaCar platform by providing the coordinates, dates, and other information of the departure and arrival locations to gather all the details of the trips, including attributes such as departure and arrival times and addresses, trip distance and duration, vehicle make and model, and pricing. Additionally, we employed web scraping techniques to collect data on train trips that occurred at the same time and on the same days in different countries, extracting attributes such as departure and arrival times and addresses, trip duration, number of changes, standard and first-class prices, and train company names for each change. This data was used to analyze the relationship between the prices set by BlaBlaCar drivers and the prices of train trips for the same routes.

We have implemented two separate Jupyter files to collect data for each platform.

For the BlaBlaCar platform, we utilized the following packages: requests for sending API requests and retrieving JSON data, pandas for converting the data into a dataframe object and a CSV file, and math functions for calculating the distance between two latitudes and longitudes. We also used the os package to interact with our operating system file environment. However, we encountered obstacles such as the API only returning the first 100 trips and not providing information on the drivers. We also attempted to web scrape the driver's information but were blocked by captchas and were unable to solve them automatically.

For the train lines platform, we used pandas to convert the data into a dataframe object and a CSV file, os to interact with the operating system file environment, time to pause code execution, selenium to open a web driver and control its functionality, and bs4 to beautify the HTML source page. We faced issues with the basic requests library not loading the required data and were blocked by captchas when using selenium. However, we were able to solve the captchas manually and used a technique of opening new tabs to avoid solving them continually. We also encountered a problem where we had to repeatedly click a "later" button to display all the trips, and implemented a function to check if the button was clickable or not.

We began gathering data on January 17th and completed the process on January 20th. Our data collection included all trips between January 20th and January 28th for both platforms. The BlaBlaCar API code was relatively quick, taking only 2-3 minutes to collect, while collecting data from the train lines platform took approximately 3 days. We defined an observational unit as a BlaBlaCar trip within a specific region and time frame, between a departure city and an arrival city.

## Data description

We have collected 2 datasets, the first one was about BlaBlaCar rides where we have collected around 10 000 rides, the second one was about Trains trips where we managed to collect around 17 000 trips.

We initially divided our data to 5 different regions but the fourth one didn't have enough data so we removed it, we finally had:

| | | | | | | |
|---|---|---|---|---|---|---|
| **Region_0** | **Paris** | Nantes | Lyon | Lille | Marseille | Reims |
| **Region_1** | **Lyon** | Toulouse | Bordeaux | Nice | Marseille | Montpellier |
| **Region_2** | **Nantes** | Rennes | Poitiers | Brest | Paris | |
| **Region_4** | **Reims** | Nancy | Strasbourg | Metz | Paris | |

*Table 1: France Regions and their Main Cities*

The bold cities are the cities where visualization is focused.

Also, we divided the day (5 am first day till 4 am the following day) to 4 time slots (Morning, Midday, Afternoon, Night) so we can visualize and analyze the data for different users depending on the time slot they usually use.

This are some descriptive analyses on the data we will use for regression on the lowest price of available afterwards:

| | Price (€) | Duration (s) |
|---|---|---|
| count | 1738 | 1738 |
| mean | **21.849252** | 12923.475259 |
| std | **13.422982** | 6949.855025 |
| min | 29 | 2400 |
| max | 107 | 47400 |

*Table 2: Descriptive Statistics on Best Trains (min prices)*

| | Price (€) | Duration (s) |
|---|---|---|
| count | 2341 | 2341 |
| mean | **43.737783** | 18653.746262 |
| std | **30.034398** | 6949.855025 |
| min | 5 | 2400 |
| max | 237.8 | 165600 |

*Table 3: Descriptive Statistics on Best Cars (min prices)*

We can see that we have respectively 1738 and 2341 best cars trains and cars for different couple of cities we have in France. We see that in average the price of best cars is equal to half of that of the trains in France. But, the price of cars has a highest standard deviation than that of the trains.

## Analysis

For the visualization part of the project, we conducted mainly 2 approaches, the first one is about the availability of the car and train rides and the second one is about the average price comparison between both type of mobility. Then we will do a regression analysis to verify our assumptions.

Now, we will show graphs for every region independently and taking into consideration the different time slots.

I. We will start by the availability's approach taking into consideration the main city as a point of departures and arrival to all the other city of the same region we created. For simplicity, we will abbreviate Time slot to TS.

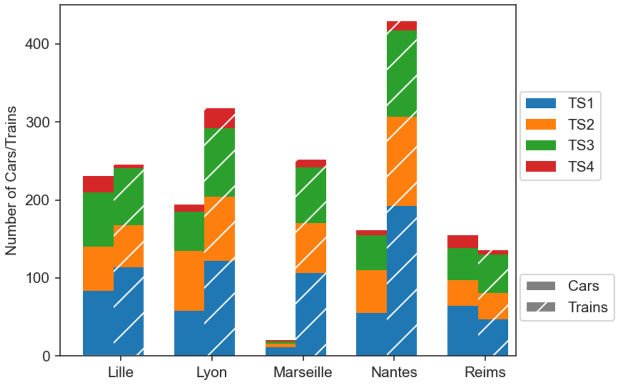


*Figure 1: Region 0 Cars/Trains availability over a week (Arrival: Paris)*

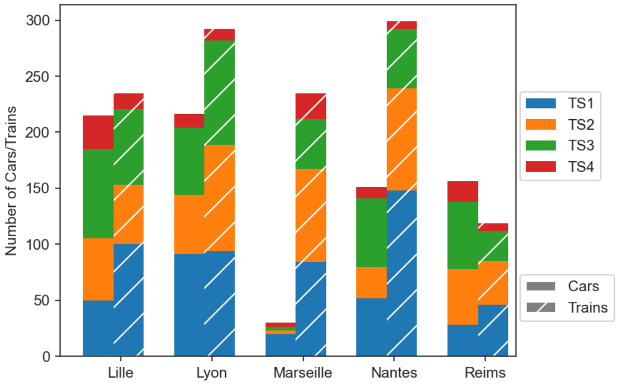


*Figure 2: Region 0 Cars/Trains availability over a week (Departure: Paris)*

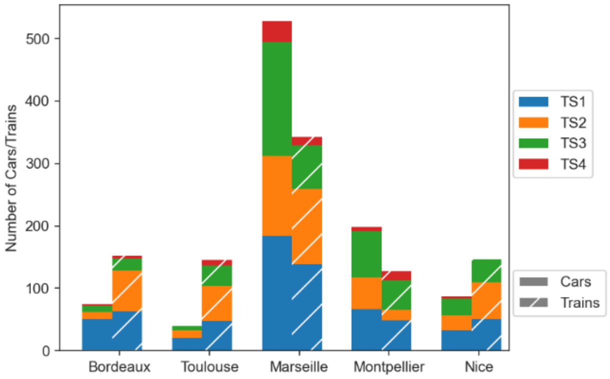


*Figure 3: Region 1 Cars/Trains availability over a week (Arrival: Lyon)*

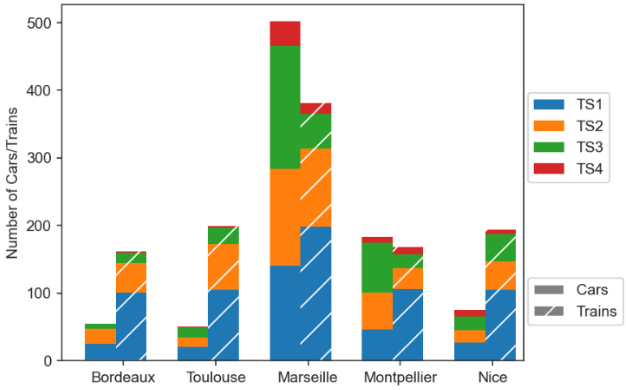


*Figure 4: Region 1 Cars/Trains availability over a week (Departure: Lyon)*

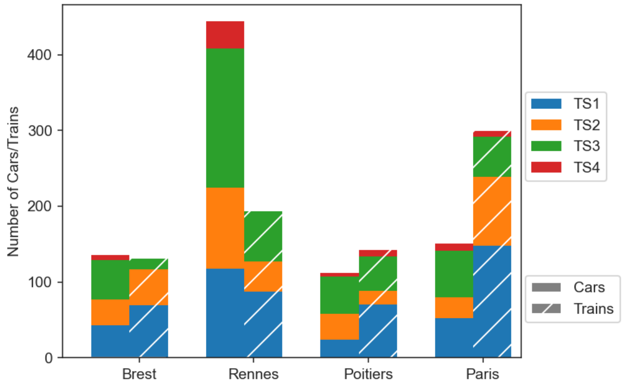


*Figure 5: Region 2 Cars/Trains availability over a week (Arrival: Nantes)*

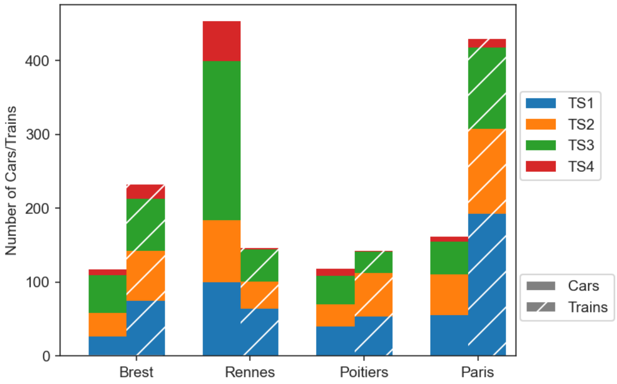


*Figure 6: Region 2 Cars/Trains availability over a week (Departure: Nantes)*

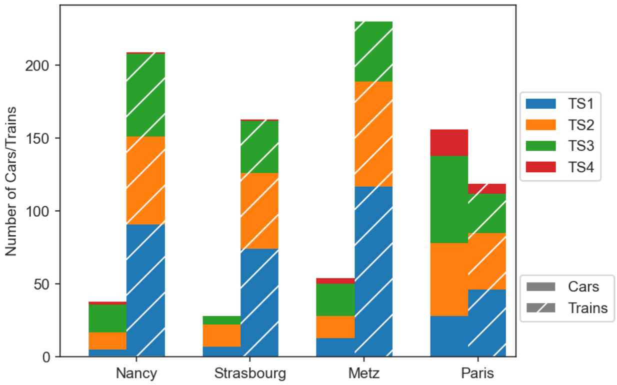


*Figure 7: Region 4 Cars/Trains availability over a week (Arrival: Reims)*

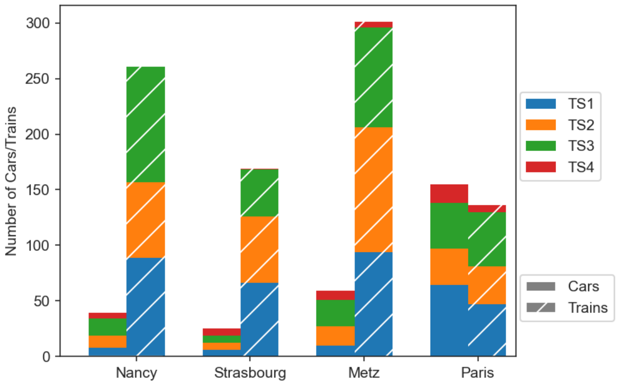


*Figure 8: Region 4 Cars/Trains availability over a week (Departure: Reims)*

We can clearly see that the departure and arrival of the same region are symmetrical. Having the same availability of trains from a city to another and vice versa can be significant for a number of reasons.

It can indicate that there is a strong demand for travel between these two cities and that the transportation infrastructure is well-established to support this demand

It can also indicate that these two cities have a strong economic or cultural connection and that there is a lot of travel between them for business or leisure purposes.

It can also suggest that the companies that operate the trains have recognized the importance of this route, and are willing to invest in and maintain the necessary resources to provide service in both directions.

In terms of logistics, it can also indicate that the companies have good resource management and that they are able to meet the demands of the route in both directions.

Having the same availability of trains in both directions can also be beneficial for passengers, as it provides them with more options and flexibility when planning their travel.

it can also be an indicator of good customer service, as it shows that the company is trying to meet the needs of their customers by providing the same services in both directions.

It's important to consider that, even though having the same availability of trains in both directions can be beneficial, it may not always be possible or practical due to factors such as traffic, infrastructure, and resource constraints.

For short distances, there are a lot of cars on the road, but there are a lot less trains. This is because cars are more convenient for shorter trips, as they offer the flexibility to go wherever you want, whenever you want.

II. Now, we will look at the average price per duration for all the trips we have over the week.

So, to do that we computed the average price of all the trips of the same day and time slot together and divided it by the respective average duration by the same principal to get a euro by hour graph for cars and train over the week. We ended with 4 points for every day.

Here are the graphs we did to visualize this concept:

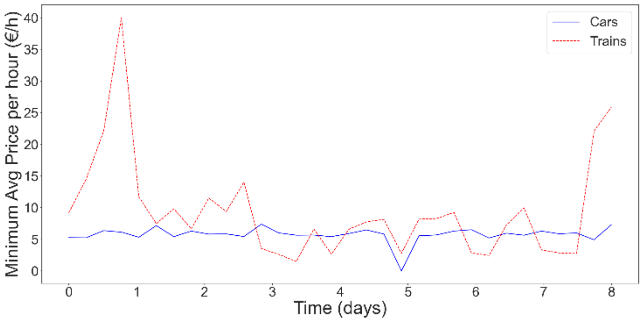


*Figure 9: Minimum Average Price of Cars/Trains over a week for region 0 (Departure: Paris)*

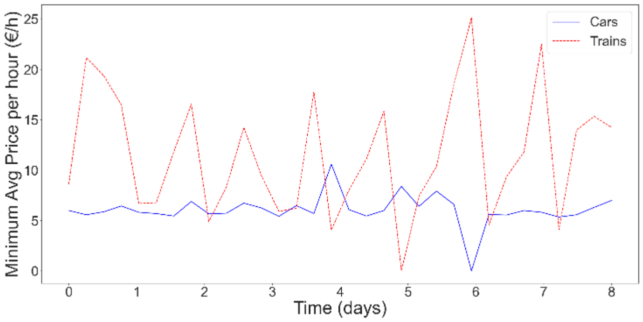


*Figure 10: Minimum Average Price of Cars/Trains over a week for region 1 (Departure: Lyon)*

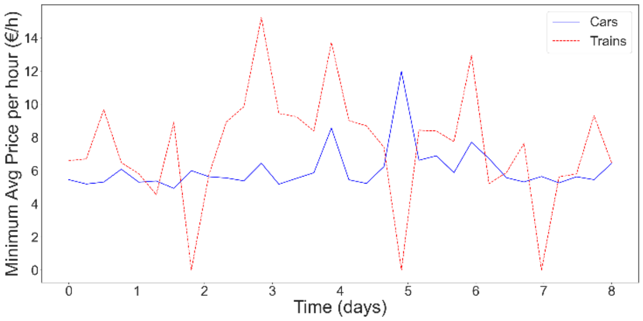


*Figure 11: Minimum Average Price of Cars/Trains over a week for region 2 (Departure: Nantes)*

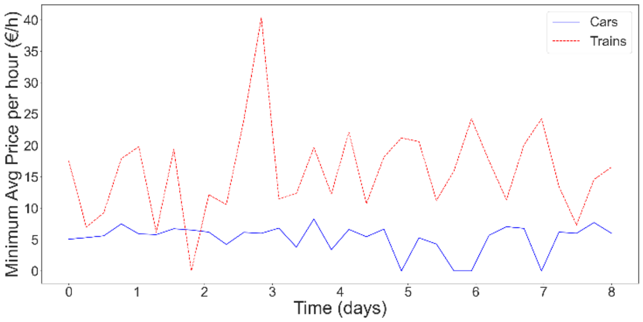


*Figure 12: Minimum Average Price of Cars/Trains over a week for region 4 (Departure: Reims)*

The price of a BlaBlaCar ride tends to remain relatively stable, with only a small amount of variation around the mean. This is because the supply and demand for carpooling rides is relatively consistent, and the price is often determined by the distance of the ride and the cost of gas. Additionally, the prices are set by individual drivers, and a lot of them tend to price their rides competitively in order to attract passengers. As a result, the price of a BlaBlaCar ride can be relatively stable.

On the other hand, the price of train rides tends to fluctuate a lot more. This is because train companies often have dynamic pricing strategies, where the price of a ticket can change based on factors such as the time of day, the day of the week, or how far in advance the ticket is purchased. Additionally, train companies may also offer sales or discounts, which can cause prices to fluctuate. This variability in pricing is more likely to be seen in long-distance train travel than in short-distance travel, as the latter is often seen as a more essential mode of transportation, and prices tend to be more stable.

Train companies may take advantage of the absence of BlaBlaCar rides to increase their prices because trains and BlaBlaCar rides are often considered substitutes for one another. When there are fewer BlaBlaCar rides available, the demand for train rides may increase as people look for alternative ways to travel. This increase in demand can lead train companies to raise their prices, as they know that people will still want to buy tickets even if the price is higher.

Additionally, train companies may also increase prices during times when BlaBlaCar rides are less available because there is less competition in the market. If people have fewer options for travel, they may be more willing to pay a higher price for a train ticket.

It's worth noting that this scenario may not be the case in all countries or regions, as the carpooling service penetration or popularity and the public transportation system may vary. It's also important to consider that Train companies are regulated by government, and there are laws and regulations to prevent them from exploiting people through prices.

- Regression is a statistical method used to analyze the relationship between a dependent variable and one or more independent variables. It is an important tool for data analysis in many fields, including economics, finance, and marketing, as it allows researchers to quantify the relationship between different variables and make predictions about future outcomes.

The regression equation we are using can be represented as follows:

$$
\begin{aligned}
& blablacar\_price\_per\_km = \\
& \quad \beta 0 \\
& + \beta 1 * log(blablacar\_duration\_in\_min) \\
& + \beta 2 * \mathbb{1}_{\{BlablacarBus\_available\}} \\
& + \beta 3 * \mathbb{1}_{\{BlablacarBus\_available\}} \\
& * train\_price\_per\_km \\
& + \beta 4 * train\_price\_per\_km \\
& + \beta 5 * \mathbb{1}_{\{Paris\_is\_dep\_city\}} \\
& + \beta 6 * \mathbb{1}_{\{Paris\_is\_dep\_city\}} \\
& * train\_price\_per\_km \\
& + \beta 7 * train\_nb\_changes \\
& + \sum_{route_i}^{nb_{routes}} \beta_{route_i} * \mathbb{1}_{\{route==route_i\}} \\
& + \sum_{region_i}^{nb_{regions}} \beta_{region_i} * \mathbb{1}_{\{region==region_i\}}
\end{aligned}
$$

*Equation 1: Regression Equation*

Where all the β are the regression coefficients.

In this equation:

"Log(blablacar_duration_in_min)" is a natural log transformation of the duration of the trip, which is likely included to account for non-linearity in the relationship between duration and price. "BlablacarBus_available" and "Paris_is_dep_city" are binary variables indicating whether the BlablacarBus service or the city of Paris is the departure city respectively.

"train_price_per_km" is the price per km of the cheapest alternative/replacing train of the same time and same route of the corresponding BlaBlaCar trips, "train_nb_changes" is the number of these trains changes, "route" and "region" are categorical variables that describe the route and region of the trip.

The two interaction terms we have in the equation:
"BlablacarBus_available*train_price_per_km" and "Paris_is_dep_city*train_price_per_km" are included to capture any potential interactions between the availability of the BlablacarBus service, departure city, and the price of train.

To be able to do this regression we followed:

- Step 1:

Preparing the dataset used for regression. We match every BlaBlaCar trip with it corresponding cheapest replacing train using both datasets. Then, we create the following columns:

  - **BlablacarBus_available**:
    Binary variable to check the presence of a *"BlaBlaCar Bus service"* as a replacement of the BlaBlaCar trip. The significance of adding this column is to check if there is competition between the train and the bus of the same platform.
  - **Paris_is_dep_city**:
    Binary variable to check if Paris is the departure city. The significance of adding this column is checking the influence of having Paris as a departure city on the price of the trip.
  - **route:**
    The trip routes written in this form *dep_city,arr_city* in order to study the influence of the route on the price of the trip.

- Step 2:

We need to mention that we did 2 types of regression, the first was without considering the clustered standard errors and the other one considering it.

The results of our regression are as follows:

| | | | |
|---|---|---|---|
| **Dep. Variable:** | blablacar_price_per_km | **R-squared:** | 0.187 |
| **Model:** | OLS | **Adj. R-squared:** | 0.177 |
| **Method:** | Least Squares | **F-statistic:** | 20.16 |
| **Date:** | Thu, 26 Jan 2023 | **Prob (F-statistic):** | 2.28E-277 |
| **Time:** | 15:43:26 | **Log-Likelihood:** | 19629 |
| **No. Observations:** | 7824 | **AIC:** | -3.91E+04 |
| **Df Residuals:** | 7735 | **BIC:** | -3.85E+04 |
| **Df Model:** | 88 | | |
| **Covariance Type:** | nonrobust | | |

*Table 4: First Regression Statistics*

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

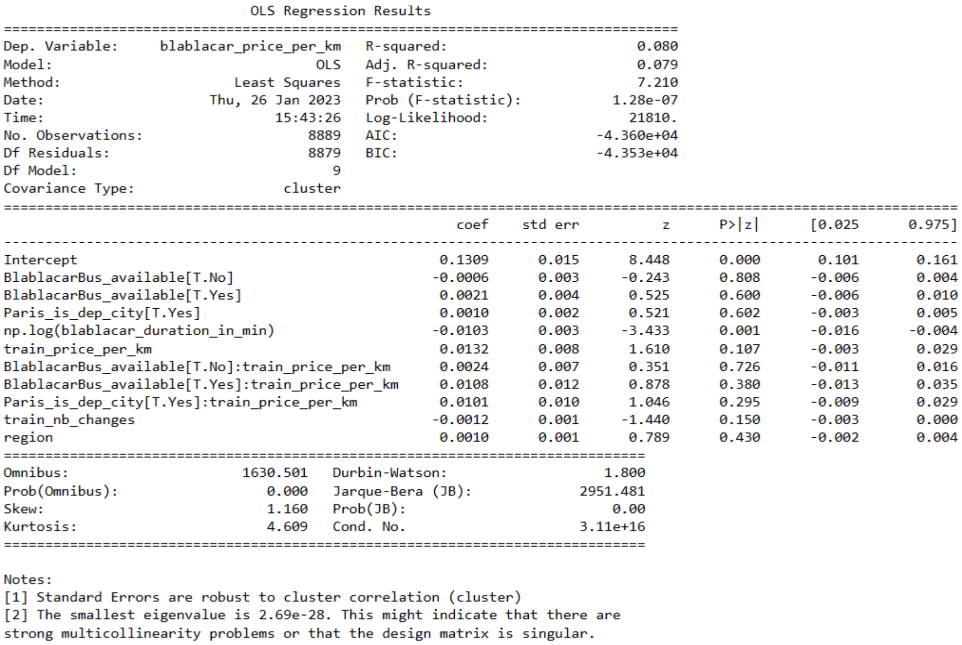
$$RSS = \sum_{n=1}^{N} e_i^2 = 3.032961275769039$$

$$TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2 = 4.183344737394302$$

*Equation 2: R2 Equation*

| | **coef** | **std err** | **t** | **P>\|t\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **Intercept** | 0.0249 | 0.012 | 2.126 | 0.034 | 0.002 | 0.048 |
| **BlablacarBus_available[T.Yes]** | 0.0014 | 0.002 | 0.684 | 0.494 | -0.003 | 0.005 |
| **Paris_is_dep_city[T.Yes]** | 0.0041 | 0.003 | 1.2 | 0.23 | -0.003 | 0.011 |
| **np.log(blablacar_duration_in_min)** | 0.0079 | 0.002 | 4.059 | 0 | 0.004 | 0.012 |
| **train_price_per_km** | 0.0154 | 0.005 | 2.924 | 0.003 | 0.005 | 0.026 |
| **BlablacarBus_available[T.Yes]:train_price_per_km** | 0.0054 | 0.013 | 0.411 | 0.681 | -0.02 | 0.031 |
| **Paris_is_dep_city[T.Yes]:train_price_per_km** | 0.0191 | 0.021 | 0.887 | 0.375 | -0.023 | 0.061 |
| **train_nb_changes** | -0.0007 | 0.001 | -1.454 | 0.146 | -0.002 | 0 |
| **region** | -1.77E-16 | 0.001 | -1.41E-13 | 1 | -0.002 | 0.002 |

*Table 5: First Regression Results*



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     blablacar_price_per_km   R-squared:                       0.080
Model:                                OLS   Adj. R-squared:                  0.079
Method:                     Least Squares   F-statistic:                     7.210
Date:                    Thu, 26 Jan 2023   Prob (F-statistic):           1.28e-07
Time:                            15:43:26   Log-Likelihood:                 21810.
No. Observations:                    8889   AIC:                        -4.360e+04
Df Residuals:                        8879   BIC:                        -4.353e+04
Df Model:                               9
Covariance Type:                  cluster
=====================================================================================================================
                                                        coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------------------
Intercept                                             0.1309      0.015      8.448      0.000       0.101       0.161
BlablacarBus_available[T.No]                         -0.0006      0.003     -0.243      0.808      -0.006       0.004
BlablacarBus_available[T.Yes]                         0.0021      0.004      0.525      0.600      -0.006       0.010
Paris_is_dep_city[T.Yes]                              0.0010      0.002      0.521      0.602      -0.003       0.005
np.log(blablacar_duration_in_min)                    -0.0103      0.003     -3.433      0.001      -0.016      -0.004
train_price_per_km                                    0.0132      0.008      1.610      0.107      -0.003       0.029
BlablacarBus_available[T.No]:train_price_per_km       0.0024      0.007      0.351      0.726      -0.011       0.016
BlablacarBus_available[T.Yes]:train_price_per_km      0.0108      0.012      0.878      0.380      -0.013       0.035
Paris_is_dep_city[T.Yes]:train_price_per_km           0.0101      0.010      1.046      0.295      -0.009       0.029
train_nb_changes                                     -0.0012      0.001     -1.440      0.150      -0.003       0.000
region                                                0.0010      0.001      0.789      0.430      -0.002       0.004
==============================================================================
Omnibus:                     1630.501   Durbin-Watson:                   1.800
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2951.481
Skew:                           1.160   Prob(JB):                         0.00
Kurtosis:                       4.609   Cond. No.                     3.11e+16
==============================================================================

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)
[2] The smallest eigenvalue is 2.69e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

*Table 6: Second Regression Results and Statistics*

**Regression Results Analysis:**

From Table 5 we could recognize the following:

- $\beta$(*train_price_per_km)* = 0.0154
- $\beta$(*BlablacarBus_available[T.Yes]:train_price_per_km)* = 0.0054
- $\beta$(*Paris_is_dep_city[T.Yes]:train_price_per_km)* = 0.0191
- $\beta(region\,) = -1.765 * 10^{-16}$

The results of the regression analysis show a positive and statistically significant relationship between the two variabes "*train_price_per_km"* and "*blablacar_price_per_km"* ($p<0.005$). An increase of 10% in the train price per km is associated with an increase of 15.4% in the BlaBlaCar price.

We could also recognize the effect of the presence of the BlaBlaCar Bus service on this route and time on the pricing of the BlaBlacar trips, which reveal no competition between the bus service and the drivers within the same BlaBlaCar platform.

Having Paris as a departure city improve the influence of the train price per km on the price of the BlaBlaCar trips. An increase of 10% in the train price per km with this condition is associated with an increase of 19.1% in the BlaBlaCar price.

The coefficient of the region variable is almost 0 which means that the pricing of the drivers of BlaBlaCar is not affected by geographical region in France.

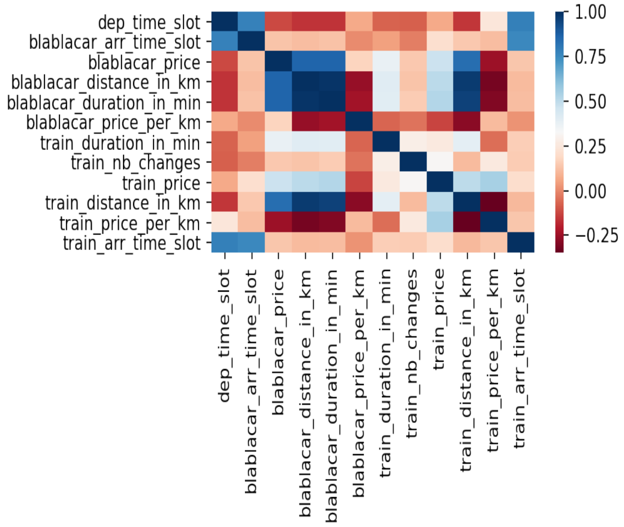


*Figure 13: Correlation Matrix*

## Conclusion

In conclusion, the results of the regression analysis show a strong positive and statistically significant relationship between the train price per km and the BlaBlaCar price per km. An increase in train prices leads to a corresponding increase in BlaBlaCar prices. Furthermore, it was found that the impact of train prices on BlaBlaCar prices is even stronger when the departure city is Paris. However, it should be noted that the influence of geographical region on BlaBlaCar prices was found to be minimal.

Then we can deduce that the BlaBlaCar drivers inspect the price of the train before pricing their trip.

Potential shortcomings of this study include the limited scope of geographical regions included in the analysis and the possibility of other factors affecting BlaBlaCar prices that were not considered like strikes or non-daily events.

In terms of future research, it would be valuable to expand the scope of the study to include more geographical regions in France and to explore other potential factors that may affect BlaBlaCar prices, such as the type of vehicle being used or the time of day of the trip which we have already collected but for the shortness of time we couldn't manage to include it in our regression. Additionally, it would be interesting to replicate this study in other countries to see if similar results are found.

## References

- ❖ https://www.blablacar.fr
- ❖ https://www.thetrainline.com