Analysis of Cycling Traffic in Paris

Project carried out as part of the

Data Analyst Bootcamp by DataScientest

September 2024 Cohort

Authors:

Amal MOHAMED

Chelsea Wright

Dirk Eisermann

Introduction

In 2021, Paris introduced the **Plan Velo: Act 2** as part of their vision to become one of the most bike-friendly cities in Europe. Since then, the city has invested significantly in upgrading its cycling infrastructure with the goal of encouraging more residents and tourists to use cycling as a mode of transportation in Paris.

As part of this initiative, the city of Paris installed cyclist counting meters (hereon referred to as counters) on various streets and cycling paths around the city. Our project aims to analyze the data collected by these counters and obtain meaningful visualizations of traffic patterns. From our findings we will present actionable insights that city planners can use to further improve the cycling infrastructure in Paris.

Data Sources

The two datasets used in this project were obtained from opendata.paris.fr. Both are free and available to the public.

The main dataset includes the hourly count of cyclists registered by the different bicycle counters across Paris. It is 80 MB in size and is updated daily. It is available here.

The second dataset contains detailed information about street addresses in Paris. We will use it to classify the counters by their arrondissement. It is available here.

Objectives

Our main objectives are to understand the cycling traffic behavior in Paris and draft business insights to build a suitable machine learning model. Specifically, this analysis will cover the following aspects:

1. **Identifying areas with the highest and lowest cycling traffic volumes and exploring cycling patterns in Paris.**
2. **Reviewing whether the cycling-friendly initiatives taken for the 2024 Summer Olympic and Paralympic games had an impact on cycling as a means of transport in Paris.**
3. **Using unsupervised machine learning models to identify the distribution of cycling traffic patterns throughout Paris.**
4. **Using a supervised machine learning model to predict afternoon cyclist counts from morning cyclist counts.**

In the following sections, we will present the results of our Exploratory Data Analysis and Data Cleaning and Preprocessing stages.

Exploratory Data Analysis (EDA)

This phase aims to explore the raw data using the Python programming language in Jupyter Notebook. We will utilize different libraries such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn to execute preliminary statistics and manipulation techniques. The exploration will be divided into two main parts:

1. Data Structure Exploration: a brief description of the two dataset and their variables.
2. Data Visualization: an initial analysis to understand cycling traffic patterns.

1. Data Structure Exploration

Main DataFrame:

After loading the file to Jupyter Notebook and displaying the first rows of the cycling traffic dataset, we noticed that the columns are in French. The first step we took was to translate and rename the columns in English.

- The dataset before renaming:

	Identifiant du compteur	Nom du compteur	Identifiant du site de comptage	Nom du site de comptage	Comptage horaire	Date et heure de comptage	Date d'installation du site de comptage	Lien vers photo du site de comptage	Coordonnées géographiques	Identifiant technique compteur	ID Photos
0	100003096-353242251	97 avenue Denfert Rochereau SO-NE	100003096	97 avenue Denfert Rochereau	4	2023-09-01T05:00:00+02:00	2012-02-22	https://filer.eco-counter-tools.com/file/10/6d...	48.83504,2.33314	Y2H21111072	https://filer.eco-counter-tools.com/file/10/6d...
1	100003096-353242251	97 avenue Denfert Rochereau SO-NE	100003096	97 avenue Denfert Rochereau	63	2023-09-01T07:00:00+02:00	2012-02-22	https://filer.eco-counter-tools.com/file/10/6d...	48.83504,2.33314	Y2H21111072	https://filer.eco-counter-tools.com/file/10/6d...
2	100003096-353242251	97 avenue Denfert Rochereau SO-NE	100003096	97 avenue Denfert Rochereau	16	2023-09-01T06:00:00+02:00	2012-02-22	https://filer.eco-counter-tools.com/file/10/6d...	48.83504,2.33314	Y2H21111072	https://filer.eco-counter-tools.com/file/10/6d...

- The dataset after renaming:

	Counter ID	Counter Name	Counting Site ID	Counting Site Name	Hourly Count	Count Date and Time	Site Installation Date	Link to Site Photo	Geo Coordinates	Technical Counter ID	Photo ID	
0	100003096-353242251	97 avenue Denfert Rochereau SO-NE	100003096	97 avenue Denfert Rochereau	4	2023-09-01T05:00:00+02:00	2012-02-22	https://filer.eco-counter-tools.com/file/10/6d...	48.83504,2.33314	Y2H21111072	https://filer.eco-counter-tools.com/file/10/6d...	tools
1	100003096-353242251	97 avenue Denfert Rochereau SO-NE	100003096	97 avenue Denfert Rochereau	63	2023-09-01T07:00:00+02:00	2012-02-22	https://filer.eco-counter-tools.com/file/10/6d...	48.83504,2.33314	Y2H21111072	https://filer.eco-counter-tools.com/file/10/6d...	tools
2	100003096-353242251	97 avenue Denfert Rochereau SO-NE	100003096	97 avenue Denfert Rochereau	16	2023-09-01T06:00:00+02:00	2012-02-22	https://filer.eco-counter-tools.com/file/10/6d...	48.83504,2.33314	Y2H21111072	https://filer.eco-counter-tools.com/file/10/6d...	tools

The main dataset contains readings from 2022 to 2024, and the DataFrame has 945853 rows and 16 columns. The renamed columns and their descriptions are listed in below table:

Column name	**Description**
Counter ID	Unique identity for each cyclist counter
Counter Name	Name of each counter named according to its specific street address and direction
Counter Site ID	Unique identity of the street address where the cyclist counter is located
Counter Site Name	Street address where the cyclist counter is located; area of the cycling traffic
Hourly Count	Number of bicycles recorded per hour; volume of the cycling traffic
Count Data and Time	Date and time when counter read was taken
Site Installation Date	Date when the counter was installed
Link to Site Photo	Website with a photo captured by the counter
Geo Coordinates	Latitude and longitude of the counter site
Technical Counter ID	Unique identity of technical meter identifier
Photo ID	Unique identity of the photo
Test Link to Site Photos	URL for test photos captured by the counter
Photo ID 1	Secondary unique identity of the photo
Site URL	Link to view the photo
Image Type	Format of the image
Month and Year of Count	Month and year where the counter read was recorded

Then we displayed overall summary information of the data to identify potential errors which could exist in our dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 945853 entries, 0 to 945852
Data columns (total 16 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   Counter ID                945853 non-null  object
 1   Counter Name              945853 non-null  object
 2   Counting Site ID          945853 non-null  int64
 3   Counting Site Name        945853 non-null  object
 4   Hourly Count              945853 non-null  int64
 5   Count Date and Time       945853 non-null  object
 6   Site Installation Date    945853 non-null  object
 7   Link to Site Photo        936139 non-null  object
 8   Geo Coordinates           945853 non-null  object
 9   Technical Counter ID      936136 non-null  object
 10  Photo ID                  936139 non-null  object
 11  Test Link to Site Photos  936139 non-null  object
 12  Photo ID 1                936139 non-null  object
 13  Site URL                  945853 non-null  object
 14  Image Type                936139 non-null  object
 15  Month and Year of Count   945853 non-null  object
dtypes: int64(2), object(14)
memory usage: 115.5+ MB
```

Findings:

1. The data has 945853 rows total, thus from the 'Non-Null' column in the above output we can see that some columns in our dataframe have missing values. However, the columns with those cases are irrelevant to our context and will not be analyzed, so they do not affect us.

2. Some columns have to be converted; the 'Count Date and Time' and 'Month and Year of Count' columns need to be converted to datetime format.

Next, we checked for any duplicate rows and found that there were none. Then we looked into the descriptive statistics of our target column (Hourly Count) to get an idea of central tendencies and variability of our dataset.

	Hourly Count
count	945853.000000
mean	78.909081
std	109.125458
min	0.000000
25%	12.000000
50%	43.000000
75%	98.000000
max	8190.000000

Findings:

1. The data is pretty spread out and might have significant outliers given the low mean (~79) and high standard deviation (~109).
2. The minimum (0) and the maximum (8190) values indicate that we may have outliers, as there is a wide range between the first quartile (Q1 = 12%) and third quartile (Q3 = 98%).
3. There is a large gap between the quartiles, signifying a high variability in our data.

Similarly, we also used the descriptive statistical summary on the non-numeric columns in our data:

	Counter ID	Counter Name	Counting Site Name	Count Date and Time	Site Installation Date	Link to Site Photo	Geo Coordinates	Technical Counter ID	Photo ID	Test Link to Site Photos	F
count	945853	945853	945853	945853	945853	936139	945853	936136	936139	936139	9
unique	101	101	69	9720	40	72	73	69	72	71	
top	100047548-104047548	Face au 25 quai de l'Oise SO-NE	Face au 25 quai de l'Oise	2024-08-19T02:00:00+02:00	2018-11-28	https://filer.eco-counter-tools.com/file/30/65...	48.89141,2.38482	Y2H20083609	https://filer.eco-counter-tools.com/file/30/65...	https://filer.eco-counter-tools.com/file/85/6f...	
freq	9715	9715	19430	98	68000	19430	19430	19430	19430	38856	9

Findings:

1. The ‘Counter Name’ column 101 unique names, so we know this is the current number of counters located around Paris.
2. The ‘Counting Site Name’ variable consists of 69 unique site names. Since there are more counter names than counting site names, we can assume there are some sites with more than one counter at that location (most likely separate counters capturing traffic in opposite directions).

Subsequently, we prepared our dataset. First, we removed unnecessary columns that were irrelevant to our scope of analysis.

- The following columns were dropped:

```
Index(['Counting Site ID', 'Site Installation Date', 'Link to Site Photo',
       'Technical Counter ID', 'Photo ID', 'Test Link to Site Photos',
       'Photo ID 1', 'Site URL', 'Image Type'],
      dtype='object')
```

Here is the dataframe after dropping the irrelevant variables and resetting the column index to Counter ID:

Counter ID	Counter Name	Counting Site Name	Hourly Count	Count Date and Time	Geo Coordinates	Month and Year of Count
100003096-353242251	97 avenue Denfert Rochereau SO-NE	97 avenue Denfert Rochereau	4	2023-09-01T05:00:00+02:00	48.83504,2.33314	2023-09
100003096-353242251	97 avenue Denfert Rochereau SO-NE	97 avenue Denfert Rochereau	63	2023-09-01T07:00:00+02:00	48.83504,2.33314	2023-09
100003096-353242251	97 avenue Denfert Rochereau SO-NE	97 avenue Denfert Rochereau	16	2023-09-01T06:00:00+02:00	48.83504,2.33314	2023-09

Then we extracted information from some columns into new columns that will be useful for our data visualization and further analyses.

- From the column ‘Data and Time’ we extracted the date, time, day of the week, month and year and created new columns. Similarly, from the column ‘Geo Coordinates’ we created separate columns for longitude and latitude.

- Below is the summary of our dataframe after the creation of the new columns and ensuring that they all are of the correct type:

```
<class 'pandas.core.frame.DataFrame'>
Index: 945853 entries, 100003096-353242251 to 300030271-353356061
Data columns (total 13 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   Counter Name            945853 non-null  object
 1   Counting Site Name      945853 non-null  object
 2   Hourly Count            945853 non-null  int64
 3   Count Date and Time     945853 non-null  datetime64[ns, UTC]
 4   Geo Coordinates         945853 non-null  object
 5   Month and Year of Count 945853 non-null  period[M]
 6   Date                    945853 non-null  object
 7   Time                    945853 non-null  object
 8   Weekday                 945853 non-null  int32
 9   Month                   945853 non-null  int32
 10  Year                    945853 non-null  int32
 11  Latitude                945853 non-null  float64
 12  Longitude               945853 non-null  float64
dtypes: datetime64[ns, UTC](1), float64(2), int32(3), int64(1), object(5), period[M](1)
memory usage: 90.2+ MB
```

Note that there are now no missing values. Our data consists of one main numeric variable (Hourly Count) and two main categorical values (Count Date and Time, Geo Coordinates), along with their subcategories.

The dataset is now ready for our initial analysis.

Second DataFrame:

The second DataFrame has 151879 rows and 5 columns. Neither statistics nor data visualization were performed on this particular dataframe. Therefore we will only provide an overview of the variables which were used.

Column name	Description	Type	Column renamed
C_AR	Arrondissement number	Integer	Arrondissement
L_ADR	Full address (street number followed by the street name)	String	Counting Site Name

We prepared this dataset to merge with the column of our main dataset ‘Counting Site Name’. We performed the following steps in order to end up with a table with the street address of the counter and its corresponding arrondissement.

1. Rename the columns.
2. Remove all the missing values.
3. Standardize the address name to be all in one format.
4. Extract street details from the column renamed ‘Counting Site Name’, such as the number, type and name of the street.
5. A merge was performed to get the arrondissement numbers and their corresponding streets.

Note: The above steps 3 and 4 were also performed on the column “Counting Site Name” in our main dataset, so that that column and the ‘Street Address’ column in this dataframe contain identical information. The column name just needs to be matched in both dataframes before merging.

- Here is the final table after merging:

	Street Address	Arrondissement
0	97 AVENUE DENFERT ROCHEREAU	14
1	105 RUE LA FAYETTE	10
2	106 AVENUE DENFERT ROCHEREAU	14
3	135 AVENUE DAUMESNIL	12
4	100 RUE LA FAYETTE	10
5	28 BOULEVARD DIDEROT	12
6	39 QUAI FRANCOIS MAURIAC	13

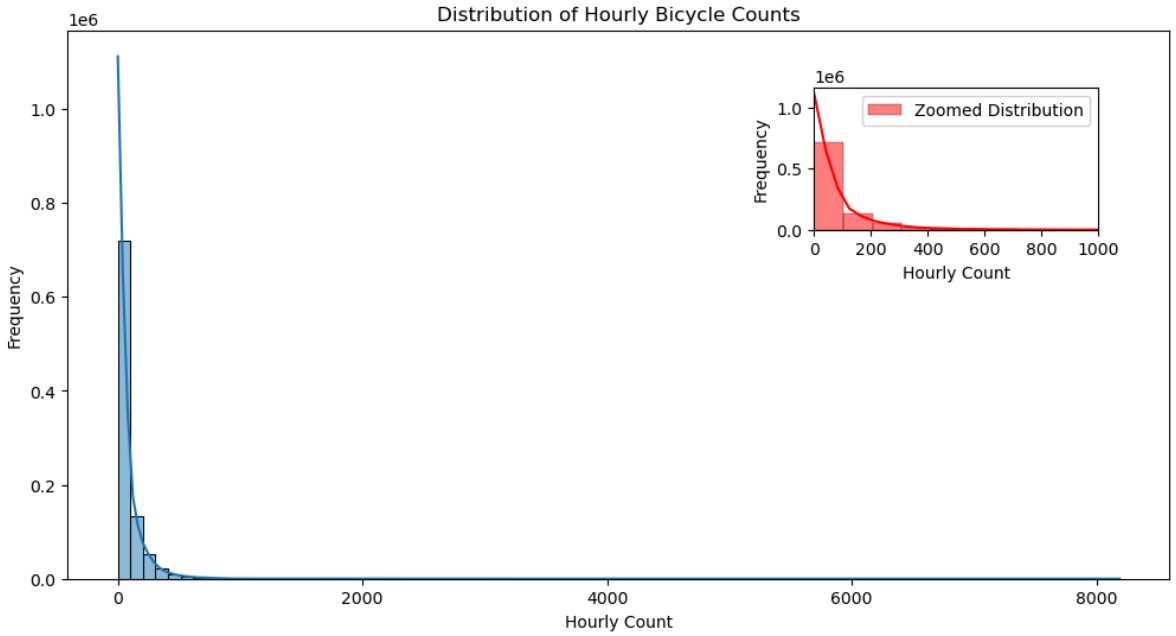
This dataframe was merged with our main dataset for the following analyses.

2. Data Visualization

Now we will explore our dataset further by using basic data manipulation to visualize the trends and patterns of cycling traffic in Paris. We will visualize the distribution of the numeric variable and categorical variables in our dataset, then visualize the relationships between the categorical variables and the numeric variable. We will interpret each graph and draft an initial conclusion with respect to our project scope and objectives.

2.1 Distribution of Hourly Counts

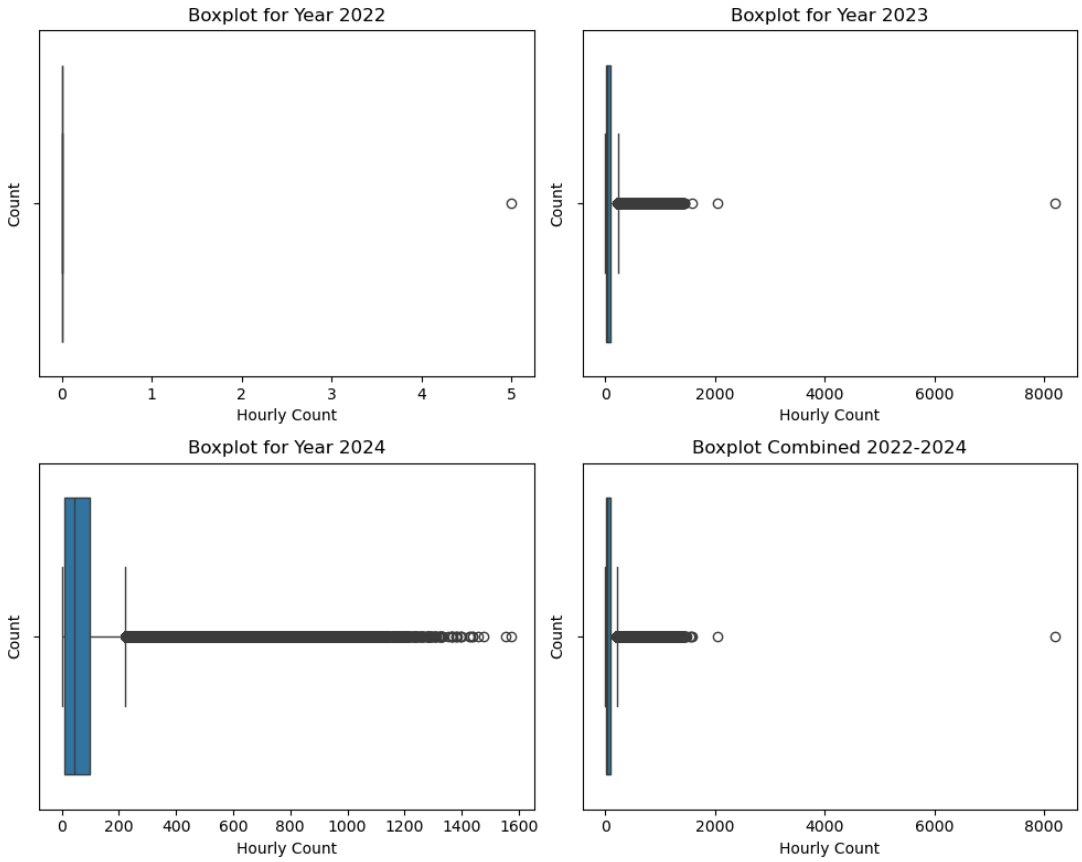
First, to demonstrate the distribution of hourly counts we used a histogram:



Interpretation:

The histogram portrays the distribution of the numeric variable 'Hourly Count' in our cycling traffic dataset. The distribution of recorded cyclist counts is heavily right-skewed due to high concentration of most counter reads between 0 and 200 approximately (as shown in the zoomed view).

Then we used boxplots to identify any possible outliers:



Interpretation:

The boxplot for 2022 shows questionably low counts, and very few hourly counts in general.

The boxplot for 2023 shows one clear outlier, and another potential outlier.

The boxplot for 2024 shows a heavy right skew, suggesting that the majority of the hourly counts are relatively low. The median is very close to the lower edge of the range, indicating that at least 50% of the data points are low values relative to the overall range.

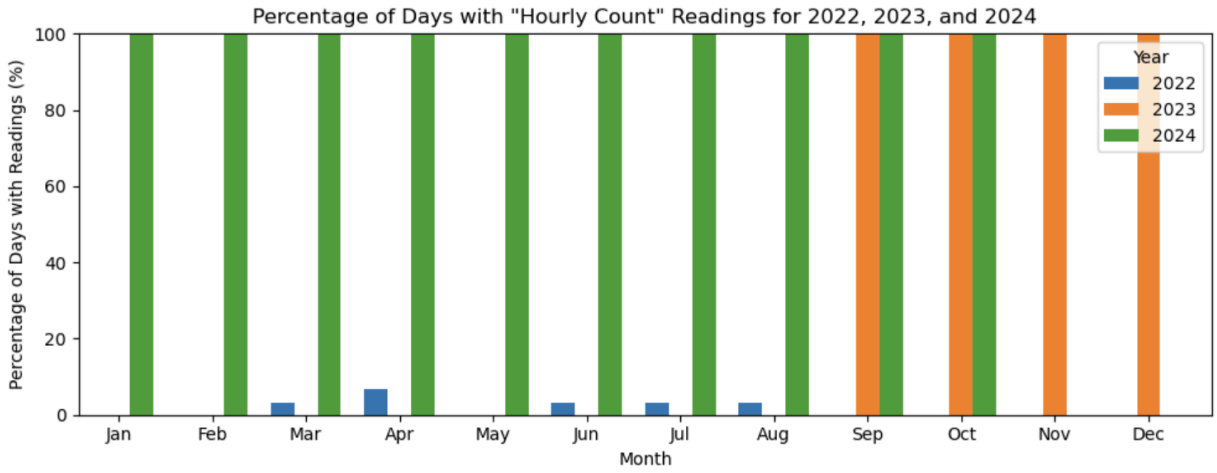
Across all three years, there are two extreme values. The values are strikingly high (above 2000 counts in one hour versus the overall mean of 79 counts per hour). These values will need to be further inspected to determine whether they are indeed outliers and then handled accordingly.

Business Insight:

The observations obtained from the histogram and boxplot will be considered during the data cleaning phase. Having clean and accurate data will optimize the accuracy of our model and impact decisions related to cycling enhancement strategies in the city.

2.2 Missing Readings

After interpreting the boxplots above, we suspected that there may not be readings taken during all the months in our 2022-2024 dataset, especially for 2022. We therefore next visualized what percentage of days in each month had hourly count readings.



Interpretation:

As the bar chart shows, there are indeed inconsistencies and gaps in time where there were no readings taken. In 2022 only five months had readings, and in those months very few days (under 10% of days in each month). Upon further inspection it was found that there were only 6 hourly count readings taken in 2022. This is far too little data to draw meaningful conclusions from, **so we therefore decide to exclude 2022 from our further analyses.**

In 2023, readings only began in September, but from then on 100% of days had readings. In 2024 all months had 100% of days with readings (only through October, because this analysis is being done in early November 2024, so the readings for that month are not included in our dataset). **We therefore decide to perform all further analyses on the time frame from September 2023 to October 2024.**

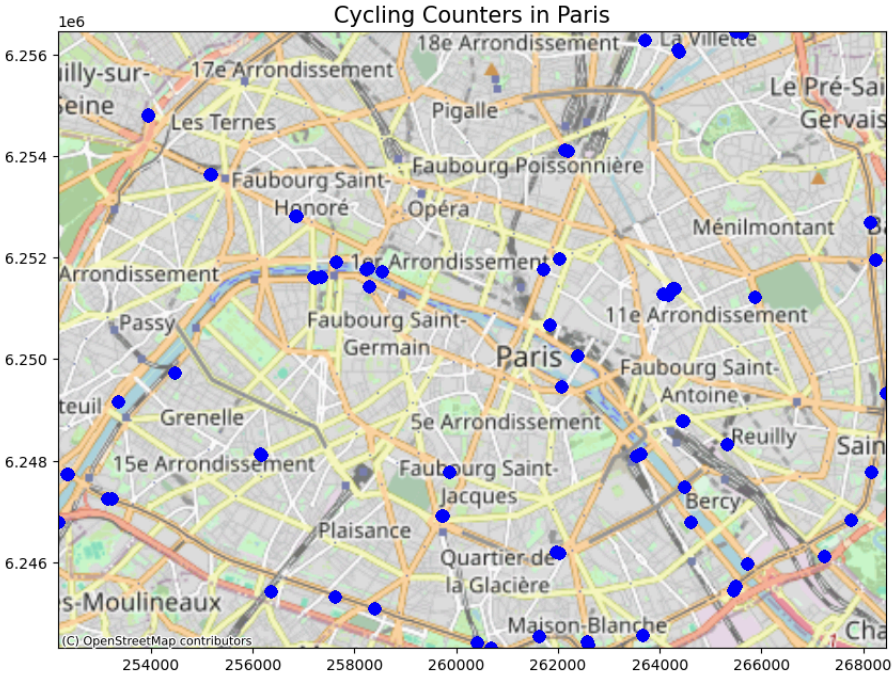
We also looked for cases of missing readings on the counter level. We wanted to know whether any counters didn't produce readings for longer periods of time (ie, were turned off). We found that this was the case for one counter.

	counter name	month_year	count
647	44 avenue des Champs Elysées SE-NO	2023-12	0
648	44 avenue des Champs Elysées SE-NO	2024-01	0
649	44 avenue des Champs Elysées SE-NO	2024-02	0
650	44 avenue des Champs Elysées SE-NO	2024-03	0

From 2023 and 2024, counter 44 didn't record any readings for 4 months. However, there are so many other counters that recorded data from that period that the overall representation of cycling traffic in Paris is not significantly affected by the data missing from this one counter for that period.

2.3 Geographic Distribution of Counters

In order to get an overall picture of where the individual counters are located in Paris, we plotted their locations on a map of the city.



For comparison, the following is a map of cycling routes in Paris (marked in pink and purple). It can be seen that the counters are mostly located along the main cycling routes.

Interpretation:

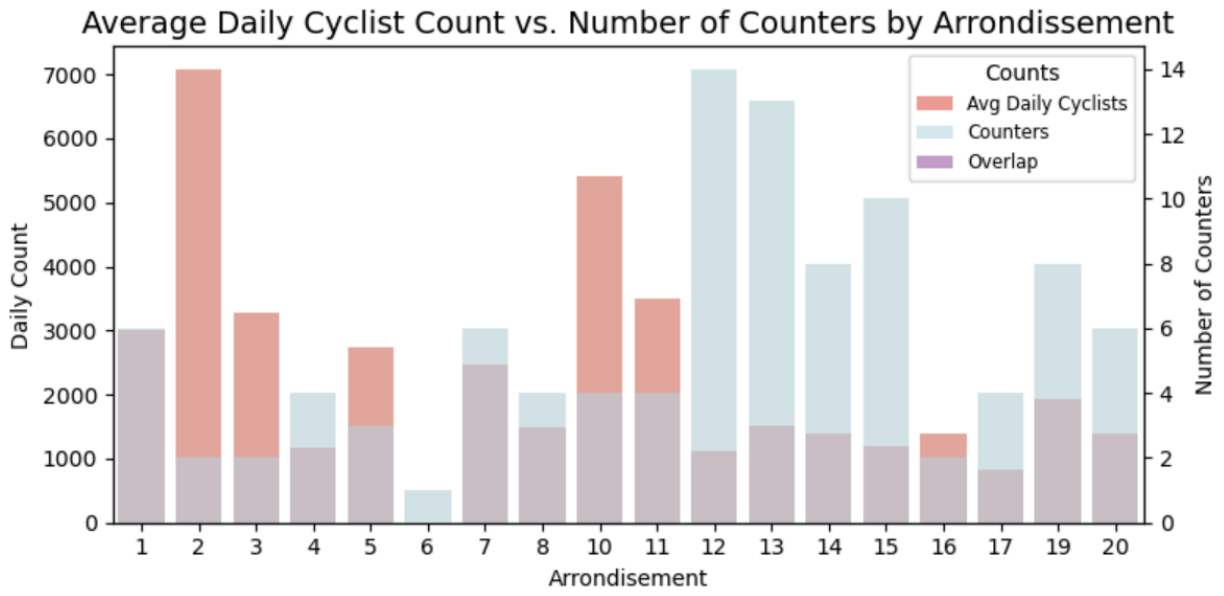
The map shows the distribution of counters located on the main roads and cycle paths in Paris. They appear to be evenly distributed, giving a good overall picture of cycling throughout the city.

Business Insight:

We conclude that this even distribution of counters reduces potential bias in our modeling.

Note: The map was done by the geopandas module. We also did an interactive map using the folium module. This map shows the counters and by clicking on the points you can see the mean of the hourly counts. This map is in HTML format, so it is a separate file. Click here to access the HTML link directly.

We also looked at the distribution of counters relative to the average daily count of cyclists for each neighborhood (arrondissement) of Paris:



Interpretation:

We can see that there is an unequal distribution of cyclists across the different areas of the city (arrondissements), and an unequal distribution of counters in the different arrondissements. Most of the cycling traffic in Paris occurs in just a few of the arrondissements (2nd, 10th, 11th, 3rd, 5th). Interestingly, these highly-trafficked areas have relatively few counters in them compared to other areas with far less cyclist traffic but many counters (for example, the 12th arrondissement).

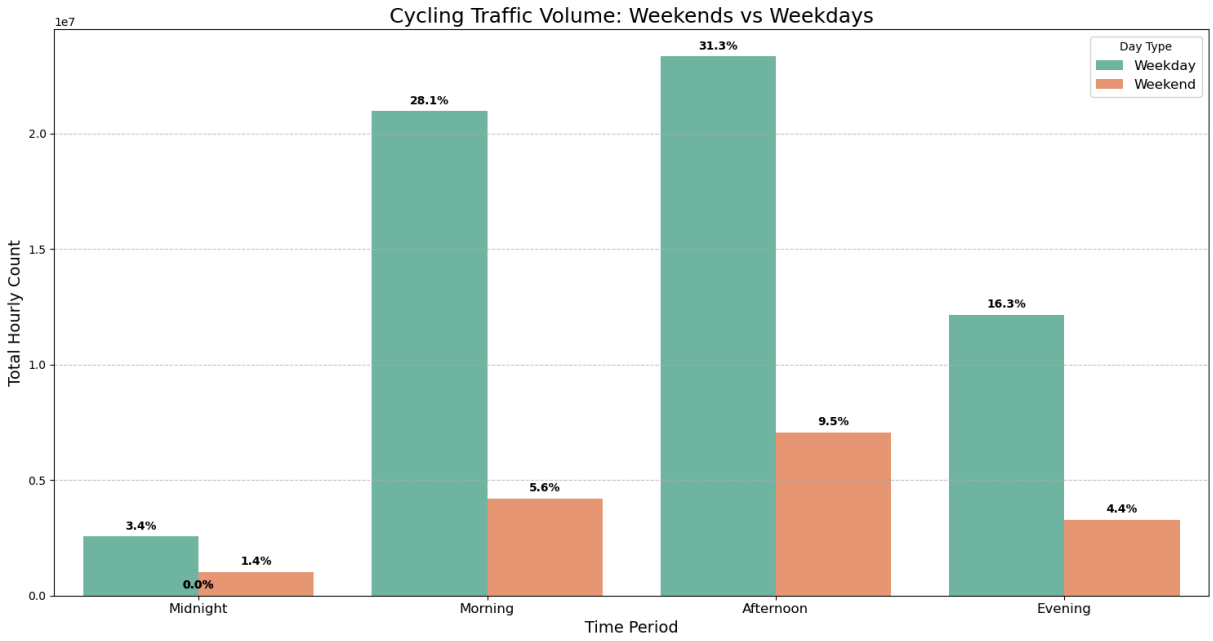
Business Insight:

It may be beneficial for the city to relocate some of the counters from low-traffic areas into higher-traffic areas. This would give a better picture of the cycling traffic in the most

highly-trafficked areas, allowing city planners to better pinpoint locations to improve cycling infrastructure (ex- where to put in or expand bicycle lanes).

2.4 Trends in Cycling Traffic

We then visualized the proportion of hourly counts made during different time periods on weekends and weekdays:



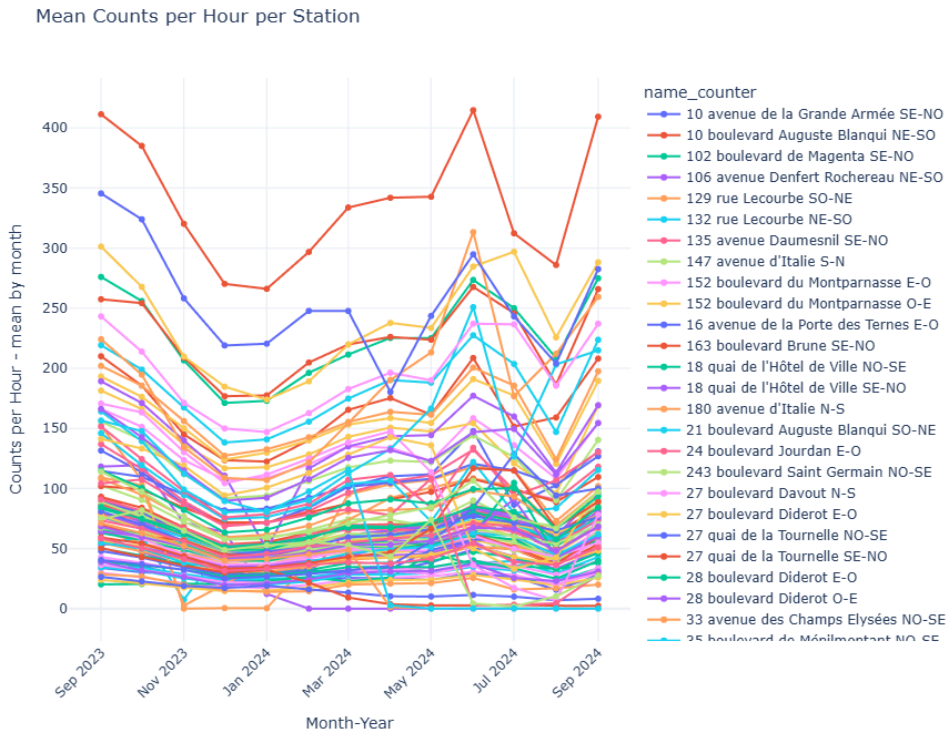
Interpretation:

The bar chart reflects the percentage of the cycling traffic volume in Paris across different periods of the day and days of week. The traffic increases significantly during weekdays, particularly in the morning and afternoon (59% of all cycling traffic in Paris takes place at these times). This is in line with common commute times (e.g, 7am to 5pm). On the contrary, it decreases during evening and midnight. The graph also shows logically minimal traffic at midnight (0.0%) both on weekdays and weekends.

Business Insight:

This reveals that Parisians use bicycles as a form of transportation more during working days than weekends. This can be used to set strategies and policies aimed at enhancing the cycling conditions during peak commute times. For example, the city could expand bike lanes during peak commuting hours, improve bike parking facilities, and implement bike-sharing programs to further encourage cycling as a primary mode of transportation.

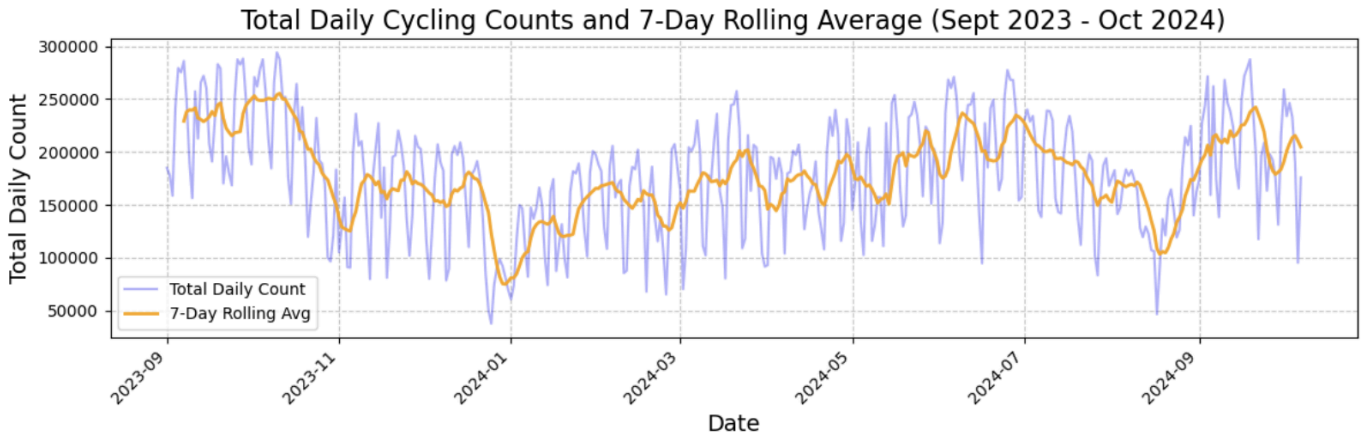
Next we wanted to get a picture of the monthly averages of the counters:



Interpretation:

This graph shows the general trend of average count per hour by month for all the counters in the city. Most of the counters register an hourly average below 200, while only a few record averages over 200. This indicates wide variance in traffic volume, with high frequency in certain counts per hour and low frequency in others.

The above graph is hard to interpret, so below we grouped all counters together per day:



Interpretation:

From the graph we can see that the average number of readings per month seems to vary across the months. From the purple line we can see a lot of sharp ups and downs, which shows the difference between cycling traffic on weekdays vs weekends. The orange line is the average of each week, which gives a smoother picture of the changes across the months.

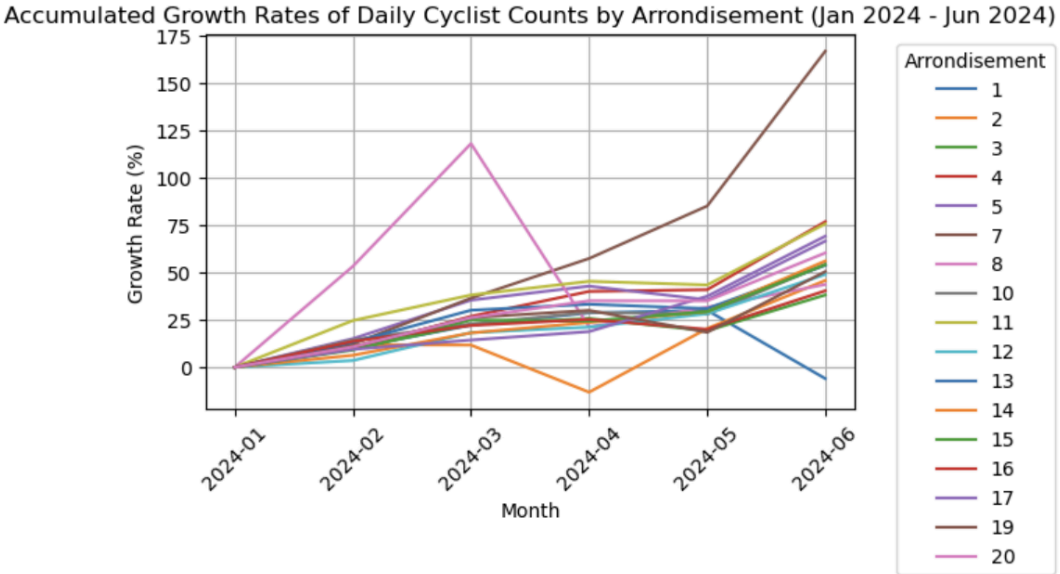
Cycling traffic in Paris starts high in September 2023, decreases during the winter months, then gradually increases through spring and summer, peaking again in June 2024. The decrease in July and August 2024 may be due to road closures or irregular cycling traffic around and during the summer Olympic and Paralympic games, which were held in Paris during that time. Alternatively, this sudden drop may also be due to Parisians leaving the city for summer holidays (Barron's, 2023).

In our later analysis we will try to cluster the counters to get better insights.

Business Insight:

During months with higher cycling traffic, considerations could be made that would benefit cyclists, such as better timing of traffic lights to support steady traffic flow, expansion of bike lanes, and increased bicycle parking. Such improvements could motivate more people to cycle rather than taking private or public transport, thus easing motor vehicle traffic and reducing air pollution in the city.

Moreover, we also briefly explored the growth rate of cycling in Paris. The rate of growth in cycling traffic was calculated from January to June 2024. Specifically, for each arrondissement, each month's total cyclist counts were compared to the counts in January 2024 for that arrondissement (accumulated growth rates). The period from September to December 2023 was excluded for this analysis because average cyclist counts across the city decreased in that period, most likely due to unfavorable weather conditions.



```
Total Accumulated Growth Rate by Arrondissement (Jan-Jun 2024):
Arrondisement
7     166.855059
4      76.831400
11     75.686591
17     69.168885
5      66.589824
20     60.344666
14     56.056378
10     54.351388
15     54.126885
13     53.804672
19     50.506870
12     48.977134
2      45.793007
8      43.668951
16     40.356985
3      38.130416
1      -6.012742
```

Interpretation:

It can be seen that roughly, all arrondissements of Paris experienced growth during this period. The average accumulated growth rate for cycling in all of Paris was 58.5%. It should be noted that increasingly favorable weather from winter into spring and then summer likely contributed significantly to these growth rates. However, since we can assume that the weather was more or less the same in all arrondissements of Paris at any given time, other factors must contribute towards the different growth rates in the arrondissements (note the vast difference between the arrondissements with the lowest and highest growth).

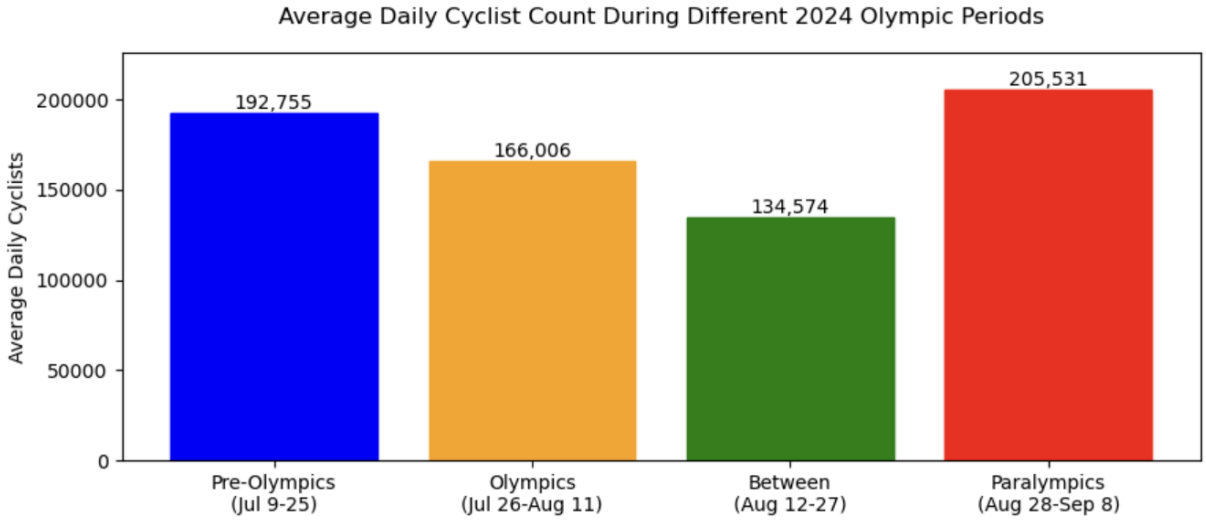
Business Insight:

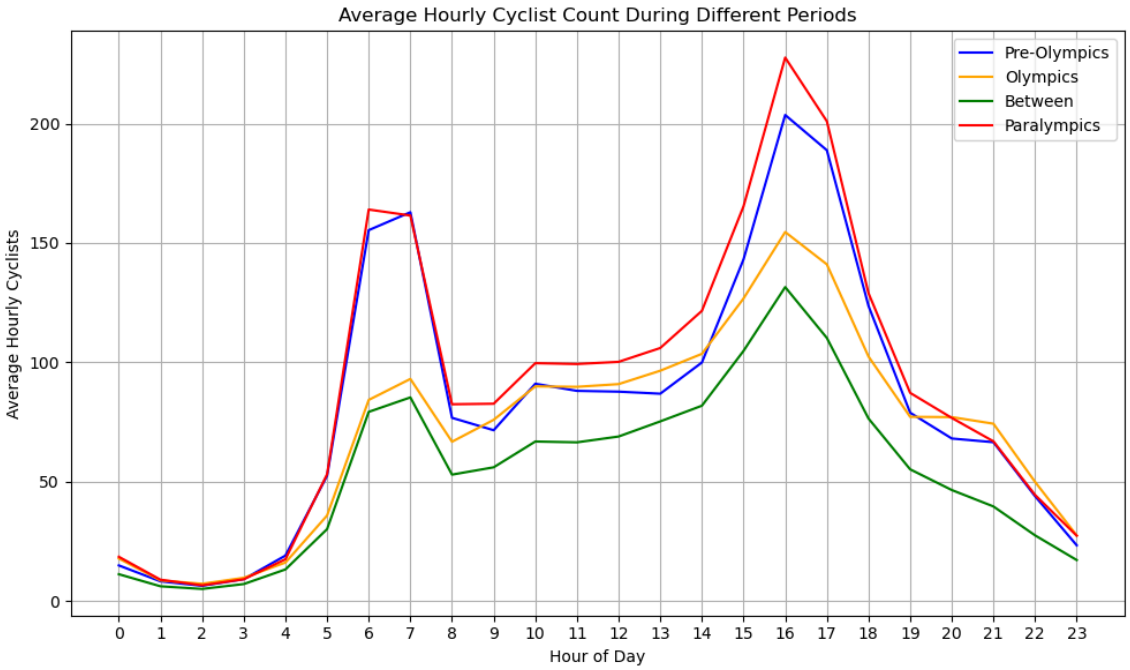
It would of course be more meaningful to draw a comparison - i.e. compare these growth rates to the growth rates for the same period in previous years. However, we are limited by the fact that our dataset didn't contain data for this period in 2023 or 2022. Having an understanding of this would be useful for assessing the effectiveness of improvement measures, and for future resource allocation decisions.

This data could at least be useful as a rough baseline for projecting cyclist counts for 2025, and hypothesizing potential reasons why some arrondissements experience higher growth rates relative to others. With the collection of future data those possible contributing factors could be more carefully examined.

2.5 Olympic and Paralympic Periods

The summer Olympic and Paralympic games were held in Paris in the summer of 2024. The Olympics took place from July 26 - August 11, 2024 (17 days), and the Paralympics took place from August 28th to September 8th, 2024 (11 days). We sought to obtain a first look at the possible effect these events had on cycling in Paris using the following two graphs:

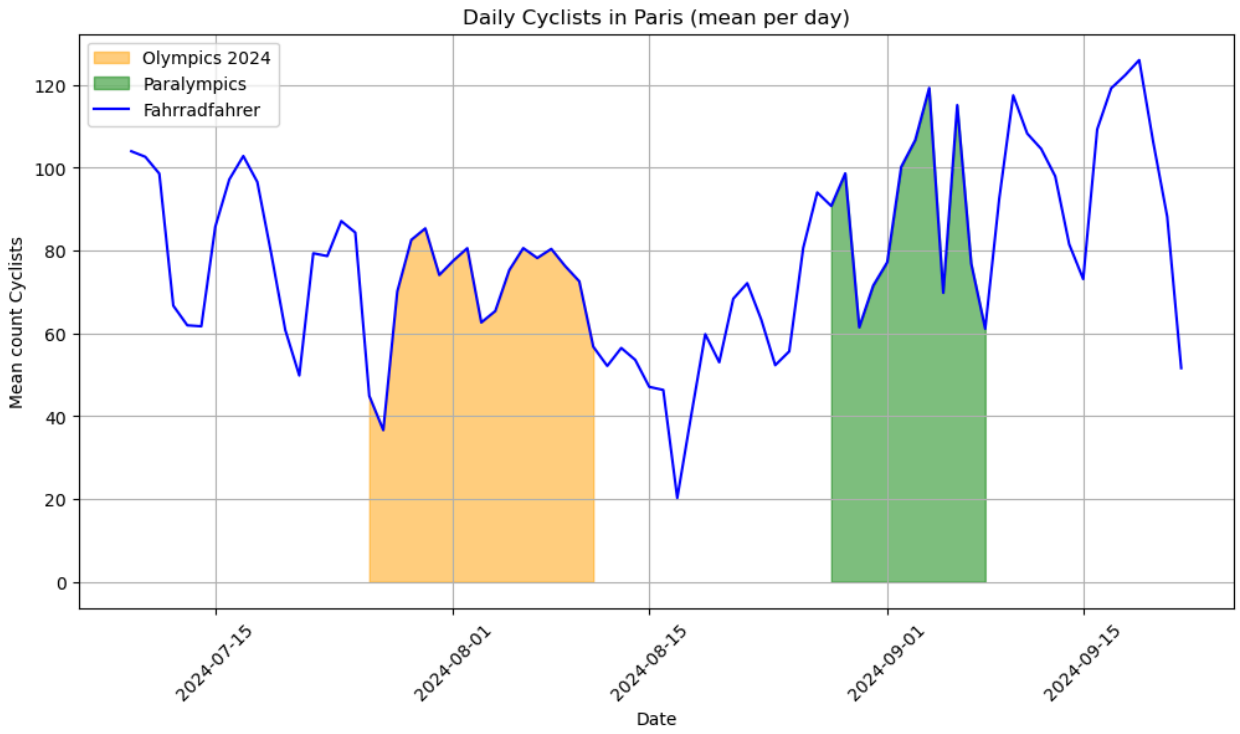




Interpretation:

In the above graphs we compared the different periods: the 17-day period before the Olympics (July 9 - July 25, 2024), the Olympic period, the 16-day period between the Olympics and Paralympics (August 12 - August 28, 2024), and the Paralympic period. The average number of daily cyclists in Paris appears to have decreased over the first 3 periods and then increased again during the Paralympics. Which hours of the day cyclists are on the roads of Paris followed the same general trend (heavier traffic during morning and evening rush hour times).

Below we visualized the average daily cyclist count per day over the entire period.



Interpretation:

From the above graph we can see that the average number of cyclists in Paris appears to increase again during the Paralympic games period as compared to the Olympic period.

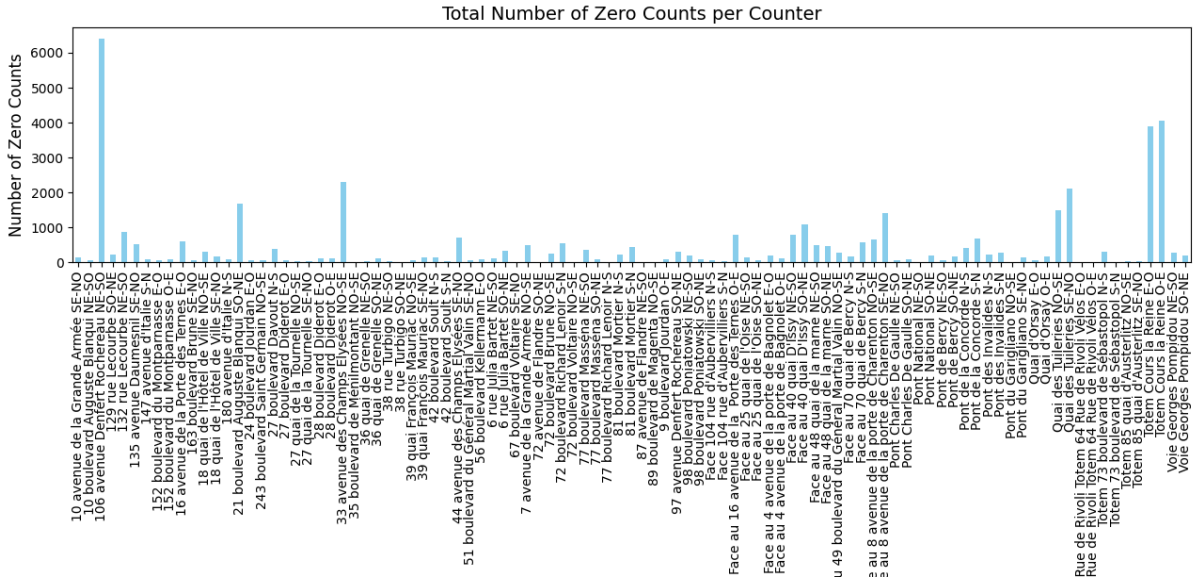
Business Insight:

The city of Paris made significant investments in improving their cycling infrastructure in preparation for the summer Olympic and Paralympic games. These preliminary results suggest that their efforts did not have the intended effect (to increase cycling as a mode of transportation in Paris), but further investigation and consideration of other contributing factors is needed.

2.6 Zero Counts

Given the presence of high values of zero counts in our dataset (so, rows where a counter recorded zero cyclists for that hour), we wanted to see why the counter reads have zero cycling activities in certain sites.

For better insight and analysis, we visualized the number of hours for which each counter recorded a zero count.



Interpretation:

There are 41582 instances (4.5% of the total data) of a counter recording an hourly count of zero. These zero counts are registered at every hour of the day and on every day of the week, but the distribution is not consistent across all the counters in the city. Some counters record significantly more zero counts than others, which suggests that these zero counts are not merely due to periodic malfunction of the counters and that there are other underlying reasons.

Business Insight:

Ensuring that all counters are operating correctly year-round is crucial for data accuracy. Furthermore, if for some reason a counter is blocked and unable to register cyclists (e.g., due to road work), it should be disabled rather than left on to register zero counts. These false zero counts introduce bias to the data. Before continuing, we will need to decide how to handle these cases.

Data Cleaning and Preprocessing

In this phase we will take what we learned in the data exploration and visualization phases to prepare an accurate dataset of cycling traffic ready for feature selection and the model development. This includes detecting and cleaning outliers and controlling for any biases within the data that could compromise its integrity.

1. Outliers

After further investigation (using boxplots and data manipulations), two outliers were identified with unusual reads of 2047 and 8190 counts per hour. They were recorded by counter ‘132 rue Lecourbe NE-SO’ on the same night at 3 am and 4 am respectively. Having this huge amount of cycling traffic at this time is highly unrealistic, especially compared to the other days at these hours. Thus we can be quite sure that both counts were caused by counter malfunction and decided to drop them from our dataset.

Check the range of the highest counter reads:

```
#the rows with highest counter reads:
updated_df.loc[(updated_df['Hourly Count'] >= 1500),
['Hourly Count','Counter Name','Counting Site Name','Date','Time','Weekday']]

#shows that two values higher than the rest of counter reads, both appears to happen in the same date, time (midnight) and weekda
```

Counter ID	Hourly Count	Counter Name	Counting Site Name	Date	Time	Weekday
100003098-101003098	1592	106 avenue Denfert Rochereau NE-SO	106 avenue Denfert Rochereau	2023-10-10	17:00	1
100050876-104050876	1554	Totem 64 Rue de Rivoli Totem 64 Rue de Rivoli ...	Totem 64 Rue de Rivoli	2024-07-23	16:00	1
100050876-104050876	1574	Totem 64 Rue de Rivoli Totem 64 Rue de Rivoli ...	Totem 64 Rue de Rivoli	2024-07-23	17:00	1
100056044-101056044	8190	132 rue Lecourbe NE-SO	132 rue Lecourbe	2023-10-22	03:00	6
100056044-101056044	2047	132 rue Lecourbe NE-SO	132 rue Lecourbe	2023-10-22	04:00	6

Verify other counters records for the same date and time:

```
# Ensure 'Date' column is in datetime format

# Create a mask to filter the data
mask = (updated_df['Date'] == pd.to_datetime('2023-10-22').date()) & \
       (updated_df['Time'] > '01:00') & (updated_df['Time'] < '05:00')

# Apply the mask
result = updated_df.loc[mask, ['Hourly Count', 'Counter Name', 'Counting Site Name', 'Date', 'Time', 'Weekday']]

display(result.sort_values(by='Hourly Count').tail(5))

#others has normal records for the night comparing to those located in 132 rue Lecourbe NE-SO.
```

Counter ID	Hourly Count	Counter Name	Counting Site Name	Date	Time	Weekday
100057445-103057445	51	Totem 73 boulevard de Sébastopol N-S	Totem 73 boulevard de Sébastopol	2023-10-22	02:00	6
100057445-104057445	53	Totem 73 boulevard de Sébastopol S-N	Totem 73 boulevard de Sébastopol	2023-10-22	02:00	6
100047536-101047536	73	102 boulevard de Magenta SE-NO	102 boulevard de Magenta	2023-10-22	02:00	6
100056044-101056044	2047	132 rue Lecourbe NE-SO	132 rue Lecourbe	2023-10-22	04:00	6
100056044-101056044	8190	132 rue Lecourbe NE-SO	132 rue Lecourbe	2023-10-22	03:00	6

2. **Missing Data**

In addition to the year 2022, which was found and handled in the first step, we detected another month with missing data. At the time of analysis, October 2024 had incomplete counter reads (last counter record dated at 09-10-2024). Therefore we will exclude this month from our dataset.

3. **Olympic Period**

The summer Olympic and Paralympic games were held in Paris in the months of July-September 2024. Road closures due to preparation and during the games, as well as unusual traffic patterns around Paris (Parisians were encouraged to take alternative routes to ease traffic during these periods) mean that the counts recorded during these months are not reflective of normal cycling activity in Paris. We therefore decided to also exclude these months from our dataset.

4. **Zero Counts**

Regarding the zero counts identified in the previous step, we used the IQR method to identify the lowest and highest extreme values. According to that, our zero values are not considered outliers.

Given the nature of our data it is indeed possible to have zero cyclists pass a counter in an hour. Though this is more plausible in the middle of the night than during the day, and zero counts were also recorded during the daytimes, it is impossible for us to determine whether a zero count is accurate or not. For that reason, and that the zero counts are not considered outliers by the IQR method, we decided to leave them in our dataset.

Data Preparation and Key Findings

Combining the insights gathered from data exploration and visualization, we cleaned and converted all the columns to their proper formats. We selected the simplest dataset to be used in the next modeling section which covers the periods from Sept 2023 to June 2024. The final dataset includes the following temporal and spatial variables: Hourly Count, Counter Site Name, Time, Weekday, Month, Year, Latitude and Longitude.

Modeling

In this section we will be using our final dataset to build different machine learning models dedicated to identifying main traffic routes and predicting traffic volume in Paris. We will first present models and their features, along with the interpretation of their results and evaluations. This will be followed by our insights and recommendations.

Part one presents our unsupervised machine learning models using two methods (K-mean and hierarchical DBSCAN). The second part is the supervised machine learning model using linear regression method.

Part 1: Clustering Analysis

In order to continue to make meaningful improvements to Paris' cycling infrastructure, it would be beneficial to identify main cycling routes around the city. Therefore, we will conduct cluster analyses to determine the extent to which the usage of various cycling routes around Paris differs. These analyses will identify clusters of counters with similar counts per day over the given time period. By identifying these clusters, we will gain a better picture of how cyclists move around Paris and where traffic is most and least concentrated. We chose to perform these analyses using two different methods to see which produces the best fitting model for our data.

The period used for this analysis will be 01.09.2023 - 30.06.2024 (justified in report 1).

1.1 Clustering the Counters using K-Means Model (4 clusters, 3 features)

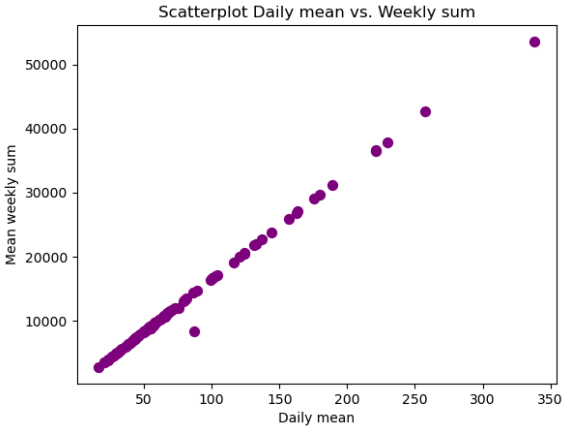
The features selected for our model are as follow:

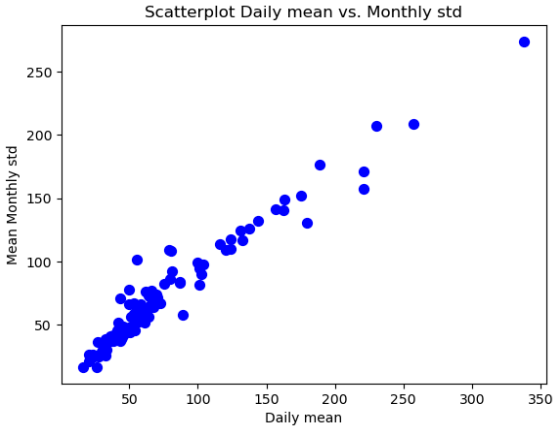
Features:

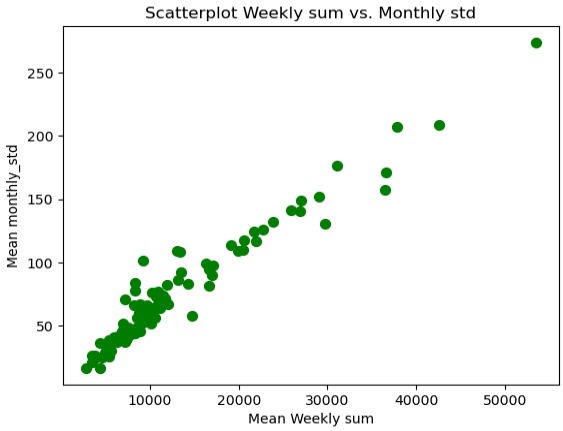
- Mean of the daily count for each counter (mean of the daily counts over the time period)
- Mean of the weekly sum for each counter (mean of the weekly counts over the time period)
- Mean of the monthly standard deviation for each counter (standard deviation of the daily counts as mean of a month)

Correlation of Feature Variables:

The graphs below show the correlation of the respective variables (features). It is clearly visible that the variables are strongly correlated with each other. We therefore assume that the three variables are well-suited for the model.







Target:

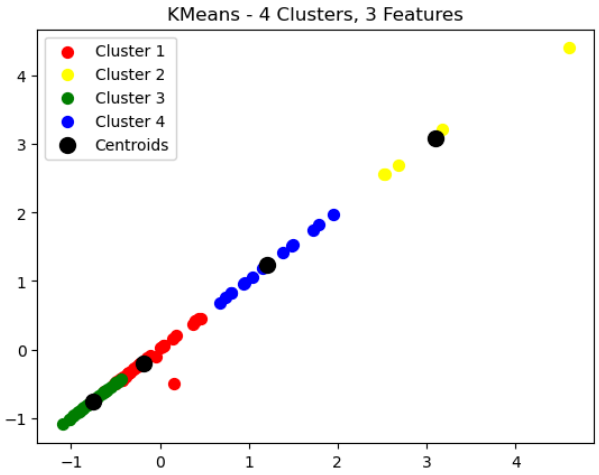
Counter Site Name; the address of the location where the traffic is being counted.

Results of K-means analysis:

The subsequent K-means cluster analysis using the 3 features introduced above shows that the model works best with the formation of four clusters. Here are the results and a visualization of the four clusters:

Cluster 3 is made up of 31 counters (those with the lowest counts per hour)
Cluster 1 is made up of 48 counters (those with the second lowest counts per hour)
Cluster 4 is made up of 14 counters (those with the second highest counts per hour)
Cluster 2 is made up of 5 counters (those with the highest counts per hour)

Scatterplot of Clusters:



Result of evaluation metrics:

- Euclidean average: 2.77
- Manhattan average: 4.79
- Silhouette Score: 0.493

Interpretation:

The evaluation metrics suggest that the K-means clustering model is usable but may benefit from further refinement. The Euclidean and Manhattan averages indicate a moderate cluster compactness, while the Silhouette Score of 0.493 suggests an acceptable but not optimal cluster separation. Potential improvements could involve testing alternative numbers of clusters, refining features, or exploring alternative clustering algorithms to achieve better-defined clusters. Specifically, the distinction between the two smaller groups might be responsible for the poor values. These two clusters overlap, and the count values are very close to each other. This inaccuracy is accepted at this point. Further analyses will demonstrate that this decision is justifiable.

The following silhouette plot shows the clustering quality for the 3-feature/4-cluster model.

Interpretation:

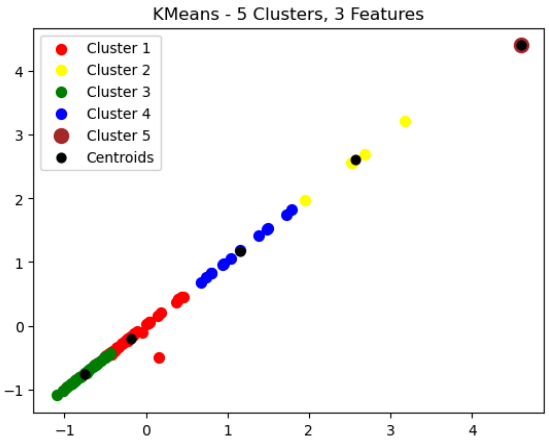
All clusters have positive silhouette coefficients (all values > 0), with the average silhouette width around 0.5 (red dashed line). This indicated that this model has moderately well-clustered data points. Cluster 2 (purple) shows the most consistent silhouette values, suggesting uniform clustering. Cluster 1 (gray) shows the most variation in silhouette values, indicating less consistent clustering. This further supports our assessment that there is room for improvement, particularly for cluster 1 where some points show silhouette values close to 0 (indicating that those points could belong to multiple clusters).

The following graph shows the clusters in the relationship between two of the features, the average daily counts and the average monthly standard deviation. As can be seen, the clustering fits very well. There are no outliers or mismatched groupings. Additionally, the issue becomes apparent that the lowest two clusters are not well-differentiated from each other.

1.2 Clustering the Counters using K-Means Model (5 clusters, 3 features)

In order to come to the conclusion that a 4-cluster model works best for our three chosen features, we also had to perform the K-means analysis with different numbers of clusters. For in brevity, we will only show the results from one of those other analyses.

The following graph and silhouette plot illustrate the result of the K-means analysis done with three features and five clusters.



Interpretation:

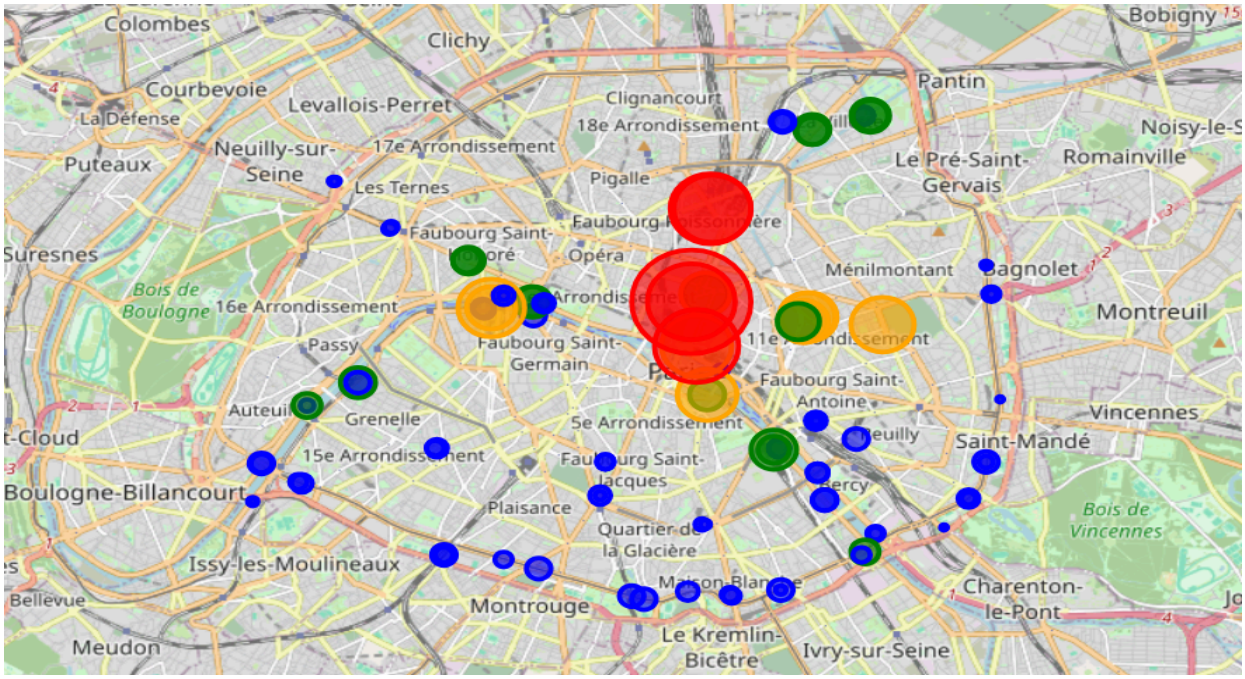
The most apparent difference is that the new cluster (5) is made up of just one point. Since one datapoint isn't a group, this suggests that this model might be forcing the data into too many clusters. The Euclidean and Manhattan averages are both higher here than in the previous model (3.655 and 6.325, respectively), and the silhouette score is also marginally lower (0.489). Given all these reasons, it is our opinion that the model with five clusters does not lead to any additional insights. Therefore, the model with 4 clusters will be maintained.

Cluster sizes:

- Size of cluster 1 : 48 counters
- Size of cluster 2 : 5 counters
- Size of cluster 3 : 31 counters
- Size of cluster 4 : 13 counters
- Size of cluster 5 : 1 counters

Visualization of 4-Cluster 3-Feature Model

So far our best fitting model is the one with four clusters and three features. Below is a map of Paris showing the locations of the counters, color-coded according to their cluster membership, and with the size of the points representing the average hourly count.



Interpretation:

The map reveals that there are some routes that are used very frequently. The first route appears to be a north-south corridor and corresponds to cluster 3 (red). Two of these counters are located near the major train stations in the northern part of Paris, so it is logical that cyclists are traveling south into the city from here. In line with this thinking, the counts remain high in central Paris (mean hourly counts of 337 and 230).

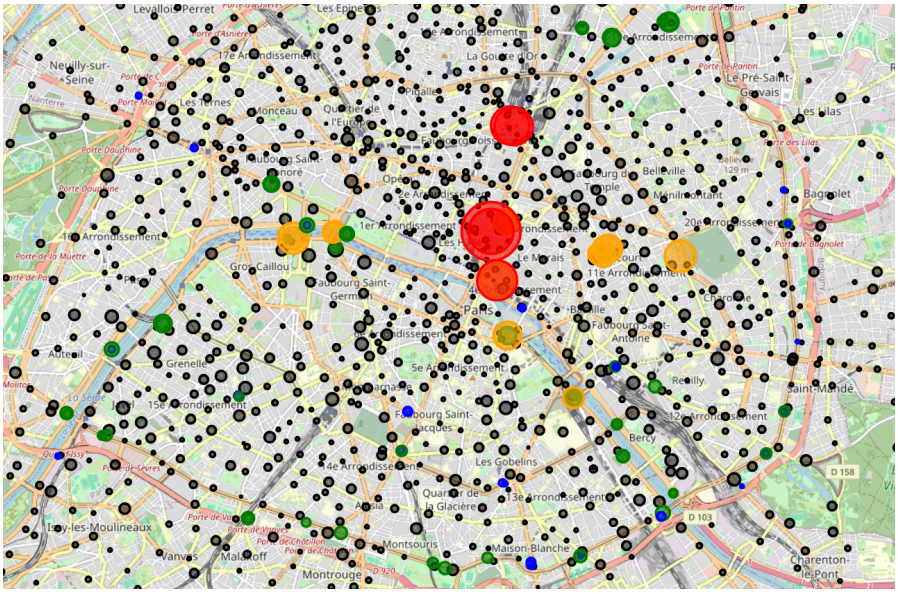
The second route appears to be an east-west corridor along the Seine, corresponding to cluster 2 (orange), with consistently high counts (mean hourly counts of 131, 163, 180, and 120).

The third route forms a loop around Paris (mean hourly counts of 66, 43, 53, etc.) and corresponds to cluster 1 (blue). Cluster 0 (green) has the lowest counts (mean hourly counts of 50, 30, and 20). These counters are located between the other main routes.

We also made an interactive map showing the counters and the clustering, which can be accessed **here**.

Map including Velib stations:

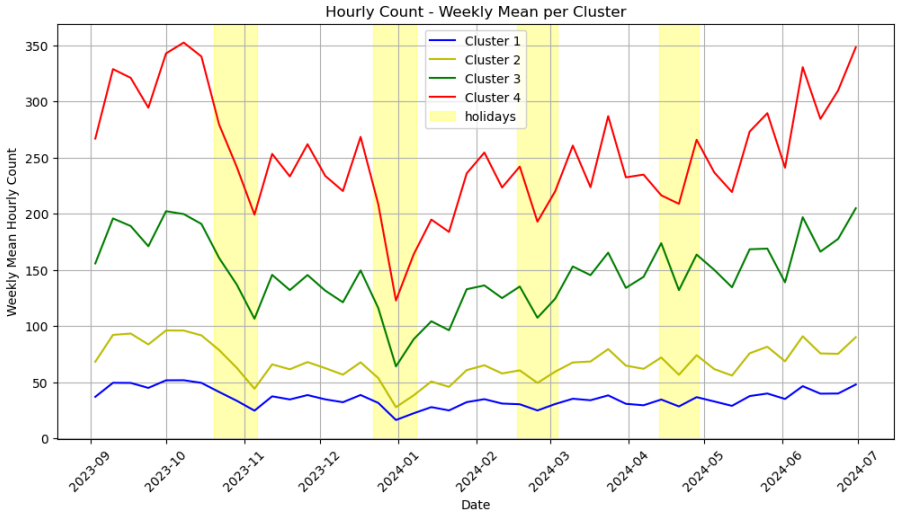
This map shows the distribution of Velib bicycle parking stations across the city (marked in black).



Interpretation:

The city of Paris has invested in bike parking facilities, many of which offer the option to charge e-bikes. There is also an app that provides real-time information on available charging points and the occupancy of bike parking stations.It is clear to see that there are many opportunities in Paris to park and charge bicycles securely. Whether additional parking stations are needed to further promote cycling cannot be determined without further analysis.

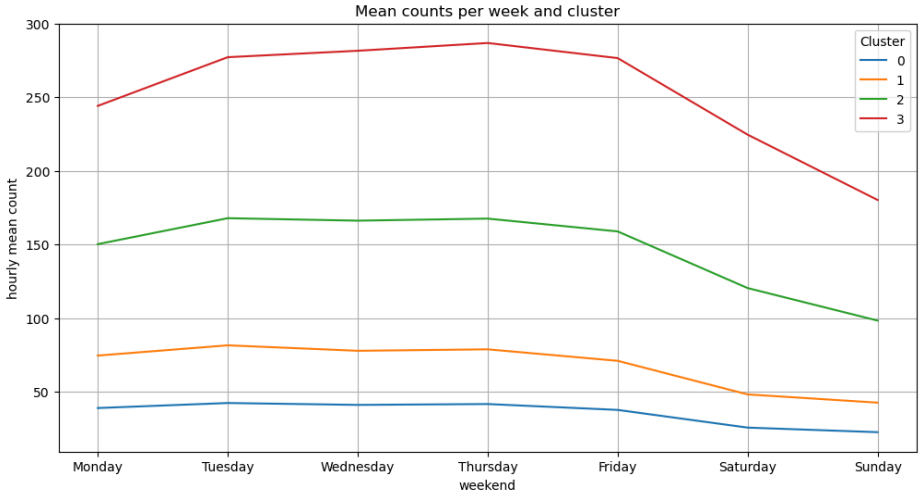
In the following graph, we calculated the average counts per cluster per week.



Interpretation:

It can be seen that all four clusters follow the same pattern, so it can be concluded that the traffic routes are used in a similar manner, regardless of the number of users. Highlighted in yellow are holidays in Paris, where we can see a reduction in counts. This further supports the idea that the residents of Paris are the primary cyclists, and not tourists.

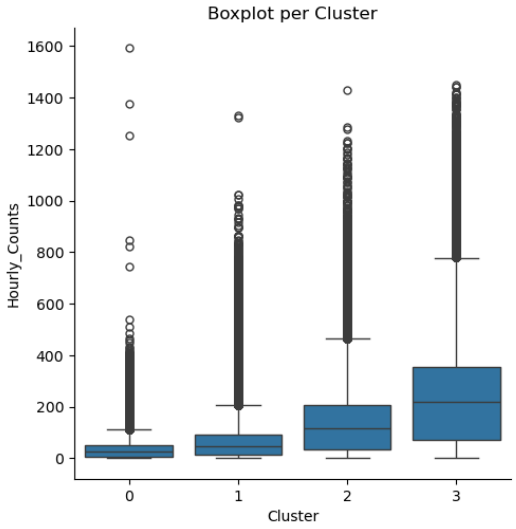
To further investigate the hypothesis that it is primarily the residents of Paris who contribute to cycling traffic, we examined the distribution of traffic throughout the week by cluster. The result is shown in the following graph.



Interpretation:

It is clear that the counts are lower on weekends across all clusters. Therefore, we can conclude that most of the traffic comes from Parisians who travel during the weekdays (for work, school, etc.).

The following boxplot and table show the key statistical information for the individual clusters found in our 4-cluster 3-feature model.



	Cluster 1	Cluster 0	Cluster 2	Cluster 3
count	344194.0	225004.0	101697.0	36014.0
mean	67.79	35.75	147.07	253.08
std	79.25	41.58	140.61	217.016
min	0.0	0.0	0.0	
25%	12.0	6.0	35.0	72.0
50%	46.0	24.0	118.0	218.0
75%	90.0	49.0	207.0	355.0
max	1330.0	1592.0	1429.0	1451.0

Interpretation:

The counters in cluster 1 count an average of twice as many cyclists as the counters in cluster 0. The counters in cluster 2, in turn, count twice as many cyclists as those in cluster 1. The counters in cluster 3 count the most cyclists, with an average of around 250 counts per hour.

From the boxplot it is clear that in all clusters there are many high values. These values are the likely reason for our model's moderate silhouette score (aka clustering quality). It is likely that these values come from peak traffic times. In the next section, we will test a new model that considers these peak traffic times to see whether their inclusion yields a better fitting model.

1.3 Clustering the Counters using K-Means Model (4 Clusters, 5 Features)

The following analysis examines whether the inclusion of additional explanatory variables (features) improves the clustering quality of our model.

To improve on the previous model we decided to add features accounting for peak and off-peak traffic times. In our exploratory analysis we identified that weekdays consistently have more cycling traffic than weekends in Paris. Therefore, we will add in two features that capture this difference.

Features:

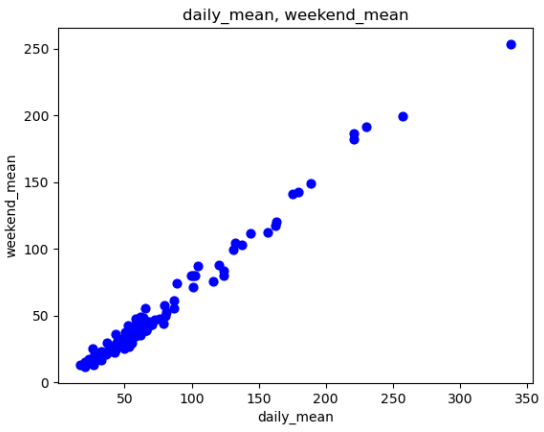
- Mean of the daily mean for each counter (mean of the daily counts over the year)
- Mean of the weekly sum for each counter (mean of the weekly counts over the year)
- Mean of the monthly standard deviation for each counter (standard deviation of the daily counts as mean of a month)

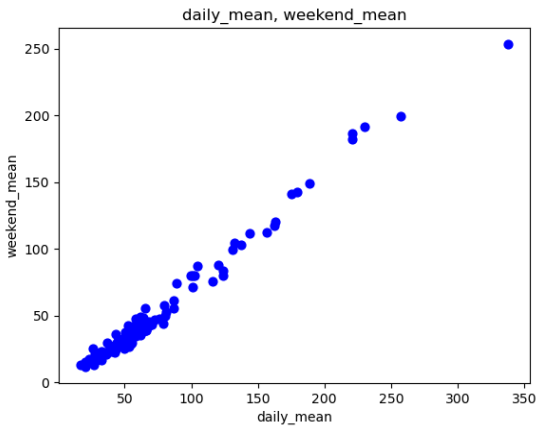
New features:

- Mean of the daily counts on weekdays
- Mean of the daily counts on weekends

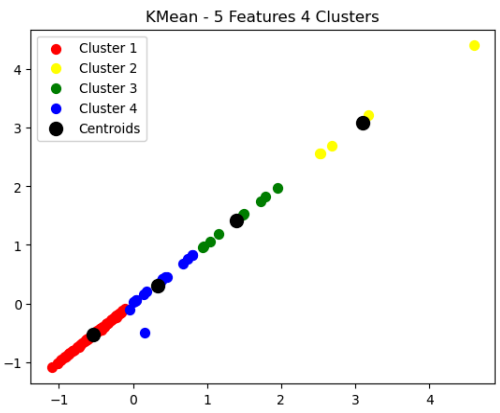
Correlation of the Feature Variables:

In the following graphs the correlation of the new features with the daily_mean feature can be seen. It is clearly visible that the variables are strongly correlated with each other. We therefore assume that the new variables are well-suited for the model.





Results:



Result of evaluation metrics:

- Euclidean average: 3.304
- Manhattan average: 7.379
- Silhouette Score: 0.5359

Interpretation:

The addition of these two new features resulted in some improvement of the model. The Silhouette score increased to 0.536 compared to 0.493 from the 3-feature model. However, the Manhattan and Euclidean averages were higher for this model, indicating poorer definition of the clusters. Since a compact model with fewer features and better clustering is preferred over a more complex one, we conclude that the 3-feature model is still the best one for our data so far.

1.4 Clustering the Counters using K-Means Model (4 Clusters, 9 Features)

Here we will again try to improve on our 3-feature 4-cluster model. In our earlier analyses, we found that cycling traffic in Paris differs greatly at different times of the day. So here we will add in features that account for these differences. Since these values are derived from the already existing variables, the additional information gain is likely to be minimal, but it is still worth investigating to see if we can improve our model.

Features:

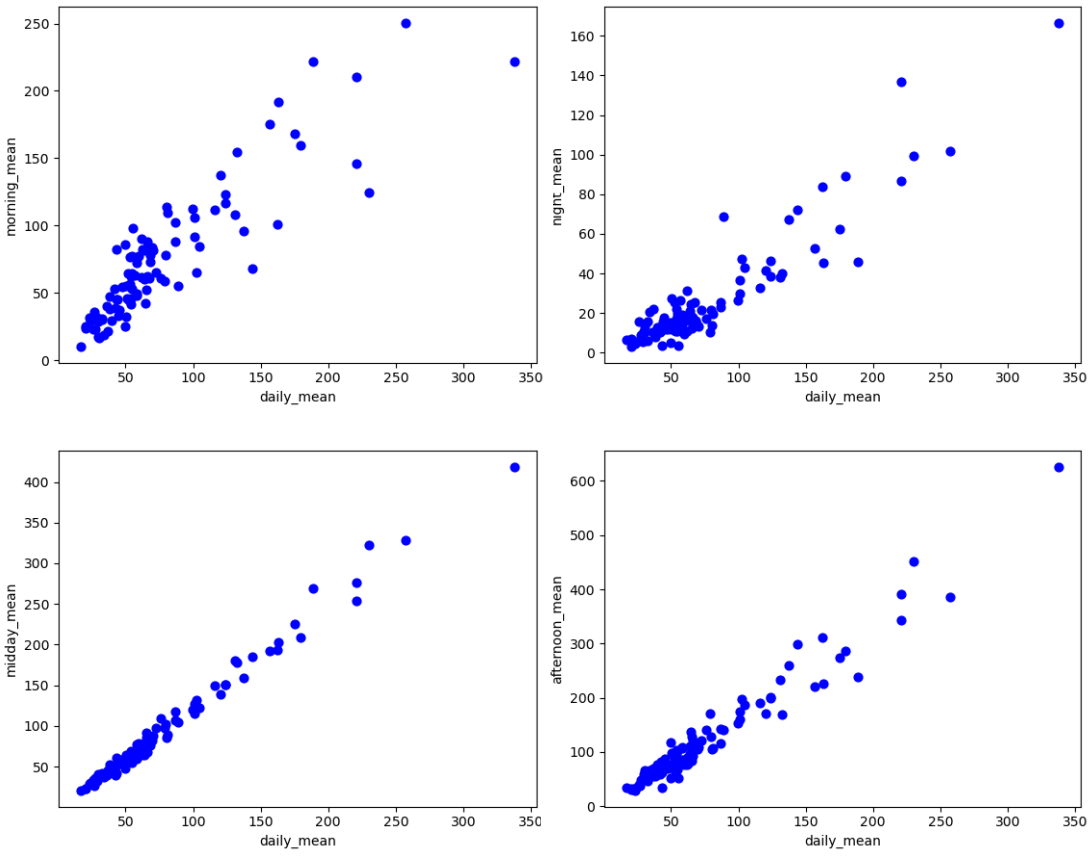
- Mean of the daily mean for each counter (mean of the daily counts over the year)
- Mean of the weekly sum for each counter (mean of the weekly counts over the year)

- Mean of the monthly standard deviation for each counter (standard deviation of the daily counts as mean of a month)
- Mean of the daily counts on weekdays
- Mean of the daily counts on weekends

New features:

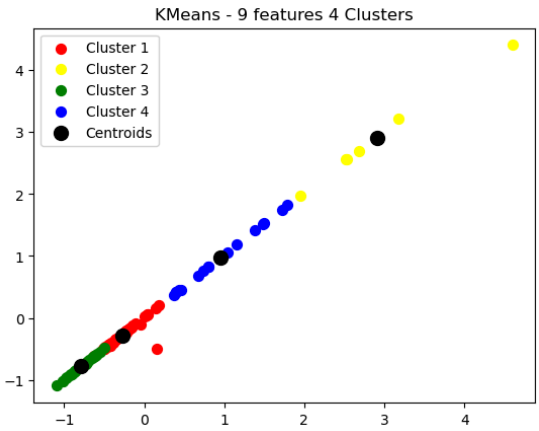
- Morning mean (3:00 - 10:00)
- Midday mean (10:00 - 15:00)
- Afternoon mean (15:00 - 20:00)
- Night mean (20:00 - 3:00)

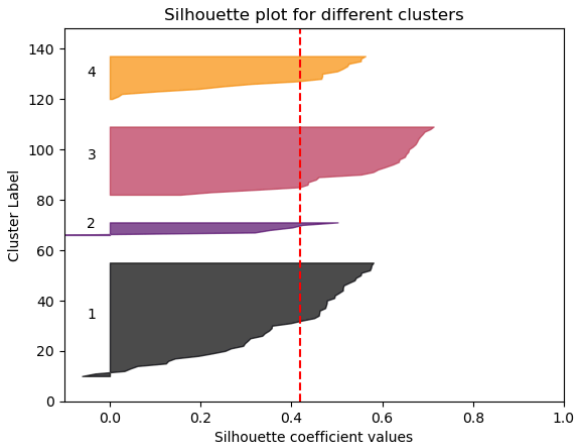
Correlation of feature variables:



Results:

The results confirm our assumption that the information gain from this model was minimal.





Result of evaluation metrics:

- Euclidean average: 4.51
- Manhattan average: 13.476
- Silhouette Score: 0.420

Interpretation:

The clustering shows almost the same result as the 3-feature model. In our opinion, these differences are negligible. Furthermore, the metrics indicate that this model is slightly worse than the compact model with three features.

A comparison of the clusters from the 3-feature model and the 9-feature model shows that a total of 9 clusters would be assigned to a different cluster (the next highest one).

- 89 counters are in the same cluster
- 9 counters are in different clusters:
- counters from 0 to 1
- 5 counters from 1 to 2
- 1 counter from 2 to 3

For all these reasons, we conclude that the 3-feature model offers the best fit for our data.

1.5 Clustering the Counters using hierarchical DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Given the right-skew of our data, we also tried an alternative modeling technique that is more robust to skewed data - the HDBSCAN.

We used two of the same features as in the initial K-means analysis: Mean Count (mean of all daily count values for each counter) and Std Count (standard deviation of all daily count values for each counter).

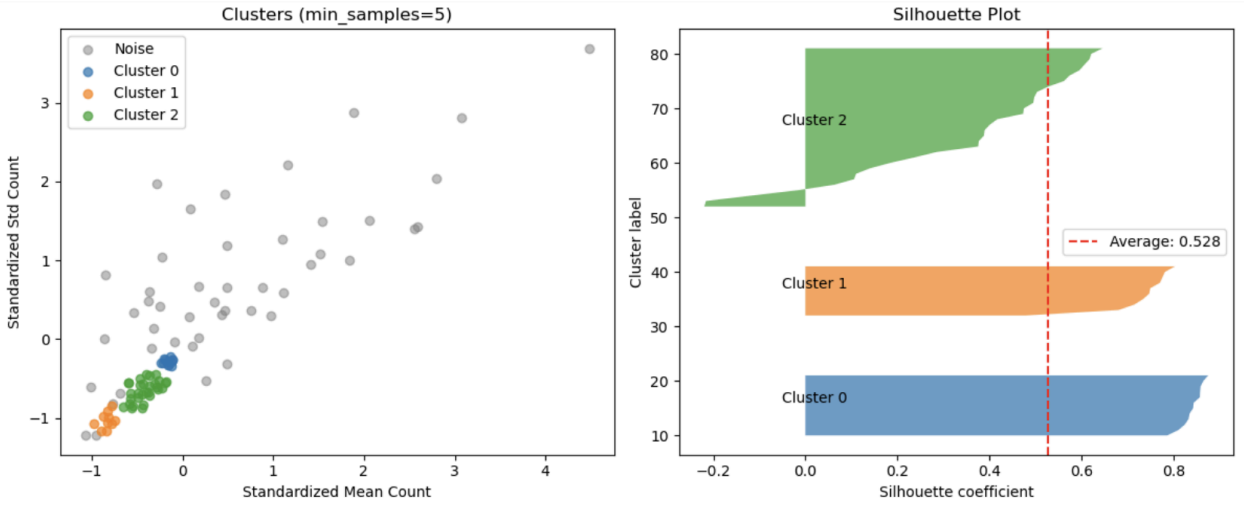
Note: A main difference between K-Means and HDBSCAN is that in K-means the number of clusters is assigned before analysis, and in HDBSCAN it is derived from the analysis. In HDBSCAN, two other parameters are assigned before analysis: the minimum number of neighboring points needed for a point to be considered part of a cluster (min samples) and the minimum number of points needed to make up a cluster (min cluster size).

Attempt 1:

Parameters set at:

- Min samples = 5
- Min cluster size = 5

Results:



Interpretation:

From the scatterplot we can see that this model identified 3 clusters, however, nearly half of all counters (49) were not assigned to a cluster and considered noise (as seen in grey). So the goodness of fit of this model has to be considered with this fact in mind. Note: Another main difference from K-means is that in HDBSCAN not all points are assigned to clusters.

Noise points are those that don't meet the inclusion criteria for any cluster, and are therefore left out (assigned to cluster -1).

The average cluster probability is 0.877, meaning that the HDBSCAN is about 88% confident in its cluster assignments. This is quite a good value, but again, has to be regarded while considering that almost half of the counters weren't included.

The overall silhouette score is 0.528, indicating reasonably good cluster separation. Points (counters) are generally well-matched to their clusters, other than some in cluster 2 that have a negative silhouette score. Overall, this HDBSCAN model performs slightly better than our k-means model, likely because it handled outliers by marking them as noise.

Next, we checked different parameter combinations to identify whether we could produce a better-fitting HDBSCAN model.

```
Top 5 parameter combinations by silhouette score:
   min_samples  min_cluster_size  n_clusters  n_noise  silhouette  avg_prob
9           10                10           2       72    0.675656  0.980544
7            7                 7           3       55    0.570232  0.926494
8            7                10           2       63    0.568920  0.941686
4            5                 5           3       49    0.527671  0.876532
5            5                 7           3       49    0.527671  0.882562

Top 5 parameter combinations by cluster probability:
   min_samples  min_cluster_size  n_clusters  n_noise  silhouette  avg_prob
9           10                10           2       72    0.675656  0.980544
8            7                10           2       63    0.568920  0.941686
6            5                10           3       49    0.527671  0.936022
2            3                 7           3       52    0.504992  0.927160
7            7                 7           3       55    0.570232  0.926494
```

Interpretation:

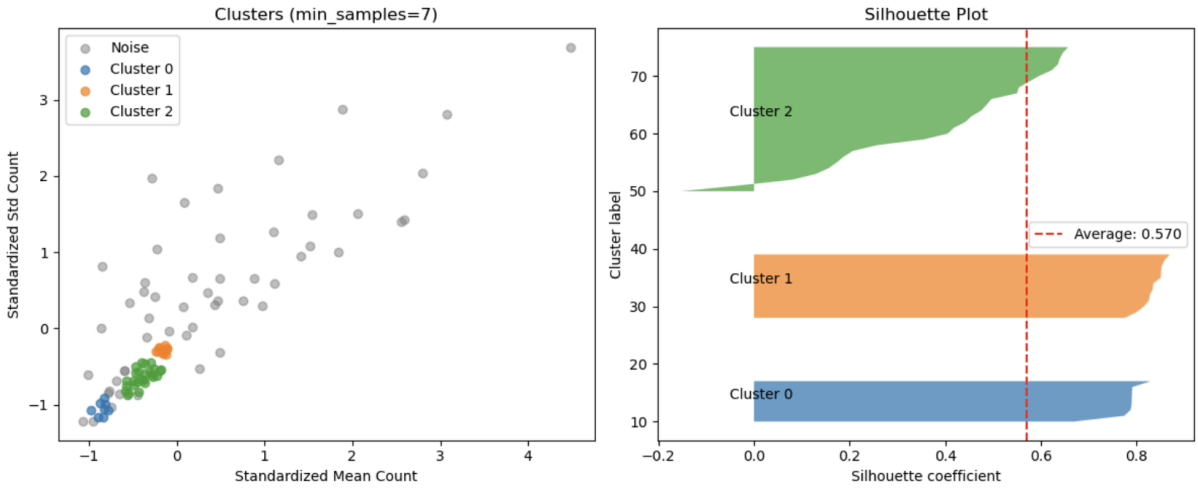
In order to optimize silhouette score and cluster probability, setting the parameters to 10 for both min samples and min cluster size would produce the best results. However, the tradeoff is that this would classify even more counters as noise (72 vs. 49 in the previous model). It would also only give us two clusters, potentially oversimplifying our overview of cycling patterns in Paris. Therefore, we decided to choose a model that still produces 3 clusters and has better performance than the previous model, namely one with the parameters each set to 7.

Attempt 2:

With starting values set at:

- Min samples = 7
- Min cluster size = 7

Results:



Here the Silhouette Score is 0.570 and average probability is 0.926. This model also identified 3 clusters, but more noise points (55) than the previous one.

Interpretation:

The scatter plot shows three clear clusters (blue, green, and orange) in the lower-left region. There are many gray points scattered across the plot, which represent the noise (cluster -1). Cluster 0 (blue) has the lowest mean counts and low variability. Cluster 1 (orange) has the highest mean counts and also low variability. Cluster 2 (green) has moderate mean counts and slightly more variability than the other two clusters.

The silhouette plot shows an average silhouette score of 0.570, which is higher than the previous analysis and means this model has better cluster separation. The shapes of clusters 0 and 1 show consistent cluster membership of those points. For cluster 2 the triangular shape means there is more variation in the silhouette scores.

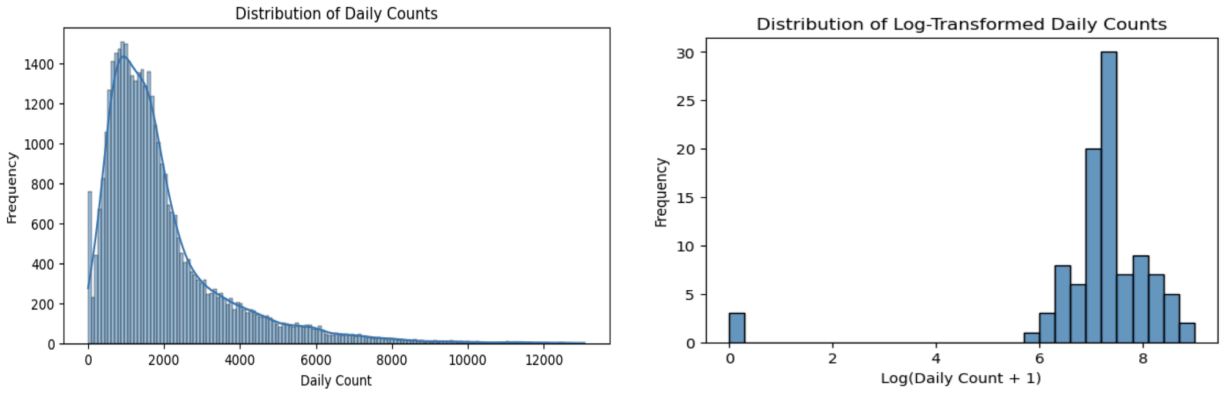
The 3 clusters likely represent low traffic routes (cluster 0) such as residential streets, medium traffic routes (cluster 1) such as main roads connecting neighborhoods, and high traffic routes (cluster 2) such as main thoroughfares. The noise points likely capture very high traffic locations, very variable locations, or very low traffic or inconsistent counters.

In the next section we will perform a log-transformation on the ‘Daily Counts’ data to see if this improves the model.

Log Transformation

Performing a log-transformation on right skewed data makes distances between high values more reasonable, so that very high counts become less extreme. We predict that this will help the clustering algorithms find more natural groupings.

Histograms of the distribution of original (left) and log-transformed (right) daily counts:



Interpretation:

The distribution of the original data is right skewed, as we already knew and sought to correct with the log-transformation. The distribution of the log-transformed data is bimodal, with a small peak near 0 and larger peak around 7-8 on the log scale. The small peak at 0 likely comes from the 41,582 instances where a counter recorded an hourly count of 0, which we identified and discussed in report 1. We decided to keep all the zero counts in our dataset for the analysis, as it would have been impossible to identify which were true and which were errors. We will still keep them for our further analyses, but it should be noted that these zero counts are likely contributing to counters being marked as noise by the clustering algorithm.

1.6 Clustering the Counters using HDBSCAN and log-transformed data

We will also add two new features for this analysis. The features used last time (mean count, mean std) are so general that they could be missing important patterns in the data. We will add weekday mean and weekend mean, as we already observed strong differences between those times of the week in our earlier analyses.

Our features will now be:

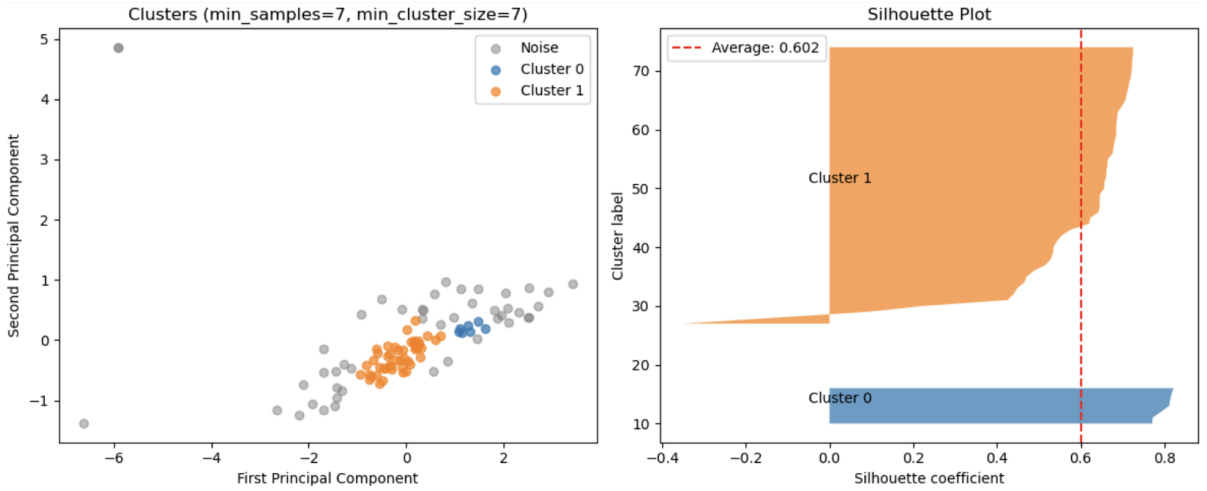
- Daily Count (log-transformed mean daily count)
- Weekend_Mean (log-transformed mean weekend count)
- Weekday_Mean (log-transformed mean weekday count)
- Count_Std (log-transformed standard deviation of counts)

Attempt 1:

We will start with the best min sample and min cluster size values from our last analysis.

- Min samples: 7
- Min cluster size: 7

Results:



Interpretation:

In the scatterplot, the clusters appear to form a continuous gradient along the first principal component (x-axis). There are two notable outlier points both with very low first principal component values (near -6). There are 46 noise counters which is an improvement from the 49 and 55 noise counters identified in the previous models.

From the silhouette plot we can see that the overall silhouette score is 0.602, which is quite good and better than that of our previous analysis (0.570). Cluster 1 (orange) appears to have slightly better silhouette scores overall than Cluster 0 (blue).

So this performance was better in terms of reducing the noise points and yielded a higher silhouette score, but it has only two clusters compared to 3 in the non log-transformed analysis without the weekday mean and weekend mean features.

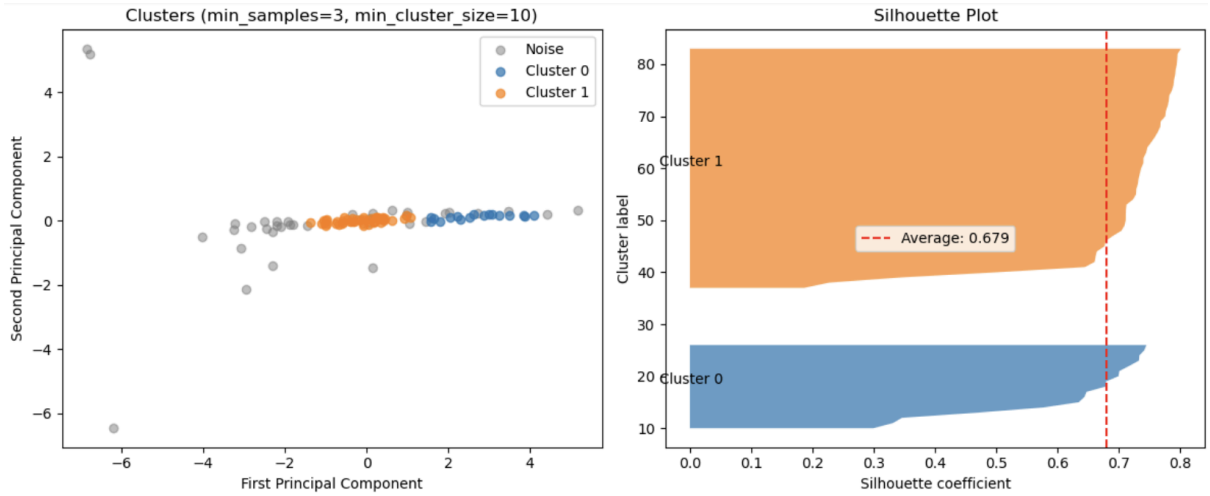
Upon further inspection it was found that setting min samples to 3 and min cluster size to 10 produces a more inclusive model.

Attempt 2:

With starting values set at:

- Min samples = 3
- Min cluster size = 10

Results:



This yields a higher silhouette score (0.627) and clusters 70 points (leaving only 31 noise points) into two clusters.

Interpretation:

In the scatterplot we can see that the data is spread mainly along the first principal component (x-axis). The two clusters appear fairly well separated and follow a somewhat linear pattern. Some noise points are clearly outliers (top left corner, eg).

In the silhouette plot we see that both clusters show strong silhouette values (most above 0.6). Cluster 1 (orange) is larger than Cluster 0 (blue), which is also visible on the scatterplot. There are no negative silhouette values, suggesting good cluster assignments.

Since the features needed to be reduced to 2D for visualization, PCA was used. Next, we determined which features contributed to the PCA loadings.

```
PCA Loadings (how much each feature contributes to each component):
              First Principal Component  Second Principal Component
Overall_Mean                   0.569745                   -0.373713
Overall_Std                    0.418953                    0.644949
Weekend_Mean                   0.517854                   -0.509825
Weekday_Mean                   0.481349                    0.429486

Variance explained by each component:
First Principal Component: 0.662
Second Principal Component: 0.191
```

Interpretation:

We can see here that the first principal component explains 66.2% of the variance. All features contribute positively and fairly evenly to it (0.48-0.57), with the highest contribution from Overall_Mean (0.57). This suggests that this component represents the general volume of bicycle traffic across all time periods.

The second principal component explains 19.1% of the variance. It shows strong contrast between Overall_Std (+0.65) and Weekday_Mean (+0.43) and Overall_Mean (-0.37) and Weekend_Mean (-0.50). This component seems to represent variability and weekday dominance.

Together they explain 85.3% of the total variance in our data, which is very good.

To interpret the scatter plot more specifically now, we can see that as you move right on x-axis the overall bicycle traffic increases, and as you move up on the y-axis the more variable counts become and the stronger there is a weekday preference.

The noise points that are far down on y-axis might be locations with more consistent counts, and/or a stronger weekend performance. The noise points that are far up on the x-axis may be points that show strong weekday/weekend contrast (so, very high weekday counts and very low weekend counts) or and/or have highly variable/inconsistent counts.

1.7 Unsupervised Methods: Key Findings

Below is an overview of our 8 tested models and their silhouette scores. The best-fitting model for each method is in bold.

- **K-Means (4 clusters, 3 features): 0.493**
- K-Means (5 clusters, 3 features): 0.489
- K-Means (4 clusters, 5 features): 0.536
- K-Means (4 clusters, 9 features): 0.420

- HDBSCAN (2 features, min_samples = 5): 0.528 (3 clusters identified)
- HDBSCAN (2 features, min_samples = 7): 0.570 (3 clusters identified)
- HDBSCAN w/ log-trans (4 features, min_samples= 7): 0.602 (2 clusters identified)
- **HDBSCAN w/ log-trans (4 features, min_samples= 3): 0.627 (2 clusters identified)**

It can be seen that all of the HDBSCAN models performed better than the previous K-means models in terms of cluster quality (silhouette score). However, it is necessary to note that the K-means scores include *all* counters while HDBSCAN scores only include non-noise points.

It can be argued that HDBSCAN modeling better suits our heavily right-skewed data (both with and without a log-transformation of the data). But the exclusion of noise counters means these models aren't giving as holistic a picture of the cycling trends as the K-means models which include all counters. The best-fitting HDBSCAN model still identified 31 counters as noise (about ⅓ of all counters).

Since our aim was to gain a comprehensive understanding of how cyclists move around Paris, we will favor inclusion of all counters over higher cluster quality. We therefore choose the K-means model with four clusters and three features as the most well-fitting for our data at this time.

With more time to work on our models, we would first further investigate the counters with a daily count of zero. An hourly count of zero is plausible, but a counter recording no cyclists for an entire 24-hour period on main cycling routes in Paris is highly unlikely and was probably due to another factor (counter malfunction, road work, etc.). These counts are therefore adding bias into our data. As we saw, even the log-transformation wasn't robust to these zero counts, and this likely caused counters to be inaccurately labeled as noise.

After handling the zero counts it would be beneficial to retest both the k-means and HDBSCAN models with the log-transformed data. Only then can a fair comparison be made between the different models for our data. Since HDBSCAN is generally more robust to skewed data, it is likely that it could produce a better model after properly handling the zero counts.

Part 2: Linear Regression

In addition to the cluster analysis, we tested whether it makes sense to combine the three features we used for clustering into a single variable using the PCA method. This seems promising due to the high correlations between the three features.

As in the cluster model, we used the following variables as the basis for the PCA: average daily counts per hour, average weekly sum of counts, and average monthly standard deviation (the same three variables we used for the cluster model).

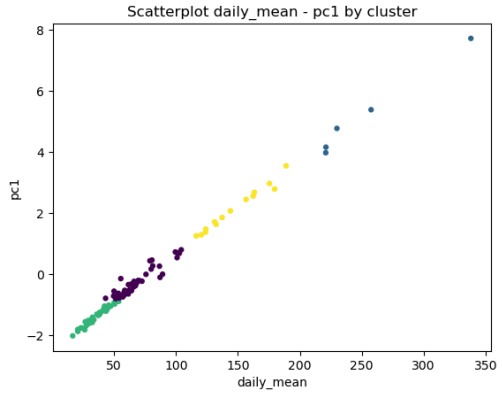
The resulting explained variance ratio was 0.9855 (98.55%), which means that the first principal component captures almost all the information in the original variables. This indicates strong correlation.

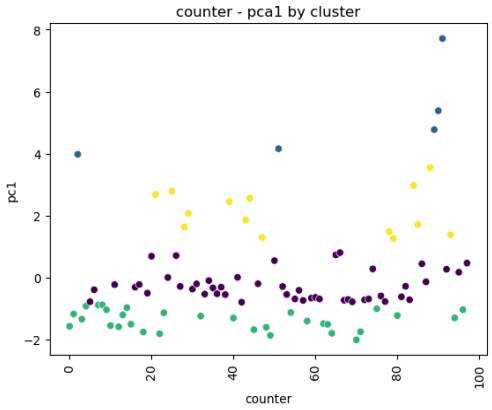
Implications:

1. **Efficient Dimensionality Reduction**: We've achieved significant reduction with minimal information loss.
2. **Interpretability**: Since most of the variance is captured by one component, understanding the data relationships becomes simpler.
3. **Potential Limitation**: If the ratio is too close to 1, it might indicate redundancy or overly homogeneous data, potentially oversimplifying the complexity.

Overall, this result suggests excellent data compression and strong underlying patterns, making it a robust PCA outcome.

The two charts clearly demonstrate that the newly created variable ('pc1') effectively represents the classification of the counting locations. The loss of information associated with using the newly created variable can be considered negligible in favor of the model's simplicity.





The newly created variable, pc1, will later play a role in linear regression. We will include it in the model as a feature variable.

1. **Basic model**

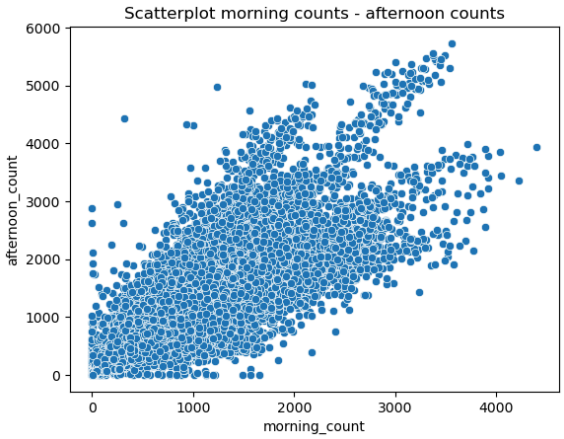
Target: Prediction of the counts in the afternoon by using the sum of the counts in the morning.

The following features were selected:

- Morning: 06:00 to 12:00
- Afternoon: 12:00 to 18:00

Correlation of the features:

Correlation Coefficient morning count - afternoon count: 0.81



Interpretation:

The correlation coefficient of 0.81 indicates a strong correlation between the two variables, and from the graph it is visible that the two are highly correlated. For this reason, a regression analysis appears to be very promising.

Feature variables:

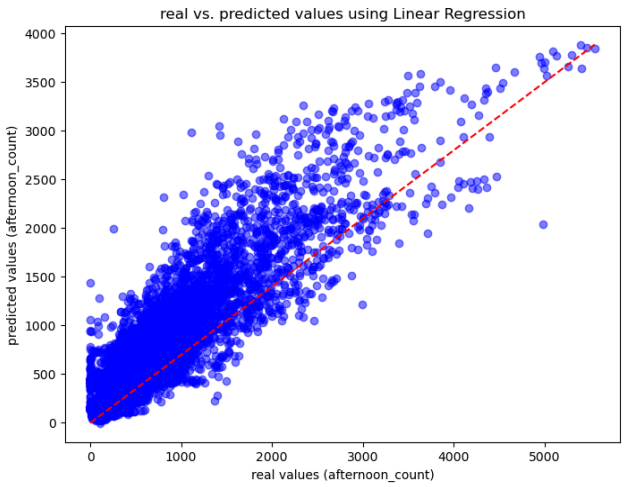
- sum counts per hour in the morning (06.00 - 12.00)
- Day of Week (‘Mon’,’Tue’,Wed’,’Thu’,’Fri’,’Sat’,’Sun’) (values: ‘0’ - no, ‘1’ - yes’)
- Holidays (values: ‘0’ - no, ‘1’ - yes’)
- PCA-Variable (‘-1’ <= value => ‘1’)

Target variable:

- sum counts in the afternoon (12.00 - 18.00)

Result of the Linear Regression:

The following graph shows the result of the regression analysis by illustrating the comparison of the 'actual' values with the predicted values.



- Model intercept: 403.457
- Model coefficient: 0.577
- R^2 Score: 0.786
- Coefficient of determination of the model on the train set : 0.783
- Coefficient of determination of the model on the test set: 0.786
- MSE: 101651.502
- MAE: 214.943
- RMSE: 318.828

Interpretation:

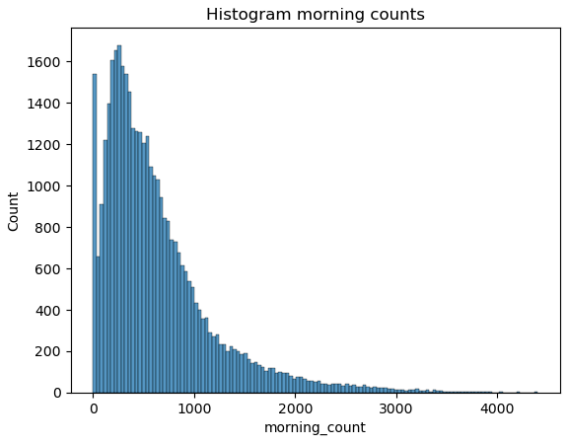
The model shows relatively strong performance, with an R^2 value of 0.786, indicating that it explains a good portion of the variance in the dependent variable. The small difference in R^2 between the training and test sets suggests that the model generalizes well to unseen data. The MAE and RMSE values indicate that the model's predictions are reasonably accurate, but there's still room for improvement, particularly in reducing the error. Overall, the model performs well, but further refinements such as feature engineering or hyperparameter tuning could enhance its predictive accuracy.

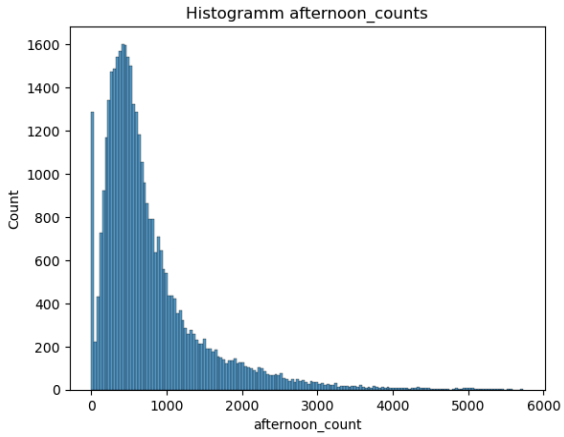
Metric	Estimated value
Intercept	403.457
morning_count	0.577
Mon	- 17.21
Tue	- 2.588
Wed	- 4.199
Thu	- 1.073
Fri	4.015
Sat	9.882
Sun	11.169
Holidays	-50.1567
pc1	193.780

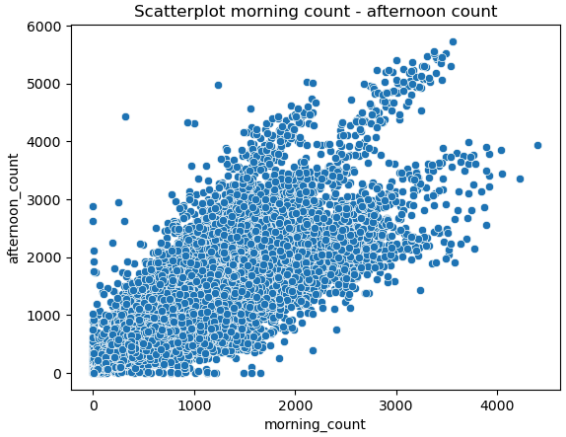
The analysis of the influencing factors reveals that the **weekdays** and **holidays** significantly impact the target variable. Weekends (Saturday and Sunday) have a positive influence, while weekdays (especially Monday to Wednesday) show a negative effect. The **morning_count** has a positive correlation with the target variable, suggesting that more morning counts lead to a higher value in the target variable. The most significant factor is **pc1**, which represents the underlying structure or clustering of the data and has a strong positive influence on the target variable. This indicates that the patterns captured by **pc1** are crucial for predicting the target variable.

2. Revised model

There are a lot of counts with the value '0'. This is the case for morning counts and afternoon counts. We believe that these instances are misleading, as these values mean that the counters did not count either in the morning or in the evening. Further examination of the dataset revealed that there are many additional sums close to "0." For the sake of time savings and simplicity, we decided to remove the lowest and highest 5% of the counts from the dataset and repeat the regression. Subsequently, we will compare both models.The following graphics show the frequency distributions of the sums for the morning and the evening.

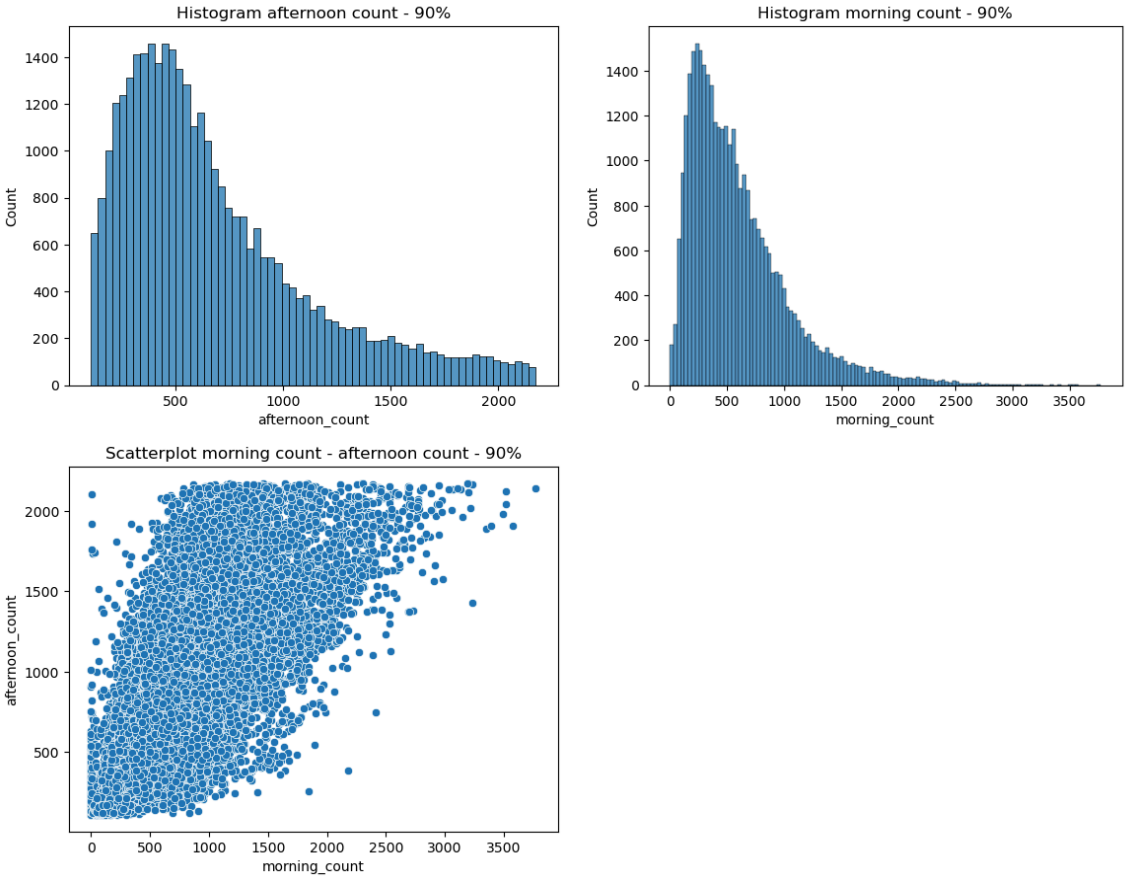




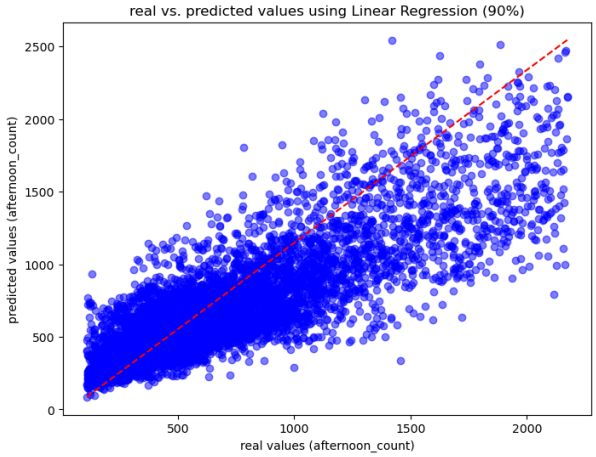


To obtain an adjusted DF, the lower 5% and the upper 5% of the values in the 'morning_count' and 'afternoon_count' columns are deleted.

After deleting the lowest and highest 5% for both variables, the following frequency distribution emerges. The third diagram illustrates the relationship between the two variables:



Result of the Regression with the same feature and target variables :



Correlation Coefficient: 0.75
Coefficient of determination of the model on the train set : 0.701
Coefficient of determination of the model on the test set 0.703

The comparison between Model 1 and Model 2 reveals several key insights:

Metric	Results Model 2 (90%)	Results Model 1 (all counts)
Model Intercept	463.780	403.457
Model Coefficient	0.4389	0.577
R^2 Score	0.703	0.786
MSE	59786.60	101651.50
MAE	178.19	214.943
RMSE	244.51	318.83

Model 2 performs better overall in terms of error metrics (MSE, MAE, and RMSE). It produces more accurate predictions, with lower errors both in absolute and squared terms. This suggests that using 90% of the data might make the model more robust, resulting in more precise predictions. However, the R^2 score in Model 2 is lower, indicating that the model explains less of the overall variance. While it may produce more accurate predictions, it seems less effective at capturing the full relationships in the data. The coefficient has decreased, suggesting a weaker relationship between the independent variables and the target variable. This could be due to changes in the dataset or model configuration.

In summary, Model 2 shows lower errors and more accurate predictions but explains less of the variance and exhibits a weaker correlation between the predictors and the target variable. It might be more robust but less explanatory.

Summary of the Changes in the Influencing Factors:

Metric	Estimated value model 1 (all counts)	Estimated value model 2 (90%)
Intercept	403.457	463.780086
morning_count	0.577	0.438865
Mon	- 17.206	-6.282793
Tue	- 2.588	9.577394
We	- 4.199	9.478100
Thu	- 1.073	11.401913
Fri	4.0149	11.396169
Sat	9.88	-17.640656
Sun	11.169	-17.930128
holidays	- 50.157	-51.852122
pc1	193.780	160.367391

- **Intercept**: Increased significantly, suggesting a shift in the baseline level of the target variable.
- **morning_count**: The relationship with the target variable weakened in Model 2.
- **Weekdays (Mon-Sun)**: The coefficients for the weekdays show more variability between the models, with some days reversing their influence (e.g., Tuesday, Wednesday, Thursday, Saturday, and Sunday). This suggests that the weekly patterns or time-based effects are less consistent or less pronounced in the smaller dataset.
- **holidays**: The effect remains relatively stable across both models, indicating a consistent negative influence of holidays on the target variable.
- **pc1**: The influence of the PCA component has decreased, suggesting that the underlying clustering structure is less important for predicting the target variable in Model 2, possibly due to the reduced data size.These changes indicate that the relationships between the predictors and the target variable are sensitive to the dataset size, and the reduced dataset in Model 2 leads to more variability in the coefficients, especially for time-based variables (weekdays).

To gain a better understanding of the relationships, the focus should be placed on the variable “pc1.” Since “pc1” is a combination of several variables, it is difficult to interpret directly. Therefore, it would be advisable to continue the analysis in the next step using the individual components of the variables that contribute to “pc1.” Another useful step would be to apply alternative regression models, such as Random Forest, as these models could identify non-linear relationships that may not be captured by linear regression. It is also recommended to examine whether modeling with different subsets of the data or across different time periods yields more accurate results, as seasonal fluctuations and short-term changes might be better captured this way. Furthermore, it would be worthwhile to explore whether there are non-linear interactions between the feature variables that are currently not sufficiently accounted for. A deeper analysis of the interactions between the days of the week and holidays could also provide valuable insights.

Conclusion

In summary, our cycling traffic analysis in Paris from September 2023 to June 2024 provided key insights. By cleaning and preparing the dataset, we identified significant traffic patterns, notably high counts on north-south and east-west corridors. The presence of extreme outliers requires further investigation.

Due to time constraints, additional models weren't tested. We recommend investigating other factors that contribute to cycling traffic (such as population growth of the city) to further refine our models.

Our findings highlight the sensitivity of models to data size and the importance of addressing zero counts. While HDBSCAN showed higher cluster quality, K-means provided a more comprehensive understanding. Further refinement will allow for better model comparison and results.

Accurate data is crucial for optimizing model performance, supporting better urban planning, and improving cycling infrastructure in Paris. By integrating additional factors and refining our models, we can achieve more reliable predictions and support effective decision-making for future developments.

References

1. City of Paris. "A new cycling plan for a 100% bikeable city." Paris.fr, link
2. World Economic Forum. "How will Paris become a fully cyclable city by 2026?" link
3. Barron's. "Where Are Parisians Going During the Olympics?" Barron's, October 19, 2023. link