# Project Guidelines

***General instructions:*** Download the *Dublinbikes* dataset at

https://data.gov.ie/dataset/dublinbikesapi. In this project, you will have the chance to work on a larger portion of the dataset, starting from "2018 Q3", at your discretion.

***Scenario:*** You are working for FUTURE-DATA, a local company specialised in data science. Dublin City Council hired your company to study the optimisation of the Dublin bike grid. Dublin City Council plans to modify the stations to improve the user experience. For example, adding new stations or even reducing the bike stands for inactive stations. FUTURE-DATA decided to investigate this by formulating multiple scenarios that should be considered by the City Council. To do so, they assigned the task to k small teams of machine learning and smart and sustainable cities experts. You are part of one of these k teams.

***Task:*** The company proposed 2 goals. The two tasks correctly carried out will give you full points. The two tasks have the same value.

**1. To identify at least 5 bike stations that could be removed or substantially reduced in size, and to identify areas where additional bike stations could be useful for increasing the user experience;**

**2. To predict the time-course for the new bike stations and for the modified stations; to identify a metric for comparing the new city bike infrastructure with the old one, making a case that your solution will actually improve the user experience.**

A few tips: The manager suggested focussing on a subset of the city for simplicity (of course, you are free to do more than that, if you like). Make sure that the data for that bike station is available if you plan on combining datasets from different quartiles. Missing or bad data-points can be a problem. So, identifying stations with good data will make your life easier (but feel free to make your life more complicated if you like the challenge).

# 1 Problem1 Analysis

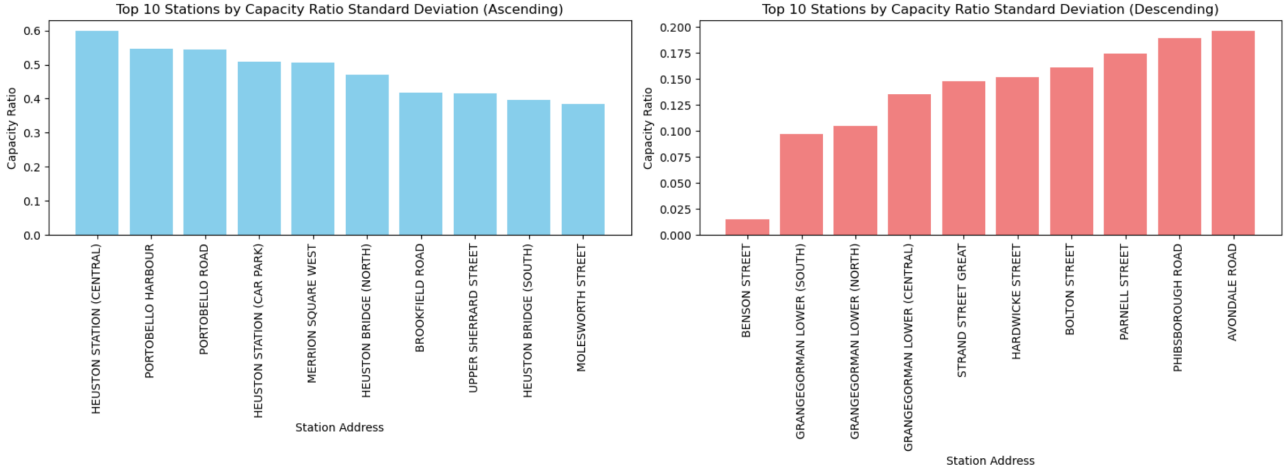
In order to identify at least five bike stations that could be removed or significantly downsized, and to identify areas where the addition of bike stations would help to enhance the user experience, I approached the problem in two ways.

1. The first way is to find the rows where the difference between the mean and the "max" is less than "max1/5" and add a column labelled "always full" after these rows. Find the rows where the difference between the mean and the "minimum value" is less than "max1/5" and add a column labelled "always empty" after these rows.

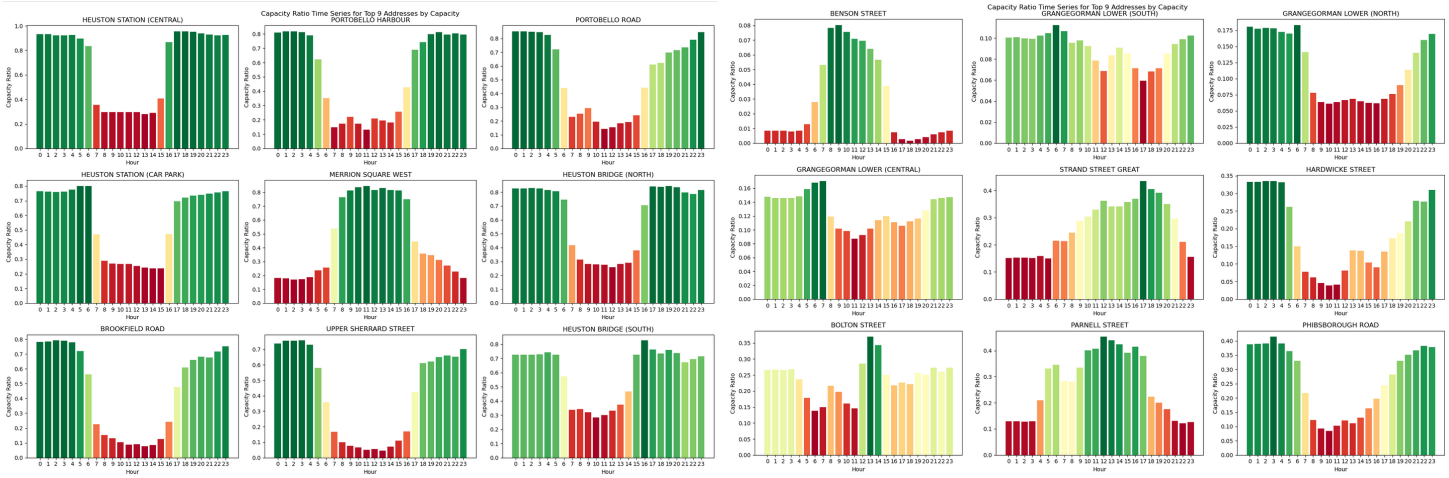
| NAME | count | mean | std | max | max1/5 | label |
|---|---|---|---|---|---|---|
| BENSON STREET | 17337.0 | 4.573283 | 12.075317 | 40.0 | 8.0 | Always empty |
| CONVENTION CENTRE | 17337.0 | 33.163985 | 9.847372 | 40.0 | 8.0 | Always full |
| DENMARK STREET GREAT | 17337.0 | 16.447079 | 4.140136 | 20.0 | 4.0 | Always full |
| ECCLES STREET EAST | 17337.0 | 22.409240 | 6.278922 | 27.0 | 5.4 | Always full |
| GRANGEGORMAN LOWER (CENTRAL) | 17337.0 | 34.871085 | 5.555260 | 40.0 | 8.0 | Always full |
| GRANGEGORMAN LOWER (NORTH) | 17337.0 | 31.810232 | 5.071726 | 36.0 | 7.2 | Always full |
| GRANGEGORMAN LOWER (SOUTH) | 17337.0 | 36.369729 | 4.131142 | 40.0 | 8.0 | Always full |
| HARDWICKE STREET | 17337.0 | 13.007383 | 3.496562 | 16.0 | 3.2 | Always full |
| MERRION SQUARE SOUTH | 11598.0 | 32.504397 | 10.308709 | 40.0 | 8.0 | Always full |
| PARNELL SQUARE NORTH | 17337.0 | 16.303628 | 4.259153 | 20.0 | 4.0 | Always full |
| RATHDOWN ROAD | 17337.0 | 33.372152 | 7.579333 | 40.0 | 8.0 | Always full |

From the picture we can see that convention centre, denmark street great, eccles street east, grangegorman lower (central), grangegorman lower (north), grangegorman lower (south), hardwicke street, merrion square south,parnell square north, rthdown road are always full. Benson Street always empty.

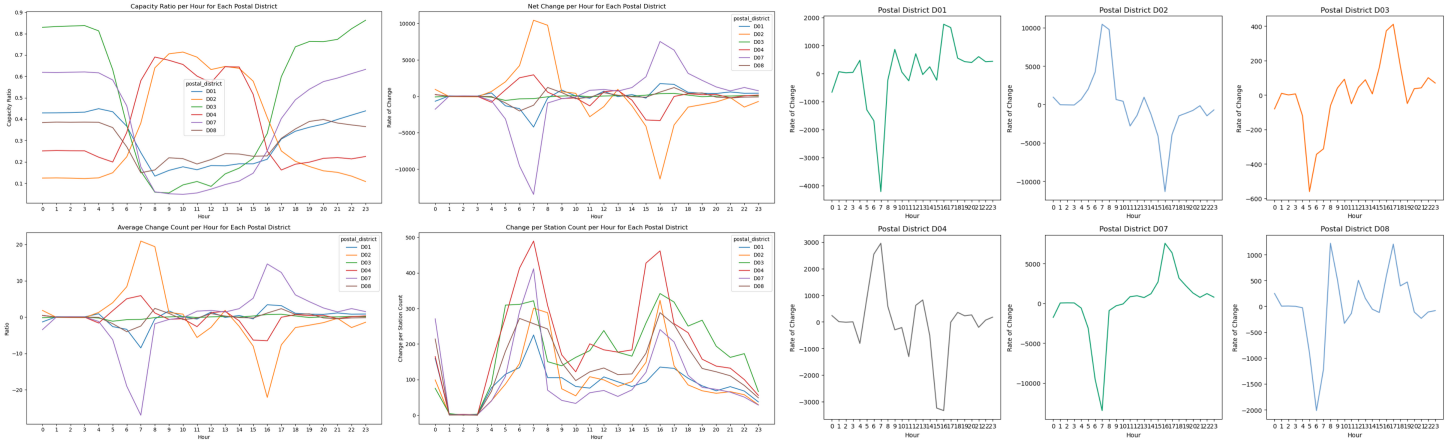
2. The second way is to capture the ebb and flow of bike use at each station. I used the standard deviation of the capacity ratio (available bikes divided by bike stands), focusing only on Mondays for consistency. Higher fluctuations in this ratio may indicate more active station activity. The ten stations with the highest and lowest volatility are listed below.




From the picture we can see that heuston station (central),portobello harbour,portobello road,heuston station (car park),merrion square west,heuston bridge(north),brookfield road,upper sherrard street,heuston bridge(south),molesworth street are always frequently high active.
Benson street,grangegorman lower (south),grangegorman lower(north),grangegorman lower(central),strand street great,hardwicke street,bolton street, parnell street,phibsborough road,avondale road are always frequently low active.
Here's how the ten stations with the highest and lowest volatility changed hourly over the course of a day.



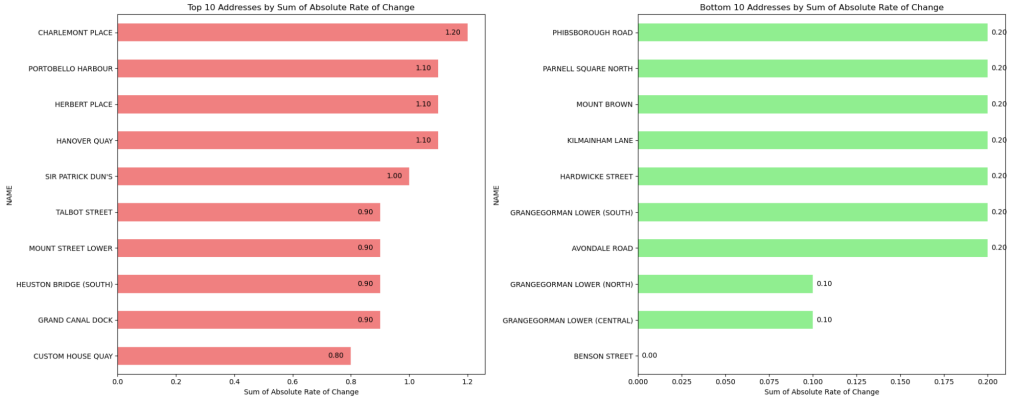

In order to know in more detail where I need to add bike stations, I have analysed bike station usage and I have grouped each station according to the corresponding Dublin postal district. Grouping stations in this way simplifies the data and allows for a focused examination of geographic trends. The dataset I used was limited to weekdays to maintain consistency, as due to weekdays everyone's presence is more regular, whereas weekends tend to vary. This clustering by area allows for the observation of different usage patterns within different areas of the city.




D4 and D2 maintain high levels of usage during working hours, suggesting that they are destinations for people's weekday activities. Conversely, zones outside of D2 and D4 show a decline in usage, suggesting that there is a flow of morning commuters from these zones to D2 and D4.

The right graph reflects the counts of change normalised by the number of available stations and shows that whilst D2 is a popular destination for cycle parking in the mornings, D4 has a higher rate of change per station. This could mean that the relative demand for cycle parking is higher in D4 and that finding free parking in D4 can be quite challenging after peak hours.

I looked again at the top ten and bottom ten stations in terms of average daily train changes.

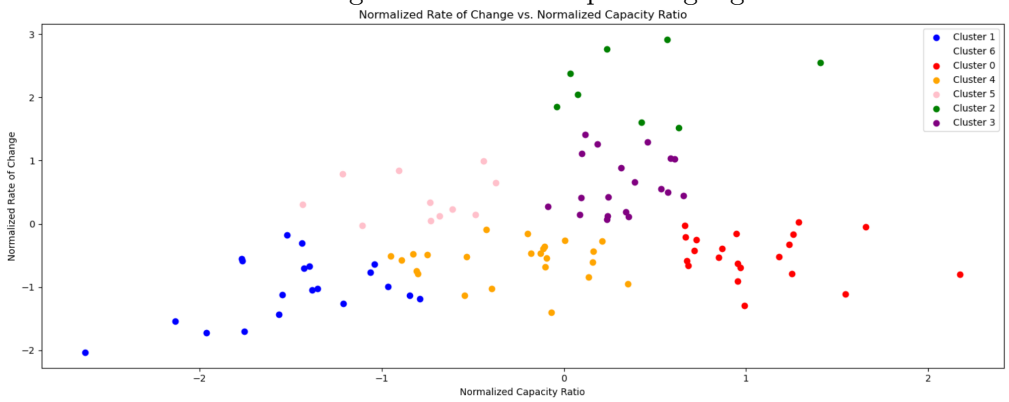



I've summarised which stations are highly active and which are low activity points as shown below.

| Stations to be significantly reduced or dismantled | | |
|---|---|---|
| Always empty or always full | Low activity | Low number of car changes |
| convention centre | benson Street | phibsborough road |
| denmark street great | grangegorman lower (south) | parnell square north |
| eccles street east | grangegorman lower (north) | mount brown |
| grangegorman lower (central) | grangegorman lower (central) | kilmainham lane |
| grangegorman lower (north) | strand street great | hardwicke street |
| grangegorman lower (south) | hardwicke street | grangegorman lower(south) |
| hardwicke street | bolton street | avondale road |
| merrion square south | parnell street | grangegorman lower(north) |
| parnell square north | phibsborough road | grangegorman lower(central) |
| rthdown road | avondale road | benson street |
| benson Street | | |

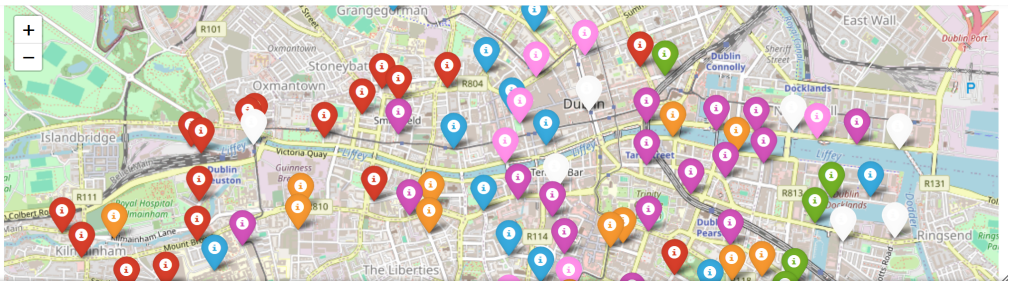
| Under-supplied stations | |
|---|---|
| High activity level | High number of car exchanges |
| heuston station (central) | charlemont place |
| portobello harbour | portobello harbour |
| portobello road | herbert place |
| heuston station (car park) | hanover quay |
| merrion square west | sir patrick dun's |
| heuston bridge(north) | talbot street |
| brookfield road | mount street lower |
| upper sherrard street | heuston bridge(south) |
| heuston bridge(south) | grand canal dock |
| molesworth street | custom house quay |

**Summary**:
In summary, I think:
**Point stations need to be removed or significantly reduced:** grangegorman lower(central),grangegorman lower(north),grangegorman lower(south),hardwicke street,parnell square north,benson Street,phibsborough road,avondale road
**Stations that require additional stations in the vicinity:** protobello harbour, heuston bridge(south)
**Validate**:
I identified low and high activity bike stations in the same way as before and used k-mean to group similarly distributed bike stations together in an attempt to highlight station features.




Cluster Characteristics:

- Cluster Red (0) — Low Capacity Ratio, Low Usage
- Clusters Yellow (4) — Low to Moderate Capacity Ratio, Low Usage
- Cluster Purple (3)— Low to Moderate Capacity Ratio, Moderate Usage
- Cluster Blue (1) — Low to Moderate Capacity Ratio, Moderate to High Usage
- Cluster White (6) — High Capacity Ratio, Moderate to High Usage
- Cluster Green (2) —Moderate to High Capacity Ratio, Moderate Usage
- Cluster Pink (5) — High Capacity Ratio, High Usage




**final conclusion**:
In conclusion, I think: The stations I need to remove are grangegorman lower(central),grangegorman lower(north),gra lower(south),hardwicke street,parnell square north,benson Street,phibsborough road and avondale road.The latitude and longitude of the station I need to add are [53.331232,-6.264900], [53.346683,-6.290708]

## Problem2 Analysis

For problem 2 I did the following:

- data preprocessing: group the raw data frame df, calculate the sum of available bikes, number of bike parking racks, station ID, latitude and longitude according to the station name, and re-index.
- Calculate capacity ratio: Based on the grouped data, calculate the capacity ratio for each station, i.e. the ratio of the number of available bicycles to the number of bicycle parking racks.
- Clustering: Use KMeans algorithm to cluster the stations, define the stations with similar locations and similar volume ratios as a class, and set the number of clusters to 40. The characteristics of the clusters are the latitude and longitude of the station, and the volume ratio.
- Cluster centre and range: Calculate the latitude and longitude range of the centre of each cluster, which is used to subsequently determine the cluster in which the station is located.
- Determine station clusters: For a given list of busy stations [station_names], find the cluster where each station is located.
- Capacity ratio of the new station: Since the new station is near a busy station, I first determine whether the new station is in the busy station's cluster, add a new station to this cluster, and calculate the expected capacity ratio of the new station.

**Result**:
Projected capacity_ratio for new station $[53.331232, -6.264900]$: 0.46253100305704564
Projected capacity_ratio for new stations $[53.346683, -6.290708]$: 0.5002242946583162
**final conclusion**:
In conclusion, I think: From the results we can see that the new stations are all over 50% utilised, proving that the addition of stations at these two locations is the right thing to do to alleviate the lack of supply at the previous stations and make it easier for people to find available bike spaces. And looking at the map and the data, we see that the stations that need to be removed significantly will not have a very bad impact if they are removed, due to the fact that they are very under-utilised and there are other stations within 500 of them, and the nearby stations are not so highly utilised that they need to be shared by this station.