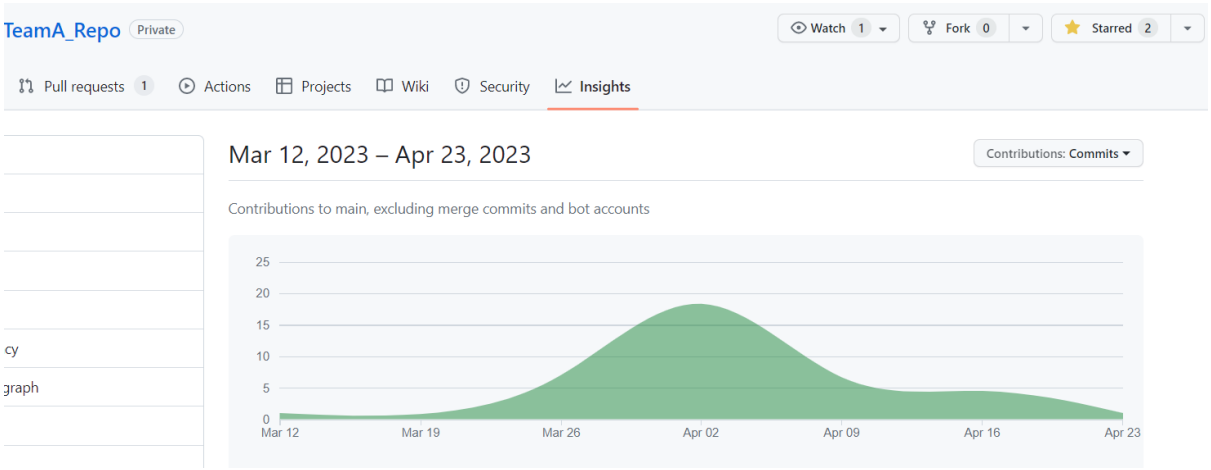
National College of Ireland

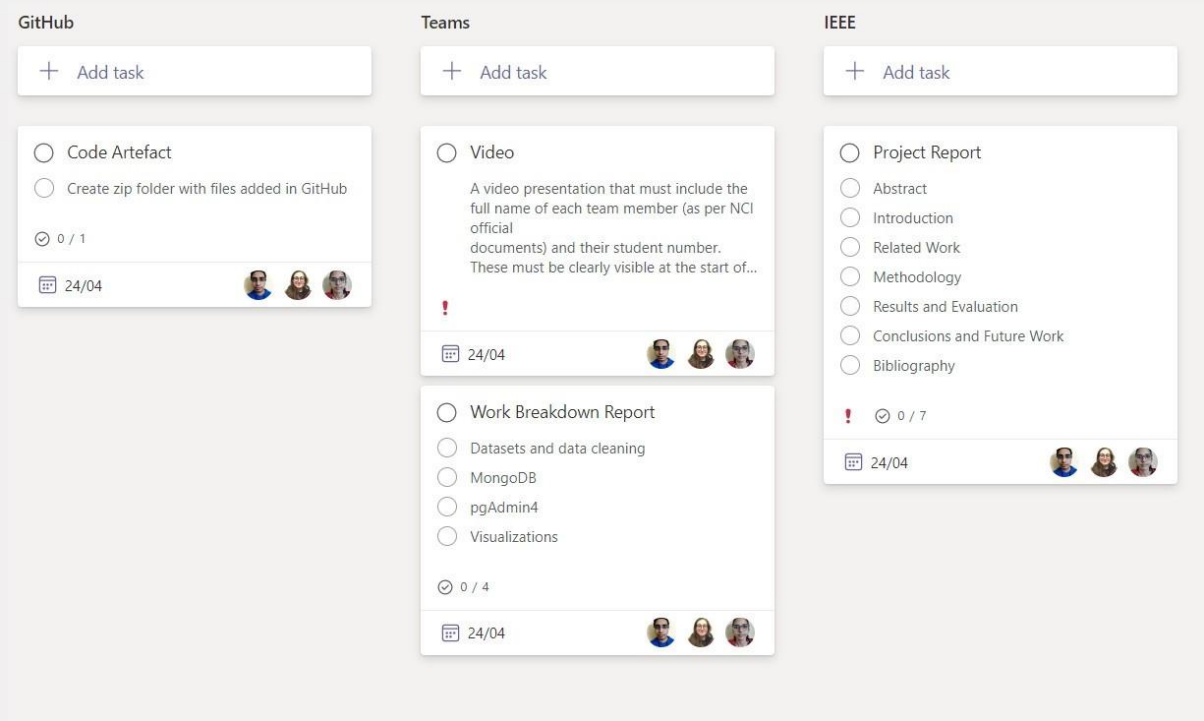
Work Breakdown Report – Giorgia Luzia Pscheidt (x22184261)

Step 1 – GitHub repository creating and dataset searches:

On March 12th, our group project started the search for ideas for the project and create the GitHub group repository. After a few days of brainstorm, we decided to focus on addressing the challenges faced by immigrants, homeless individuals, and renters in Ireland since it was an issue that all 3 of you faced.



Step 2 – We used Planner from Microsoft Teams to organize and keep track of our project status:



Step 3 – Datasets:

I searched for a dataset that satisfied the requirement of 2 semi-structured and 1 structured dataset. After an extensive search, I eventually discovered a suitable combination that could provide valuable insights into the housing crisis in Ireland. Specifically, I was able to locate two .json files and one csv file on the website https://data.gov.ie/dataset?theme=Housing.

Step 4 – Data cleaning and validation:

The data cleaning and validation is an import step, since it ensures we will use appropriate data for our study. For each dataset, I took one 1 sample and in this sample I checked how the data was being pushed for the origin file. The first dataset, emigration_and_imigration.csv, I checked how many values were being pushed, then using the function *info()*, I verified of the datatypes were correct, then I checked if we had any duplicated or missing values. After that, I changed the name of the column for a more user-friendly. Then, I filtered the rows and left only the ones from 2020 and 2021. After these changes, a new .csv file was creates. The second data, *rent.json*, set I went through the same process as the previous dataset, but this one I faced it an issue when I was checking the data type, the data was coming as string, but I knew the year should be a integer in the rent cost should be a float number. So what was happening, when the fire was loaded it was duplicating the name of the columns as an entry, however this was not caught in the function *duplicated()*. To fix this, the row showing the column name was drop it and the data type was changed to numeric. Since this data set was extremely long, I filtered the year, same as the previous dataset, and the data that we weren't going to use in our analysis. For the last data set, *homeless.json*, I run the same procedure as the previous ones, however I didn't find any error, inconsistencies, or inaccuracies.

Step 5 – Uploaded the datasets to MongoDB:

I assisted in debugging and reviewing code to develop a MongoDB database and store datasets.

Step 6 - Visualisation and Insights:

Together with Saheli, we created the visuals using Python's Plotly Express library.

- Our objective with the scatter plot was to display the relationship between rent cost, immigrant, and homeless adults, with marker size representing the number of homeless adults. Also, we adjusted the plot colour as blind-friendly, as suggested in class. It displayed rented location names using a hover tooltip. We changed the title and axis labels to a better understanding.
- We created a line chart by grouping 'immigration_year' and 'homeless_region', then using 'rent_cost' we calculate the average for each group.
- The box plot was created to show the rent costs by region using data from a dataframe. With that, we were able to see difference between the average of rent cost per region.
- A histogram was created to display rent costs by homeless region using data from a Pandas DataFrame called "housing_df". The histogram had 20 bins and used a color map to represent the homeless region.
- We created a grouped bar chart by grouping data by region and calculating the sum of homeless male and female adults for each region in a Pandas dataframe. Our goal with this chart was to identifying gender disparities in the number of homeless adults.
- We used OpenCageGeocode API to obtain latitude and longitude for rent locations, which were added to a new dataframe and concatenated with the original. Our goal with this chart was to show in the map how the rent cost is distributed throw the country.