# VAISHNAVI VISHNU UDANSHIV

# Student No: R00224406

**For the module Data9005 as part of the**

**Master of Science in Data Science and Analytics, Department of Mathematics, 2022/23**

# TOPIC: Homeless People Data In Ireland between 2016-2022

Declaration of Authorship

I, Vaishnavi Udanshiv, declare that the work submitted is my own.

§ I declare that I have not obtained unfair assistance via use of the internet or a third party in the completion of this examination.

§ I acknowledge that the Academic Department reserves the right to request me to present for oral examination as part of the assessment regime for this module.

§ I confirm that I have read and understood the policy and procedures concerning academic honesty, plagiarism and infringements.

§ I understand that where breaches of this declaration are detected, these will be reviewed under MTU (Cork) policy and procedures concerning academic honesty, plagiarism and infringements, as well as any other University regulations and policies which may apply to the case. I also understand that any breach of academic honesty is a serious issue and may incur penalties.

§ EXAMINATION/ASSESSMENT MATERIAL MAY, AT THE DISCRETION OF THE INTERNAL EXAMINER, BE SUBMITTED TO THE UNIVERSITY'S PLAGIARISM DETECTION SOLUTION

§ Where I have consulted the published work of others, this is always clearly attributed

§ Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this work is entirely my own work

§ I have acknowledged all main sources of help

Signed: Vaishnavi Udanshiv

Date: 06 Mar 2023

**Introduction:** The report aims to summarize my work on visualization on data of homeless People in Ireland between the year 2016-2022. The report include detailed exploratory Data analysis and visualization on the available variables.

**Where did you get the idea/code for this concept/implementation?**

I have been here in Ireland for about 6 months now and I am seeing a lot of people struggling with the accommodation due to the increase in demand. I was curious to find out more about the accommodation crisis in Ireland and how it is affecting the lives of people. Through this analysis I was sure that I would be able to find the trend of homelessness in Ireland for the past couple of years and would be able to answer key questions such as Which region in Ireland is facing this issue the most? What is the age group that's most affected by this?and what type of accommodation do they access in such situations? This is where I got the idea to implement visualization on this topic.

**Has anyone else published visualizations (on the internet etc.) for this dataset?**

I can not see any single visualization published on the internet which gives complete overview of the homelessness in Ireland. I could see visualizations for dec 2014 to 2016 And some basic graphs on different sectors. But i couldn't find any visualization which Gives the overall trend and information about homelessness from year 2016 to year 2022 which includes the latest information.

**What is novel about your approach in this assignment?**

My aim through this assignment was to understand, learn and implement the complete data visualization lifecycle from scratch. When I was looking for the data, I found one dataset on kaggle which had several csv files. I didn't want data where I can directly get a csv or excel file and start data cleaning and visualization over it. Therefore I chose to get the data from the official website of the government of ireland. The data was in the form of a pdf. There were pdfs for each month and year. The pdf had different tables for the different components like age, accommodation type, total adults, families based on regions etc. So it didn't have a direct single data table to extract and use. It had multiple tables and the column names weren't aligned to the data properly. So I started with the first step in data analysis that is data scraping. I scraped the data from each pdf and exported it into csv formats. After all the csv files were ready, I imported them into a list of dataframes using python. Once I had a list of all the dataframes I renamed column names for each data frame so that the data will get merged properly. Once the column names were changed I merged all the data frames in a single dataframe. Individually every data frame had around 9 rows and 14 columns. When I combined the main dataframe now had 785 rows and 17 columns. Two columns were not part of the data and were added accidentally so I dropped them. One extra column with the name 'Perios' I added explicitly by extracting the year and month from the csv file names and concatenating that period with each dataframe. With this my main data frame was ready and I exported it into the csv file. Did some basic formatting using Microsoft Excel and again imported the new csv file in the notebook and then started with exploratory data analysis and data cleaning. As a part of cleaning, I checked for null values, duplicate values and removed them if they existed. Checked for the data types of each column. Performed some summary statistics operations and then started with the visualization. One more thing that I think is novel about my approach is I pretty much tried to include various types of visualizations to get hands on various concepts.

**Detail at least one complex concept/implementation in your work?**

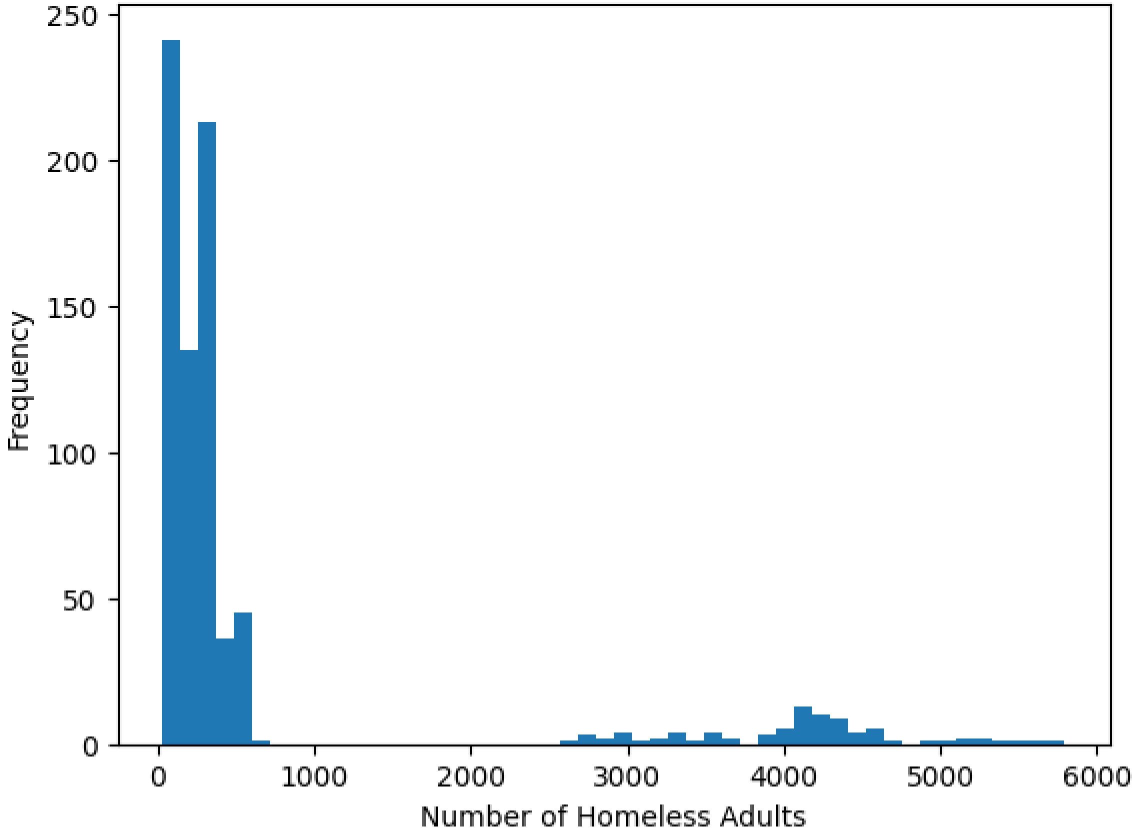
I think data preparation was one of the most complex implementation parts in my work which took the majority of the time. As the data wasn't getting extracted in proper format sometimes, I faced issues with the data types which were not creating the visualizations properly. Apart from that maps were the second most complex implementation. I used an opencage geocoder to get the longitude and latitude of the region and plotted the number of homeless adults for each region on the map with the help of folium library. Apart from that I tried creating one interactive visualization using plotly which states the number of homeless adults have accessed other accommodations in each region.

**Discuss at least one theoretical consideration/choice you made.**

Theoretical consideration that i made was that dublin being most popular and populated region I assumed it would have the most number of homeless people. On analysis this theory proved to be True. As from 2016 to 2022 the number of homeless people in dublin region are far more Than other regions when compared. The second theoretical consideration that i made is the Homeless people would be more between 25-44 age group considering a lot of people Migrating to Dublin for either work or studies in this age group. Which was again see to be True during the analysis. When compared to other age groups people between 25-44 are Homeless in huge numbers followed by people between 45-64 age group. And there are Much lower number of homeless people with the age 65 and above. Another theoretical consideration was that number of homeless people have been increased drastically between year 2021 and 2022 which again can be significantly seen through time series Plot.

**What aspect of data analysis did you investigate (e.g. pattern recognition, distribution comparison, statistics etc.)?**

I started the visualization with frequency distribution of column Homeless Adults which denotes the number of adults who are homeless. From the graph below we can clearly see that the frequency of adults between 0 to 1000 is highest compared to other numbers. So we can conclude from this graph that number of people that are homeless in maximum region throughout past 7 years is somewhere between 0 to 1000.

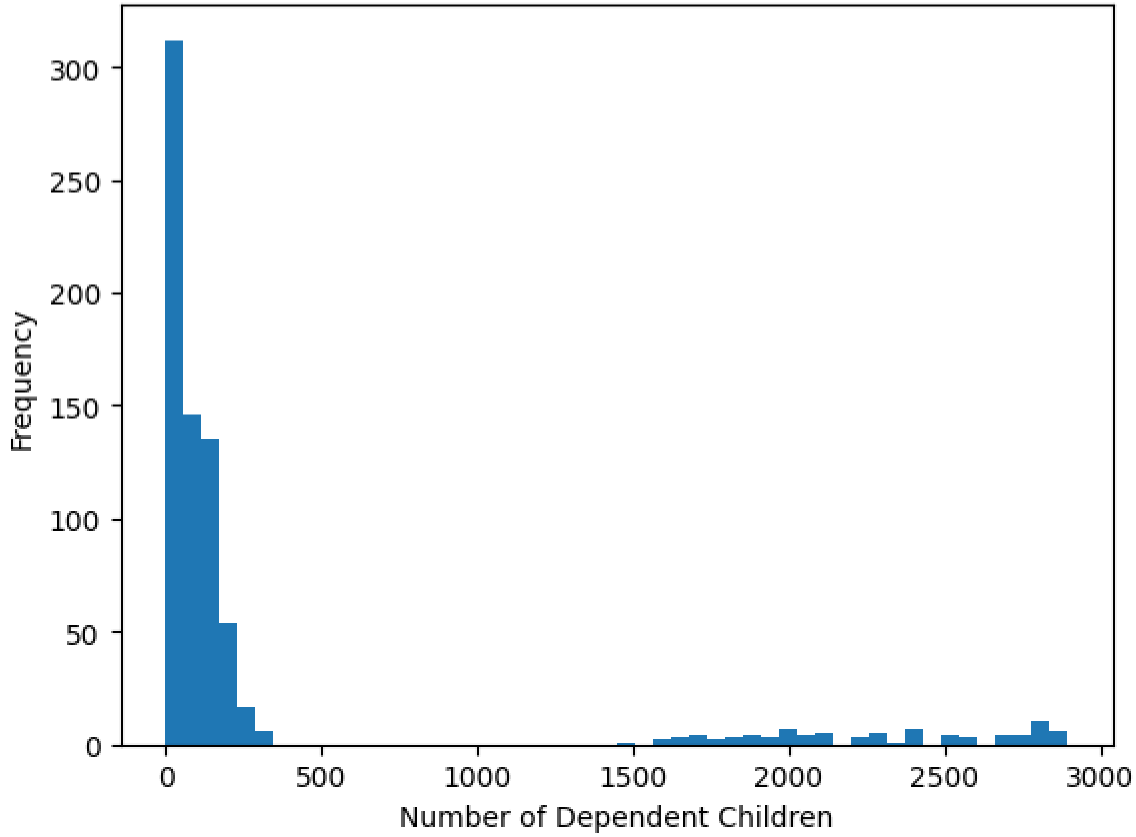


2. Summary statistics of column dependent child:
count      755.000000

```
mean      332.495364
std       709.700150
min         0.000000
25%        34.000000
50%        74.000000
75%       162.500000
max      2894.000000
```

3. Frequency distribution of number of dependent children

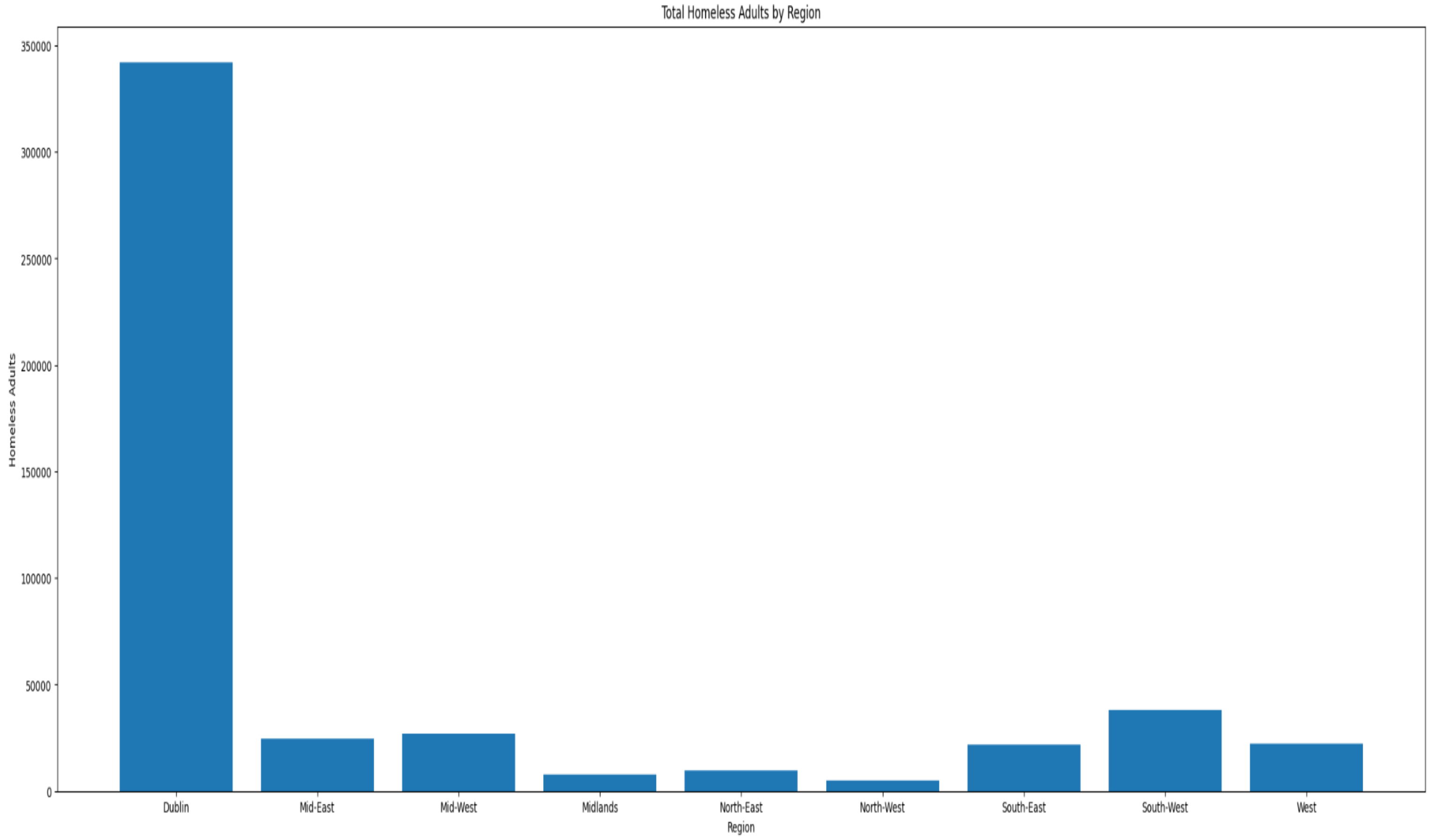
From the graph below we can conclude that the maximum number of dependent children who are homeless lies between the range of 0 to 500.



4. Count of homeless adults for each regions

```
      Region  Homeless Adults
0     Dublin           341834
1   Mid-East            24596
2   Mid-West            26915
3   Midlands             7565
4 North-East             9751
5 North-West             5187
6 South-East            21687
7 South-West            38241
8       West            22134
```

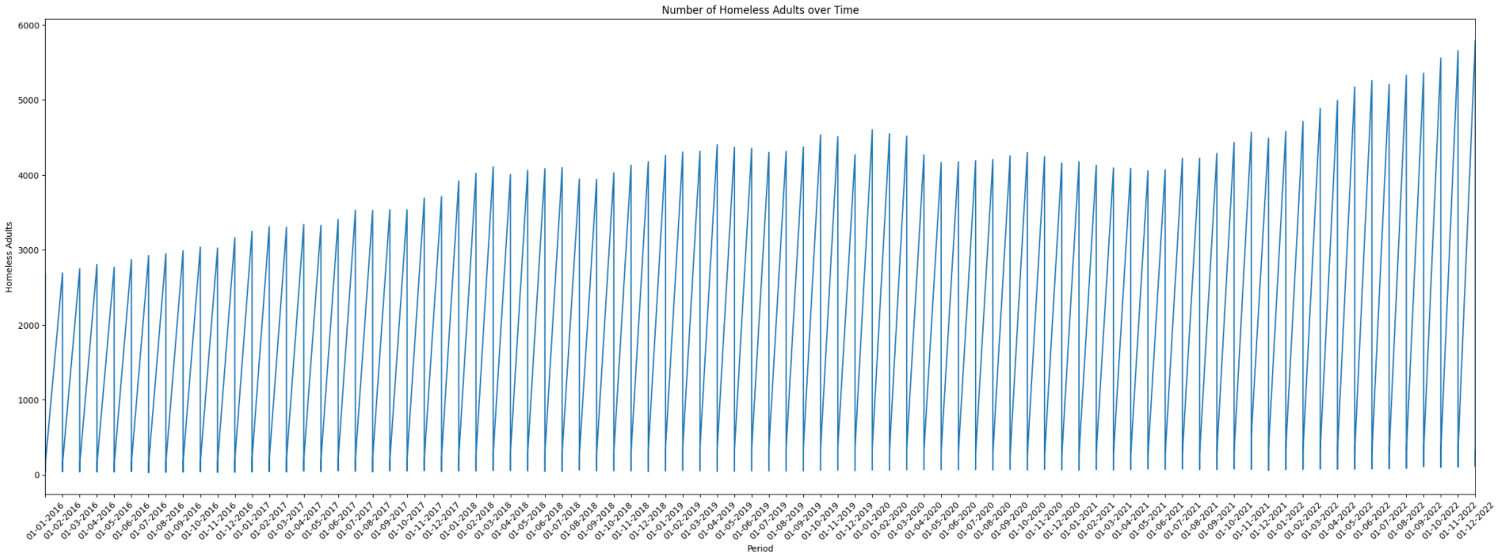
5. Bar plot to show the above counts



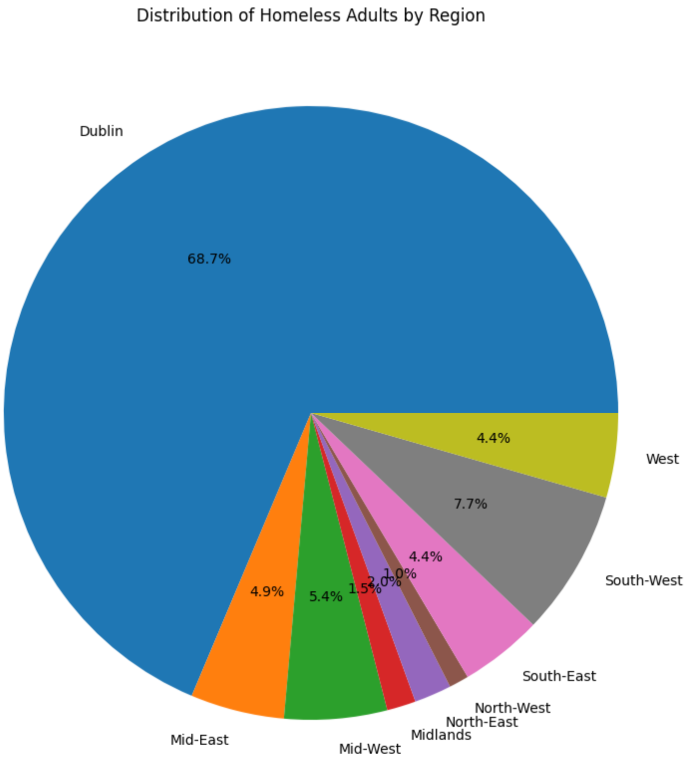
From the above graph it is evident that Dublin has more homeless adults in comparison to other regions and the second region which is south west.

6. Time Series analysis

Time series analysis shows the increase in the number of homeless adults over the tenure of 2016 to 2022 in Ireland and we can see an extensive increase in number over the duration.



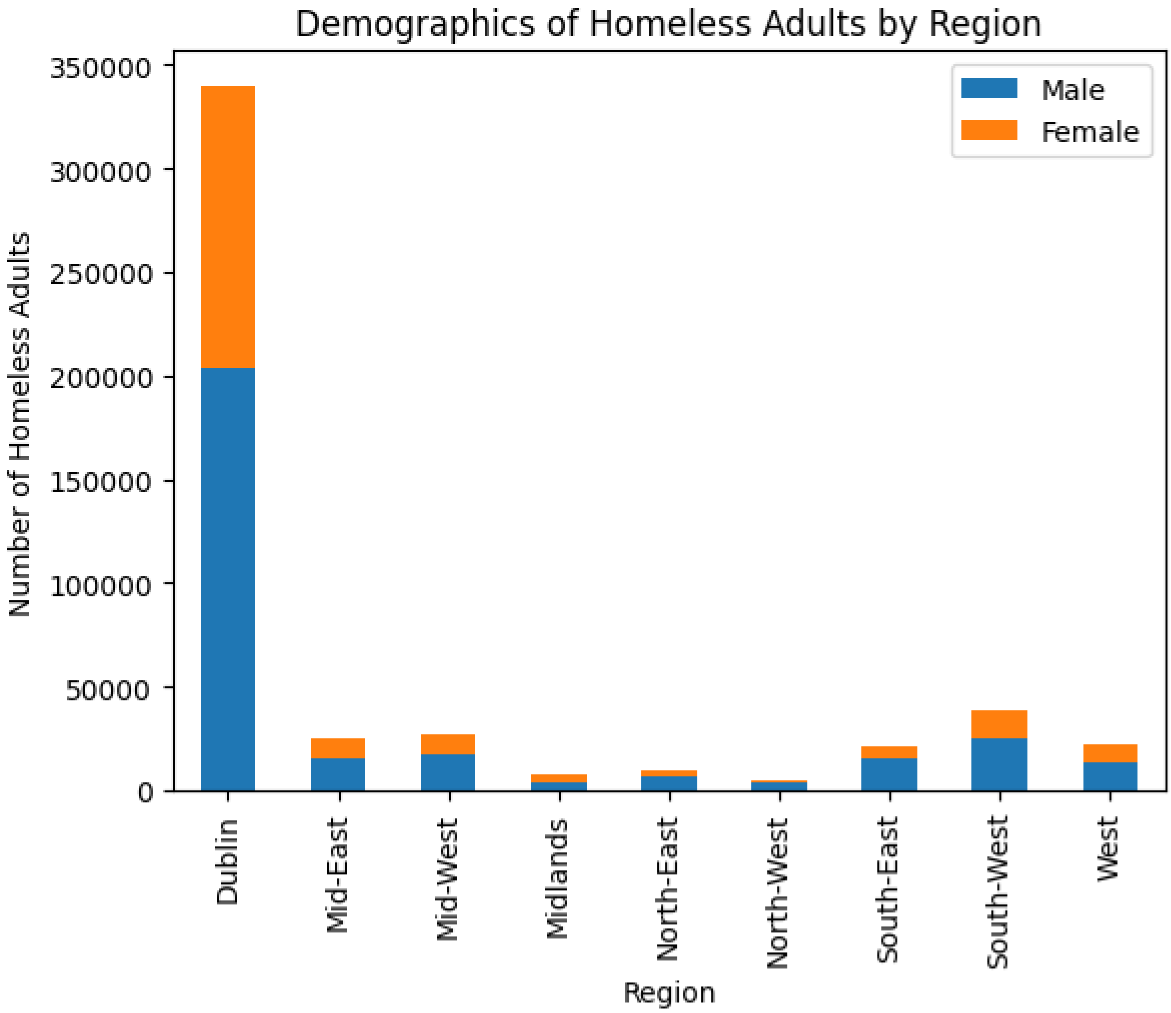
7. Distribution of homeless adults in each region with percentages and we can see Dublin stands with the highest percent and north west region with lowest percentage of homeless people.



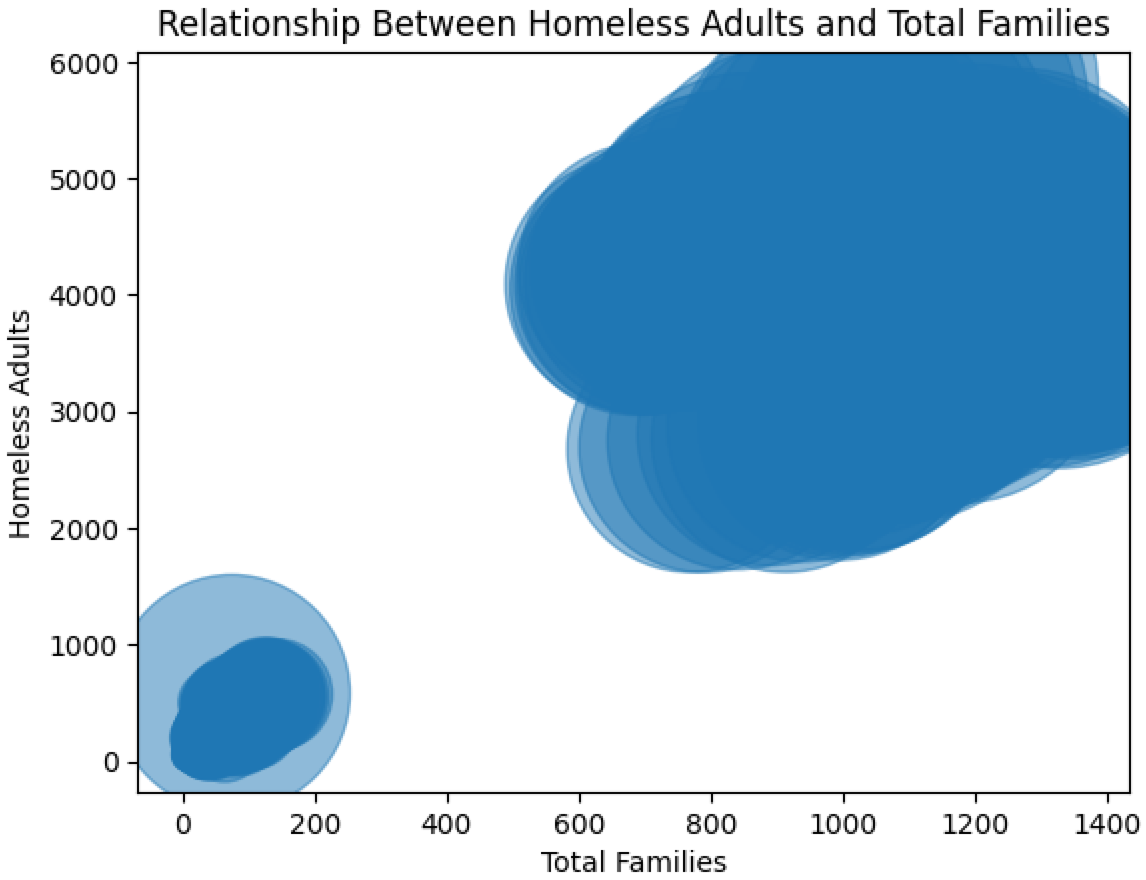
8. Gender wise Distribution of homeless adults in all regions.

The below graph concludes that the distribution based on gender is not much varied.

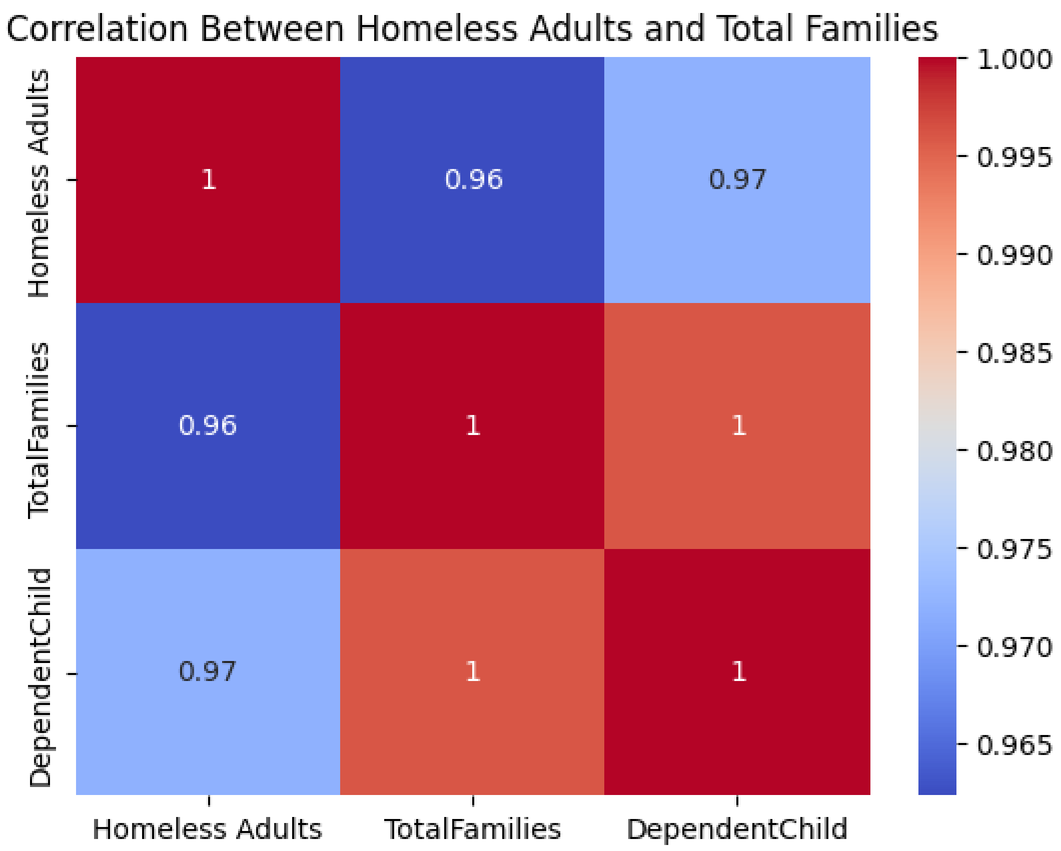
The male to female ratio in homeless adults is almost similar.



9. Relationship between homeless adults and total families.
We can see that there is a linear relationship between these two columns which can help us to Assume that homeless adults can be from the same family.

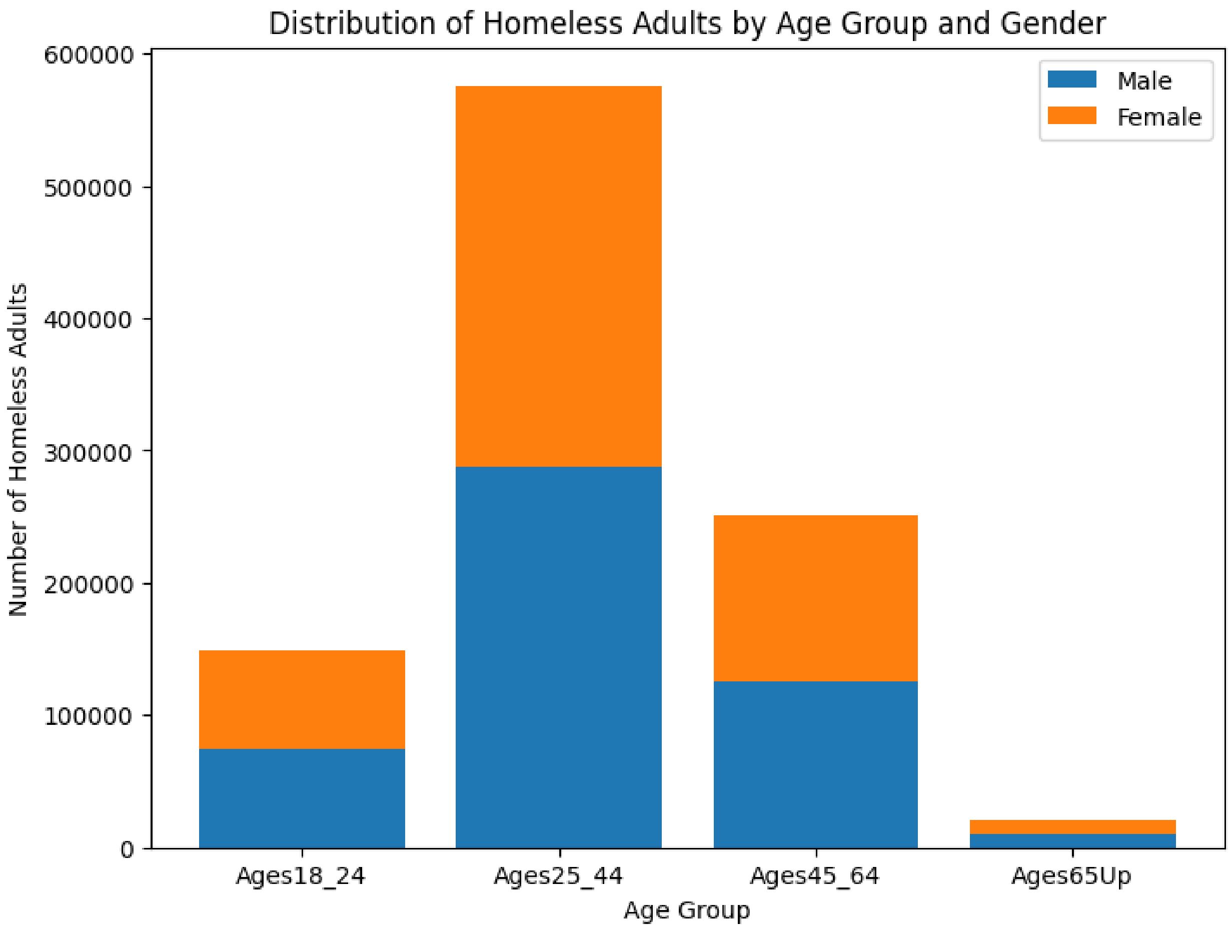


10. This can be again visualized with the help of correlation matrix and heat map which shows the positive relationship between these two variables. I have also added dependent children as another variable which again denotes that when the total number of families increases homeless adults and children increases in number .



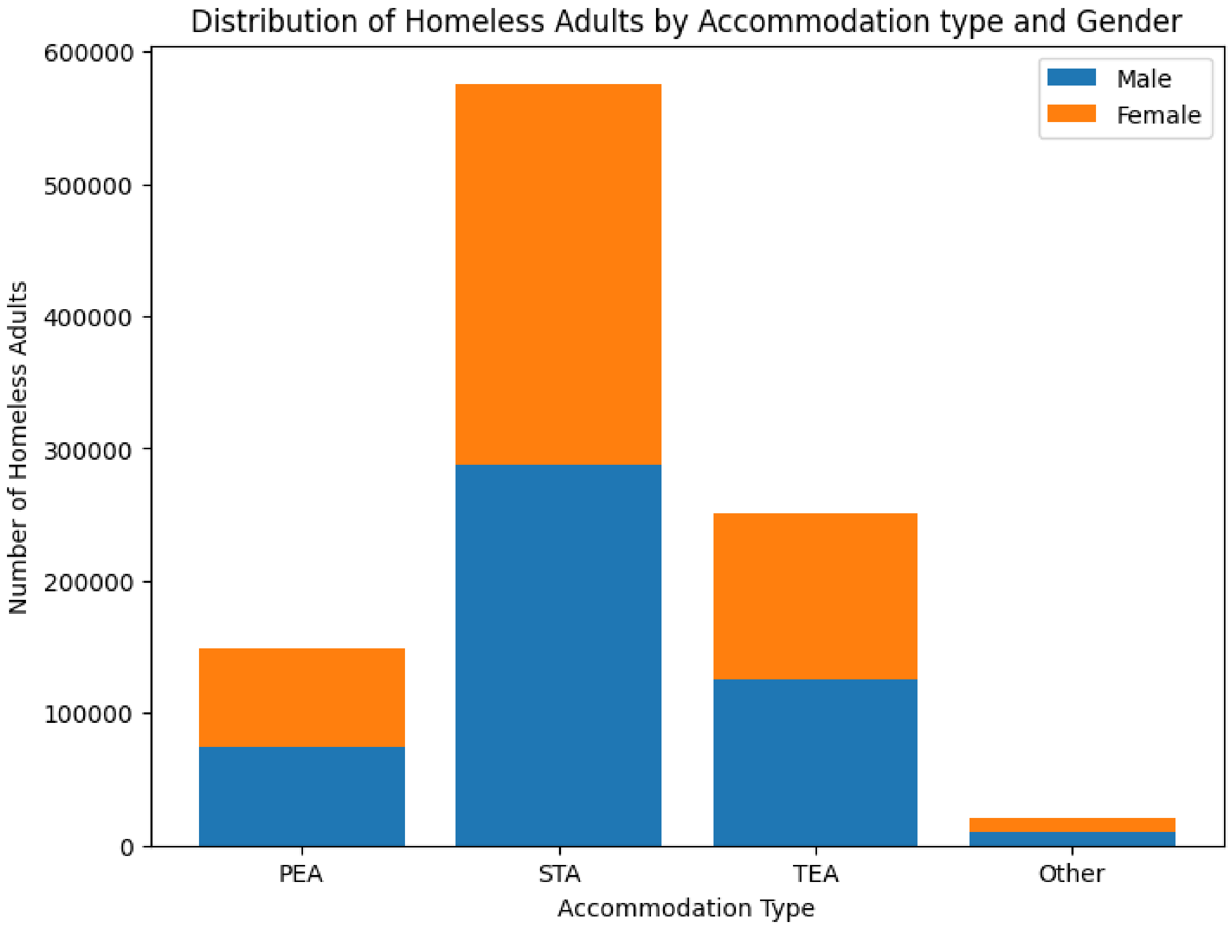
11. Distribution of homeless adults by Gender in each age group

The below graph demonstrates that the homeless adults are maximum between the age group of 25-44 and the male to female ratio is equal in the distribution. Lowest number of homeless adults falls in the age group of above 65.

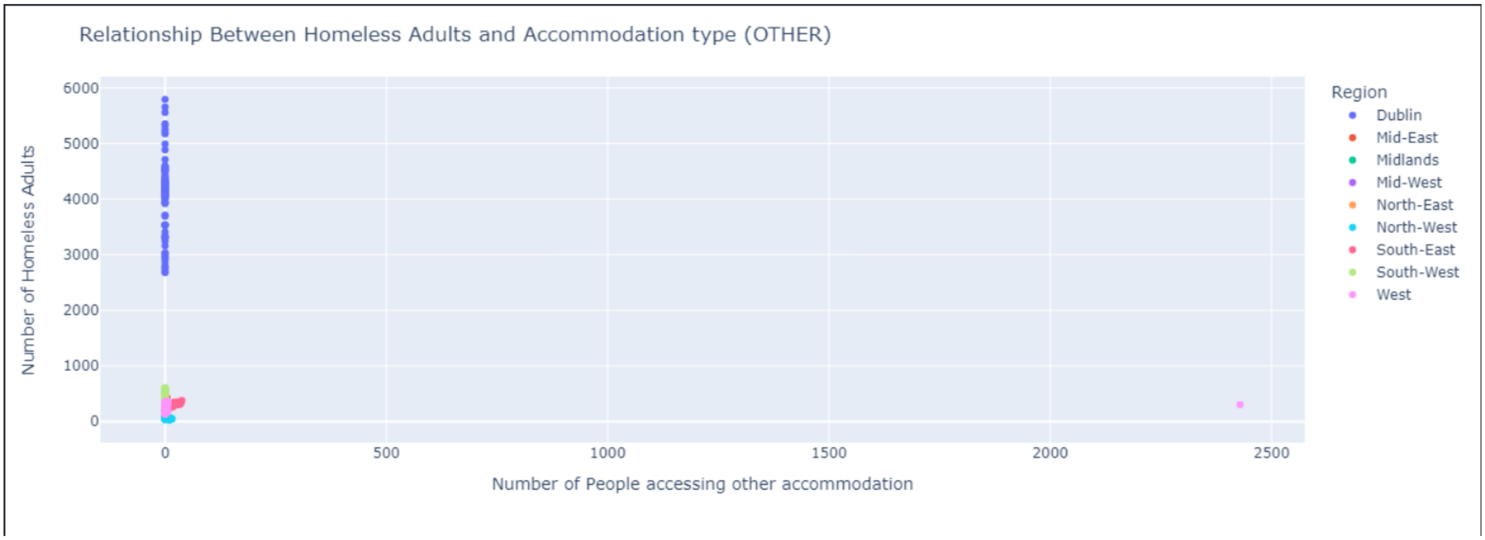


12. Gender wise distribution of homeless adults accessing different type of accommodation

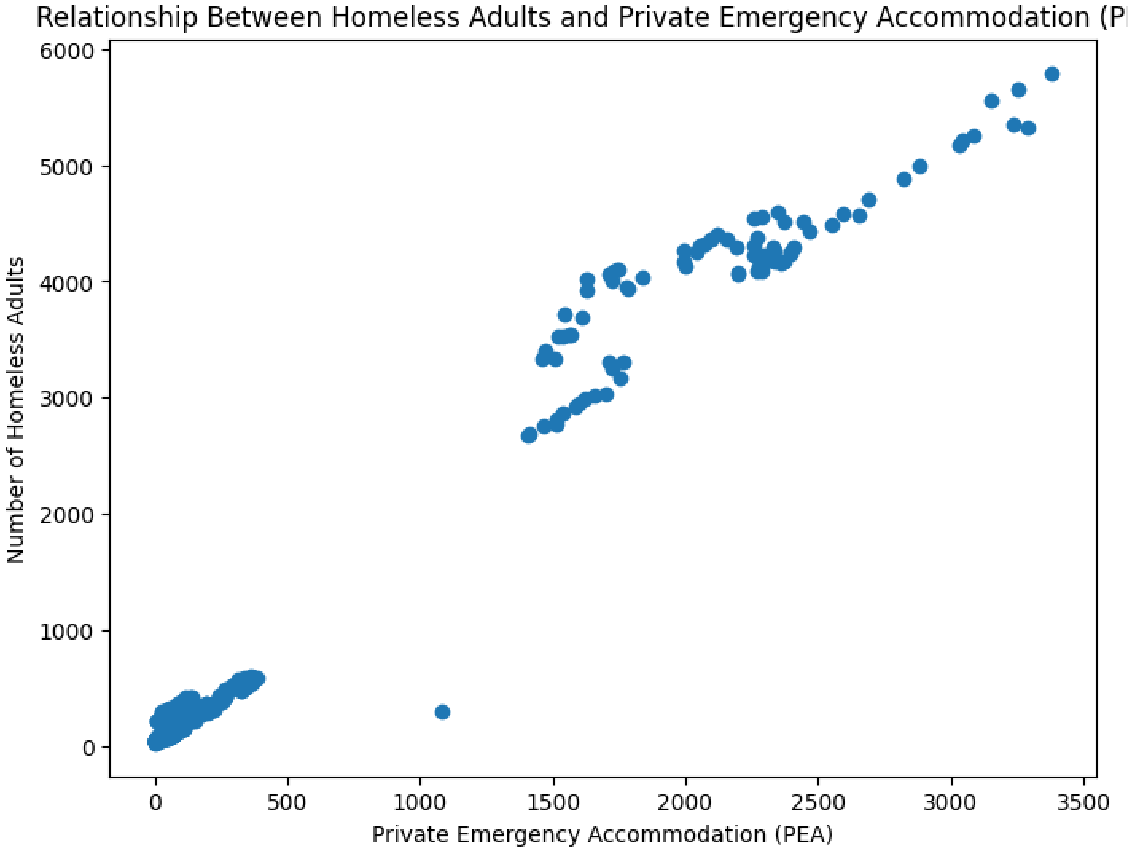
We can clearly see from the graph that maximum people are accessing Supported Temporary Accommodation (STA) which includes accommodation including hostels which received on site professional support.

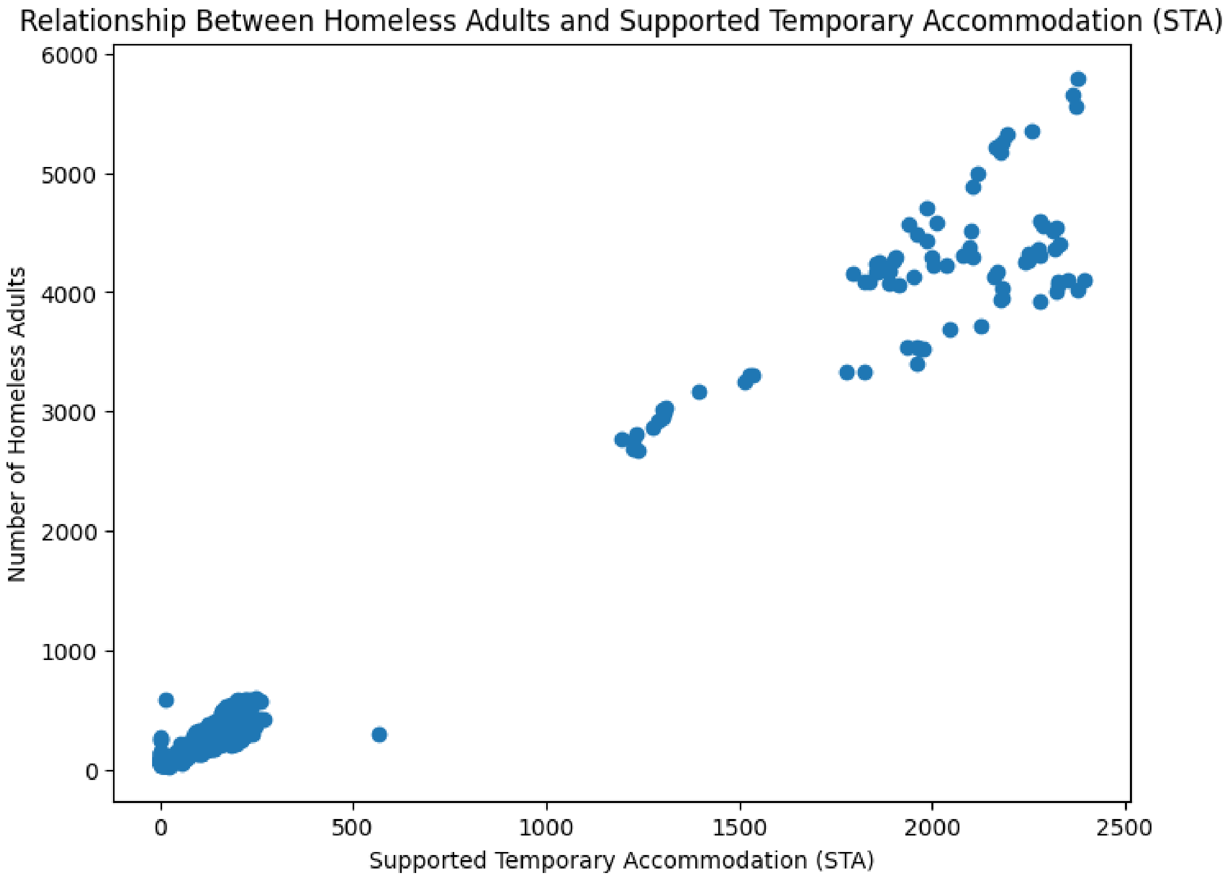


13. Very few people are accessing other accommodation and the number of people accessing it based on region can be visualized through below interactive scatter plot which shows the actual number of people accessing other accommodations when clicked on the points.

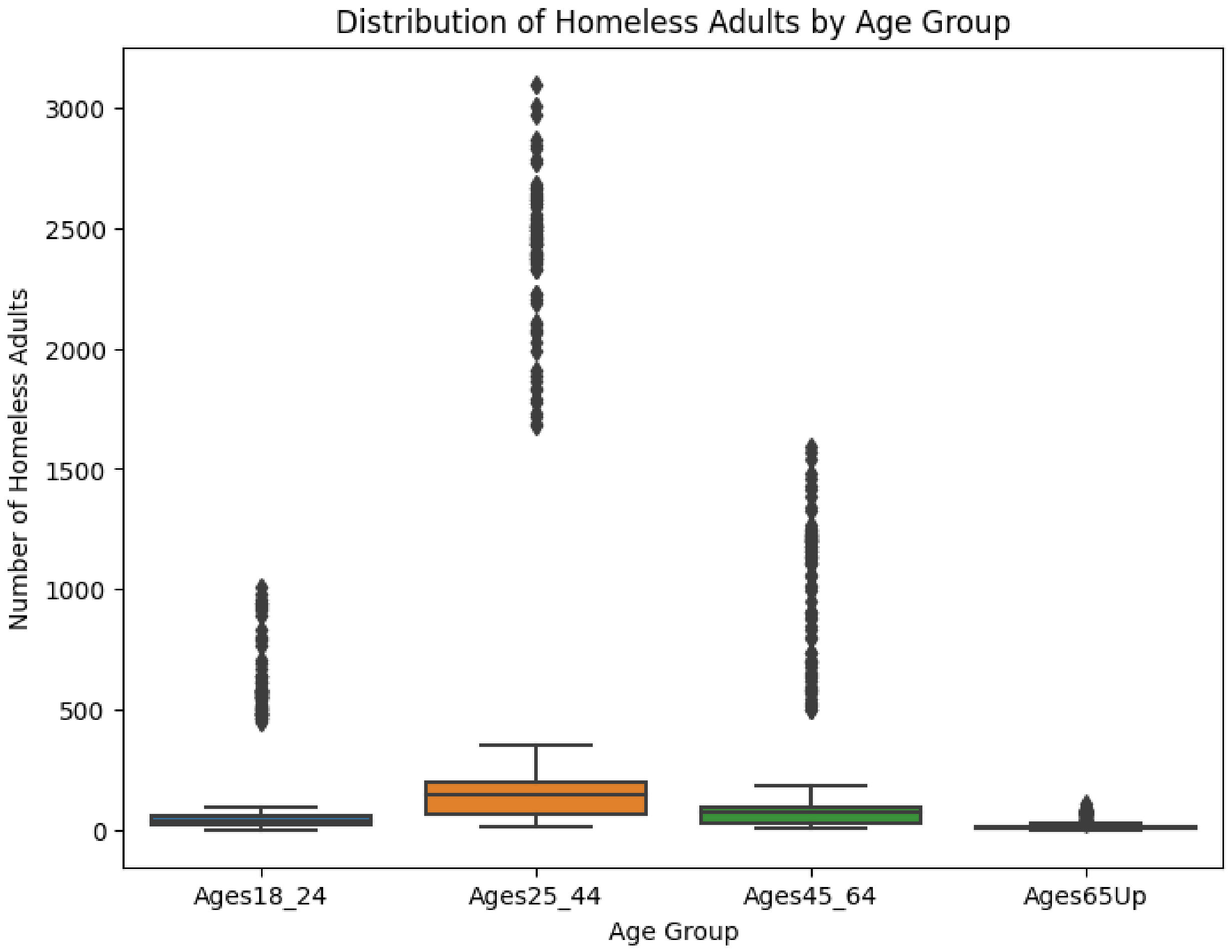


14. To visualize the relationship between homeless adults and people accessing Private Emergency Accommodation and Supported Temporary Accommodation I have plotted the below graphs which shows that they have a linear relationship between them and shows the demand for these accommodations increases with the increase in homeless people.

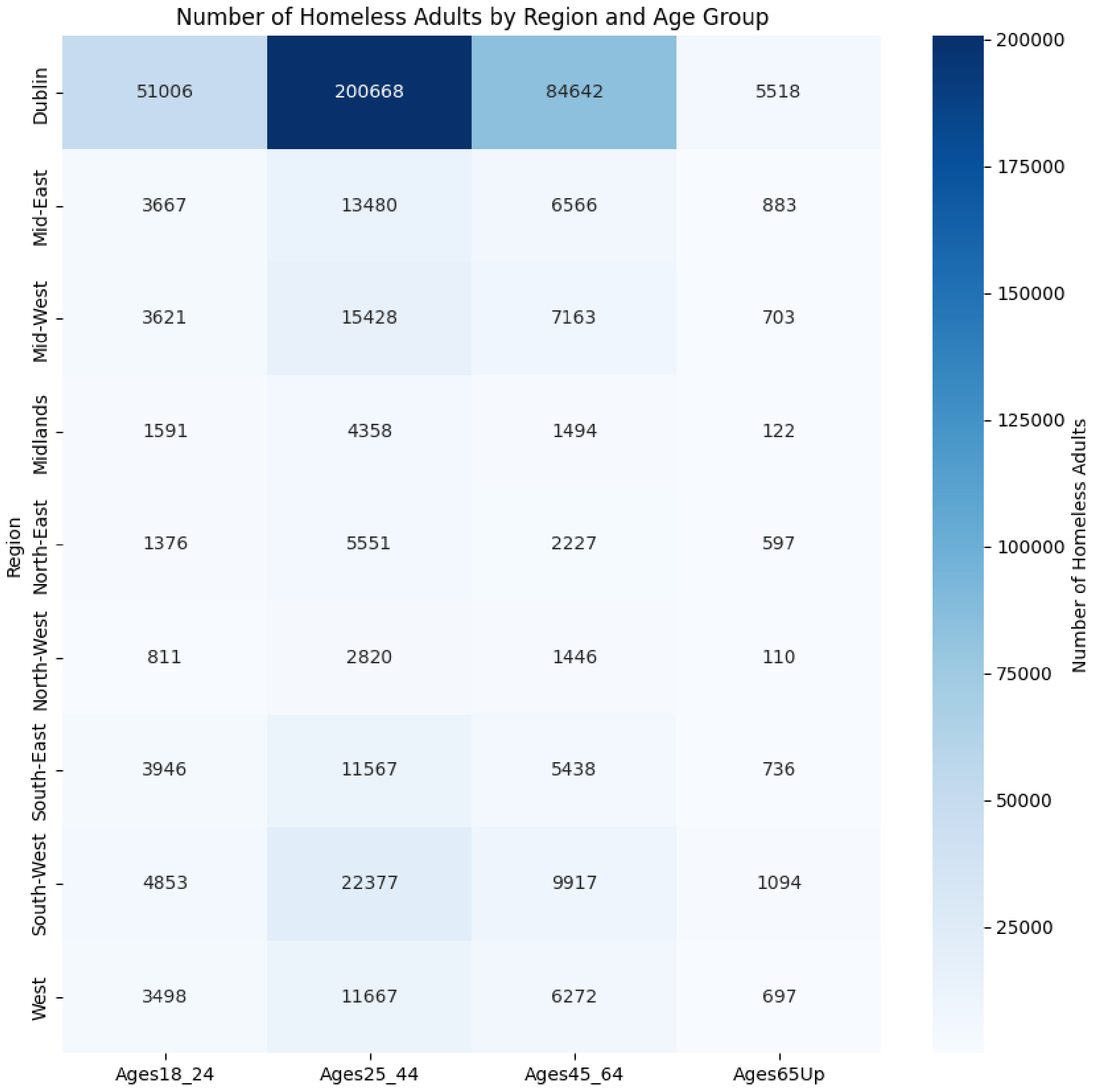




15. I have also created a boxplot which shows the distribution of the number of adults by age group which gives a clear picture on the number of homeless people falling in each category and their distribution  lies in which region in the boxplot with the outliers existing.



16. Lastly I have created a heat map which shows the number of homeless people in each region and age group.



16. I have also tried including maps with the help of folium and plotly library. Folium library helped me to create a map with the number of homeless people in each region marked with the location icon.

Ireland
Belfast
Lisburn
Armagh
Newry
Éire / Ireland
Galway
D
Limerick
Waterford
Cork
St Davi

Given more time how can your work be improved?

Few things I realized while implementing that could have been improved if I had more time.

1. Instead of region which is a very generalized are I could have taken Counties present in each region
Which could have given much larger perspective in understanding the distribution and I could

Have worked on implementing better maps by using counties instead of regions. As county wise data
Wasn't available for all the years in the sheets I couldn't include it but with more time
I could have tried finding county wise data.

2. Maps could have been more efficient and better and could have shown the distribution based on
Different components in detail.

3. I could have include columns which states the citizenship wise distribution as it wasn't available
for all the years in the sheets and population to homeless people ratio for each county.

Conclusion: We can conclude from the visualization dublin has the highest issue of homeless in the age group between 24-44 and necessary actions needs to be taken to overcome this issue

Thank You