**Relationship Between Hospital’s Accessibility via Subway Trains and Inpatient Demographic Groups in New York City**

Yizhou Guo, Shujun Jin

CS 5891-05 & CS 3891-05

Vanderbilt University

**Abstract**

* **Objective:** This study investigates the relationship between hospital accessibility via subway networks and inpatient demographics in New York City (NYC). It focuses on the role of subway connectivity in influencing hospital visits and demographic diversity, identifying disparities in healthcare accessibility.
* **Materials and Methods:** NYC subway station data and inpatient demographic data were used to construct weighted station networks, with weights derived from custom metrics and geographical distances. Regression analyses were performed to evaluate the relationship between hospital accessibility (connectivity) and inpatient characteristics (e.g., diversity and patient volume).
* **Results:** While diversity scores showed a positive relationship with connectivity, the trend lacked statistical significance. No significant linear relationship was observed between subway connectivity and total inpatient volumes. Demographic analysis revealed significant racial diversity in NYC hospitals, contrasting with more homogenous populations in rural areas.
* **Conclusion:** Subway connectivity appears to align with greater diversity in hospital demographics, but its overall impact on patient volumes remains inconclusive. Further research integrating additional factors, such as hospital capacity and specialty, is necessary to better understand healthcare accessibility.

**1 Introduction**

Accessibility to healthcare is essential for equitable outcomes, especially in urban areas like New York City (NYC), where public transportation is central to mobility. The subway system plays a crucial role in connecting residents to medical care, yet disparities in accessibility persist, influenced by geographic, socioeconomic, and demographic factors. Vulnerable populations, such as older adults, low-income individuals, and racial minorities, are often disproportionately affected, relying heavily on public transit.

This study investigates the relationship between NYC hospitals' proximity to subway networks and inpatient demographics, focusing on patterns of age, gender, and race. We constructed two subway network models: one weighted by geographical distance and another by custom metrics combining distance and transit frequency. These models were used to compare network characteristics, such as closeness and betweenness centrality. Additionally, inpatient data was reorganized to match the network model, enabling key demographic metrics for hospitals, including patient volume and diversity, to be analyzed.

Using regression analysis, we evaluated the correlation between subway connectivity and inpatient characteristics, assessing whether better-connected hospitals serve more diverse or larger patient populations. This study provides insights into the role of transit systems in shaping healthcare access and highlights potential disparities, offering actionable recommendations for improving equity in healthcare delivery across NYC.

**2 Materials and Methods**

Based on insights from our status report, we extended our previous methods and analyses while incorporating feedback to refine the focus of our study. Specifically, we emphasized two objectives: (1) analyzing the differences between the subway station network weighted by our custom metric and the network weighted by geographical distance, and (2) examining the relationship between the number of inpatient visits at a hospital and the network metrics (e.g. connectivity) of the stations closest to that hospital.

**2.1 Data Sources**

1. **Subway System Network**:
   * **Datasets**:
     + [MTA Subway Stations | State of New York](https://data.ny.gov/Transportation/MTA-Subway-Stations/39hk-dx4f/about_data)
     + [MTA Subway Service Delivered: Beginning 2020](https://data.ny.gov/Transportation/MTA-Subway-Service-Delivered-Beginning-2020/bg59-42xi/about_data)
   * The *MTA Subway Stations* dataset provided station ID, daytime lines, and geographical coordinates (longitude and latitude), which we used to construct the subway network. Nodes represent subway stations, and edges represent connections between stations. The network is modeled as an undirected single-mode graph.
   * The *MTA Subway Service Delivered* dataset was used to adjust edge weights based on train frequency. The weight calculation formula is:
2. **Inpatient Demographics**:
   * **Dataset**: [New York State Hospital Inpatient Discharges 2022](https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/5dtw-tffi/about_data)
   * Selected columns include hospital service area, facility ID, facility name, race, gender, and age group. This data was used to group patients by hospital and analyze demographic characteristics such as age, gender, and race.

**2.2 Network Analysis**

For analyzing the subway system network, we used NetworkX in python to visualize the network and extract key features. The python code was designed to visualize and analyze the network in accordance with the algorithms above. The visualized network contains all nodes and edges, with edges colored by train routes.

The code workflow is summarized as follows:

1. Data Loading: The CSV files containing subway stations and service delivered were loaded using the pandas library.
2. Grouping: Group nodes by subway routes, with repetition allowed.
3. Distance Calculation: Use GTFS longitude and latitude data to calculate the distance between each pair of nodes for each subway route.
4. Determine Route: Use Kruskal's algorithm to find the Minimum Spanning Tree (MST). Since no MTA subway lines have more than 2 terminal stations, MST is the running route.
5. Adjusting Edge Weights: Combine distance data and service delivered data to calculate the adjusted weights for each edge.
6. Visualization and Analysis: Use NetworkX to visualize the network, where edges are colored by route. Then, use built-in functions to calculate closeness centrality and betweenness centrality.

**2.3 Demographic Analysis**

For analyzing the inpatient demographics data, we used python to generate pie charts. The Python code was designed to process the inpatient dataset by loading the CSV data file and filtering by different "Hospital Service Areas." Each area's data was grouped by facility, and demographic characteristics such as age group, gender, and race were extracted for analysis.

The code workflow is summarized as follows:

1. Data Loading: The CSV file containing inpatient demographic data was loaded using the pandas library.
2. Filtering by Region: The data was filtered based on the "Hospital Service Area" column to separate out regions such as New York City, Capital/Adirond, Hudson Valley, Long Island, and others.
3. Grouping and Visualization:
   * For each region, the data was grouped by "Permanent Facility Id" to analyze demographics at the facility level.
   * Pie charts were generated for age group, gender, and race distributions using matplotlib to visualize the percentage breakdown for each category.
   * The pie chart titles included the facility name and the total number of patients for clarity and context.
4. Overall Regional Summaries: In addition to individual facilities, pie charts were created to display the total distribution of age, gender, and race for entire regions. This provided a comparative overview of demographic characteristics across different parts of New York State.

After constructing our network using the methods stated above and generating preliminary results based on the inpatient demographic pie charts, we focused on finding the differences between our station network and the network weighted by geographical distance and examining the relationship between the number of inpatient visits at a hospital and the network metrics of the stations closest to that hospital.

**2.4 Combining Networks and Demographics**

To explore the relationship between inpatient demographic data and the subway station network, we reorganized the inpatient demographic dataset into a more structured and concise format. This process focused on extracting data specific to hospitals in New York City to align with the scope of our station network analysis.

Using a Python program, we generated key demographic metrics for each facility, including total number of patients, number of patients aged 70 or older, number of patients identified as white, additional demographic characteristics (e.g., age group and racial composition), and so forth.

The reorganized dataset provided a comprehensive summary of inpatient demographics for NYC hospitals. This structured data was then used in a regression analysis to examine correlations between the subway station network metrics (e.g., connectivity and centrality) and inpatient demographic patterns.

The code workflow for generating the new inpatient file is summarized as follows:

1. Data Loading: The original inpatient demographic dataset is loaded using the pandas library into a DataFrame. The file is read from the specified path containing detailed demographic and hospital information.
2. Filtering by Region: The dataset is filtered to include only rows corresponding to hospitals located in New York City, as identified by the "Hospital Service Area" column.
3. Defining Helper Functions:
   * Age Group Count: A function is created to calculate the number of patients falling within a specified age range (e.g., "0 to 17", "70 or Older") by matching the "Age Group" column.
   * Category Count: A function is implemented to count the occurrences of specific values in a column (e.g., counting males, females, or racial groups).
4. Grouping by Facility Name: The data is grouped by the "Facility Name" column to aggregate information for each hospital in New York City. For each group, key statistics are calculated:
   * Total Patients: Total number of records for the hospital.
   * Age Distribution: Counts of patients within specific age ranges (0–17, 18–29, 30–49, 50–69, 70+).
   * Gender Distribution: Counts of male, female, and unidentified gender patients.
   * Race Distribution: Counts of patients by race categories (e.g., White, Black/African American, Multi-racial, Other Race).
5. Result Compilation: Aggregated metrics are stored in a new DataFrame, where each row corresponds to a hospital and each column represents a specific demographic metric. The resulting DataFrame is saved to a new CSV file for use in further analysis. The file includes structured information for demographic characteristics by facility, simplifying downstream tasks such as regression analysis.

Next, we examined the relationship between our refined inpatient data and our station network using the steps summarized as below:

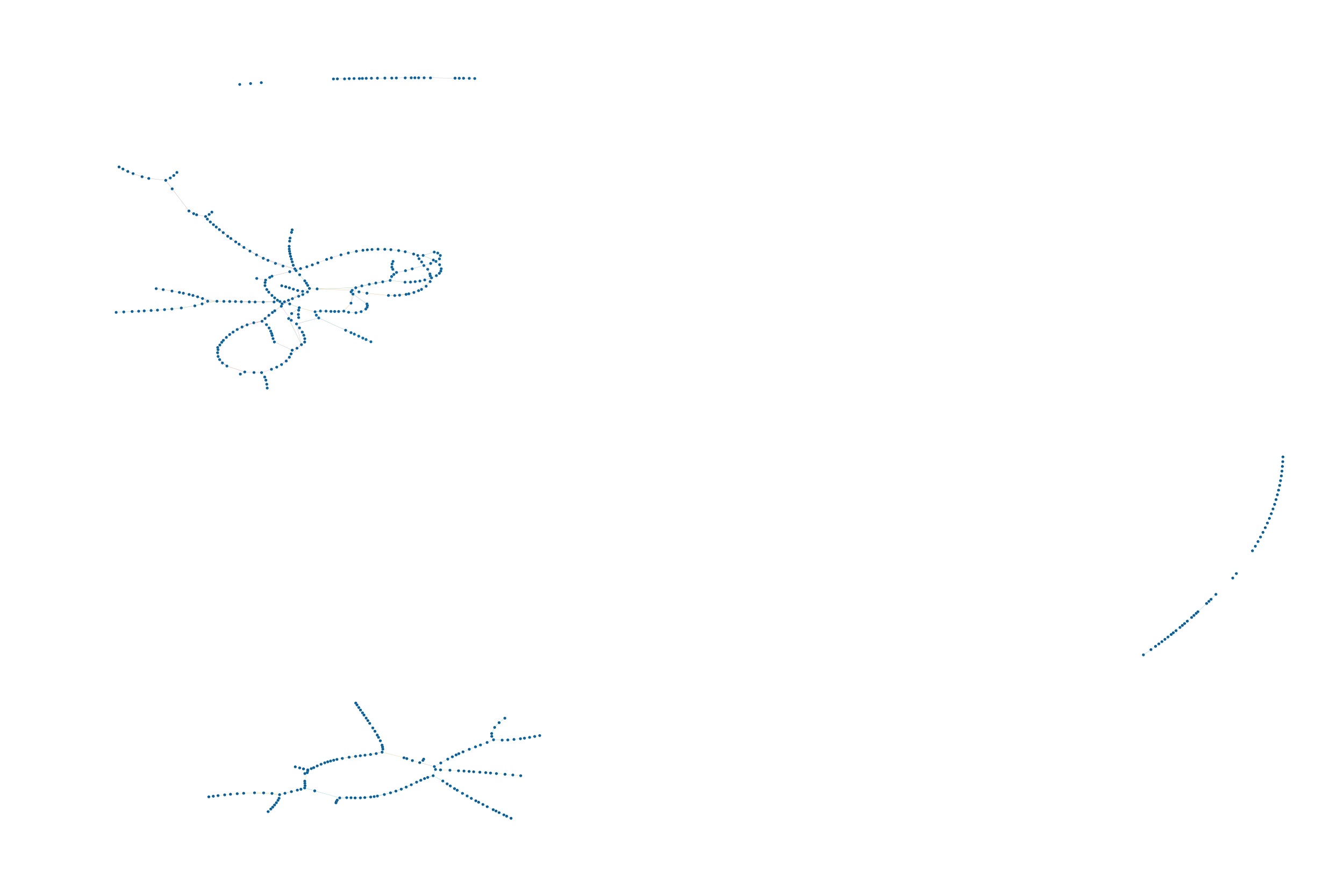
1. Determine Diversity Score and Connectivity Score: For each hospital in New York City, calculate Shannon Index and use it as the diversity score. Then, use the following formula to calculate connectivity score:
2. Apply Linear Regression: Use Linear Regression to assess the relationship between the calculated diversity score and connectivity score.

In the next section, we will include the results from our network analysis, demographics analysis, and the combined analysis.

**3 Results**

We have included key findings from our status report that were critical to our study while adding new results from our analysis of differences between the subway station network weighted by our custom metric and the geographically-based station network. Additionally, we examined the relationship between the station network and inpatient demographic data.

**3.1 Network Analysis**



**Figure 1:** *Network uses geographical distance as weight*



**Figure 2:** *Network uses custom formula as weight*

**3.1.1 Differences in Networks**

We constructed networks with two types of edge weight. The first network (Figure 1) uses only geographical distance between stations, and the second network (Figure 2) uses our custom formula as introduced in the Data Sources section. While the majority of the networks remain the same, including the closeness and betweenness centrality scores, the network with custom metrics visualizes express route (only stop at selected stations) and local route (stop at every station) in a more clear way.

**3.1.2 Closeness Rank (Top 10):**

230 Broadway-Lafayette St : 0.056880870966474074

167 W 4 St-Wash Sq : 0.054873813207146536

231 Grand St : 0.05483264755582835

232 2 Av : 0.0542224919821359

25 DeKalb Av : 0.053410244203083074

26 Atlantic Av-Barclays Ctr : 0.053410244203083074

227 34 St-Herald Sq : 0.0531770965383188

101 Delancey St-Essex St : 0.0529459755102638

233 Delancey St-Essex St : 0.05185662943733181

226 42 St-Bryant Pk : 0.051709882696794614

**3.1.3 Betweenness Rank (Top 10):**

230 Broadway-Lafayette St : 0.1435817269150603

174 Jay St-MetroTech : 0.10231818863397797

175 Hoyt-Schermerhorn Sts : 0.08749028749028749

231 Grand St : 0.0864694629606909

232 2 Av : 0.08382216224321484

233 Delancey St-Essex St : 0.08281342439237174

234 East Broadway : 0.08182922340817075

235 York St : 0.08098951730530674

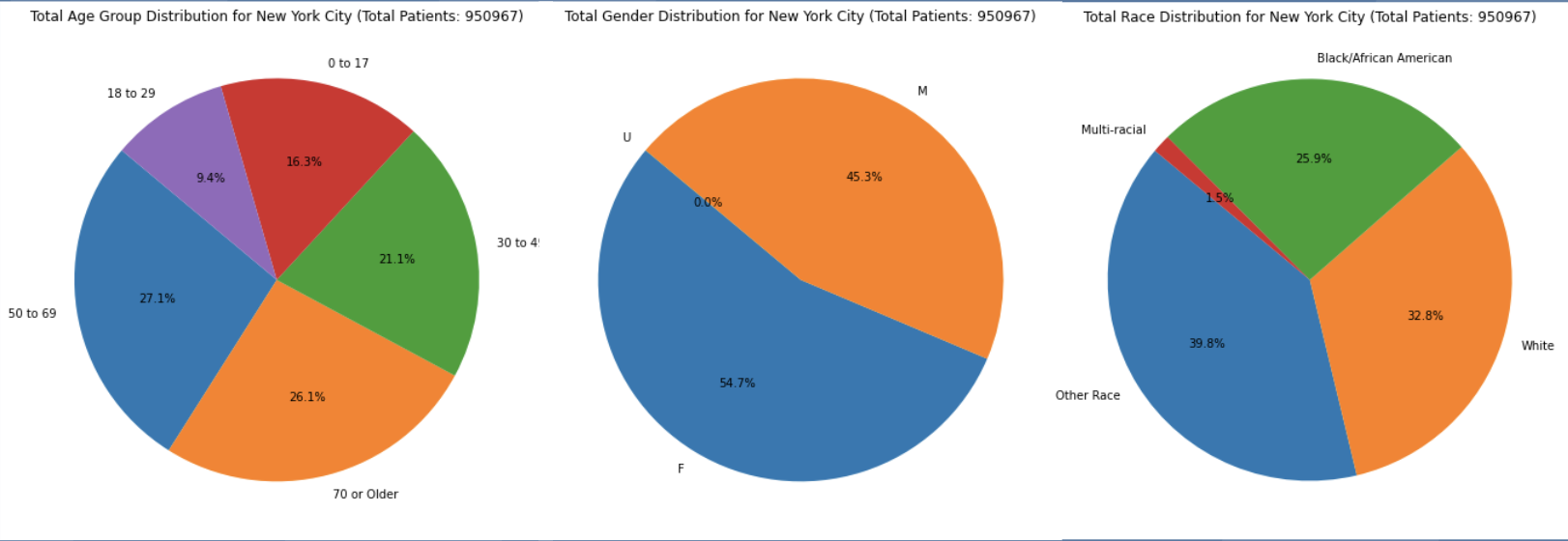
160 59 St-Columbus Circle : 0.06396372633214759

225 47-50 Sts-Rockefeller Ctr : 0.06391911532262397

**3.2 Demographic Analysis**

To ensure a cohesive final report, we have integrated key findings from our status report while tailoring the analysis to focus on the relationship between inpatient data and the subway station network. Detailed charts and specific numerical breakdowns from the status report have been retained there for reference, while key findings are summarized as below.

Building on our previous analysis, we observed the following trends in the inpatient demographic characteristics of hospitals in New York City:



**Figure 3:** *Demographic Data for New York City*

**Age Distribution**: Older adults (50 to 69 and 70+) form a significant portion of the inpatient population, while younger age groups contribute smaller but notable shares.

**Gender Distribution**: Gender representation is relatively balanced, with a slightly higher proportion of female inpatients compared to males.

**Race Distribution**: NYC demonstrates substantial racial diversity among inpatients, with no single racial group overwhelmingly dominant.

Comparing the data in NYC with the data in other regions in New York State (numerical details have been included in the status report), we have the following findings:

**Age Trends**: Regions outside NYC, such as “Capital/Adirond” and “Long Island,” show higher proportions of older adults, reflecting an aging population in these areas.

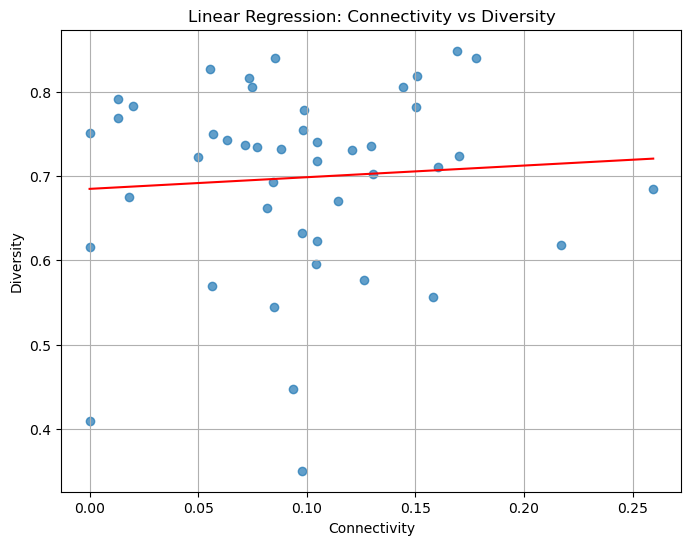
**Gender Patterns**: Gender distribution is consistent across all regions, with females slightly outnumbering males statewide.

**Racial Composition**: NYC has greater racial diversity compared to regions like “Capital/Adirond” and “Southern Tier,” where the White population is predominant.

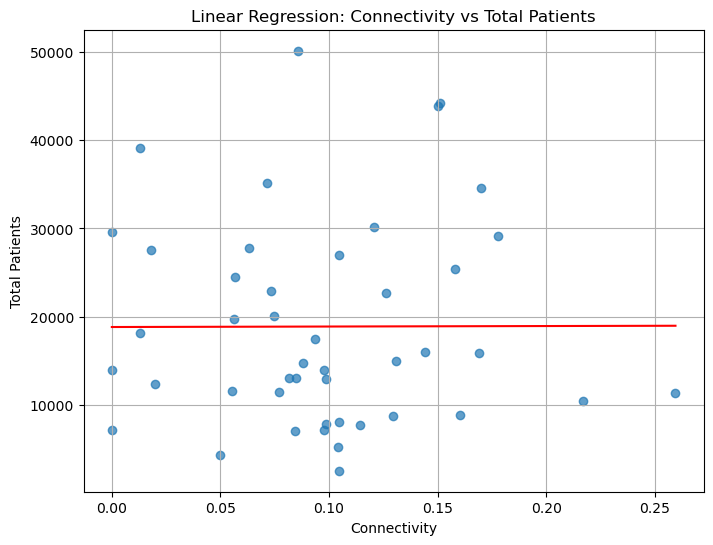
NYC’s diverse inpatient demographics suggest broader accessibility and varied healthcare needs driven by the city’s dense and heterogeneous population. In contrast, rural and suburban areas with higher proportions of older adults may face challenges in addressing the healthcare needs of aging populations, particularly due to limited infrastructure. Statewide, similar gender engagement patterns in healthcare utilization indicate consistent access among male and female populations.

**3.3 Relationship Between Networks and Demographics**

The plotted graphs are shown below as Figure 4 and Figure 5. The regression functions and p-values are

****

**Figure 4:** *Linear Regression between Connectivity and Diversity*



**Figure 5:** *Linear Regression between Connectivity and Number of Inpatient Visits*

**4 Discussion**

The results of our linear regression analysis reveal insights into the relationship between subway connectivity and hospital inpatient demographics. We found the diversity score is positively related to connectivity score by a factor of 0.14, suggesting that hospitals with better access to the subway system are more likely to serve a diverse patient population, which aligns with our expectation. However, the p-value of 0.649 suggests that there is only a slight trend but no statistical significance, implying other factors may influence the relationship.

The relationship between connectivity and total patients remains unclear, as indicated by a p-value of 0.986. This result indicates minimal evidence of a linear association between the two variables, challenging the hypothesis that better-connected hospitals necessarily experience higher patient volumes.

**5 Conclusion**

This study underscores the potential of subway systems to enhance hospital accessibility and promote demographic diversity in urban healthcare. While hospitals with better subway connectivity tended to serve more diverse populations, statistical analysis revealed that other factors, such as hospital capacity and specialization, likely outweigh connectivity in determining inpatient volumes.

Comparisons with rural regions highlight disparities, as suburban and rural hospitals served older, less diverse populations, often constrained by limited infrastructure. These findings emphasize the need for integrated urban planning and healthcare policy to improve accessibility for vulnerable groups, particularly in underserved areas.

Future studies should incorporate additional factors, such as socioeconomic data, hospital specialties, and geographic barriers, to refine the understanding of transportation's role in healthcare equity. Enhanced models may better guide interventions aimed at creating inclusive, accessible urban healthcare systems.

**6 Threats to validity**

First, populations in certain areas can have different density and diversity compared to other areas. This can cause bias and potentially result in outliers in the regression model. Second, a hospital’s capacity can influence the total patient visits per year. Third, a hospital’s specialty can influence the diversity of patient demographics. For example, a children’s hospital is unlikely to have patients ages 70 or older. While we try to eliminate side effects as much as we can, these factors can still threaten the validity of the study.

**7 References**

[1] American Hospital Association. (2017, November 15). *Social determinants of health series: Transportation and the role of hospitals*.<https://www.aha.org/ahahret-guides/2017-11-15-social-determinants-health-series-transportation-and-role-hospitals>

[2] Metropolitan Transportation Authority. (2023). *Subway and bus ridership for 2023*. <https://new.mta.info/agency/new-york-city-transit/subway-bus-ridership-2023>