

Understanding The Toronto Demographic Seeking Shelter Services*

A Data-Driven Analysis of Shelter Service Count by Population and Service Type

Aakash Vaithyanathan

December 14, 2024

This paper aims to study the various factors that influence the daily service count of demographics that seek shelter services across Toronto. By examining the demographics, the type of shelter and the various services they offer, we hope to better understand the key factors that influence the daily shelter service count. Our study utilizes a negative binomial model to estimate the daily count accurately. We find that the most popular demographic that seek shelters are Men with the primary shelter type being an emergency shelter offering short term stay. We believe the results from this study can help us in understanding the key factors that influence the shelter service count and provide valuable information in better supporting shelters and develop policies to improve the homelessness situation in Toronto.

Table of contents

1	Introduction	3
2	Data	4
2.1	Overview	4
2.2	Measurement	5
2.3	Outcome variable — Service Count	6
2.4	Predictor variables	8
2.4.1	Demographic	8
2.4.2	Shelter Service Type	9
2.4.3	Shelter Types	10

*Code and data is available at: <https://github.com/aakash2002/quantitative-analysis-canadian-shelters>.

3	Model	11
3.1	Model overview	11
3.2	Model setup	11
3.3	Model selection	12
3.4	Model diagnostics and validation	13
4	Results	14
4.1	Model Summary	14
4.2	Credibility Intervals for Explanatory Model	15
4.2.1	Role of Demographics in daily shelter count	16
4.2.2	Role of Shelter Types in daily shelter count	16
4.2.3	Role of Overnight Service Type in daily shelter count	16
5	Discussion	16
5.1	Why are men more likely to seek shelter services	16
5.2	Daily volume at Emergency shelter types	17
5.3	Need for improved infrastructure and services at standard shelters	17
5.4	Limitations	18
5.5	Future Steps	18
5.5.1	Improvement to our proposed model	18
5.5.2	Amalgamation of various data sources	18
5.5.3	Conducting surveys to capture missing information from our dataset	18
6	Appendix	20
6.1	Surveys and Sampling Methodology	20
6.1.1	Survey	21
6.1.2	Closing Remarks	23
6.2	Dataset Description	25
6.3	Model details	25
6.3.1	Posterior predictive check	25
6.3.2	Leave-One-Out (LOO) Comparison	26
6.3.3	Model summary	26
6.3.4	QQ-Plot For Explanatory Model	27
6.3.5	Pareto K-Diagnostics	27
6.3.6	Residual Plot for explanatory model	27
6.3.7	Markov Chain Monte Carlo Convergence Diagnostic	27
6.3.8	R-Hat Convergence Diagnostic	29
	References	30

1 Introduction

Homeless Shelters are the backbone support system to the growing problem of homelessness in Toronto and other parts of Canada. In 2023, the City of Toronto declared homelessness as a significant emergency that needs immediate attention (Zakari Omar 2023). Aside from not being able to afford proper meals, or a roof over their heads, the homeless population also are at increased risk of health concerns due to compromised immune systems and lack of hygienic circumstances either on the streets or at crowded shelters that they call home. Since 2018, visits to non-emergent emergency room by homeless people has increased by over 24% (Nicole Ireland 2023). Before more policies and findings are allocated to address this problem, it is important to understand the population that most experiences this issue and what factors play a key contributing role.

This study estimates the daily number of service counts (our outcome variable) in shelters by examining the demographic group, the type of shelter and the services they offer. By analyzing these relationships, the study evaluates how changes in these predictors affect daily shelter service count. The research question this paper aims to answer is: “What factors influence the preference for shelters that the Toronto demographics seek”. To help answer this question, we use a Negative Binomial Model with an interaction term between the demographics and the shelter type. We are interested in examining if the interaction between the above mentioned predictors can better model the relationship between the predictors and the outcome variable.

Our findings suggest that Men are the majority group that tend to shelter services with Emergency based shelters being their primary shelter type. These type of shelters allow individuals to stay for a short duration of about a month and allow re-admissions at a later date (Toronto Shelter & Support Services 2024c). In addition, as of November 2024, our study identified Emergency based shelters as the most popular shelters with increasing population at these centers. Our study highlights the need for addressing these issues before any potential health outbreaks occur at these populous shelters.

The remainder of this paper is structured as follows: The [data](#) discusses the data set used in the study, appropriate cleaning steps carried out and brief overview of the measurements. The [model](#) describes the details of the model used in this study and a justification for our choice. It also highlights the limitation of our proposed model. The [results](#) presents the our findings in this study and what it means. We then discuss these findings in detail under the [discussion](#) section and conclude our study with some identified [limitation](#) and proposed [future steps](#).

2 Data

2.1 Overview

In order to understand the factors influencing how many people seek shelter services, we decided to use the data provided by Open Data Toronto (Gelfand 2022). The data set is maintained and published by Toronto Shelter & Support Services (Toronto Shelter & Support Services 2024a) under the Open Government License - Toronto. The data set contains information regarding the various shelter types across Toronto offering different types of overnight shelter services along with the daily count and the demographic of the population. In addition, it also includes information about the type of capacity it offers which is either a bed based or room based service. The data set also contains daily statistics about the available funding beds/rooms, the current used beds/rooms and the rate of occupancy of beds/rooms. Each row also specifies the date when it was collected.

Since the data supports both bed based and room based shelter services, rows of data that specific bed based shelter details won't include any data for room based services. As a result, our data will only contain information for one of the following types. To address this, we decided to focus on the bed based shelter capacity in our study. The motivation for why is because a key factor we're interested in studying is the trends in the various demographic groups that seek shelter services and room based shelter services are primarily focused towards families or couples (Toronto Shelter & Support Services 2024b) which may skew our analysis.

This data set is updated daily and has the following supported formats: CSV, JSON and XML. It includes data between the year 2021 to 2024. In our study, we decided to use the 2024 data set which includes the daily shelter service information from January 1, 2024 to November 26, 2024.

Alternatively, another data set that could have been used is the Toronto Shelter System Flow (Toronto Shelter & Support Services 2024d) that is also maintained and published by Toronto Shelter & Support Services under the Open Government License - Toronto. The reason why this data set was not chosen is because it primarily focuses on the the count based data of various groups and doesn't contain categorical details into the shelter types and services that are necessary for this study.

A sample of the cleaned data set is described in Table 1.

Table 1: Sample rows from Cleaned Toronto Shelter Service Data

Demographic	Shelter Type	Service Type	Service Count
Mixed Adult	Emergency	Shelter	8
Men	Emergency	Shelter	42
Men	Emergency	Shelter	30
Mixed Adult	Emergency	Warming Centre	19
Mixed Adult	Emergency	Warming Centre	21

The data set contains 44300 observations and after removing room based shelter observations and any missing values we are left with 31484 data points for analysis. The process of cleaning the data was done using the R (R Core Team 2024a) and Janitor (Firke 2024) package. The key steps taken in data cleaning includes standardizing column names to only contain lower case letters and underscores. Next, we select the rows of data that correspond bed based capacity. Following this, we select the desired columns for our study and remove all missing values. Since the independent variables include categorical types, we convert these to factors and rename the variables to be less ambiguous. As mentioned previously, a key factor we’re interested in observing is the role of demographic group in the count of users seeking shelter services. Therefore, we re-level our data to use ‘Men’ as our point of reference for comparison. Finally, we save our data as a parquet file, compressing the file size significantly using the Arrow (Richardson et al. 2024) package. In addition to saving the cleaned data set, we also split our data into training and testing set for analysis by the statistical models we later build in this study. This is done using the Caret (Kuhn and Max 2008) package. The training and testing data set parquet files are also saved. All the graphs shown in the dataset are made using ggplot2 (R Core Team 2024b) package. All the tables in this study are created using kable (Xie Y 2024).

2.2 Measurement

This data set outlines daily list of active overnight shelter services part of the Shelter Management Information System (SMIS) database under Toronto Shelter and Support Services. The daily entries recorded by various shelters across Toronto are connected to the SMIS database and reflected in it’s system by 4 AM every day except for weekends. As of April 23, 2024, hotel-based overnight service types that is prioritized towards refugee claimants and operated by the Canadian Red Cross Society (Canadian Red Cross 2024) is also included in this data set. However, since our study focuses on the bed based service types, during the above described data cleaning process, these data points will not be included. Nonetheless, it is important to acknowledge this as it highlights the coverage of this data set and more accurately reflects the various shelter systems available in Toronto.

Each shelter system records the details of the daily overnight service that it offers. The daily service count in the shelter is the total count for the recorded day and is stored as an integer in the data set. Categorical fields like service type, shelter type, demographic are stored as character type in the data set. Each row is representative of the daily service unique to the specific demographic, service type and shelter type along with the date when it was recorded into the system. Each shelter also includes a unique shelter ID and an organization ID for easy identification. It is important to keep in mind that since the SMIS database is updated at 4 AM the following data, at any given point in time, the data set may not be indicative of the current state of the services offered by various shelters.

2.3 Outcome variable — Service Count

The service count attribute represents the total count of individuals that seek overnight shelter services for the given day. This information is made available from the data set and gives an overview of the daily service workload across shelters. Figure 1 describes the distribution of the service count by various demographic groups. Figure 2 describes the distribution of the service count by shelter types across Toronto and Figure 3 describes the distribution of the service count by the various overnight services that shelters in Toronto offer.

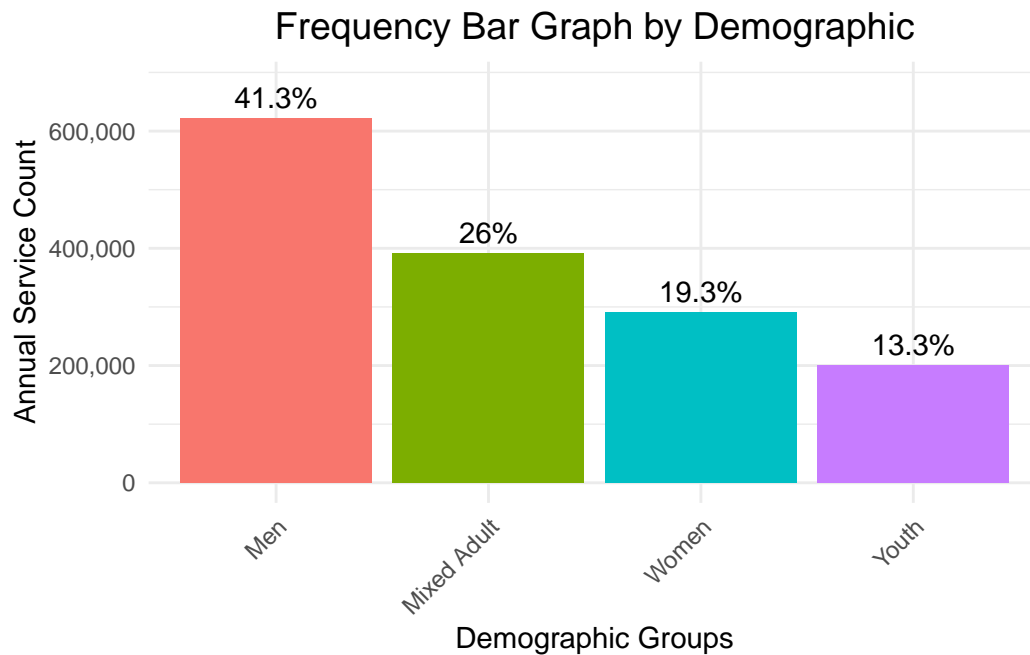


Figure 1: Distribution Shelter Service Count by Demographic for the year 2024

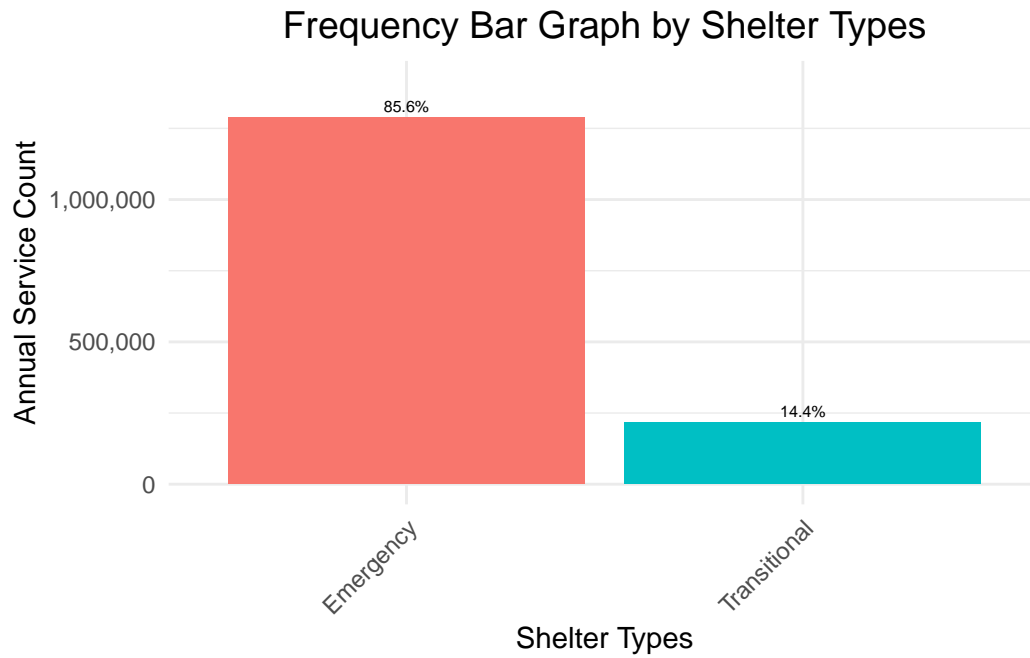


Figure 2: Distribution Shelter Service Count by Shelter Type for the year 2024

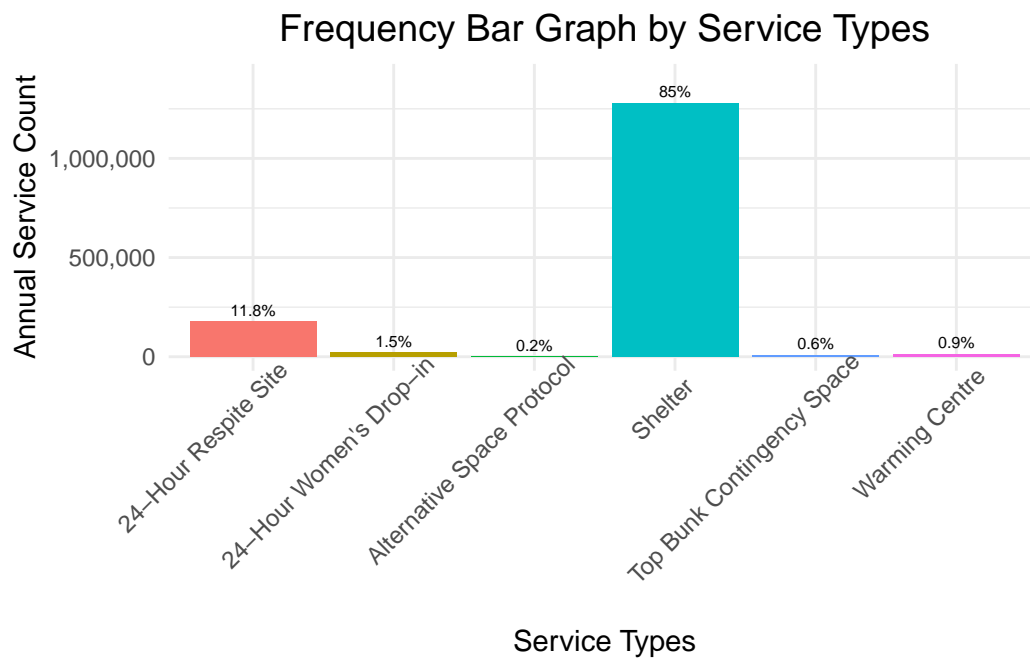


Figure 3: Distribution Shelter Service Count by Service Type for the year 2024

2.4 Predictor variables

2.4.1 Demographic

The demographic attribute specifies the population group seeking shelter service. In our data set, we have Men, Women, Mixed Adults and Youth as the demographic groups. It is important to note that ‘Families’ is another possible demographic group. However, as mentioned in the [data overview](#) section, our study focuses on the bed based shelter services and families are primarily part of the room based services. Hence, the cleaned data doesn’t represent this demographic group. Furthermore, the ‘Mixed Adult’ group represents any gender identify individual seeking shelter service. Figure 4 shows the distribution of each group in our data set.

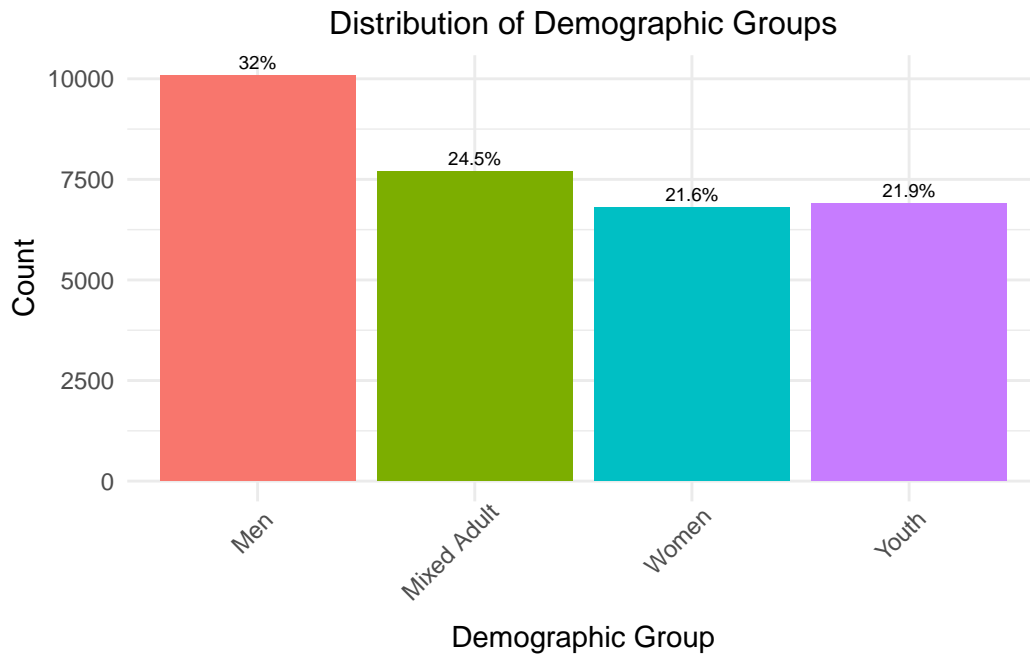


Figure 4: Distribution of demographic groups seeking shelter services for the year 2024

2.4.2 Shelter Service Type

The shelter service type attribute specifies the overnight service the specific shelter offers. The data set contains 8 distinct types of services offered by various shelters. In our study, since we are focused on the bed based service, ‘Motel/Hotel’ type of shelters won’t be included in our cleaned data set. Furthermore, since the data set is for the year 2024, our cleaned data set does not appear to contain any observations of shelters that offer COVID-19 recovery sites. Figure 5 shows the distribution of each shelter type in our data set.

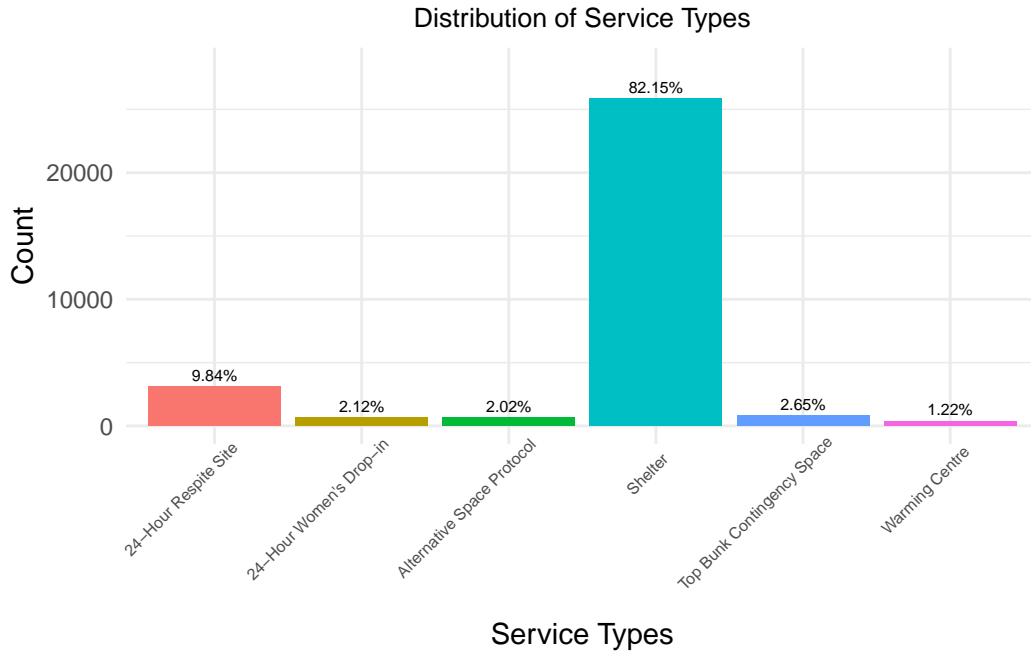


Figure 5: Distribution of shelter service types offered for the year 2024

2.4.3 Shelter Types

The shelter type attribute specifies the type of the shelter operated in Toronto. In our data set, we have Emergency and Transitional as the available shelter types. Emergency shelters are intended for individuals seeking short term stay before returning to their permanent housing. Transitional shelters are intended for individuals seeking long term stay until they are able to address their specific housing needs. Figure 6 shows the distribution of each shelter type in our data set.

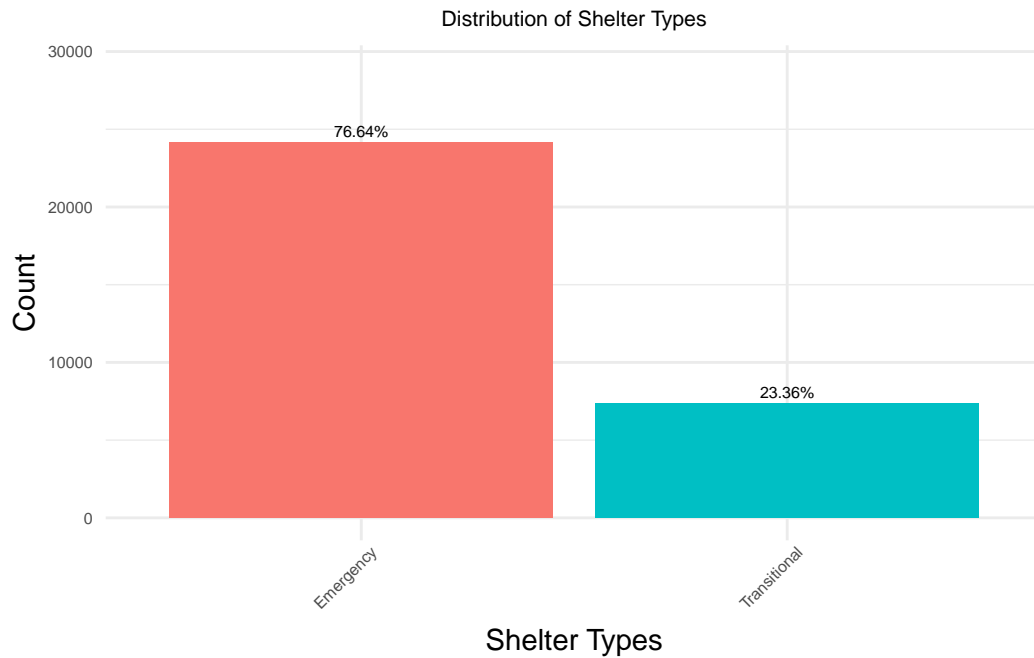


Figure 6: Distribution of shelter types for the year 2024

3 Model

3.1 Model overview

In order to study the count of people seeking shelter services, we decided to build a negative binomial regression framework. Since in our study, we are interested in understanding the count of people, a regression model is preferred. The reason for why a negative binomial model is chosen over other alternatives such as poisson model which is also used to model count is due to the nature of the data distribution. From Table 2, we can see that our variance is significantly larger than our mean. When this happens, we say that there exist overdispersion in our data set (Alexander 2024). The poisson model assumes equal mean and variance which is clearly not true for our data set. Hence, using a poisson model would significantly under fit our data. Furthermore, a linear regression model isn't a viable option either since it's used for modeling continuous data where as we are interested in modeling the count data for specific predictor variables. Thus, we have chosen the negative binomial regression model as our chosen model for this study as these are specifically designed to model count data with overdispersion making them appropriate for our situation.

Table 2: Summary Statistics of Shelter Service Count

Statistic	Value
Mean	47.77849
Variance	1926.32567

3.2 Model setup

The setup for the Negative Binomial Regression model is as follows:

$$y_i | \lambda_i \sim \text{Negative Binomial}(\lambda_i, \phi) \quad (1)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 \times \text{demographic}_i \times \text{shelter_type}_i + \beta_2 \times \text{service_type}_i \quad (2)$$

$$\text{where,} \quad (3)$$

$$\beta_1 \sim \text{Normal}(0, 2.5) \quad (4)$$

$$\beta_2 \sim \text{Normal}(0, 2.5) \quad (5)$$

$$\phi \sim \text{Exponential}(1) \quad (6)$$

$$(7)$$

In the above model:

- λ_i is the expected $(service_count)_i$, modeled through a log link function.
- β_0 is the intercept term which denotes the count when all coefficients are 0.
- β_1 is the coefficient for the interaction term **demographic** and **shelter type**.
- β_2 is the coefficient for the **service type**.
- ϕ is the **dispersion parameter** that controls the degree of overdispersion in the negative binomial distribution. All coefficients $(\beta_0, \beta_1, \beta_2)$ are assigned the default prior of $Normal(0, 2.5)$ from the **rstanarm** package (Goodrich B, et al. 2024). This choice of priors are called weakly informative priors and are designed to stabilize computation offering moderate regularization. This allows the coefficients to scale naturally to the variability in the data.

The prior distribution used for ϕ , the overdispersion parameter is *Exponential*(1). By including this distribution, we specify that our data includes overdispersion and allow the model to adjust for the data variability. Alternative distributions that were experimented were *Exponential*(2) and *Exponential*(0.5) to control the degree of variability allowed by the model. However, we found there was no significant difference between the distributions and decided to proceed with *Exponential*(1) for our model setup.

All the predictor variables are categorical in our model and are converted to factors during the data cleaning process. Since we know categorical variables have a distinct set of values that they take in our data set, by converting them to factors, we assign each of them an integer value to take (Wickham et al. 2024). This allows them to be represented as integers when fitting our above described model.

3.3 Model selection

As discussed above, due to the presence of overdispersion in our data set, we decided to choose the negative binomial model over the poisson model. For our model selection process, we built a total of three bayesian models. The first is the poisson model and the other two are negative binomial models. One of the negative binomial model includes an interaction term between demographic and shelter type to explore whether there is a relationship between the type of shelters the population tends to seek.

From the above section, in addition to checking for the presence of overdispersion, we also compared the three models predictive power. This is done using the Leave-One-Out Cross Validation (LOO) Comparison method. The LOO comparison allows us to compare the Expected Log Pointwise Predictive Density (ELPD) between each of the models and identify which model has the best predictive power. The model with its ELPD difference closest to 0 is the model with best predictive power (Alexander 2024). We use the `loo_compare()` function using the **rstanarm** package to compare the ELPD of the above three models. This function creates a matrix of the three models and returns the pairwise ELPD difference between each. The best model is output at the top of the matrix and will always have it's ELPD difference as 0 (Vehtari A et al. 2024). From Table 6, we can see that our negative binomial model with

interaction has the best predictive capabilities while address the overdispersion that exists in our data set. Therefore, we have decided to use this model in our study.

3.4 Model diagnostics and validation

Aside from performing the LOO comparison on the above models, we also asses their actual predictive power and variability. The two metrics we use for this is calculating the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for our models. During the data cleaning steps outlined in [data overview](#) section, we mentioned that we split our data set into training and testing set using the caret package (Kuhn and Max 2008). Since we have sufficient data points in our study, by using an 80-20 split, we were able to have sufficient training and test data. Each of the described models is fitted on the training data and the diagnostics are evaluated on the testing data. From Table 3, we observe that the negative binomial model without interaction had the least MAE of the three models. However, the difference between each of the three models is negligible. Similarly from Table 4, we observe that the negative binomial model with interaction has the least RMSE of all models however the difference between them is negligible. It is important to understand what these statistics mean. MAE measures the absolute difference in errors between the actual and predicted service count without addressing any outliers that may exist in our study. For our model, the difference in predicted and actual service count is about 25 people. The RMSE calculates the square root of the averaged errors between the predicted and actual data. The presence of outliers can impact the calculated error value as the squared difference between the predicted and actual observations can be large. In our model, the negative binomial model with interaction has an RMSE of about 40.

A key limitation in the model is the significant over dispersion that exists along with any potential outliers in our data set. These results explain why our model demonstrates a high MAE and RMSE statistics. However, we cannot conclude on the models predictive power simply by observing the MAE and RMSE and must explore other diagnostics methods to determine the models predictive power. Figure 8 outlines the predictive power of each of the three model by plotting the posterior predictive check.

By observing these diagnostics methods, we conclude that our negative binomial model with interaction is the optimal explanatory model to use for the purposes of our study.

Additional model convergence plots using Markov Chain Monte Carlo algorithm, \hat{R} plot for model estimates, and other diagnostic plots like Pareto K, QQ-Plot and Residual Plot for the explanatory model is specified in the [appendix](#) section of the report.

Table 3: Model Comparison: Mean Absolute Error (MAE)

Model	MAE
Poisson Model	24.75762
Negative Binomial (Basic)	24.59819
Negative Binomial (Interaction)	24.71875

Table 4: Model Comparison: Root Mean Squared Error (RMSE)

Model	RMSE
Poisson Model	39.57011
Negative Binomial (Basic)	39.69422
Negative Binomial (Interaction)	39.31156

4 Results

The key findings from our study are discussed in the below sections.

4.1 Model Summary

The summary table of the model summary was generated using `modelsummary` (Arel-Bundock 2022), `broom.mixed` (Bolker and Robinson 2024) and `kableExtra` (Zhu 2024) packages. In this study, one of the factors we are interested in exploring is the distribution of the various demographics that seek shelter services. For this purposes, as outlined in the data cleaning steps, we re-leveled our demographics column to use Men as the reference column. Table 7 outlines the estimates for the various categorical values that influence the service counts in various shelters across Canada. The intercept term represents the service count for the day when the predictor variables are at the reference level. The model summary automatically selects the first appearing categorical factor value as the reference in our data set. However, since we have explicitly specified ‘Men’ as the reference level for demographic, the model selects this as the reference. Therefore in our model, our intercept has a value of 4.2 which means that when the demographic for the day is men, for a shelter that is an emergency type and offers 24 hour respite site overnight service, we have about approximately 5 service counts in that shelter for the day for the specific group.

4.2 Credibility Intervals for Explanatory Model

In addition to the model summary, we can look at the credibility interval for our model which was constructed using the `broom` (Robinson, Hayes, and Couch 2024) package. The 95% credibility interval outlines the probability of confidence that each predictor variable takes in our model. From Figure 7, we can see that our intercept term that specifies the daily shelter service count for the day is positively correlated with the predictors in our model as it lies greater than 0 and has a value of about 4.25.

Since the credibility interval outlines the certainty of the estimates for various predictors, we can see that the model shows high uncertainty for “Alternative Space Protocol” and “Top Bunk Contingency Space” shelter service types. A possible explanation for this is because from Figure 5, we know that these make up for about 2 % and 2.8 % of our data set distributions and lack of sufficient observations explains why the model has high variability in predicting these service types with certainty. Furthermore, we see that the model has moderately high uncertainty when estimating for Transitional shelter type. An equivalent justification can be made for this as from Figure 6, we know that transitional shelter types make up for only 23.4 % of our observations so the model experiences some degree of uncertainty. For the rest of the predictors, the credibility interval is relatively small indicating that the model is certain in predicting the estimates and their effects on the daily service count.

From Table 7 and Figure 7, we can make the following key observations

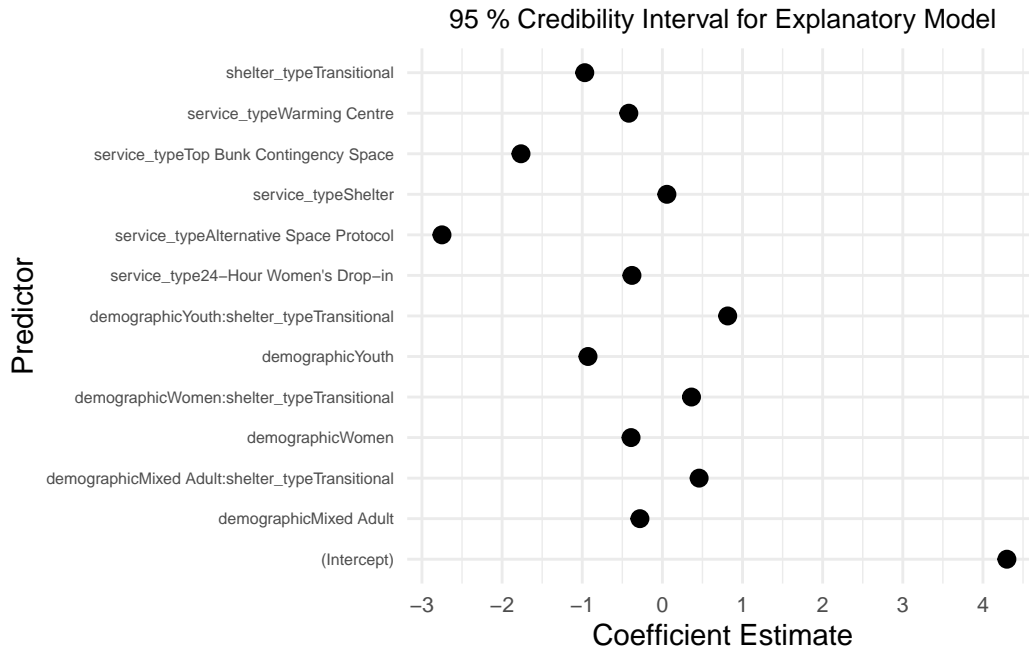


Figure 7: 95 percent credibility interval for coefficients

4.2.1 Role of Demographics in daily shelter count

We observe that the “Men” demographic group is the most popular group that seeks daily overnight shelter services for the year 2024. Youth take up 39.5 % of daily shelter usage of men or equivalently, they have 60 % lower shelter usage than Men. Women take up about 67.64 % of shelter usage of men or equivalently have about 32% less shelter usage than men. Mixed Adult group takes up about 76% of the shelter usage of men or equivalently have about 24% lower rate of shelter usage compared to Men. These observations are in line with our findings from Figure 1.

4.2.2 Role of Shelter Types in daily shelter count

We observe that the “Emergency” shelter types is the most shelter type that offers overnight services for the year 2024. Transitional shelter types take up about 38% of daily shelter usage than that of the emergency shelter type or equivalently, have about 62% lower rate of daily shelter count than Emergency type. This observation is in line with our findings from Figure 2.

4.2.3 Role of Overnight Service Type in daily shelter count

We observe that the “Shelter” service type is the most populate service that people seek for the year 2024. Standard shelters are about 105 % of the daily shelter usage than a 24-hour respite shelter service type. This observation is in line with our findings from Figure 3.

5 Discussion

5.1 Why are men more likely to seek shelter services

This study highlights that the demographic plays a key role in the type of population group that seek overnight shelter services. From our model, we estimate that Men are the most densely populated group that seek daily overnight shelter across Toronto. As of November 2024, over 600,000 people at various shelters across Toronto constitute Male gender which is 16% more than the Mixed Adult population group, 21.4 % more than the Women population and 31.1 % more than the Youth population group that seek shelter services. This finding is in line with a study conducted by Housing Infrastructure Canada (City of Toronto 2024a) in the year 2022 that highlighted that from 2020 to 2022, emergency shelters saw an increase in Men population by 17% while women and youth shelters saw a 5% and 17% increase respectively. The same organization conducted a follow up to this study the following year and highlight that Toronto made up for the largest increase in available beds in shelters of about 300 beds (City of Toronto 2024b).

A key reason why men are more likely to seek shelter services than women is because the readmission of men into various shelters is a lot higher than women (Glenda Laws 2013). Emergency shelters are designed to house individuals for short term duration. This allows Men to move across various shelters and upon completing their stay in a shelter, readmit themselves in the future. Another reason for why men are the more dominant group is due to hidden homelessness. A study by Canada Mortgage Housing Corporation (Schwan K. et al., n.d.) highlights that women and youth are most likely to experience hidden homelessness.

This highlights the need for effective measures to improve the situation of shelter services across Canada and prioritizing the male population that may be unable to be referred to a transitional shelter and seek multiple readmission into emergency shelters.

5.2 Daily volume at Emergency shelter types

Our study shows that as of November 2024, over 1.5 million people that reside in shelters utilize emergency based shelters. These shelters are designed to accommodate individuals for short term stay of traditionally less than a month. However, these shelters allow for self referrals and readmission which allow people to move around various shelters across Toronto. This is significant as it shows the intensity of usage of emergency based shelters and the volume of people that seek it. Due to the sheer volume of people, these types of shelters are unable to offer sufficient spacing between individual beds. This is problematic as it often means that these shelters are packed with people in close proximity to one another. This comes at a huge cost as these shelters are poised for breakout due to the lack of spacing among the people (Out of the cold foundation 2021).

It is important to note that our study does not consider the age groups of the population at these shelters and elderly people with compromised immune system can find themselves are an increased danger of contracting infections and other diseases. This calls for the need for adequate budgeting and allocation of resources to emergency shelters due to their high popularity among the shelter seeking community.

5.3 Need for improved infrastructure and services at standard shelters

Aside from standard shelter services, other alternative shelter services like 24 hour respite sites or warming centers for instance are temporary stay purposes for only about few days. For the population seeking long term stay at shelters, these aren't the adequate places to be in. Our study shows that as of November 2024, over 1 million people reside in traditional shelters with the rest of the shelter service types amounting to less than 250,000 people. This signifies the importance of the type of shelters that are most popular further reinforcing the need for improved budgeting and resource allocation to prioritize support at the standard shelter services that experience large volume of people.

5.4 Limitations

Despite being able to extract key insights about the daily shelter service count based on the demographic, shelter type and the service it offers, there exist some weaknesses in our study. Aside from the limitations discussed for our model in the [model](#) section, this study only looks at the bed based shelter services and doesn't investigate the issues and trends in room based shelters that primarily house families and youth. In order to better allocate resources and protocols for the homeless population, it is crucial to study the trends in room based shelters and the struggles they experience. Another limitation in our study is the lack of observations relating to hidden homelessness. As highlighted by Canada Mortgage Housing Corporation (Schwan K. et al., n.d.), women and youth population experience hidden homelessness the most and this is a potential reason why our data set contains significant observations relating to the Men demographic group.

5.5 Future Steps

In order to identify deeper insights into the shelter systems, the following are some potential directions for future steps of this research.

5.5.1 Improvement to our proposed model

Our model currently only considers the demographic, shelter type and its service. However, factors like the month of the year or location could provide significant insights into specific cities or months of the year when homelessness and shelter services reach their all time high. These could better capture any hidden relationships in the model and also provide valuable information in providing increased support to the critical months and cities.

5.5.2 Amalgamation of various data sources

In addition to having data pertaining to daily shelter usage, we could look into data sets relating to deaths experienced in various shelters which is another dataset provided by Open Data Toronto (Gelfand 2022) or other appropriate data sources. A combined analysis on these data sources could provide richer contextual information and could help provide concrete next steps on improving shelter conditions or addressing the lack of adequate spacing.

5.5.3 Conducting surveys to capture missing information from our dataset

Lastly, many data sets only offer quantitative information about the shelters which although critical are often only one piece of the puzzle. In order to propose appropriate next steps to the highlighted problems, we also need to look at the qualitative aspect of the situation.

One way to do this is by conducting surveys into the shelter systems in Toronto which can signify the critical pain points and concerns the shelters experience on a day to day basis and how their struggles change across the various months of the year. These two findings need to be provided in tandem when proposing any solutions to some of the discussion questions we talked about in this paper.

6 Appendix

6.1 Surveys and Sampling Methodology

The below section discusses a potential survey they we would carry out to better understand the scope of our problem domain.

6.1.0.1 Motivation for the Survey

As discussed in the [future steps](#), we argue that any data set used can only capture the quantitative aspects of the problem domain. In order to better understand our research question, we need qualitative information about some of our key predictors and other relevant information. Our aim is by gathering responses from the below survey response from various shelters across Toronto, we can better understand our problem and offer solutions to problems that exist in the current policies that shelters experience today.

6.1.0.2 Survey Introduction

Thank you for taking the time to participate in this survey! We are collecting data from various shelters across Toronto to better understand the key factors that influence the count of daily overnight services at these shelters and what demographic of population is most at risk for homelessness presently. Your responses will help create improved policies and regulations for the people experiencing homelessness in Toronto. We hope to design systems that offer the best support these groups of people.

We would like to let you know that your identity will be kept confidential and the responses from this survey will only be used for research purposes. The survey is expected to take about 10 to 15 minutes.

If you have any questions, please contact:

- **Name:** Aakash Vaithyanathan
- **Email:** aakash.vaithyanathan@mail.utoronto.ca

6.1.1 Survey

6.1.1.1 Section 1: Shelter Information

1. Shelter Name

Text response:

2. Unique Shelter ID (if available)

Text response:

3. City Located In (Select one)

- ☐ Toronto
- ☐ Mississauga
- ☐ Scarborough
- ☐ Oakville
- ☐ Etobicoke
- ☐ North York

6.1.1.2 Section 2: Shelter specific information

4. Demographics of people at shelter (Select all that apply):

- ☐ Men
- ☐ Women
- ☐ Youth
- ☐ Families / Couples
- ☐ Mixed Adults (Any gender identity)

5. Services Offered (Select all that apply):

- ☐ 24 hour Respite Site
- ☐ 24 hour Women's Drop In
- ☐ Alternative Space Protocol
- ☐ Top Bunk Contingency Space
- ☐ Warming Center
- ☐ Shelter

6. Shelter Type (Select one):

- ☐ Emergency
- ☐ Transitional

6.1.1.3 Section 3: Statistics about shelter

7. **Average daily count of services at shelter** (Select one):

- ☐ Between 1 and 10
- ☐ Between 10 and 50
- ☐ Between 50 and 100
- ☐ Over 100

8. **Average monthly re-admissions to shelter** (Select one):

- ☐ Between 1 and 10
- ☐ Between 10 and 50
- ☐ Between 50 and 100
- ☐ Over 100

6.1.1.4 Section 3: Capacity and Infrastructure

9. **Daily occupancy rate of beds / rooms** (Select one):

- ☐ Below 50%
- ☐ Between 50% and 75%
- ☐ Between 75% to 90%
- ☐ Over 90%

10. **Available spacing between individual beds** (Select one):

- ☐ Less than 2 feet
- ☐ Between 2 to 5 feet
- ☐ Over 5 feet

11. **How often are people transferred to a different shelter** (Select one to the best of your knowledge):

- ☐ Not very often (Once in a quarter)
- ☐ Often (Few times a quarter)
- ☐ Very often (More than once a month)

6.1.1.5 Section 4: Medical Emergencies

12. **Have people been taken to a hospital during their stay at the shelter?** (Select one)

- ☐ Yes
- ☐ No

13. How did you deal with this situation? (optional)

Text response:

6.1.1.6 Section 5: Challenges With Residents

14. What are some challenges you've experienced with the people residing at shelters? (Select all that apply):

- ☐ Experience mental health challenges
- ☐ Experience physical symptoms (cold, body pain, nausea, etc)
- ☐ Aggressive behavior
- ☐ Substance abuse (eg: drugs, alcohol)
- ☐ Other (Please elaborate):

15. If you selected any option for the above, please describe how did you handle this situation? (optional):

Text response:

6.1.1.7 Section 6: Additional Information

16. Do you have additional comments regarding the shelter and its ability to staff large volumes of people? (optional):

Text response:

6.1.2 Closing Remarks

Thank you for completing this survey! Your responses have been recorded. Thank you for taking the time in improving the shelter experience for the homeless population of Toronto. We greatly appreciate your support.

If you would like to receive a copy of your responses, please enter your email below:

Email:

If you have any questions regarding the survey, please contact the listed email at the start of the survey.

6.1.2.1 Survey Data Cleaning Process

Before we proceed to analyzing our survey results, it is crucial to perform data cleaning steps to ensure the consistency and validity of the data gathered. Some of the cleaning processes we need to do is to remove missing entries and duplicate submissions from the survey results. Another step we would need to take is to determine how we plan to handle text based responses. One strategy for this could be to look at the similarity in the responses and extracting semantics from the individual response. These semantics could be represented as categorical variables which could be appended to our revised model for future study.

6.1.2.2 Survey Sampling Strategy

To ensure our survey results are a valid representation of the population, we will employ a stratified sampling approach. This style of sampling ensures that we can gather sufficient samples from population with lesser proportion of data points like ‘youth’ for demographic or ‘transitional’ shelter types. A random sampling approach wouldn’t gather these under represented population well in our data set and could skew any inference we make.

For stratified sampling, we can gather survey results from various strata like shelter type, demographic and shelter service which are the key influencing factors in determining the daily overnight shelter service count. These results could help us better understand the qualitative aspects of what influence shelter service counts. To use such a sampling strategy at a large scale, we would require sufficient budget to gather samples from the above described strata. If funding is limited, we could instead employ cluster sampling and focus on the Emergency shelter type given it’s large proportion of overnight service counts compared to Transitional shelter type. The results from the survey could serve as a preliminary step towards improving shelter experiences for the homeless population.

6.2 Dataset Description

Table 5 provides a brief description of the important variables used in this study in our cleaned data set.

Table 5: Attributes and Description Table

Attribute	Description
Demographic	Demographic of the population (eg: Male, Female, Youth, etc)
Shelter Type	Type of shelter (Emergency/Transitional)
Shelter Service	Type of service provided by shelter (eg: Warming Center, Shelter, etc)
Service Count	Total count of the demographic type seeking shelter services

6.3 Model details

As described in the [model](#) section of our study, we have 3 proposed models to use of which the negative binomial model (with interaction) was chosen. The following sections provide some details about the model diagnostics and predictability.

6.3.1 Posterior predictive check

Figure 8 shows the graph for the posterior predictive check between the poisson model and the negative binomial model (with interaction).

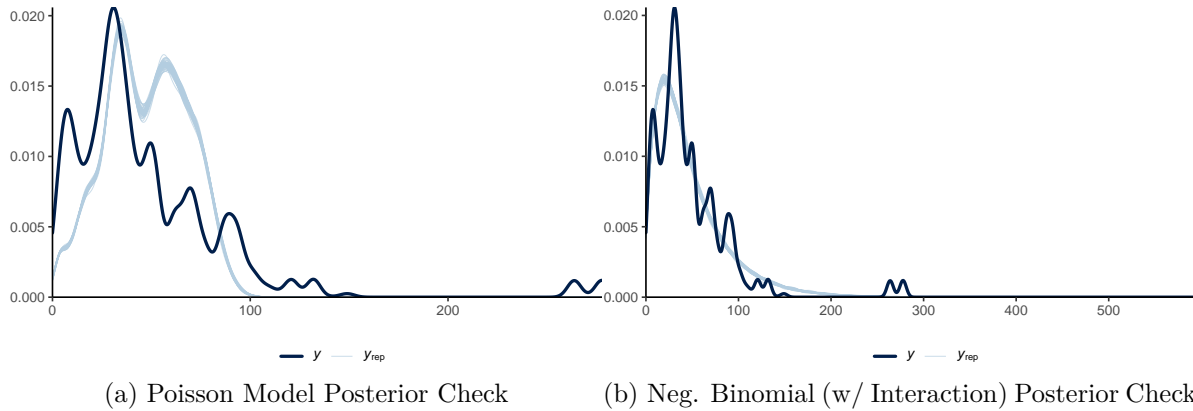


Figure 8: Comparing posterior prediction checks for the Poisson model and the Negative Binomial Model With Interaction

6.3.2 Leave-One-Out (LOO) Comparison

Table 6 compares each of the models loo values. The model with higher (less negative) loo value is one with the best predictive power. From the table below, we can see that our negative binomial model with interaction has the best predictive power of the proposed models.

Table 6: Model Comparison using LOO

	elpd_diff	se_diff
neg_binomial_model_interaction	0.00	0.00
neg_binomial_model_basic	-473.33	29.27
poisson_model	-223601.09	3585.48

6.3.3 Model summary

The below table describes the summary statistics for the explanatory model chosen in our study.

Table 7: Negative Binomial Model Summary with Interaction Terms

term	estimate	std.error
(Intercept)	4.2989384	0.0172791
demographicMixed Adult	-0.2802075	0.0143124
demographicWomen	-0.3909558	0.0139626
demographicYouth	-0.9281969	0.0144416
shelter_typeTransitional	-0.9683637	0.0185924
service_type24-Hour Women's Drop-in	-0.3797025	0.0342079
service_typeAlternative Space Protocol	-2.7504061	0.0398229
service_typeShelter	0.0564118	0.0159593
service_typeTop Bunk Contingency Space	-1.7642895	0.0315280
service_typeWarming Centre	-0.4184851	0.0402914
demographicMixed Adult:shelter_typeTransitional	0.4586099	0.0291345
demographicWomen:shelter_typeTransitional	0.3632607	0.0298635
demographicYouth:shelter_typeTransitional	0.8151307	0.0260808

6.3.4 QQ-Plot For Explanatory Model

Figure 9 plots the QQ Plot for the explanatory model chosen in our study.

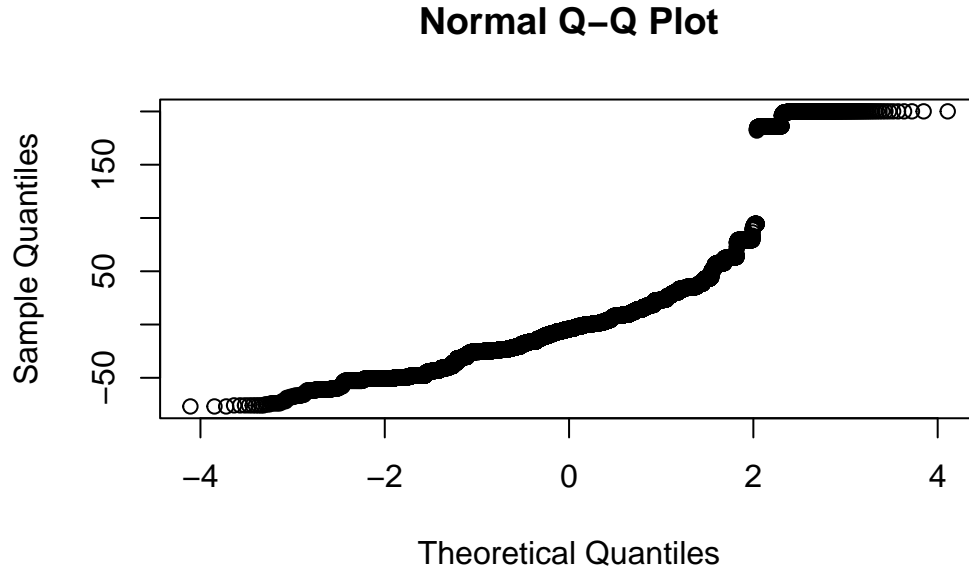


Figure 9: QQ-Plot for explanatory negative binomial (interaction) model

6.3.5 Pareto K-Diagnostics

Figure 10 plots the Pareto K-Diagnostics for our proposed model. A k-value of less than 0.7 indicates that our proposed model fits the data without any issues in convergence.

6.3.6 Residual Plot for explanatory model

Figure 11 shows the residual plot for our explanatory model. From the below graph we can see that residuals are spread randomly along the intercept line indicating our model is a good fit for the data set.

6.3.7 Markov Chain Monte Carlo Convergence Diagnostic

Figure 12 shows the convergence of our predictors in our explanatory model using the Markov Chain Monte Carlo (MCMC) convergence algorithm.

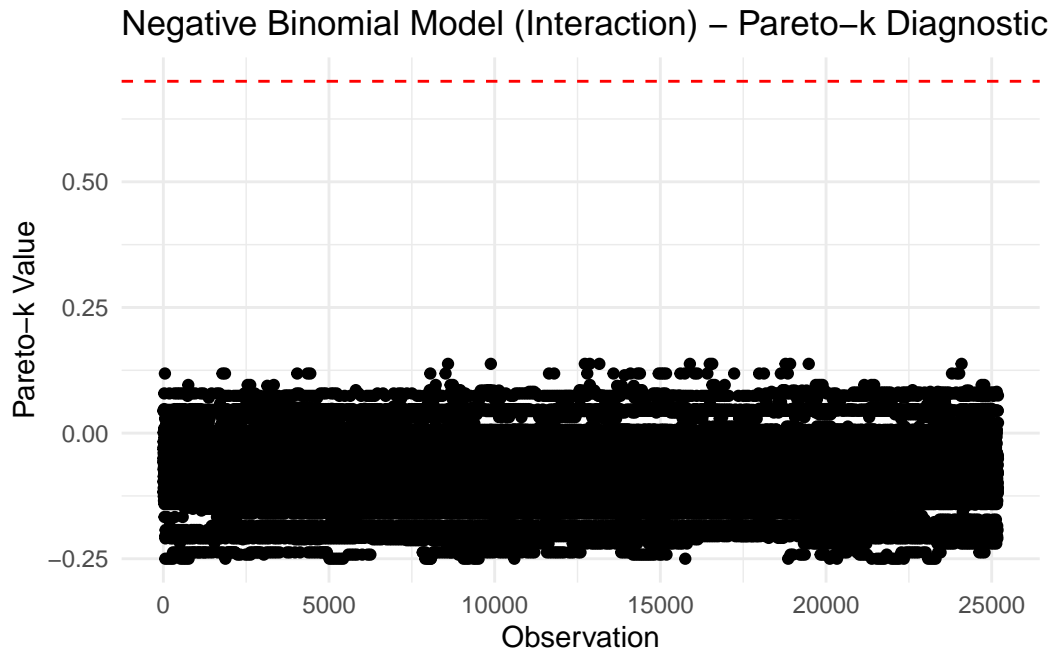


Figure 10: Graph of pareto K-Diagnostics for explanatory model

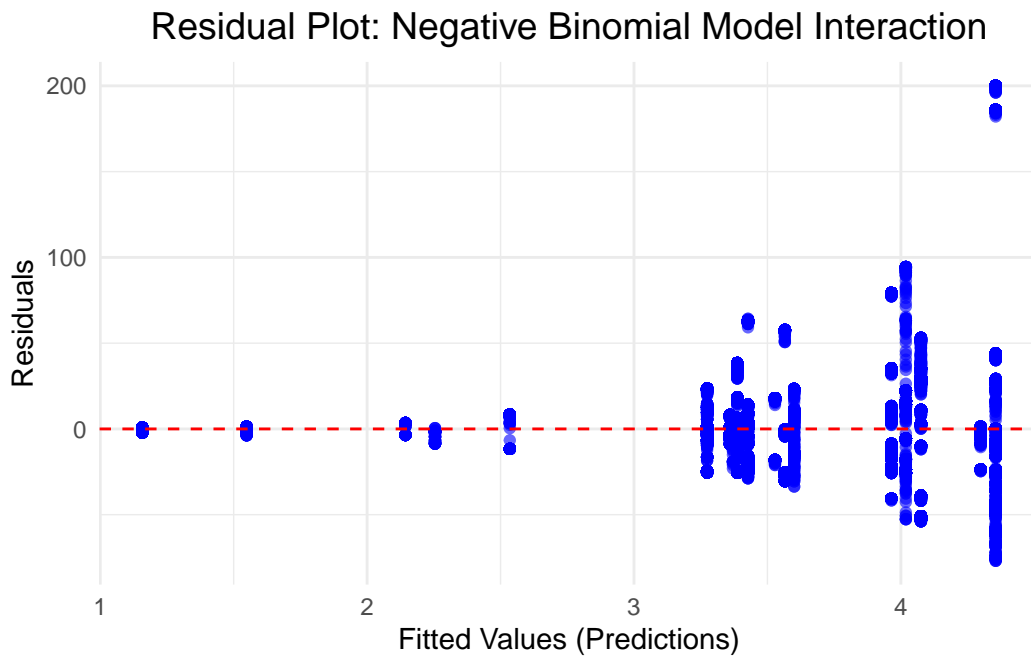


Figure 11: Residual Plot for explanatory model

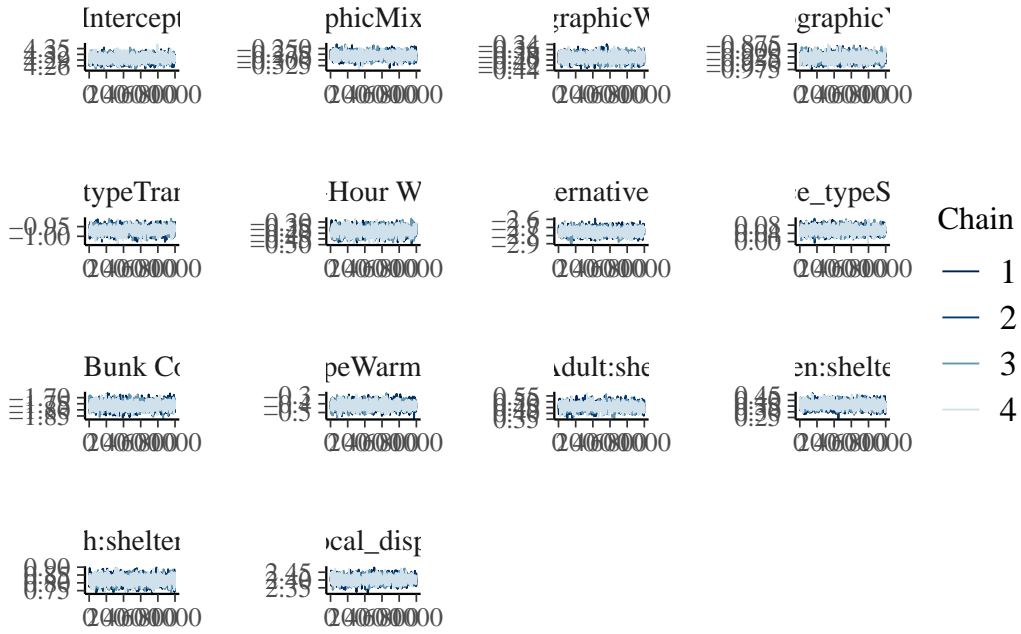


Figure 12: Checking the convergence of the predictors using MCMC algorithm

6.3.8 R-Hat Convergence Diagnostic

Figure 13 shows the variability in the convergence of the predictor chains using the MCMC convergence algorithm of our predictors. \hat{R} values of around 1.01 indicate model has converged successfully without any extreme variability.

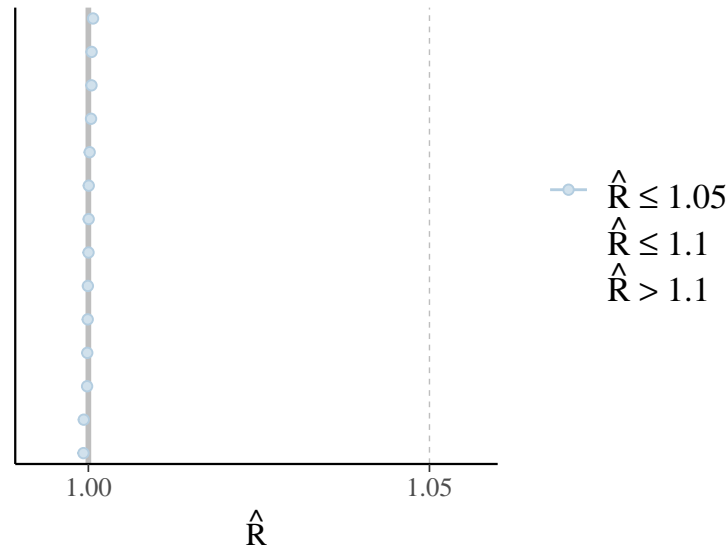


Figure 13: Checking the R-hat variability for predictors

References

- Alexander, Rohan. 2024. *Telling Stories with DataScience*. <https://tellingstorieswithdata.com>.
- Arel-Bundock, Vincent. 2022. “modelssummary: Data and Model Summaries in R.” *Journal of Statistical Software* 103 (1): 1–23. <https://doi.org/10.18637/jss.v103.i01>.
- Bolker, Ben, and David Robinson. 2024. *Broom.mixed: Tidying Methods for Mixed Models*. <https://CRAN.R-project.org/package=broom.mixed>.
- Canadian Red Cross. 2024. *Canadian Red Cross*. <https://www.redcross.ca/>.
- City of Toronto. 2024a. *Toronto Shelter System*. <https://housing-infrastructure.canada.ca/homelessness-sans-abri/reports-rapports/shelter-cap-hebergement-2023-eng.html>.
- . 2024b. *Toronto Shelter System*. <https://housing-infrastructure.canada.ca/homelessness-sans-abri/reports-rapports/shelter-cap-hebergement-2022-eng.html>.
- Firke, Sam. 2024. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. <https://CRAN.R-project.org/package=janitor>.
- Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. <https://CRAN.R-project.org/package=opendatatoronto>.
- Glenda Laws. 2013. *Emergency Shelter Networks in Urban Area*. https://www.tandfonline.com/doi/abs/10.2747/0272-3638.13.2.99?casa_token=3AdUNOVOJX8AAAAA:jWGbxbc1lv-SbkivfYQEKy-S93pUbYbgvPXEWJ6Lo1ywWaIuzShSWAv2KpdgyjN16l0mW2j9OyAZJxA.
- Goodrich B, et al. 2024. *Goodrich b, Gabry j, Ali i & Brilleman s. (2024). Rstanarm: Bayesian Applied Regression Modeling via Stan. R Package Version 2.32.1* <https://McStan.org/Rstanarm>. <https://mc-stan.org/rstanarm>.

- Kuhn, and Max. 2008. “Building Predictive Models in r Using the Caret Package.” *Journal of Statistical Software* 28 (5): 1–26. <https://doi.org/10.18637/jss.v028.i05>.
- Nicole Ireland. 2023. *Growing Number of Homeless People Turning to ERs for Shelter and Warmth in Ontario, Study Says*. <https://www.cbc.ca/news/canada/toronto/homeless-people-emergency-room-shelter-1.7042041>.
- Out of the cold foundation. 2021. *Packed Toronto Homeless Shelter System Poised for Breakdown*. <https://outofthecold.org/packed-toronto-homeless-shelter-system-poised-for-breakdown-report/>.
- R Core Team. 2024a. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- . 2024b. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoş Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to ‘Apache’ ‘Arrow’*. <https://CRAN.R-project.org/package=arrow>.
- Robinson, David, Alex Hayes, and Simon Couch. 2024. *Broom: Convert Statistical Objects into Tidy Tibbles*. <https://CRAN.R-project.org/package=broom>.
- Schwan K. et al., year = 2018, title = The Pan-Canadian Women’s Housing & Homelessness Survey. Toronto, ON: Canadian Observatory on Homelessness. n.d.
- Toronto Shelter & Support Services. 2024a. *Daily Shelter & Overnight Service Occupancy & Capacity*. <https://open.toronto.ca/dataset/daily-shelter-overnight-service-occupancy-capacity/>.
- . 2024b. *Daily Shelter & Overnight Service Usage*. <https://www.toronto.ca/city-government/data-research-maps/research-reports/housing-and-homelessness-research-and-reports/shelter-census/>.
- . 2024c. *Toronto Shelter System Flow*. https://www.toronto.ca/wp-content/uploads/2018/12/9547-A1600035_TSS_FinalDraft_V3_Dec4_Blue_SimpleAccessible_updated2.pdf.
- . 2024d. *Toronto Shelter System Flow*. <https://open.toronto.ca/dataset/toronto-shelter-system-flow/>.
- Vehtari A et al. 2024. *Loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models*. <https://mc-stan.org/loo/>.
- Wickham et al. 2024. *R for Data Science*. <https://r4ds.had.co.nz/factors.html>.
- Xie Y. 2024. *A General-Purpose Package for Dynamic Report Generation in r_*. *R Package Version 1.48*. <https://yihui.org/knitr/>.
- Zakari Omar. 2023. *Toronto Creates Another Winter Plan to Address Rising Homelessness. Is It Enough?* <https://torontoobserver.ca/2023/11/14/toronto-creates-another-winter-plan-to-address-rising-homelessness-is-it-enough/>.
- Zhu, Hao. 2024. *kableExtra: Construct Complex Table with ‘Kable’ and Pipe Syntax*. <https://CRAN.R-project.org/package=kableExtra>.