The analysis of occupancy rates of beds in the Shelter Support and Housing Administration division's Shelter Management Information System*

Fengyuan Tang

27 April 2022

Abstract

In this paper, my goal is to analyze the variables that have significant effects on the proportion of actual bed capacity that is occupied for the reporting date. It was found that contributing factors include the programs' locations, gender, age, household size of the service user group, the type of overnight service provided, the program area. Furthermore, the number of beds showing as available for occupancy, rooms that a program has been approved to provide, rooms showing as occupied by a shelter user, rooms that are showing as available for occupancy that are not occupied, and rooms that are not currently available also have significant influences on the response variable. This matters because it provides updated information about the ways to effectively improve occupancy rates of beds in the shelter and overnight service programs administered by SSHA.

1 Introduction

The Daily Shelter & Overnight Service Occupancy & Capacity provides a list of active overnight shelters and allied services in the Shelter Support and Housing Administration division's Shelter Management Information System database (Toronto Open Data Portal 2021). It provides the daily updated information about shelters and overnight service programs administered by SSHA, including the program's operator, location, classification, occupancy and capacity (Toronto Open Data Portal 2021). This dataset was conducted in 2021 (Toronto Open Data Portal 2021). Based on the current events relating to the daily shelters and overnight service occupancy as well as capacity, we wanted to focus on the occupancy rates of beds in this system.

In the data section, I would examine all variables and possible data sets that are similar. I would also contain graphs which help people to understand what the variables look like. Furthermore, I would convey the features of this data, including summary statistics and relationships between the variables. In the model section, I conduct some explanatory data analysis as well as conduct the EDA. This includes the numerical Summaries and the graphical summaries. The first step is choosing a starting model, which is also called the full model. The second step is to ensure there is no multicollinearity in the model, so I build new models. Then, I check the 2 conditions to make sure that i can use residual plots to analyze the model. After that, I use residual plots to identify potential violations against model assumptions (linearity, normality, constant variance, and uncorrelatedness). The next step is to explore model transformations to correct assumption violations and fit a new model with transformed variables. When I begin to reduce the model, I conduct automated selection and manual selection. Model comparisons determine which model is better so I use ANOVA to compare these models. Then, I apply the diagnostic plots on both the manual reduced model and the auto reduced model. The results section displays the findings, including summary statistics, tables, graphs, images and statistical analysis.

The results interpret that contributing factors include the city of the location of the program (LOCATION_CITY),the gender, age, household size of the service user group (SECTOR), the type of overnight

*Code and data are available at: https://github.com/FengyuanTang/Sta304finalpaper service provided (OVERNIGHT_SERVICE_TYPE), the PROGRAM_AREA, which includes base shelter and overnight services system, or a temporary response program. Furthermore, the number of beds showing as available for occupancy (T_CAPACITY_ACTUAL_BED), rooms that a program has been approved to provide (T_CAPACITY_FUNDING_BED), rooms showing as occupied by a shelter user (T_OCCUPIED_BEDS), rooms that are showing as available for occupancy that are not occupied as of the occupancy date (T_UNOCCUPIED_BEDS), rooms that are not currently available in a program (T_UNAVAILABLE_BEDS) also have significant impacts on the proportion of actual bed capacity that is occupied for the reporting date. Moreover, I would include discussions of some interesting points and weaknesses of our paper.

The importance is that, as people have been impacted by COVID-19, we believe that there would be some important changes in the daily shelters and overnight service programs. Our aim is to analyze the factors that have significant effects on the proportion of actual bed capacity that is occupied for the reporting date. Therefore, I build models to find its important contributing factors. In this way, it promotes the efficiency of overnight shelters and allied services in the system. Furthermore, it could development government policies and make the society more satisfied. Moreover, there would be an improvement of living standards for those people who have a demand for shelters and overnight services.

2 Data

In this paper, we focused on analyzing the contributing factors of the proportion of actual bed capacity that is occupied for the reporting date. We used R programming language (R Core Team 2020) tidy-verse (Wickham et al. 2019), janitor (Firke 2021), readxl (Wickham and Bryan 2019), knitr (Xie 2021), ggplot2 (Wickham 2016), dplyr (Wickham et al. 2021), patchwork (Pedersen 2020), car (Fox and Weisberg 2019) and readr (Wickham, Hester, and Bryan 2022).

The data-set is called "Daily Shelter & Overnight Service Occupancy & Capacity" (Toronto Open Data Portal 2021). The variables I used include LOCATION_CITY, which is the city of the location of the program (Toronto Open Data Portal 2021). SETOR is defined as the means of categorizing homeless shelters based on the gender, age and household size of the service user group(s) served at the shelter location (Toronto Open Data Portal 2021). PROGRAM_MODEL is a classification of shelter programs as either Emergency or Transitional (Toronto Open Data Portal 2021). OVERNIGHT_SERVICE_TYPE identifies the type of overnight service being provided (Toronto Open Data Portal 2021). PROGRAM_AREA indicates whether the program is part of the base shelter and overnight services system, or is part of a temporary response program (Toronto Open Data Portal 2021). CAPACITY_ACTUAL_BED shows the number of beds showing as available for occupancy in the Shelter Management Information System (Toronto Open Data Portal 2021). CAPACITY_FUNDING_BED displays the number of beds that a program has been approved to provide (Toronto Open Data Portal 2021). OCCUPIED_BEDS illustrates the number of beds showing as occupied by a shelter user in the Shelter Management Information System for this program for this date (Toronto Open Data Portal 2021). UNOCCUPIED_BEDS is the number of beds that are showing as available for occupancy that are not occupied as of the occupancy date (Toronto Open Data Portal 2021). This is calculated as CAPACITY_ACTUAL_BED minus OCCUPIED_BEDS (Toronto Open Data Portal 2021). UNAVAILABLE_BEDS shows the number of beds that are not currently available in a program (Toronto Open Data Portal 2021). Specifically, this is calculated as CAPACITY_FUNDING_BED minus CAPACITY_ACTUAL_BED (Toronto Open Data Portal 2021). The response variable is called OCCUPANCY_RATE_BEDS, which displays the proportion of actual bed capacity that is occupied for the reporting date (Toronto Open Data Portal 2021).

Since the data-set is not tidy, I filter the missing values in some important variables and only keep the data that is "Bed Based Capacity". Specifically, the variable CAPACITY_TYPE is defined as whether the capacity for this program is measured in rooms or beds (Toronto Open Data Portal 2021). The project creates a histogram for the proportion of actual bed capacity that is occupied for the reporting date. This is calculated as OCCUPIED_BEDS divided by CAPACITY_ACTUAL_BED (Toronto Open Data Portal 2021). I also make plots about some possible predictors, including CAPACITY_ACTUAL_BED, CAPACITY_FUNDING_BED, OCCUPIED_BEDS, and UNOCCUPIED_BEDS. Then, I choose my starting model by using the results of EDA and my common sense. Drawing scatter-plots between yi and y_hat and that between numerical predictors can be used to check Condition 1 and 2. The Residual vs. Fitted, Residual vs. Predictors and Residual QQ Plot can decide whether each regression modelling assumption is satisfied. Since power transform fails to work if variable contains 0, we add 0.0000001 instead. After applying the box-cox transformation, I use mutate to create transformed variables and fit a new model. This is called candidate model 1, and model comparisons can determine which model is better. Same methods are applied, including checking two conditions, creating residual plots, and so on.

2.1 Summary statistics

It shows that the proportion of actual bed capacity has a minimum of 7.14, and a maximum of 100. The average is 95.62. The number of rooms showing as available for occupancy has a minimum of 1, and a maximum of 235. The average is 34.73. The number of rooms that a program is has been approved to provide has a minimum of 2, and a maximum of 235. The average is 36.5. The number of rooms showing as occupied has a minimum of 1, and a maximum of 234. The average is 33.85. The number of beds that are showing as available for occupancy that are not occupied has a minimum of 0, and a maximum of 39. The average is 0.88. The number of beds that are not currently available in a program has a minimum of 0, and a maximum of 180.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.14   96.00  100.00   95.64  100.00  100.00
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   17.00   27.00   34.73   45.00  235.00
```

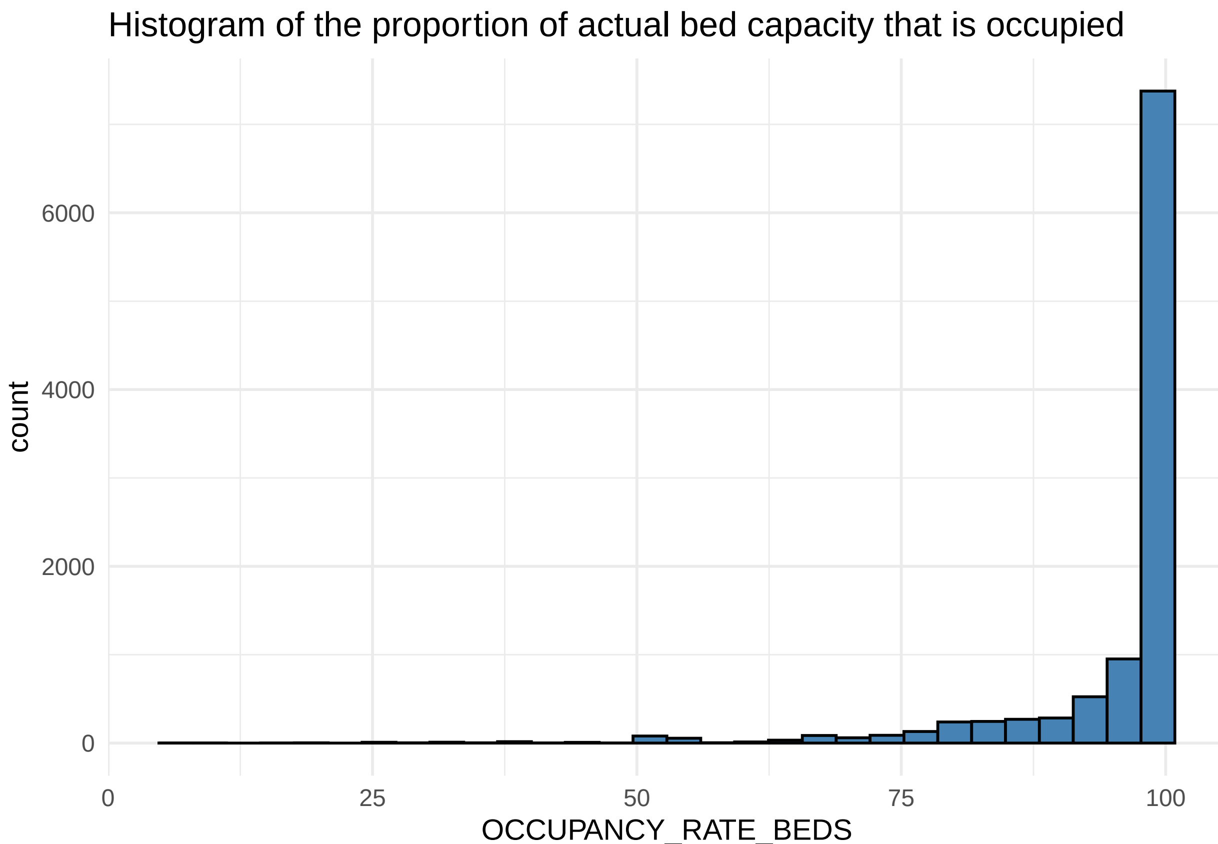
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   19.00   28.00   36.49   50.00  235.00
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   16.00   25.00   33.85   44.00  234.00
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.8817  1.0000 39.0000
```

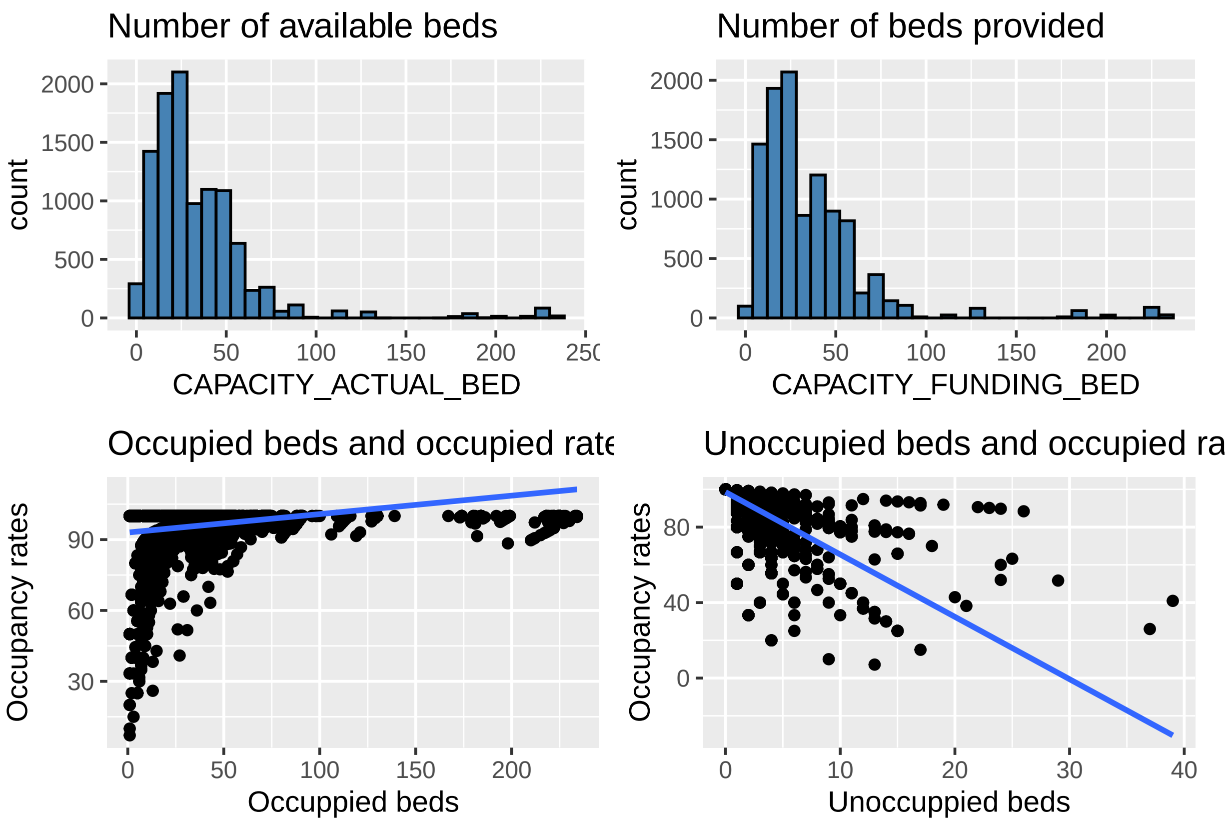
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.762   1.000 180.000
```

2.2 Exploratory Data Analysis



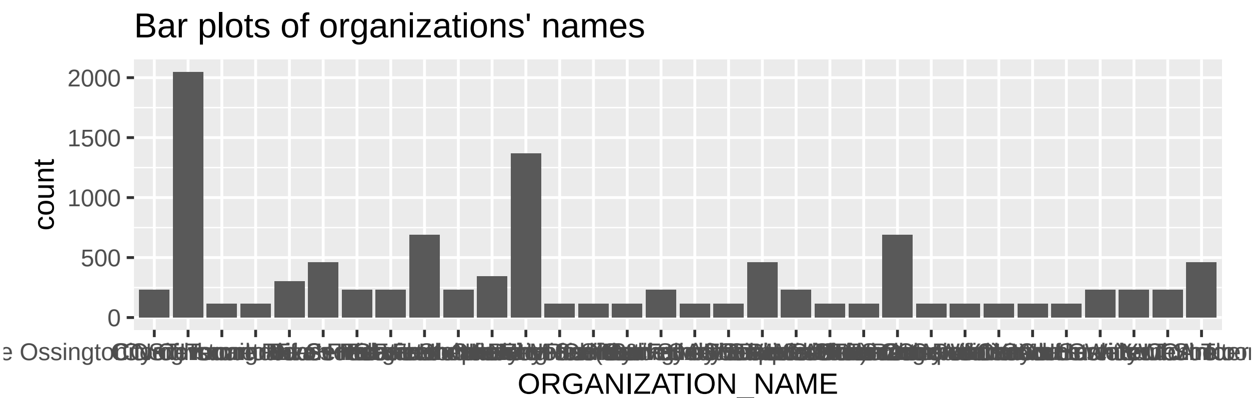
I

should conduct the exploratory data analysis in order to build the starting model. Thus, Figure **??** is a histogram of the response variable. The graph is extremely right skewed. It will be much better if it is normal distribution. This problem can be fixed later by using model transformation.



From Figure **??**, **??**, **??**, and **??**, I infer that variables “CAPACITY_ACTUAL_BED”, “CAPACITY_FUNDING_BED”, “OCCUPIED_BEDS” and “UNOCCUPIED_BEDS” could have influence on the response variable. Therefore, I include them in the starting model.

Then, I create bar plots for some categorical variables



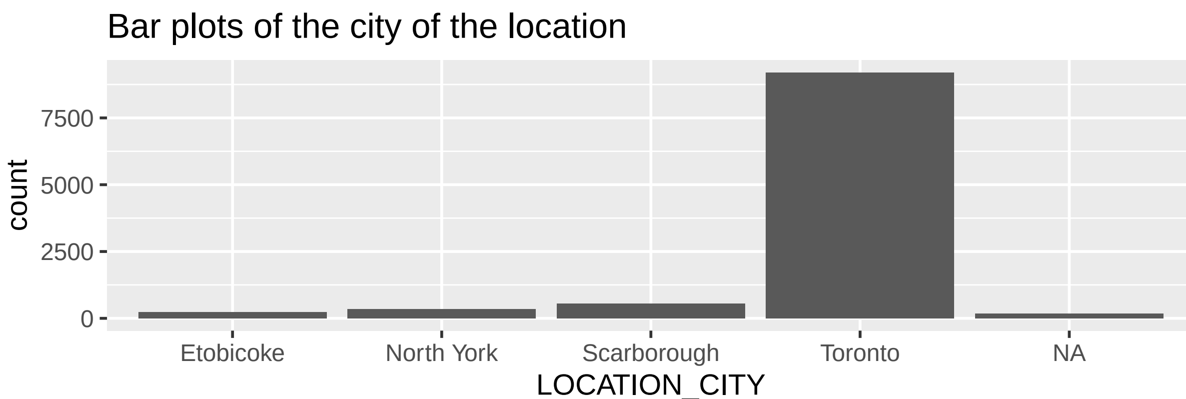
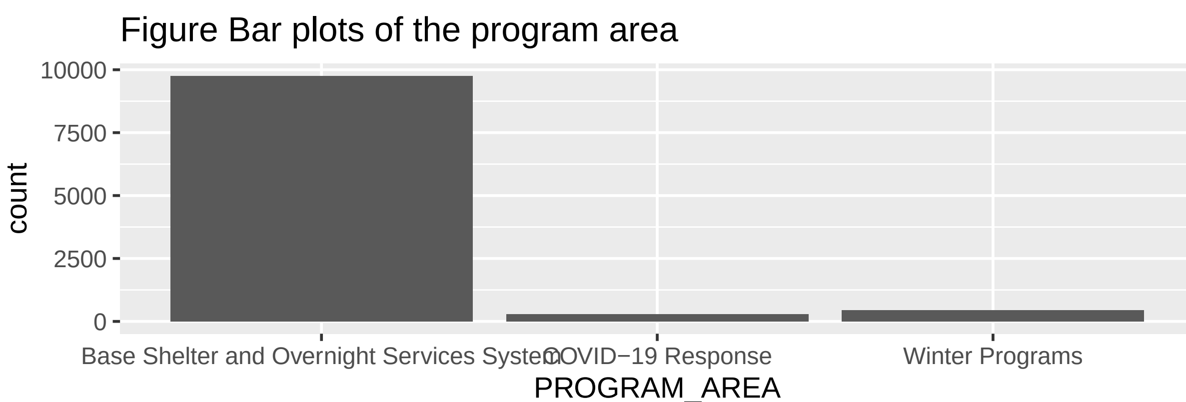
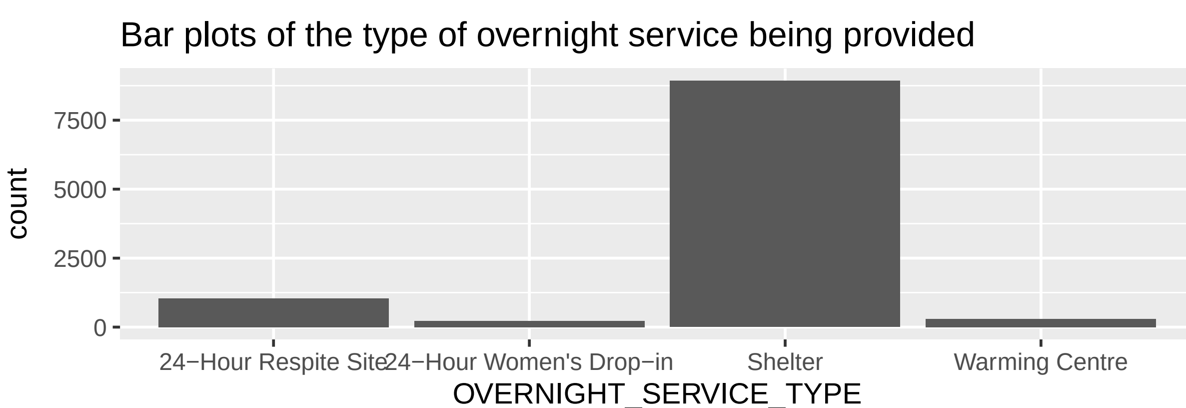


Figure **??** and Figure **??** show that the data volume is unbalanced. Thus, the credibility is weak and we do not need to include them.





From the website, PROGRAM_AREA displays whether the program is part of the base shelter and overnight services system, or is part of a temporary response program (Toronto Open Data Portal 2021). Figure **??** illustrates that most programs are in the type of Base Shelter and Overnight Services System. These programs are intended to be regular, year-round, and permanent (Toronto Open Data Portal 2021). Also, the variable OVERNIGHT_SERVICE_TYPE shows the type of overnight service provided (Toronto Open Data Portal 2021). This includes Shelter, 24-Hour Respite, Motel/Hotel, Interim Housing, Warming Center, 24-Hour Woman's Drop-in, Isolation/Recovery Site (Toronto Open Data Portal 2021). From Figure **??**, it shows that shelters are the most type of overnight services provided in this system.

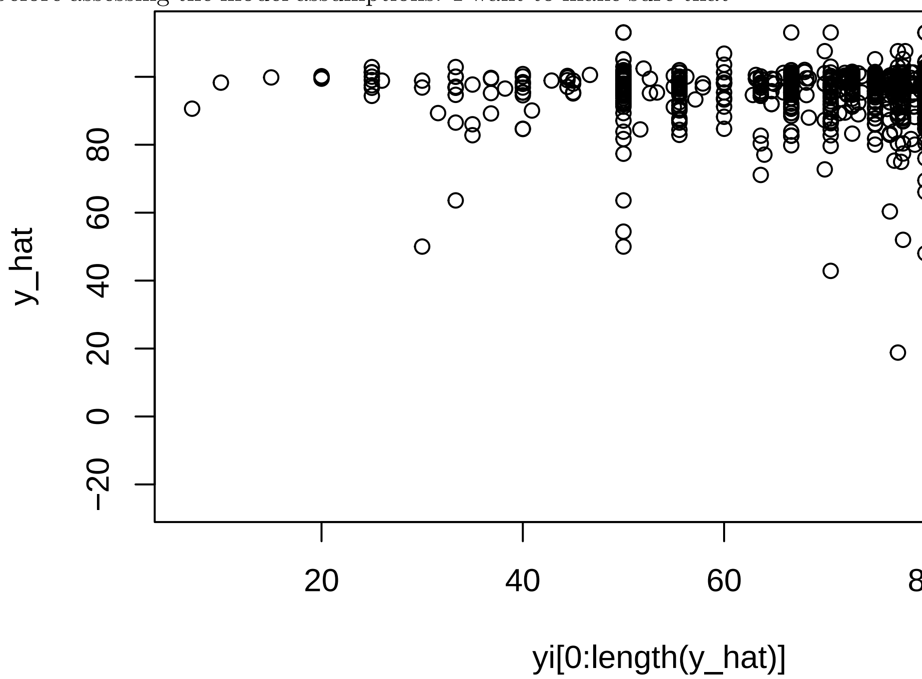
3 Model

The second step is to create my starting model by using common sense and results of EDA. Also, predictors can't have overlapping information.

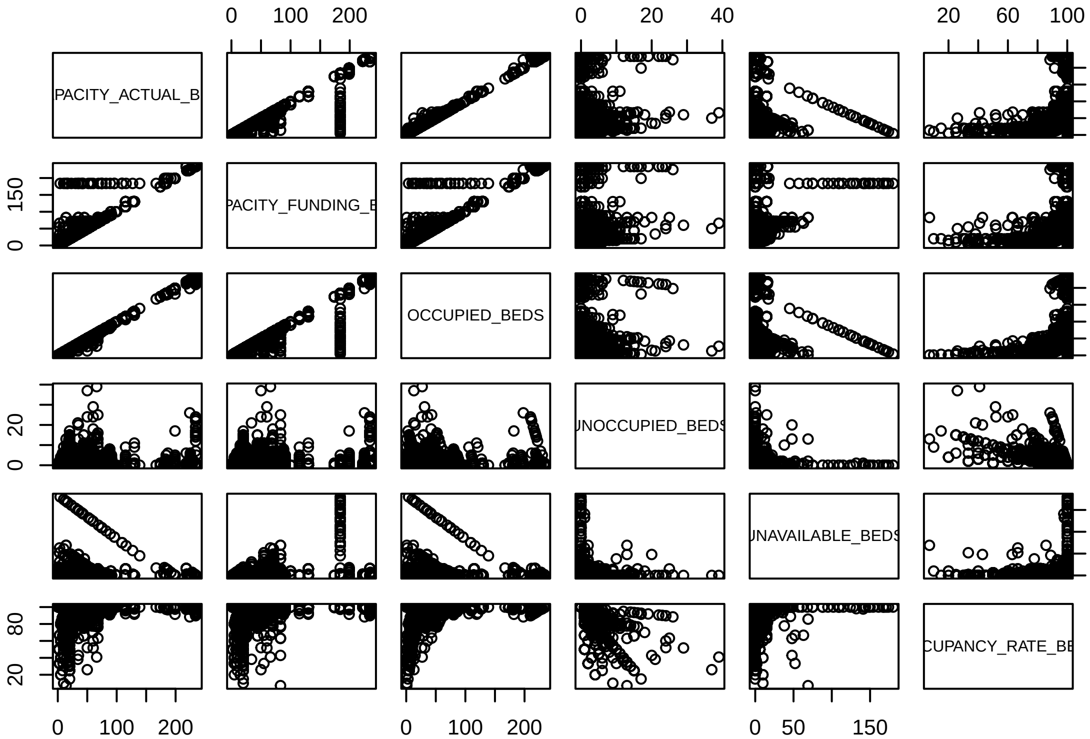
```
##
## Call:
## lm(formula = OCCUPANCY_RATE_BEDS ~ LOCATION_CITY + SECTOR + PROGRAM_MODEL +
##     OVERNIGHT_SERVICE_TYPE + PROGRAM_AREA + CAPACITY_ACTUAL_BED +
##     CAPACITY_FUNDING_BED + OCCUPIED_BEDS + UNOCCUPIED_BEDS +
##     UNAVAILABLE_BEDS, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.089  -0.548   0.829   2.666  66.421
##
## Coefficients: (2 not defined because of singularities)
##                                                Estimate Std. Error t value
## (Intercept)                                   97.231860   1.032721  94.151
## LOCATION_CITYNorth York                       -4.429749   0.607316  -7.294
## LOCATION_CITYScarborough                      -2.842847   0.561329  -5.064
## LOCATION_CITYToronto                          -3.573494   0.460063  -7.767
## SECTORMen                                      3.402597   0.877223   3.879
## SECTORMixed Adult                              3.240465   0.894177   3.624
## SECTORWomen                                    4.383045   0.885533   4.950
## SECTORYouth                                    0.811962   0.884321   0.918
## PROGRAM_MODELTransitional                     -1.588452   0.161343  -9.845
## OVERNIGHT_SERVICE_TYPE24-Hour Women's Drop-in  0.655644   0.533457   1.229
## OVERNIGHT_SERVICE_TYPEShelter                 -0.458066   0.276850  -1.655
## OVERNIGHT_SERVICE_TYPEWarming Centre          -0.893423   0.717104  -1.246
## PROGRAM_AREACOVID-19 Response                 -2.793813   0.511582  -5.461
## PROGRAM_AREAWinter Programs                    0.289538   0.624026   0.464
## CAPACITY_ACTUAL_BED                           -3.184641   0.034786 -91.548
## CAPACITY_FUNDING_BED                           0.010830   0.008837   1.225
## OCCUPIED_BEDS                                  3.246875   0.033768  96.153
## UNOCCUPIED_BEDS                                      NA         NA      NA
## UNAVAILABLE_BEDS                                     NA         NA      NA
##                                               Pr(>|t|)
## (Intercept)                                    < 2e-16 ***
## LOCATION_CITYNorth York                       3.23e-13 ***
## LOCATION_CITYScarborough                      4.17e-07 ***
## LOCATION_CITYToronto                          8.77e-15 ***
## SECTORMen                                     0.000106 ***
## SECTORMixed Adult                             0.000292 ***
## SECTORWomen                                   7.55e-07 ***
```

```
## SECTORYouth                                        0.358549
## PROGRAM_MODELTransitional                           < 2e-16 ***
## OVERNIGHT_SERVICE_TYPE24-Hour Women's Drop-in 0.219082
## OVERNIGHT_SERVICE_TYPEShelter                      0.098043 .
## OVERNIGHT_SERVICE_TYPEWarming Centre               0.212838
## PROGRAM_AREACOVID-19 Response                      4.84e-08 ***
## PROGRAM_AREAWinter Programs                        0.642669
## CAPACITY_ACTUAL_BED                                 < 2e-16 ***
## CAPACITY_FUNDING_BED                               0.220432
## OCCUPIED_BEDS                                       < 2e-16 ***
## UNOCCUPIED_BEDS                                          NA
## UNAVAILABLE_BEDS                                         NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.512 on 10309 degrees of freedom
##   (172 observations deleted due to missingness)
## Multiple R-squared:  0.5498, Adjusted R-squared:  0.5491
## F-statistic: 786.7 on 16 and 10309 DF,  p-value: < 2.2e-16
```

Here are the two conditions we need to check before assessing the model assumptions. I want to make sure that

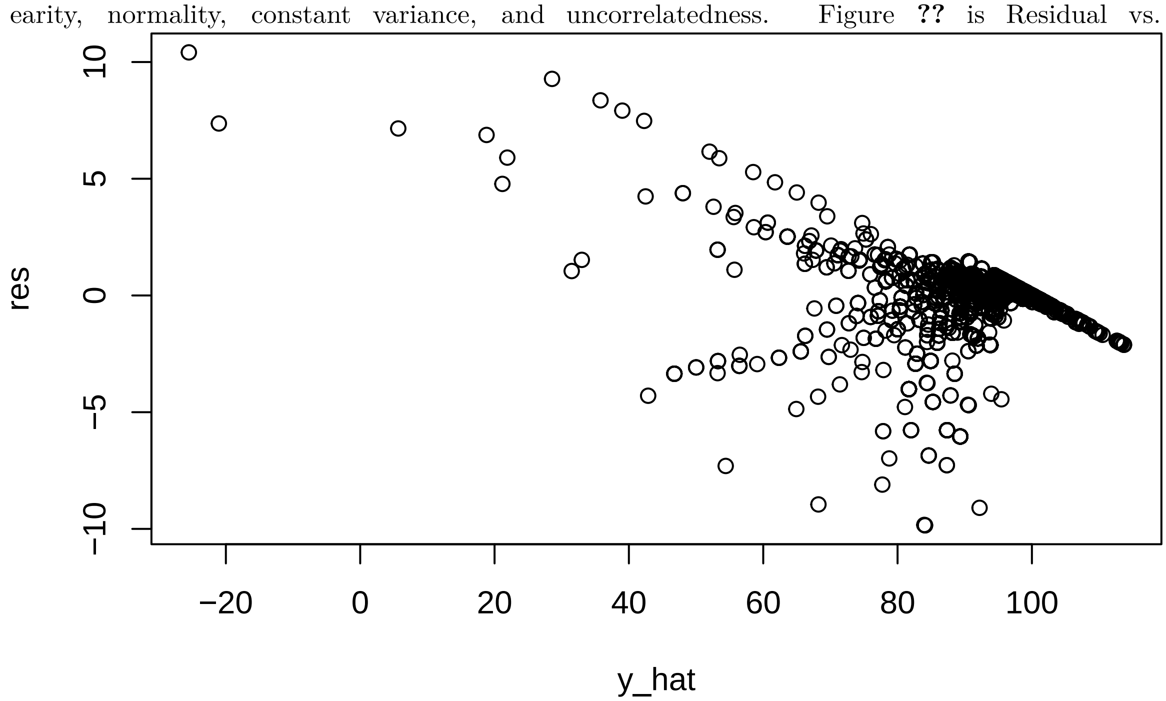


we can use residual plots to analyze the model.



Now, I check the two conditions, which are Figure ?? and Figure ??. I draw a scatter-plot between Yi and Y_hat to check condition 1. It shows that there is a strong pattern between them. Thus, Condition 1 is satisfied. Next, we draw scatter plots between numerical predictors. This graph displays that there is no or linear relationship between these predictors. Therefore, Condition 2 is satisfied.

After that, I use plots to identify potential violations against model assumptions, which are linearity, normality, constant variance, and uncorrelatedness. Figure ?? is Residual vs. Fitted.



In this graph, the linearity holds, but the independence and constant variance can be improved.

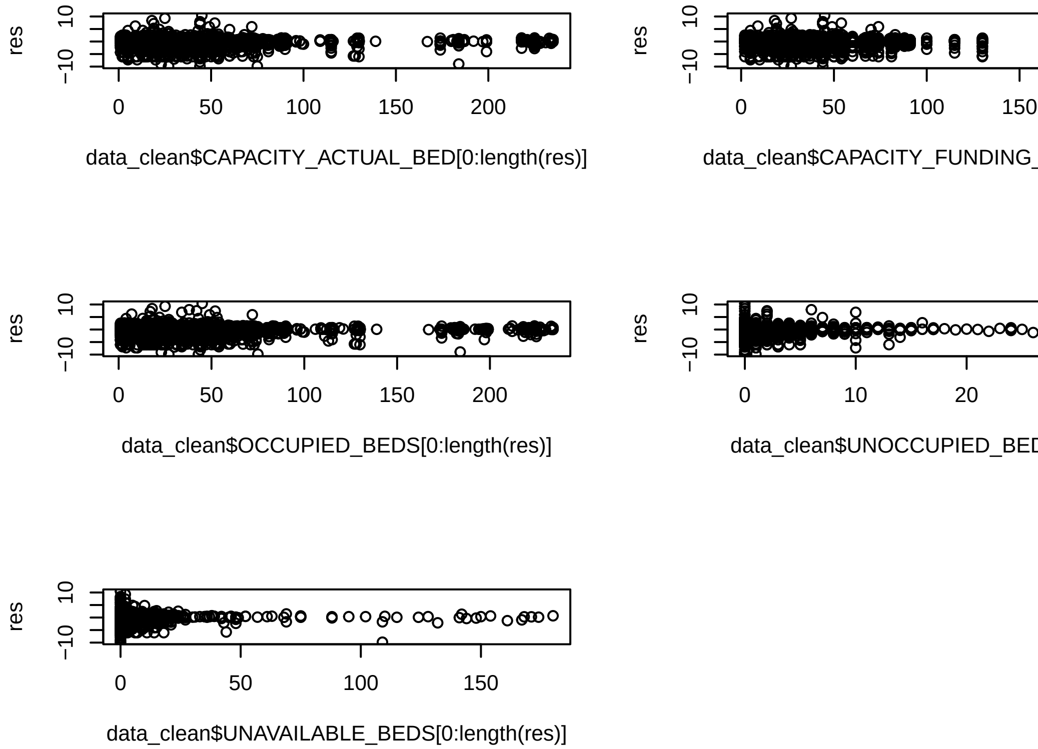


Figure ?? is Residual vs. Predictors.

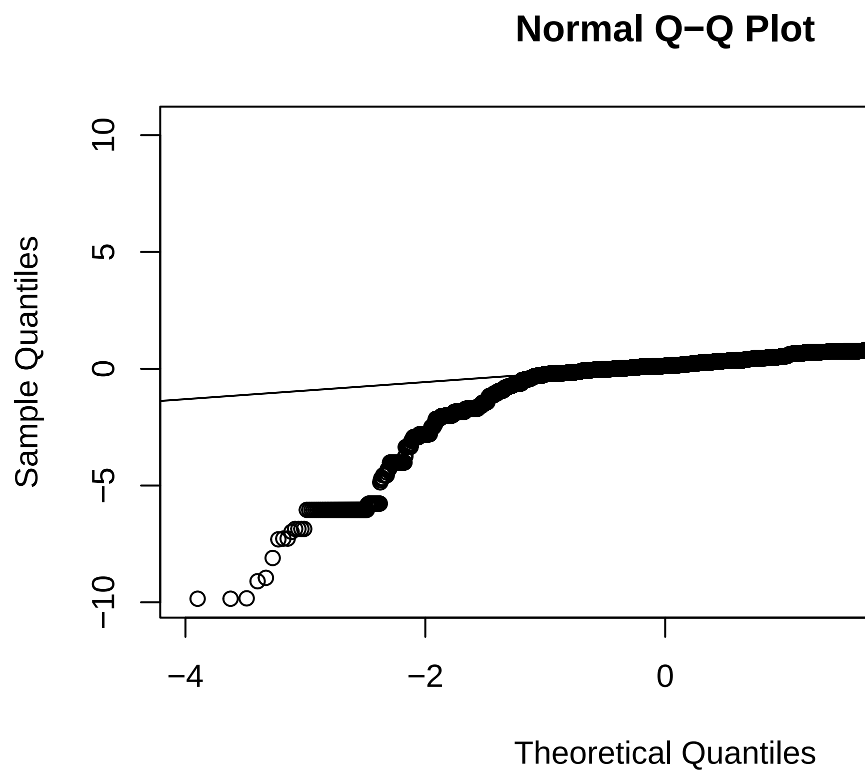


Figure ?? is Normal Quantile-Quantile (QQ) plots.
To summarize, the linearity should hold. However, the constant variance, independence and normality may be violated and can be improved by using model transformations. There is a severe deviation in the Normal QQ plot.

The next step is to explore model transformations to correct assumption violations. Since the power transform fails to work if any variable contains 0, one way to fix this problem is to add 0.0000001 to this variable. After that, I apply box-cox transformation for numerical variables.

```
## bcPower Transformations to Multinormality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.4590        0.46       0.4510       0.4670
## Y2    0.3753        0.38       0.3661       0.3845
## Y3    0.4709        0.47       0.4630       0.4788
## Y4   -0.1140       -0.11      -0.1185      -0.1096
## Y5   -0.1993       -0.20      -0.2043      -0.1943
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                    LRT df       pval
## LR test, lambda = (0 0 0 0 0) 25358.92  5 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                LRT df pval
## LR test, lambda = (1 1 1 1 1) NaN   5   NA
```

Then, I create the transformed variables.

The next step is to fit a new model with transformed variables.

```
##
## Call:
## lm(formula = OCCUPANCY_RATE_BEDS ~ LOCATION_CITY + SECTOR + PROGRAM_MODEL +
##     OVERNIGHT_SERVICE_TYPE + PROGRAM_AREA + T_CAPACITY_ACTUAL_BED +
##     T_CAPACITY_FUNDING_BED + T_OCCUPIED_BEDS + T_UNOCCUPIED_BEDS +
##     T_UNAVAILABLE_BEDS, data = T_data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.865  -0.695   0.300   1.627  55.688
##
## Coefficients:
##                                                 Estimate Std. Error  t value
## (Intercept)                                    94.258742   0.646097  145.889
## LOCATION_CITYNorth York                        -2.926808   0.344139   -8.505
## LOCATION_CITYScarborough                       -1.220282   0.321532   -3.795
## LOCATION_CITYToronto                           -1.584899   0.261095   -6.070
## SECTORMen                                       3.126610   0.503591    6.209
## SECTORMixed Adult                               2.705678   0.509992    5.305
## SECTORWomen                                     3.717832   0.507566    7.325
## SECTORYouth                                     2.430745   0.507221    4.792
## PROGRAM_MODELTransitional                       0.037665   0.095516    0.394
## OVERNIGHT_SERVICE_TYPE24-Hour Women's Drop-in   1.171274   0.302438    3.873
## OVERNIGHT_SERVICE_TYPEShelter                  -0.160162   0.157648   -1.016
## OVERNIGHT_SERVICE_TYPEWarming Centre            0.689692   0.408355    1.689
## PROGRAM_AREACOVID-19 Response                  -3.290231   0.277578  -11.853
## PROGRAM_AREAWinter Programs                     0.266227   0.355685    0.748
## T_CAPACITY_ACTUAL_BED                         -46.537773   0.306342 -151.915
## T_CAPACITY_FUNDING_BED                          2.622009   0.190429   13.769
## T_OCCUPIED_BEDS                                43.397663   0.263865  164.469
## T_UNOCCUPIED_BEDS                               0.400910   0.021217   18.896
## T_UNAVAILABLE_BEDS                              0.014244   0.003596    3.962
##                                               Pr(>|t|)
```

```
## (Intercept)                                      < 2e-16 ***
## LOCATION_CITYNorth York                          < 2e-16 ***
## LOCATION_CITYScarborough                        0.000148 ***
## LOCATION_CITYToronto                            1.32e-09 ***
## SECTORMen                                       5.55e-10 ***
## SECTORMixed Adult                               1.15e-07 ***
## SECTORWomen                                     2.57e-13 ***
## SECTORYouth                                     1.67e-06 ***
## PROGRAM_MODELTransitional                       0.693344
## OVERNIGHT_SERVICE_TYPE24-Hour Women's Drop-in 0.000108 ***
## OVERNIGHT_SERVICE_TYPEShelter                   0.309680
## OVERNIGHT_SERVICE_TYPEWarming Centre            0.091259 .
## PROGRAM_AREACOVID-19 Response                    < 2e-16 ***
## PROGRAM_AREAWinter Programs                     0.454182
## T_CAPACITY_ACTUAL_BED                            < 2e-16 ***
## T_CAPACITY_FUNDING_BED                           < 2e-16 ***
## T_OCCUPIED_BEDS                                  < 2e-16 ***
## T_UNOCCUPIED_BEDS                                < 2e-16 ***
## T_UNAVAILABLE_BEDS                              7.49e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.687 on 10307 degrees of freedom
##   (172 observations deleted due to missingness)
## Multiple R-squared:  0.8557, Adjusted R-squared:  0.8554
## F-statistic:  3395 on 18 and 10307 DF,  p-value: < 2.2e-16
```

Now, i begin to reduce our model. Specifically, I apply automated selection because I want to know which one can be removed.

```
## Start:  AIC=26967.62
## OCCUPANCY_RATE_BEDS ~ LOCATION_CITY + SECTOR + PROGRAM_MODEL +
##     OVERNIGHT_SERVICE_TYPE + PROGRAM_AREA + T_CAPACITY_ACTUAL_BED +
##     T_CAPACITY_FUNDING_BED + T_OCCUPIED_BEDS + T_UNOCCUPIED_BEDS +
##     T_UNAVAILABLE_BEDS
##
##                          Df Sum of Sq    RSS   AIC
## - PROGRAM_MODEL           1         2 140137 26966
## <none>                                140135 26968
## - T_UNAVAILABLE_BEDS      1       213 140349 26981
## - OVERNIGHT_SERVICE_TYPE  3       407 140542 26992
## - LOCATION_CITY           3      1042 141178 27038
## - PROGRAM_AREA            2      1949 142084 27106
## - SECTOR                  4      2109 142244 27114
## - T_CAPACITY_FUNDING_BED  1      2578 142713 27154
## - T_UNOCCUPIED_BEDS       1      4855 144990 27317
## - T_CAPACITY_ACTUAL_BED   1    313772 453907 39102
## - T_OCCUPIED_BEDS         1    367776 507911 40262
##
## Step:  AIC=26965.78
## OCCUPANCY_RATE_BEDS ~ LOCATION_CITY + SECTOR + OVERNIGHT_SERVICE_TYPE +
##     PROGRAM_AREA + T_CAPACITY_ACTUAL_BED + T_CAPACITY_FUNDING_BED +
##     T_OCCUPIED_BEDS + T_UNOCCUPIED_BEDS + T_UNAVAILABLE_BEDS
##
##                          Df Sum of Sq    RSS   AIC
```

```
## <none>                                 140137 26966
## + PROGRAM_MODEL          1          2 140135 26968
## - T_UNAVAILABLE_BEDS     1        215 140352 26980
## - OVERNIGHT_SERVICE_TYPE  3       407 140544 26990
## - LOCATION_CITY          3       1057 141194 27037
## - PROGRAM_AREA           2       1950 142087 27104
## - SECTOR                 4       2115 142252 27112
## - T_CAPACITY_FUNDING_BED  1      2581 142718 27152
## - T_UNOCCUPIED_BEDS      1       5230 145367 27342
## - T_CAPACITY_ACTUAL_BED   1    314210 454347 39110
## - T_OCCUPIED_BEDS        1     367913 508051 40263
```

It shows the predictors that are removed, which is “PROGRAM_MODEL”. The AIC in the last model is 26000.22.

```
## Analysis of Variance Table
##
## Model 1: OCCUPANCY_RATE_BEDS ~ LOCATION_CITY + SECTOR + OVERNIGHT_SERVICE_TYPE +
##     PROGRAM_AREA + T_CAPACITY_ACTUAL_BED + T_CAPACITY_FUNDING_BED +
##     T_OCCUPIED_BEDS + T_UNOCCUPIED_BEDS + T_UNAVAILABLE_BEDS
## Model 2: OCCUPANCY_RATE_BEDS ~ LOCATION_CITY + SECTOR + PROGRAM_MODEL +
##     OVERNIGHT_SERVICE_TYPE + PROGRAM_AREA + T_CAPACITY_ACTUAL_BED +
##     T_CAPACITY_FUNDING_BED + T_OCCUPIED_BEDS + T_UNOCCUPIED_BEDS +
##     T_UNAVAILABLE_BEDS
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1  10308 140137
## 2  10307 140135  1    2.1142 0.1555 0.6933
```

The P value here is 0.5843, which is large. This means that the auto reduced model is better than the full model. Therefore, I can create diagnostic plots for Auto_reduced_model.

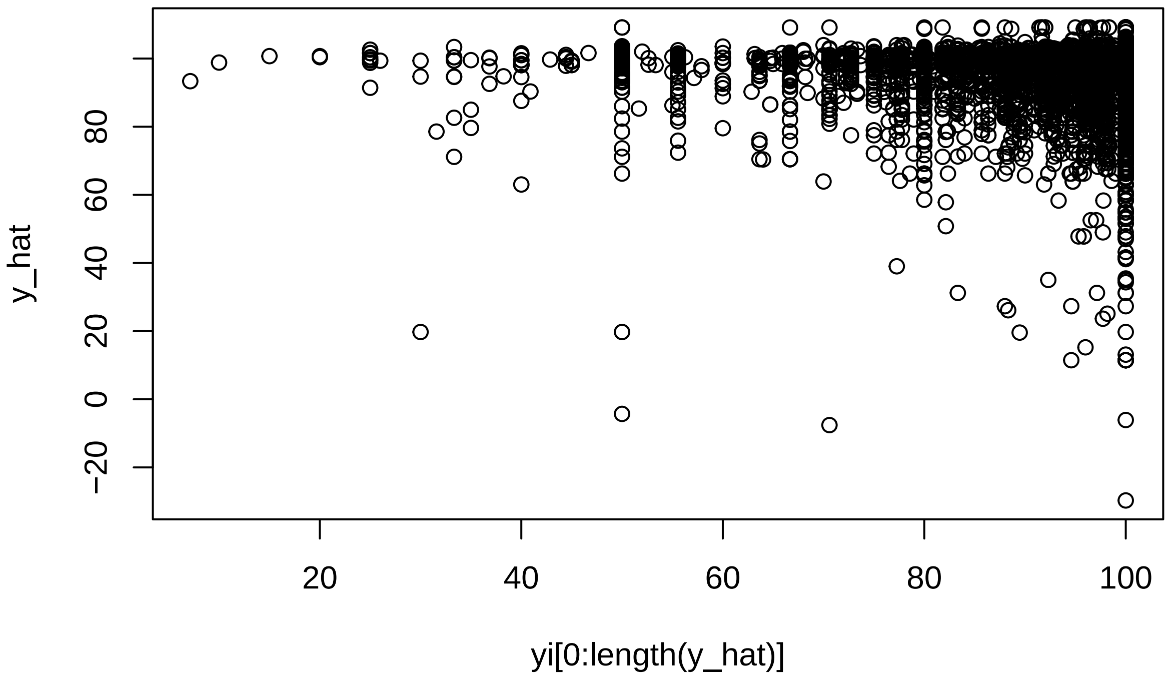
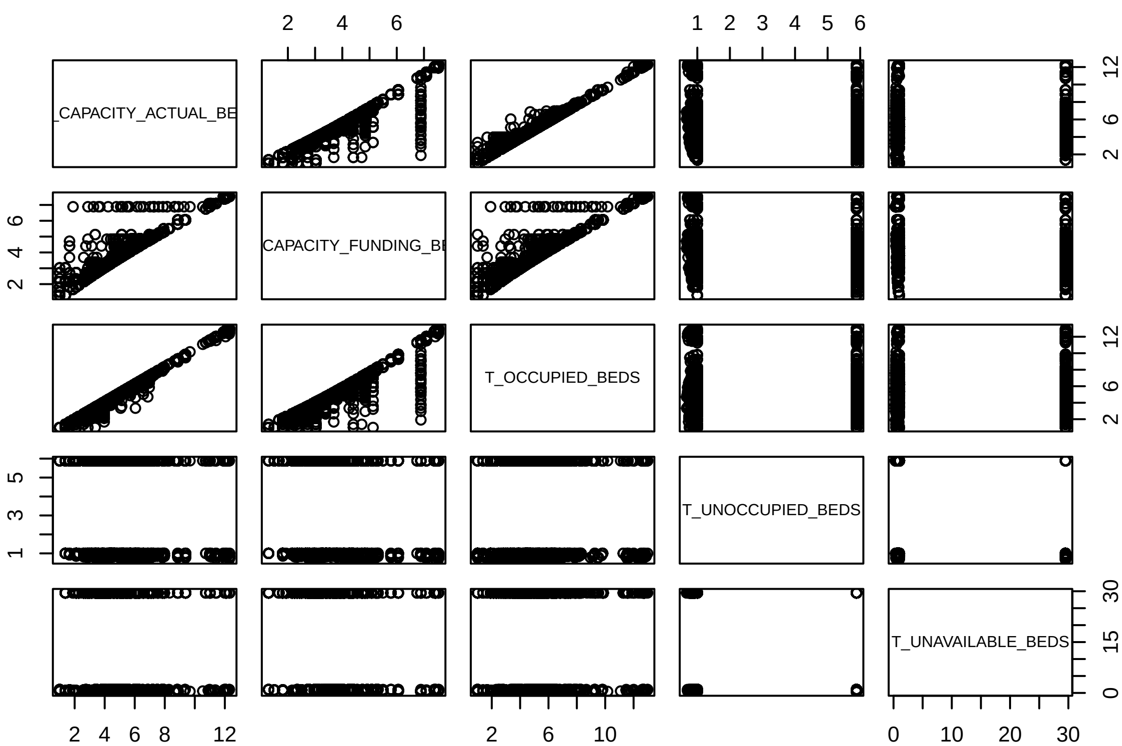


Figure 1: Condition 1: draw a scatter plot between yi and y_hat.

4 Results



Same as before, Figure 1 and Figure ?? display that both Condition 1 and 2 are held in the reduced model.

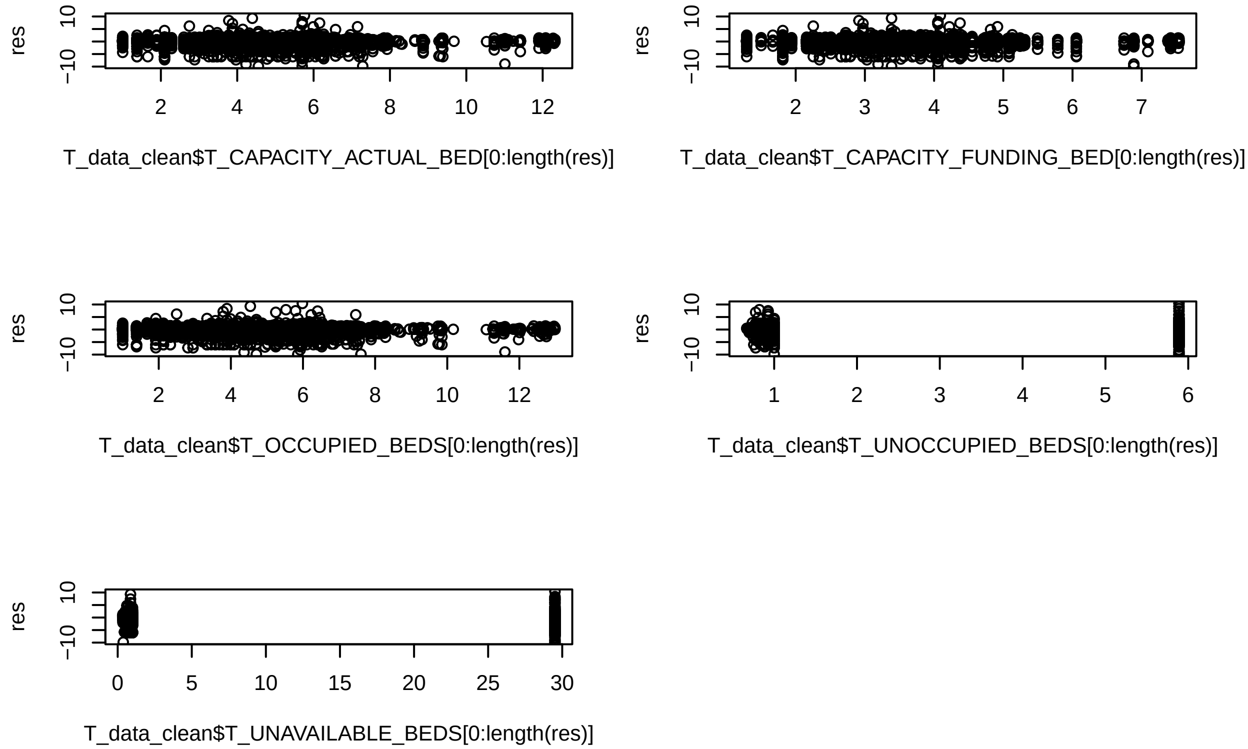
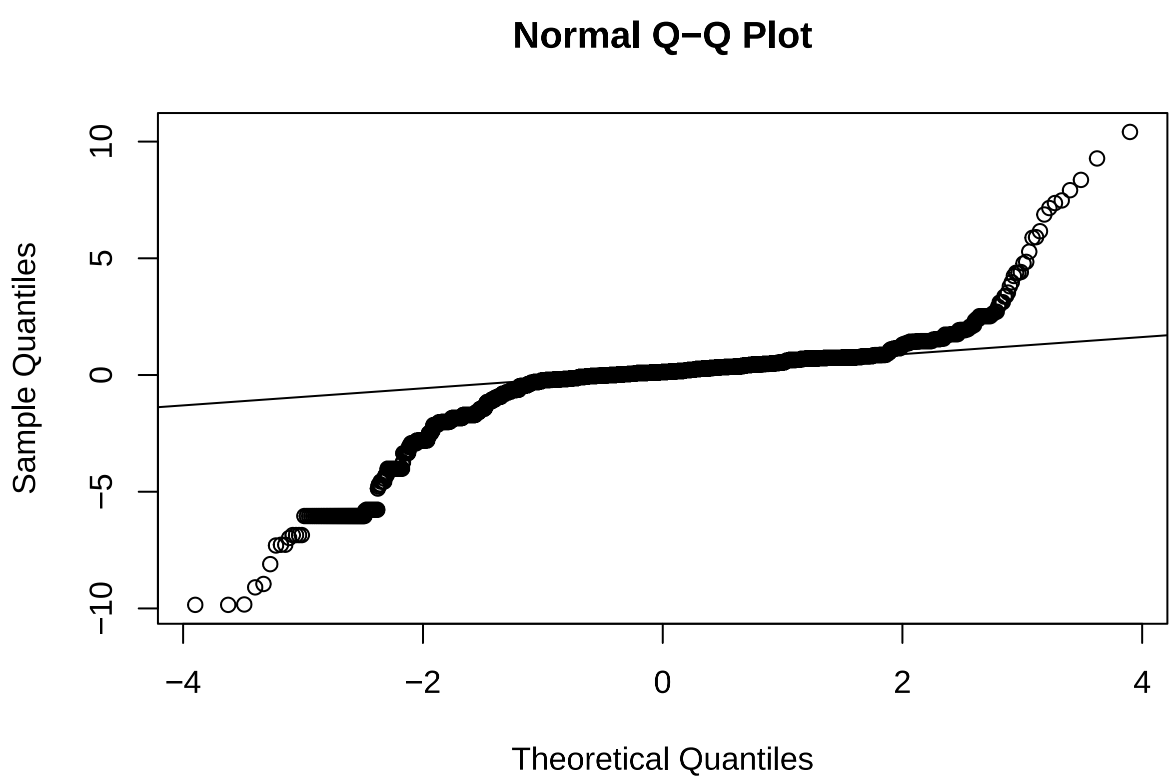


Figure 2: Residual vs. Predictors



From Figure 2 and Figure ??, the linearity and the constant variance should hold. However, the independence and normality may be violated.

The distribution of my response variable is right skewed, which may cause problems. Moreover, it is observed that there is a positive relationship between the number of rooms showing as occupied and the proportion of actual bed capacity that is occupied for the reporting date. It also displays that the number of rooms that are showing as available for occupancy that are not occupied as of the occupancy date has a negative relationship with the proportion of actual bed capacity that is occupied. There is a strong pattern between yi and y_hat, and there is no/linear relationship between predictors, so the two conditions hold. According to the following graphs, the linearity holds. Regarding to independence, there appear to be some evidence of grouping in residual plots. Constant variance is difficult to tell since some residuals vs. predictors plots contain points that are far away. For normality, there is lifting in the tails. After fitting a new model with transformed variables, we apply automated selection to determine which variables can be removed. To be specific, predictor removed is PROGRAM_MODEL".

The results means LOCATION_CITY, SECTOR, OVERNIGHT_SERVICE_TYPE, PROGRAM_AREA, T_CAPACITY_ACTUAL_BED, T_CAPACITY_FUNDING_BED, T_OCCUPIED_BEDS, T_UNOCCUPIED_BEDS, T_UNAVAILABLE_BEDS are important predictors. The Auto reduced model has ANOVA p-value of 0.5843. This means the auto reduced model is better. For both Auto_reduced_model, the linearity holds. There appear to be some evidence of grouping. The constant variance can be improved. There is lifting in the tails.

In conclusion, I will state some of the possible ways for the program to increase its beds' occupied rates. This includes increasing the number of beds showing as available for occupancy and the beds showing as occupied by a shelter user. Moreover, it is also useful to decrease the number of beds that are showing as available for occupancy that are not occupied as of the occupancy date, and beds that are not currently available.

5 Discussion

5.1 First discussion point

In this paper, I aim to analyze the variables that have significant effects on the proportion of actual bed capacity that is occupied for the reporting date. The variables I used include LOCATION_CITY, SETOR, PROGRAM_MODEL, OVERNIGHT_SERVICE_TYPE, PROGRAM_AREA, CAPACITY_ACTUAL_BED, CAPACITY_FUNDING_BED, OCCUPIED_BEDS, UNOCCUPIED_BEDS, UNAVAILABLE_BEDS (Toronto Open Data Portal 2021). The response variable is OCCUPANCY_RATE_BEDS (Toronto Open Data Portal 2021). To begin with, I conduct the exploratory data analysis in order to build the starting model. After creating my starting model by common sense and results of EDA, I check the two conditions before assessing the model assumptions. After that, I use plots to identify potential violations against model assumptions, which are linearity, normality, constant variance, and uncorrelatedness. The next step is to explore model transformations to correct assumption violations. After that, I apply box-cox transformation for numerical variables and create transformed variables. The next step is to fit a new model with transformed variables by applying automated selection. Then, I apply same analysis on my reduced model, which is better than the full model. Overall, It was found that contributing factors include the programs' locations, gender, age, household size of the service user group, the type of overnight service provided, the program area. Furthermore, the number of beds showing as available for occupancy, rooms that a program has been approved to provide, rooms showing as occupied by a shelter user, rooms that are showing as available for occupancy that are not occupied, and rooms that are not currently available also have significant influences on the response variable.

5.2 Second discussion point

From this paper, we learn that increasing the number of beds showing as available for occupancy and the beds showing as occupied by a shelter user is an effective way to increase its beds' occupied rates. Moreover, it is also useful to decrease the number of beds that are showing as available for occupancy that are not occupied as of the occupancy date, and beds that are not currently available.

5.3 Third discussion point

Furthermore, it's important to find the patterns and trends of the proportion of actual room capacity that is occupied for the reporting date. This report shows its contributing factors, so program managers can utilize this to make disciplines that can solve this issue more effectively. In this way, people can have a better living standards, and the economic development will be improved.

5.4 Weaknesses and next steps

The limitations include that the EDA part shows the response variable and predictors don't follow normal distributions. Specifically, there is unbalanced data volume. This may cause the results to be incredible and may be the reasons why the QQ plots contain violations. Also, I didn't checkthe leverage points, so it's possible there do contain pointsthat aren't credible. Moreover, violations still exist in residual plots of my reduced model. For the next step, we should try to avoid these problems and focus on other possible factors that can influence the proportion of actual bed capacity that is occupied for the reporting date.

References

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots*.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Toronto Open Data Portal, City of. 2021. *Open Data Dataset*. https://open.toronto.ca/dataset/daily-shelter-overnight-service-occupancy-capacity/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Jennifer Bryan. 2019. *Readxl: Read Excel Files*. https://CRAN.R-project.org/package=readxl.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2022. *Readr: Read Rectangular Text Data*.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.