Homelessness Report

February 2023

Objective: Study what's driving homelessness in big cities

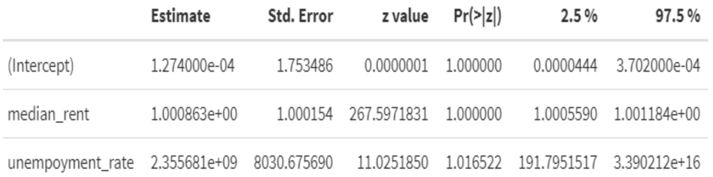
Key findings & Methodology

36 of samples were collected across USA cities in 2019

The report found that more than 81,729 people experienced homelessness in New York on a single night in 2019. Some of the more densely populated cities like NY and CA do have higher counts of homelessness.Figure1 shows the ten top homelessness cities in USA. Study results illustrated the main drivers of homeless ness in big cities are median rent and unemployment rate.

Negative Binomial(NB) model and Poisson model are compared. Residual plot below indicated that NB model are preform better. Count data often have an exposure variable, which indicates the number of times the event could have happened. Thus, population variable is incorporated into model with the use of the offset option.

Uses p-value 0.05 as threshold for significant findings. The statistically significant terms are the intercept, median rent, unemployment rate. The estimates are exponentiate the model coefficients. The output indicates that the incident rate of homelessness for higher median rent are 1 times the incident rate for the lower median rent group holding the other variables constant, and incident rate of homelessness for higher unemployment rate are 2.35 times the incident rate for the low unemployment rate group holding the other variables constant.



	Estimate	Std. Error	z value	Pr(>|z|)	2.5 %	97.5 %
(Intercept)	1.274000e-04	1.753486	0.0000001	1.000000	0.0000444	3.702000e-04
median_rent	1.000863e+00	1.000154	267.5971831	1.000000	1.0005590	1.001184e+00
unempoyment_rate	2.355681e+09	8030.675690	11.0251850	1.016522	191.7951517	3.390212e+16

Hypothesis Testing: Wilcoxon rank sum test and Pearson's Chi-squared test was performed to compare between high rent and low rent group across all regions. Table 1 Summary table illustrates descriptive statistics, p-value significant suggest s these two distributions differ We can conclude that a significant difference exists between high and low rent group associated with homelessness.

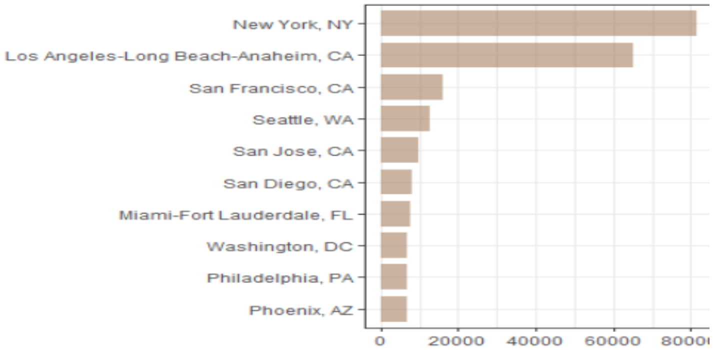


Figure1: Homelessness Count

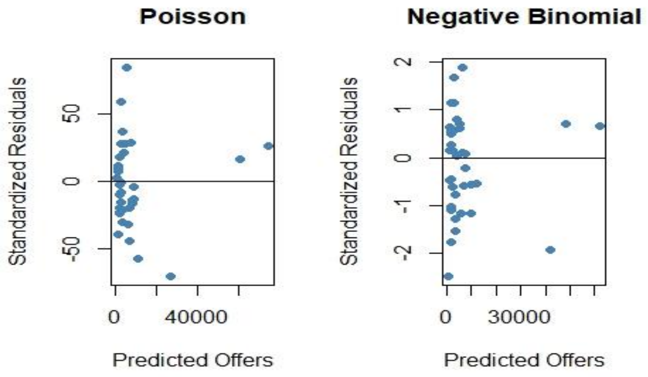
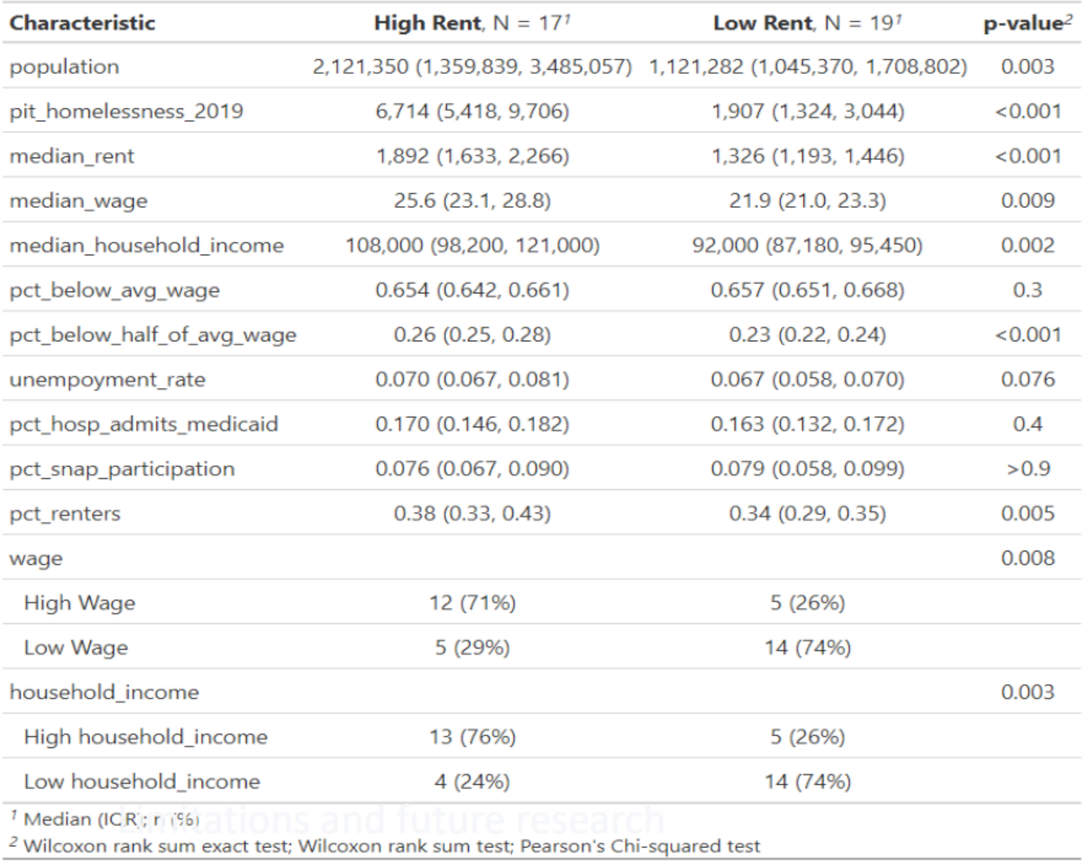


Table 1 Summary table illustrates descriptive statistics, p-value significant suggests these two distributions differ.



Characteristic	**High Rent**, N = 17[1]	**Low Rent**, N = 19[1]	**p-value**[2]
population	2,121,350 (1,359,839, 3,485,057)	1,121,282 (1,045,370, 1,708,802)	0.003
pit_homelessness_2019	6,714 (5,418, 9,706)	1,907 (1,324, 3,044)	<0.001
median_rent	1,892 (1,633, 2,266)	1,326 (1,193, 1,446)	<0.001
median_wage	25.6 (23.1, 28.8)	21.9 (21.0, 23.3)	0.009
median_household_income	108,000 (98,200, 121,000)	92,000 (87,180, 95,450)	0.002
pct_below_avg_wage	0.654 (0.642, 0.661)	0.657 (0.651, 0.668)	0.3
pct_below_half_of_avg_wage	0.26 (0.25, 0.28)	0.23 (0.22, 0.24)	<0.001
unempoyment_rate	0.070 (0.067, 0.081)	0.067 (0.058, 0.070)	0.076
pct_hosp_admits_medicaid	0.170 (0.146, 0.182)	0.163 (0.132, 0.172)	0.4
pct_snap_participation	0.076 (0.067, 0.090)	0.079 (0.058, 0.099)	>0.9
pct_renters	0.38 (0.33, 0.43)	0.34 (0.29, 0.35)	0.005
wage			0.008
High Wage	12 (71%)	5 (26%)	
Low Wage	5 (29%)	14 (74%)	
household_income			0.003
High household_income	13 (76%)	5 (26%)	
Low household_income	4 (24%)	14 (74%)	

[1] Median (IQR); n (%)

[2] Wilcoxon rank sum exact test; Wilcoxon rank sum test; Pearson's Chi-squared test

Limitations and future research

- Homelessness is a complex issue caused by a variety of factors including economic factors, family relationships, mental illness, lack of affordable housing, drug abuse, and alcoholism. Social forces such as addictions, family breakdown, and mental illness are compounded by structural forces such as lack of available low-cost housing, poor economic conditions, and insufficient mental health services [1]. Based on the limited data we collected, we did not address any of those issues mentioned above.

- Observed counts are need be adjusted to optimum conditions because it is not possible to be at every site of every annual period due to population fluctuations, weather conditions.

- Collecting more **Demographic information** include age, sex, education, and race etc. for further analysis. In addition, we should focus on collect more data over the time period and combine census data with **population density** to get more sense of the trending. Need to adjust for size of cities and its density.

- Collect more **data related to the policies** we want to validate and or promote for hypothesis testing;

1. Mago, V.K., Morden, H.K., Fritz, C. et al. Analyzing the impact of social factors on homelessness: a Fuzzy Cognitive Map approach. BMC Med Inform Decis Mak 13, 94 (2013). https://doi.org/10.1186/1472-6947-13-94

Appendix R code

```
########################################################################
#   Assessment of what's driving homelessnesss in these big citites
########################################################################

#read data and load packages;
data<- read.csv("what_is_driving_homelessness.csv")
setwd("C:/Users/SSE6/Downloads")
library(ggplot2);library(dplyr);library(plotly);library(tidyverse);  library(psych); library(biostat3);library(AER);
library(gtsummary);library(kableExtra);

set.seed(1)
#Conducting a Poisson regression will allow you to see which predictor variables (if any) have a statistically significant effect on the response variable.
summary(data)

# randomly assign new groups based on the median value of rent,wage,pop,income and to see if the variance across the groups
new_data<-data %>% select(-metro_id) %>%
  mutate(rent=ifelse(median_rent> median(median_rent),"High Rent","Low Rent"),
      wage=ifelse(median_wage> median(median_wage),"High Wage","Low Wage"),
household_income=ifelse(median_household_income>median(median_household_income),"High household_income","Low household_income"),
pop=ifelse(population> quantile(population,0.75),"High","Low"))

#check out variance between groups;
new_data %>%
  filter(pop=="High") %>%
  select(pit_homelessness_2019) %>%
  var() / new_data %>%
  filter(pop=="Low") %>%
  select(pit_homelessness_2019) %>%
  var()

new_data %>%
  filter(household_income=="High household_income") %>%
  select(pit_homelessness_2019) %>%
  var() / new_data %>%
  filter(household_income=="Low household_income") %>%
  select(pit_homelessness_2019) %>%
  var()


#summary table
new_data %>% select(-region_name) %>%
  tbl_summary(by=rent)  %>%
  add_p()


#check distribution and overall summary for homelessness
describeBy(data$pit_homelessness_2019)
hist(data$pit_homelessness_2019)

#The variance is much greater than the mean, which suggests that we will have over-dispersion in the model.
data %>%
  summarize(mean=mean(pit_homelessness_2019),var=var(pit_homelessness_2019))

#get top 10 cities for homelessness
data1 <- data %>%
  group_by(region_name) %>%
  summarise(homneless_mean = mean(pit_homelessness_2019)) %>%
  arrange(desc(homneless_mean)) %>%
  top_n(10)
```

Appendix R code

```
# plot the top 10 cities
 data1 %>%
  mutate(region_name = fct_reorder(region_name, homneless_mean)) %>%
  ggplot(aes(x=region_name, y = homneless_mean)) +
  geom_bar(stat="identity", fill="#A68060", alpha=.6, width=.8) +
  coord_flip() +
  xlab(" ") + ylab("Figure1: Homelessness Count") +
  theme_bw()

#fit the poisson model
 p_model <-glm(pit_homelessness_2019 ~ offset(log(population)) + median_rent + median_wage +
          pct_below_avg_wage + pct_below_half_of_avg_wage +median_household_income+
          unempoyment_rate + pct_hosp_admits_medicaid + pct_snap_participation,
          family = poisson(link = "log"), data =  new_data )
#view model output
summary(p_model)

#Over dispersion means the assumptions of the model are not met, hence we cannot trust its output;
dispersiontest(model)



#fit negative binomial regression model
summary(nb_model <- glm.nb(pit_homelessness_2019 ~ offset(log(population)) + median_rent + median_wage +
                 pct_below_avg_wage + pct_below_half_of_avg_wage +median_household_income+
                 unempoyment_rate + pct_hosp_admits_medicaid + pct_snap_participation,
                data =  new_data ))
par(mfrow = c(1, 2))
#Residual plot for Poisson regression
p_res <- resid(p_model)
plot(fitted(p_model), p_res, col='steelblue', pch=16,
    xlab='Predicted Offers', ylab='Standardized Residuals', main='Poisson')
abline(0,0)

#Residual plot for negative binomial regression
nb_res <- resid(nb_model)
plot(fitted(nb_model), nb_res, col='steelblue', pch=16,
    xlab='Predicted Offers', ylab='Standardized Residuals', main='Negative Binomial')
abline(0,0)


pchisq(2 * (logLik(nb_model) - logLik(p_model)), df = 1, lower.tail = FALSE)



# define the NB models to compare
cand.models <- list( )
cand.models[[1]] <- glm.nb(pit_homelessness_2019 ~ offset(log(population)) + median_rent + median_wage +
                      pct_below_avg_wage + pct_below_half_of_avg_wage +median_household_income+
                      unempoyment_rate + pct_hosp_admits_medicaid + pct_snap_participation,
                     data =  new_data )
cand.models[[2]] <- glm.nb(pit_homelessness_2019 ~ offset(log(population)) + median_rent + median_wage +
                      pct_below_avg_wage + pct_below_half_of_avg_wage +median_household_income+
                      unempoyment_rate + pct_hosp_admits_medicaid,
                     data =  new_data )
cand.models[[3]] <- glm.nb(pit_homelessness_2019 ~ offset(log(population)) + median_rent + median_wage +
                      pct_below_avg_wage + pct_below_half_of_avg_wage +median_household_income+
                      unempoyment_rate,
                     data =  new_data )
```

Appendix R code

```
cand.models[[4]] <- glm.nb(pit_homelessness_2019 ~ offset(log(population)) + median_rent + median_wage +
                  pct_below_avg_wage + pct_below_half_of_avg_wage +median_household_income,
                 data =  new_data)

cand.models[[5]] <- glm.nb(pit_homelessness_2019 ~ offset(log(population)) + median_rent + median_wage +
                  pct_below_avg_wage,
                 data =  new_data)
cand.models[[6]] <- glm.nb(pit_homelessness_2019 ~ offset(log(population)) + median_rent + unempoyment_rate ,
                 data =  new_data)

cand.models[[7]] <- glm.nb(pit_homelessness_2019 ~ offset(log(population)) + median_rent,
                 data =  new_data)

cand.models[[8]] <- glm.nb(pit_homelessness_2019 ~ offset(log(population)) ,
                 data =  new_data)

# name the models
model.names <- c("1", "2", "3", "4", "5", "6", "7", "8")

names(cand.models) <- model.names
# calculate and combine AIC, AIC weights, and BIC
results <- data.frame(models = model.names)
results$bic.val <- unlist(lapply(cand.models, BIC))
results$bic.rank <- rank(results$bic.val)
results$aic.val <- unlist(lapply(cand.models, AIC))
results$aic.delta <- results$aic.val-min(results$aic.val)

results$aic.likelihood <- exp(-0.5* results$aic.delta)
results$aic.weight <- results$aic.likelihood/sum(results$aic.likelihood)
# sort models by AIC weight
results <- results[rev(order(results[, "aic.weight"])),]
results$cum.aic.weight <- cumsum(results[, "aic.weight"])


#Results for Final Count Regression Model (Negative Binomial)
results


#model 6 appears best across all model compassion

summary(cand.models[[6]])
coef(summary(cand.models[[6]]))


table<-exp(cbind(estimate = coef(summary(cand.models[[6]])),
      confint(cand.models[[6]])))
kable(table, escape=F, align=c("l",rep("r",8))) %>%
  kable_styling(full_width = F,position="left")
```