Final Project: Predicting Chronic Homelessness in Los Angeles County Using A Logistic Regression Model

Author: William Huang

Discussants: Stack Overflow was consulted to learn how to create certain plots in this study.

Introduction

Metropolises across the world have faced the homeless crises for generations. According to the United Nations Human Settlements Program, 1.6 billion people worldwide live in inadequate housing and 100 million people are devoid of any housing at all. Following the COVID-19 pandemic, a loss of housing has been on the rise in the U.S. as a result of the termination of COVID relief packages. With this newfound risk of a rise in homelessness, it is imperative that society puts a priority in reversing this pressing issue. In other words, an effective solution must be derived to bring people out of houselessness, mitigating chronic homelessness rates worldwide.

As a resident of Los Angeles (LA), California, I have witnessed firsthand the extent of the homeless problem. Walking through the streets in my city, encountering the homeless is a regular occurance. Because of my personal connection to this issue, this report targets the homeless problem in LA county. Specifically, I am interested in discovering the key predictors behind the likelihood of a LA resident in becoming chronically homeless (the condition of being homeless for at least a year).

To that end, this report aims to develop a computational model that can predict the likelihood of a given homeless person in becoming chronically homeless in LA county. Using this model, we can gain insights on what predictors contribute most (cause the greatest increase in probability) to a person becoming chronically homelessness in LA. Understanding the major determinants of homelessness in LA may allow society to develop preventative measures to catalyze a reversal of this generational problem not only in LA but nationwide

This study utilizes a dataset from the Los Angeles County Homeless County Data Library. Collected by Paul Beeman from UCLA, the data contains information recorded by the Homeless Management Information System (HMIS) records and surveys conducted between 2011 and 2017 on sheltered and unsheltered homeless individuals (inhabit sidewalks, vehicles, and embankments). Within the dataset, specific information was collected on 27 different characteristics about the study's participants For example, *gender* (male, female, transgender, unknown), *ethnicity* (African America, European American, Latino, Other Ethnicity, Unknown), and other well-known predictors for homelessness were gathered. For more information on the dataset and a detailed codebook, please visit this link from the data library's website.

Looking through Google Scholar, I did not find previous analyses using this dataset. However, studies have been done on similar datasets with information on unsheltered and sheltered homeless to analyze the difference in demographics between the two populations. The proposed method is not a replication of published work.

Results

Data wrangling: Preparation of Data for Data Visualization and Analyses

The dataset with information on unsheltered and sheltered homeless in Los Angeles County from 2011 to 2017 was first loaded into RStudio Containing 43,761 cases and 28 variables, I began looking into the dataset to see which variables were unnecessary and whether or not I could filter out missing/unhelpful data. I removed the variables Survey_Year and ID as these predictors would have little effect on determining whether or not a homeless person will be permanently houseless. I also removed Unemployment_Looking as this variable is repetitive with Unemployment_Not_Looking (one will be 0 when the other is 1); therefore, leaving both in could cause multicollinearity. Then, I removed the variable Birth_Year because the dataset already contained the age of the participants. The dataset also contains three different variables pertaining to race:

- Ethnicity: European American, African American, Latino, Other Ethnicity, Unknown
- Race_full: Raw data input
- Race_Recode: European American, African American, Latino, Other Ethnicity, Unknown

In this project, I decided to select ethnicity for two reasons. First, the variable race_full contains the exact race participants inputed in the survey, which is less helpful than races that are grouped together. Ethnicity was preferred over race_recode because of the added distinction of the race Latino.

As the dataset contained individual aged 0-100, I filtered out every participant below the age of 18 as I am interested in predicting the likelihood of **adults** in becoming homeless. Finally, as there were "NA" and "Unknown" values in a few variables (Times_Homeless_Past_Year, Times_Homeless_3yrs, Gender, Current_Stint_Duration, Ethnicity), I filtered out those values. Because all the individuals are homeless and all predictors have categorical levels, I determined that there were no outliers.

```
# loading the dataset into RStudio
homeless_data <- na.omit(read_excel("/Users/williamhuang/Documents/Yale/Classes/s&ds230/Final_Project/2

# data cleaning
homeless_cleaned <- homeless_data |>
    filter(Times_Homeless_Past_Year %in% c("1 time", "2 to 3 times",
        "4 or more times")) |>
    filter(Times_Homeless_3yrs %in% c("1 time", "2 to 3 times",
        "4 or more times")) |>
    filter(Age >= 18) |>
    filter(!(Gender %in% c("Unknown", "NA"))) |>
    filter(!(Current_Stint_Duration == "Unknown")) |>
    filter(!(Ethnicity == "Unknown")) |>
    select(-Survey_Year, -Birth_Year, -Race_Full, -Race_Recode,
        -Unemployed_Looking, -1)
```

Visualize the data

First, I was interested in investigating how the number of chronically homeless changed at various ages. Analyzing the bar chart, I noticed that there were significantly more people in the targeted population in their middle ages (40-60).

```r
# Age Groups by 10 from 18 to 100
homeless_cleaned$Age_Group <- cut(homeless_cleaned$Age, breaks = seq(18,
    100, by = 10), labels = FALSE)

# Plot histogram
ggplot(homeless_cleaned, aes(x = Age_Group, fill = factor(Chronic))) +
    geom_bar(position = "dodge", alpha = 0.7, binwidth = 1) +
    labs(title = "Distribution of Chronically Homeless by Age Group",
        x = "Age Group", y = "Count") + facet_wrap(~Chronic,
    scales = "free_y") + theme(axis.text.x = element_text(angle = 45,
    hjust = 1))
```

After analyzing age, I decided to visualize the relationship between the the dependent variable (Chronic) and all 23 predictors in the homeless_cleaned dataset. Specifically, I am interested in analyzing how the amount of chronically homeless individuals changes with respect to every level of each variable. For example, are there more chronically homeless females as opposed to transgenders or males? Please see the appendix for the implementation of the code and all 23 bar charts.

Looking at the bar charts, I noticed that the majority of chronically homeless individuals were Latino, African American, and European American. Furthermore, there were a significantly more chronically homeless females and males than transgenders. However, for many of these bar charts, the distribution was nearly identical for both individuals who were chronically homeless and individuals who weren't. Furthermore, the bar charts indicated that fewer individuals who abused alcohol were chronically homeless than individuals who did. Intuitively, this does not make sense. This phenomenon could be attributed to the fact that there were more individuals who participated in the study who did not abuse alcohol. Therefore to gain a better understanding of the true predictors of chronic homelessness, I began my analyses. (To view the summary

of my model, please see the appendix.)

Analyses: Building a Multivariable Logistic Regression Model

The overall goal of my project is to identify the largest contributors to a person becoming chronically homeless. Therefore I decided to build a multivariable logistic regression model, as I am interested in analyzing how the probability that a person in chronically homeless (a binary variable) changes as a function of various predictors.

Before building my model, I first implemented cross-validation to mitigate overfitting in my logistic regression model. Using a 80%/20% split for training/testing, I created the two respective dataframes.

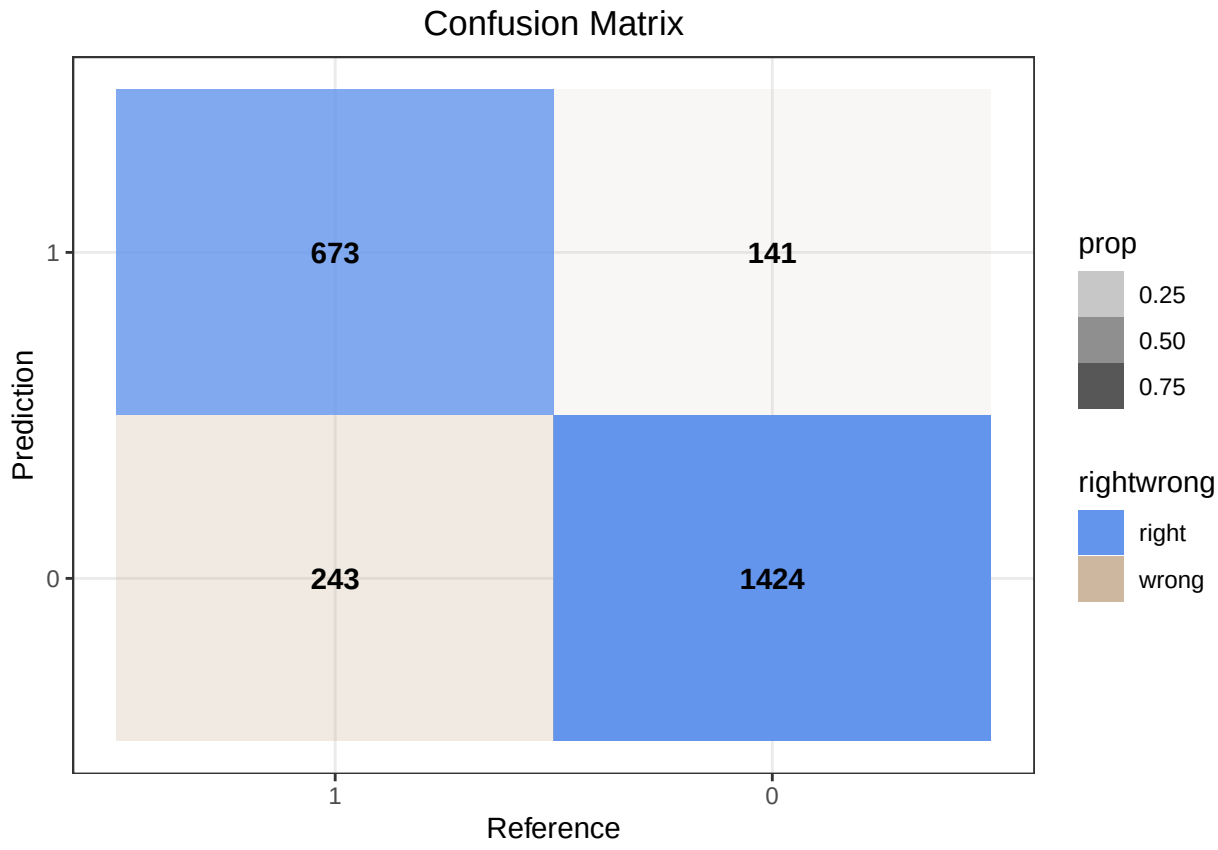
To build the multivariable logistic regression model, I used a methodology called backwards elimination. In essence, I first built a model containing all 23 predictors in train_data. However, many predictors had p-values > 0.05, meaning that they were insignificant in predicting whether a given homeless person would be chronically homeless. To build the optimal logistic regression model, I adopted the backward elimination strategy to remove the predictor with the largest p-value each time and keep the remaining predictors to refit the new model. Although it may appear that the predictors like Times_Homeless_3yrs is insignificant because a few of its levels have large p-values, the predictor as a whole is significant because at least one of its levels has a p-value less than 0.05. Finally, I factored binary predictors so that R would interpret them as categorical variables.

```
# Split data into train/test
set.seed(42)
train_indices <- sample(1:nrow(homeless_cleaned), 0.8 * nrow(homeless_cleaned))
train_data <- homeless_cleaned[train_indices, ]
test_data <- homeless_cleaned[-train_indices, ]

# Multivariable logitic regression model
lr_fit <- glm(Chronic ~ Age + Gender + Ethnicity + factor(Veteran) +
    Times_Homeless_3yrs + Times_Homeless_Past_Year + Current_Stint_Duration +
    factor(Physical_Sexual_Abuse) + factor(Physical_Disability) +
    factor(Mental_Illness) + factor(Alcohol_Abuse) + factor(Drug_Abuse) +
    factor(Drug_Alcohol_History) + factor(HIV_Positive) + factor(Unemployed_Not_Looking),
    data = homeless_cleaned, family = "binomial")

# Removed variables: factor(chronic_time) +
# factor(chronic_condition) + factor(adult_with_child) +
# SPA + factor(part_time) + factor(full_time)
```

Once the multivariable logistic regression model was built, I conducted a thorough analysis of its efficacy. First, I calculated the number of true positives, true negatives, false positives, and false negatives. Then to visually decipher the difference between each group, I generated a confidence matrix with the help of Stack Overflow that illustrates the total number in the reference groups. As they are significantly more true positives and true negatives than false positives and false negatives, I moved to the next step of my analysis.

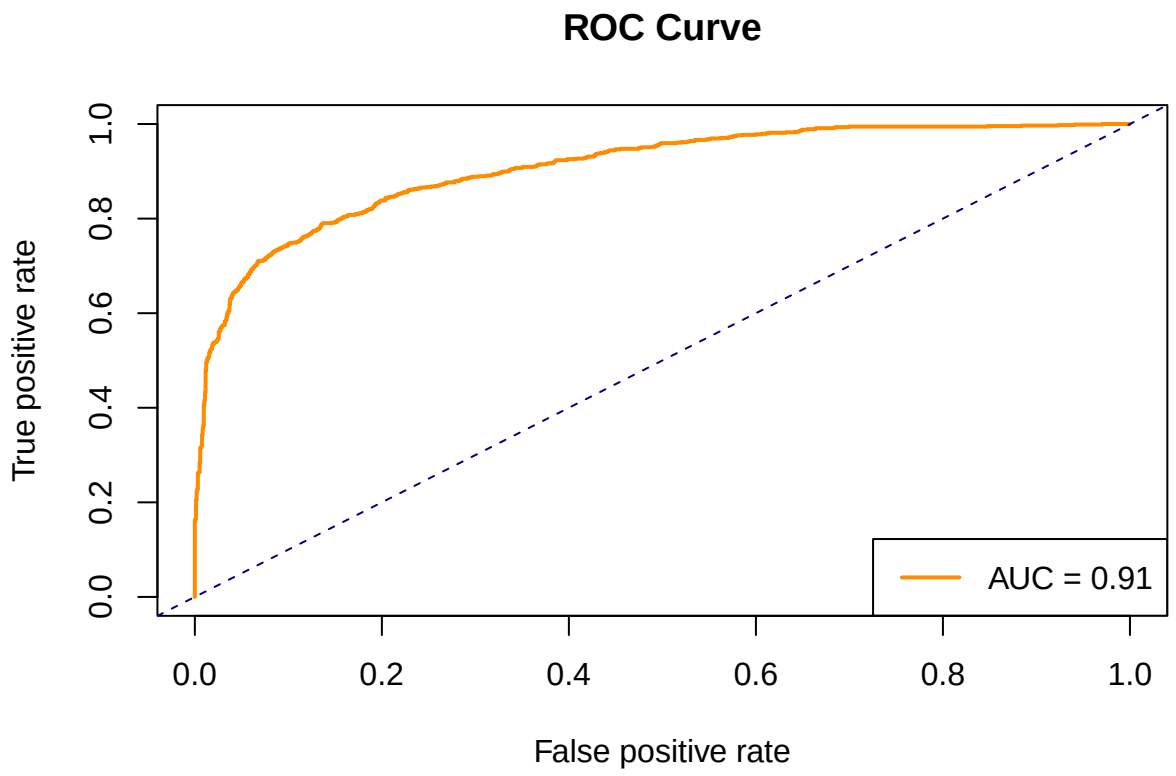
```
# Predicted vs actual of chronic homelessness
actual_labels <- test_data$Chronic
y_pred_all_features <- predict(lr_fit, newdata = test_data, type = "response")
predicted_labels <- as.numeric(y_pred_all_features >= 0.5)
```

```r
# Determining the number of TP, FP, TN, and FN
positives_negatives <- data.frame(confusionMatrix(factor(predicted_labels),
    factor(actual_labels))$table)

# Table of TP, FP, TN, FN rates
table <- positives_negatives |>
    mutate(rightwrong = ifelse(positives_negatives$Prediction ==
        positives_negatives$Reference, "right", "wrong")) |>
    group_by(Reference) |>
    mutate(prop = Freq/sum(Freq))

# Confusion matrix with transparency of each group
# according to the proportion relative to the total number
# of predictions
ggplot(data = table, mapping = aes(x = Reference, y = Prediction,
    fill = rightwrong, alpha = prop)) + geom_tile() + geom_text(aes(label = Freq),
    vjust = 0.5, fontface = "bold", alpha = 1) + scale_fill_manual(values = c(right = "cornflowerblue",
    wrong = "bisque3")) + theme_bw() + xlim(rev(levels(table$Reference))) +
    ggtitle("Confusion Matrix") + theme(plot.title = element_text(hjust = 0.5))
```

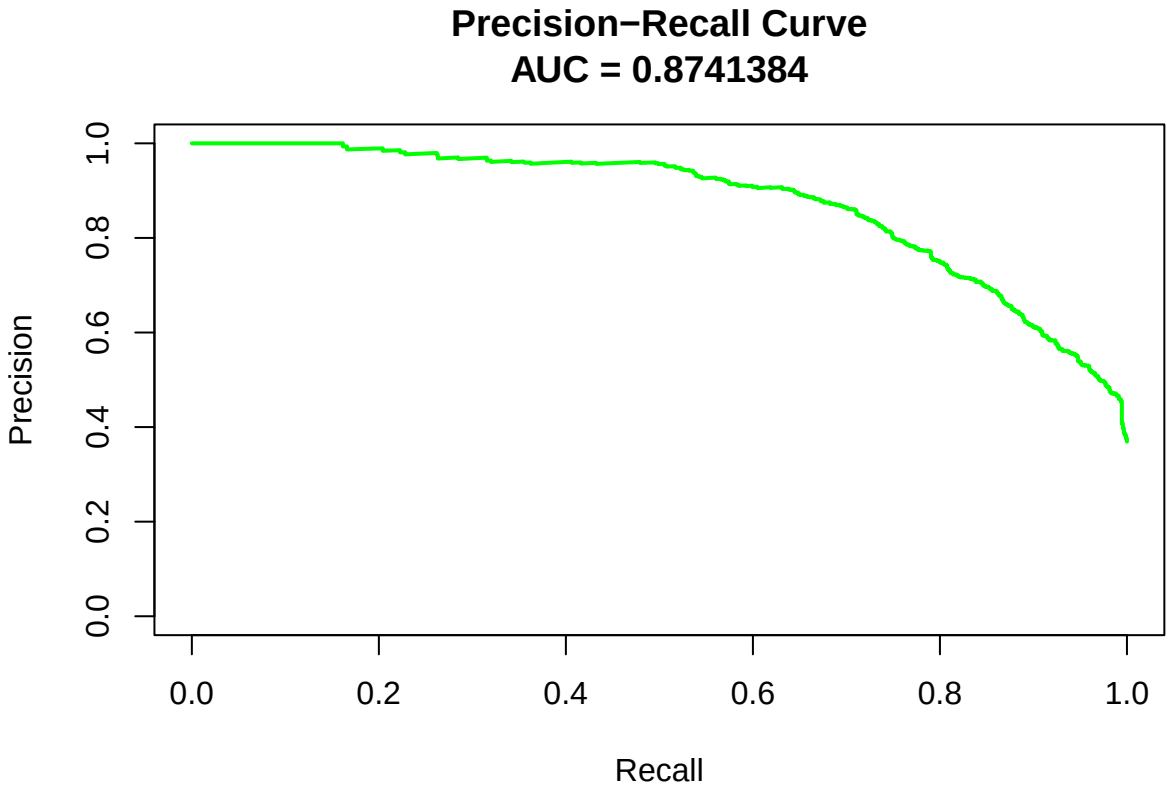


Two common graphical evaluative metrics for classification, the Receiver Operating Characteristic (ROC) Curve and the Precision-Recall (PR) Curve, were generated. The ROC curve evaluates the relationship between true positive and false positive rate, while the PR curve demonstrates the model's ability to capture positive cases.

```r
# Combines the predicted probabilities with the binary
# value from the test_data
roc_pred_all_features <- prediction(y_pred_all_features, test_data$Chronic)

# Calculates true positive and false positive rate
roc_perf_all_features <- performance(roc_pred_all_features, "tpr",
    "fpr")

# Plotting the ROC curve
plot(roc_perf_all_features, col = "darkorange", main = "ROC Curve",
    lwd = 2)
abline(a = 0, b = 1, lty = 2, col = "navy")
legend("bottomright", legend = paste("AUC =", round(performance(roc_pred_all_features,
    "auc")@y.values[[1]], 2)), col = "darkorange", lwd = 2)
```




```r
# Plotting the PR curve
pr_curve <- pr.curve(scores.class0 = y_pred_all_features, weights.class0 = actual_labels,
    curve = TRUE)
plot(pr_curve, main = "Precision-Recall Curve", col = "green",
    lwd = 2)
```



To evaluate these two curves, we can calculate the Area Under the Curve (AUC). With a AUC of 0.91 for the ROC curve and a AUC of 0.874 for the PR curve, we can conclude that the model is performing significantly better than random chance (AUC = 0.5). In fact, it is prediction level is much closer to perfect discriminatory (AUC = 1).

Finally, I calculated the accuracy, precision, recall, and F1 score of the model. As each metric was very high (> 0.7), I concluded that my multivariable logistic regression model is trustworthy.

```
conf_matrix <- table(Actual = actual_labels, Predicted = predicted_labels)

# Logistic regression model evaluative metrics
accuracy <- sum(diag(conf_matrix))/sum(conf_matrix)
precision <- conf_matrix[2, 2]/sum(conf_matrix[, 2])
recall <- conf_matrix[2, 2]/sum(conf_matrix[2, ])
f1_score <- 2 * (precision * recall)/(precision + recall)

cat("Accuracy:", accuracy, "\n")
```

```
## Accuracy: 0.8452237
```

```
cat("Precision:", precision, "\n")
```

```
## Precision: 0.8267813
```

```r
cat("Recall:", recall, "\n")
```

```
## Recall: 0.7347162
```

```r
cat("F1 Score:", f1_score, "\n")
```

```
## F1 Score: 0.7780347
```

As I am interested in predicting a given homeless person's likelihood of becoming chronically homeless, I built a logistic regression function using R's builtin predict() function that predicts this probability given an input of the 15 parameters in my model. To help make the process more robust, I developed another function that converts a user's inputs into a dataframe in the global environment. With this function, we can input characteristics on a hypothetical individual (race, gender, whether or not they use drugs, whether or not they consume alcohol, etc) and determine their likelihood of becoming homeless.

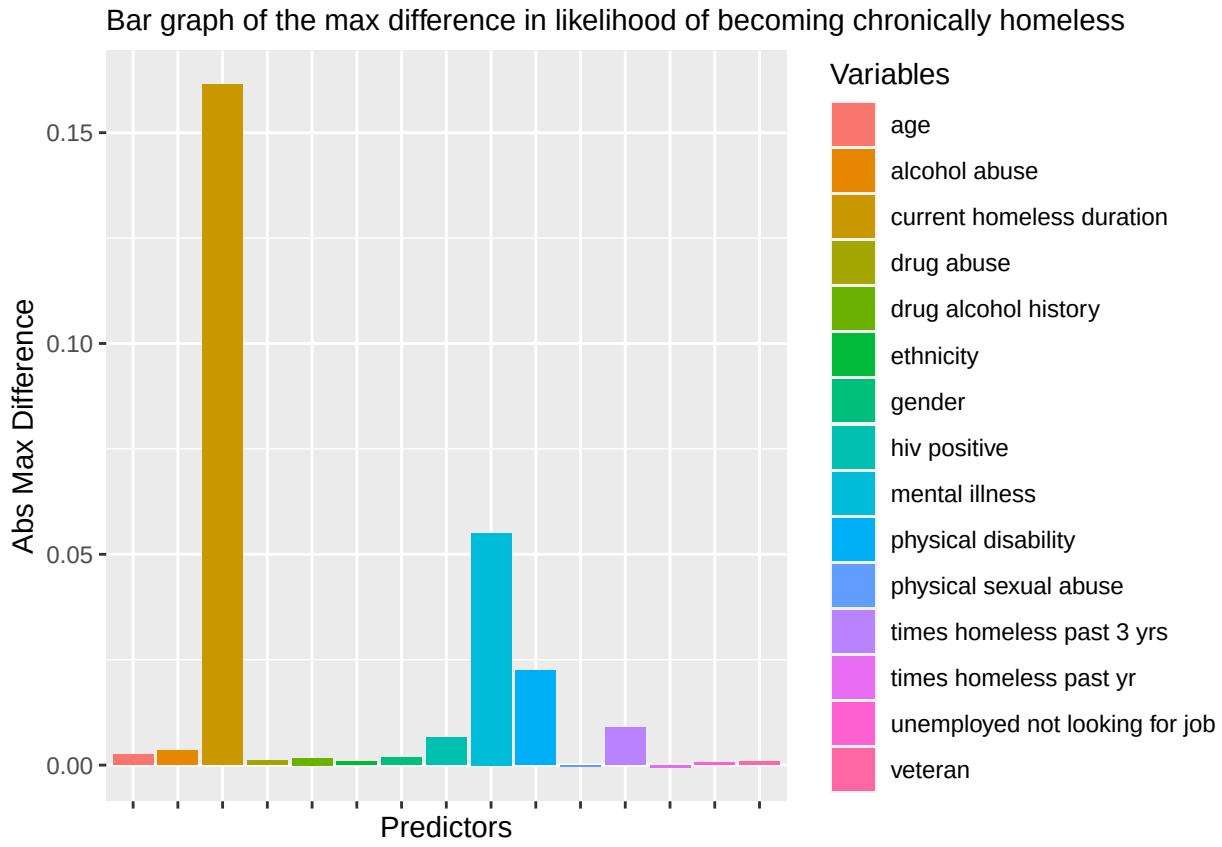
```r
# Function that converts user's inputs into a dataframe
input_data <- function(Age, Gender, Ethnicity, Veteran, Times_Homeless_3yrs,
    Times_Homeless_Past_Year, Current_Stint_Duration, Physical_Sexual_Abuse,
    Physical_Disability, Mental_Illness, Alcohol_Abuse, Drug_Abuse,
    Drug_Alcohol_History, HIV_Positive, Unemployed_Not_Looking) {
    df <- data.frame(Age = Age, Gender = Gender, Ethnicity = Ethnicity,
        Veteran = Veteran, Times_Homeless_3yrs = Times_Homeless_3yrs,
        Times_Homeless_Past_Year = Times_Homeless_Past_Year,
        Current_Stint_Duration = Current_Stint_Duration, Physical_Sexual_Abuse = Physical_Sexual_Abuse,
        Physical_Disability = Physical_Disability, Mental_Illness = Mental_Illness,
        Alcohol_Abuse = Alcohol_Abuse, Drug_Abuse = Drug_Abuse,
        Drug_Alcohol_History = Drug_Alcohol_History, HIV_Positive = HIV_Positive,
        Unemployed_Not_Looking = Unemployed_Not_Looking)
    # sends the df into the global environment
    assign("input_test", df, envir = .GlobalEnv)
}

# Probability function
homeless_pred <- function(model, data) {
    predict(model, data, type = "response")
}
```

Findings: Deriving the Most Impactful Predictors of Chronic Homelessness

To find the main contributors of chronic homelessness, I used two evaluative metrics. The first test I conducted was finding the magnitude in difference between the probability predicted by my model when a variable was at its most "extreme case" compared to the "least extreme case" while keeping other predictors constant. For example, the difference in probability was calculated when the predictor veteran was a 1 compared to when veteran was 0. The remaining 14 predictors were kept constant to ensure that the difference is due solely to the change in a single predictor. This was repeated for all 15 predictors (please see the appendix for the calculation of differences). Once the differences were calculated, I created a column graph illustrating the absolute value of the differences.

```r
# Dataframe of the differences between the largest and
# smallest probabilities of each variable
diff_df <- data.frame(Variables = c("age", "gender", "ethnicity",
```

```
    "veteran", "times homeless past 3 yrs", "times homeless past yr",
    "current homeless duration", "physical sexual abuse", "physical disability",
    "mental illness", "alcohol abuse", "drug abuse", "drug alcohol history",
    "hiv positive", "unemployed not looking for job"), Differences = c(age_diff,
    gender_diff, ethnicity_diff, veteran_diff, times_homeless3_diff,
    times_homeless_pastyr_diff, current_stint_dur_diff, physical_sexual_abuse_diff,
    physical_disability_diff, mental_illness_diff, alcohol_abuse_diff,
    drug_abuse_diff, drug_alcohol_hist_diff, hiv_pos_diff, unemployed_not_looking_diff))

# Generating the column graph
ggplot(diff_df, aes(Variables, Differences, fill = Variables)) +
    geom_col() + ggtitle("Bar graph of the max difference in likelihood of becoming chronically homeless
    xlab("Predictors") + ylab("Abs Max Difference") + theme(plot.title = element_text(size = 11),
    axis.text.x = element_blank())
```

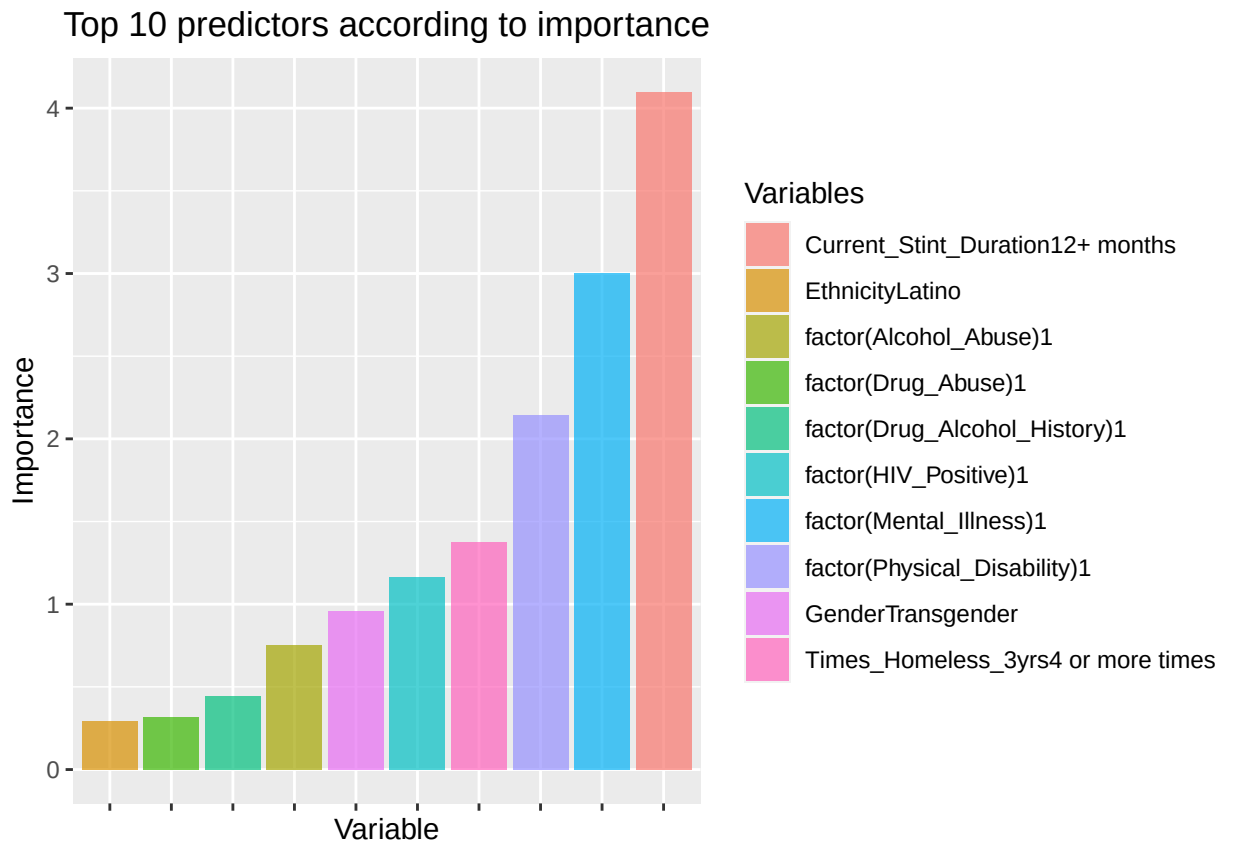


This column graph shows that the top three predictors (Current duration of homelessness, mental illness, aand physical disability) have significantly larger differences of their maximum - minimum probabilities.

The second method to evaluate the contribution of each predictor is through an importance plot. To calculate importance, I extracted the coefficients of each predictor from the regression model. This is because in logistic regression, the coefficients is the log-odds change in the response variable for a change in the predictor. As such, there is a correlation between larger coefficients and the importance of the predictor.

Therefore, I extracted the abs(coefficients) from the regression model and picked the 10 largest ones. Using this, I created a bar chart of the importance of these 10 predictors.

```r
# Extracting coefficients from regression model to
# calculate importance
coefs <- coefficients(lr_fit)
importance <- abs(coefs)
variables_names = names(coefficients(lr_fit))

# Dataframe of importance of each predictor
vi_df <- data.frame(Variables = variables_names, Importance = importance)[2:24,
    ]

# Sorting the vi_df from largest to smallest and extracting
# the first 10
top_10_vi <- head(vi_df[order(-vi_df$Importance), ], 10)

# Plotting a bar chart of the largest 10 importance
# predictors from smallest to largest
ggplot(top_10_vi, aes(x = reorder(Variables, Importance), y = Importance,
    fill = Variables)) + geom_bar(stat = "identity", alpha = 0.7) +
    labs(title = "Top 10 predictors according to importance",
        x = "Variable", y = "Importance") + theme(plot.title = element_text(hjust = 0.5),
    axis.text.x = element_blank())
```



Similar to the column graph of the difference in max-min probabilities of each predictor, the importance plot shows that the top 3 predictors of chronic homelessness is when the current duration of homelessness is larger than 12 months, followed by when the homeless person has a mental illnesss and physical disability.

Conclusion

The multivariable logistic regression model in this study is capable of making accurate predictions on whether unsheltered and sheltered homeless people in the county of Los Angeles will become chronically homeless. With an accuracy of 0.85, precision of 0.83, recall of 0.73, and F-1 score of 0.78, we demonstrate quantitatively the reliability of the model. Furthermore to prevent overfitting, the study uses cross-validation and ensures that the p-values of all predictors are significant.

Using this model, we discovered that the primary determinants of chronic homelessness are the duration of the person's current homelessness, whether they have a mental illness, and whether they have a physical disability in that order. With this knowledge, governments can develop preventative measures such as policies that will aid current homeless individuals in securing housing earlier and programs that will help people who are disabled or mentally ill. This could potentially reduce the number of chronically homeless people in Los Angeles County.

In the future, I hope to improve my logistic regression model by adding interaction terms. Furthermore, I am interested in whether using a machine learning model will improve the predictions of chronic homelessness.

Reflection

In total, I spent around 30 hours on this project. The data cleaning process went very smoothly, as the dataset contained the variables I was interested in and few extraneous cases. Furthermore, I found enjoyment in analyzing my logistic regression model as I had to venture beyond topics learned in class to generate confusion matrices, ROC curves, and PR curves. Finally, it was satisfying to see that my column graph of the differences in max-min probabilities of each predictor determined the same three predictors as significant contributors to chronic homelessness as the importance plot.

However, the whole process was not completely smooth. In particular, I struggled in developing data visualizations for my cleaned data, as it was hard to generate graphs beyond bar chars for the data when most variables were binomial. Therefore, I focused on generating visualizations dependent on the counts of variables. Finally, I also attempted creating plots of my logistic regression model against a distribution of chronic homelessness. However, it was difficult to create plots because my regression model had many predictors.

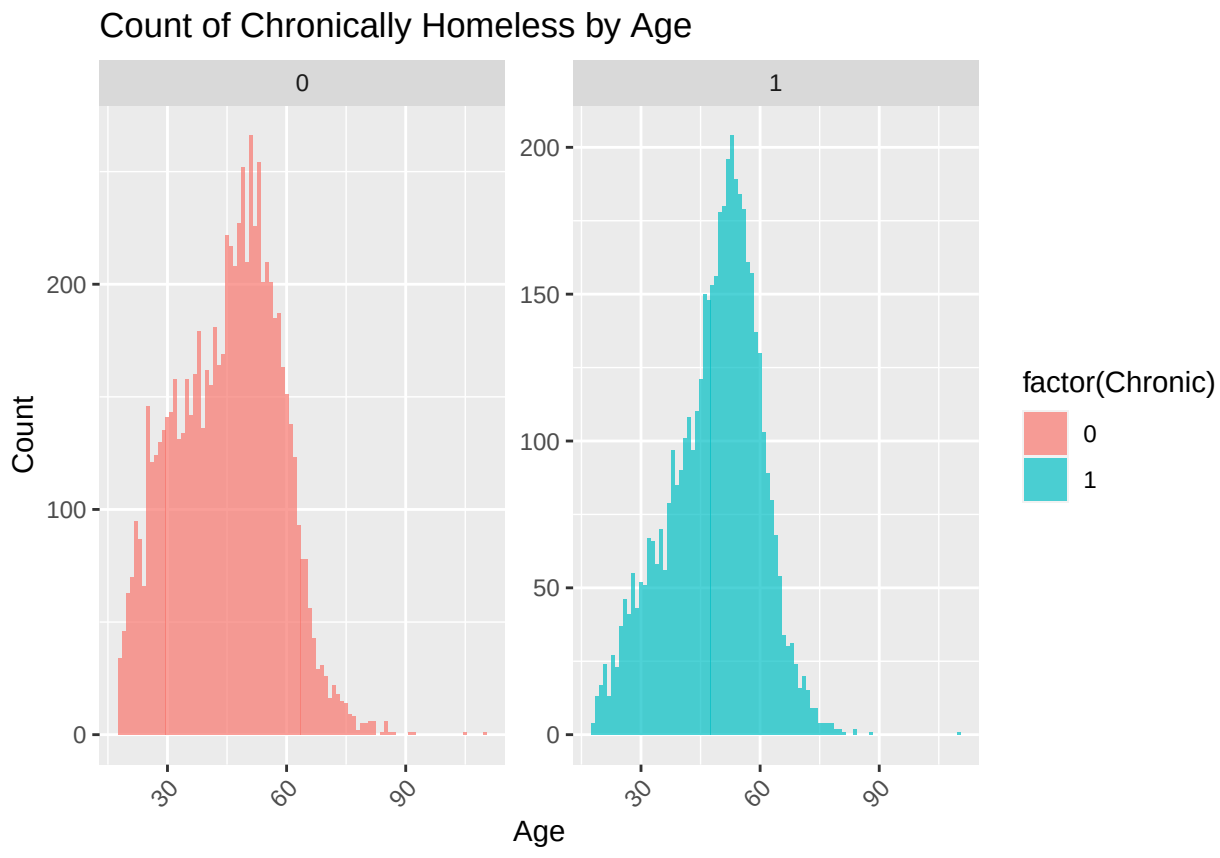
Appendix

Visualization of the Count of Chronic Homeless for Every Level of Each Variable

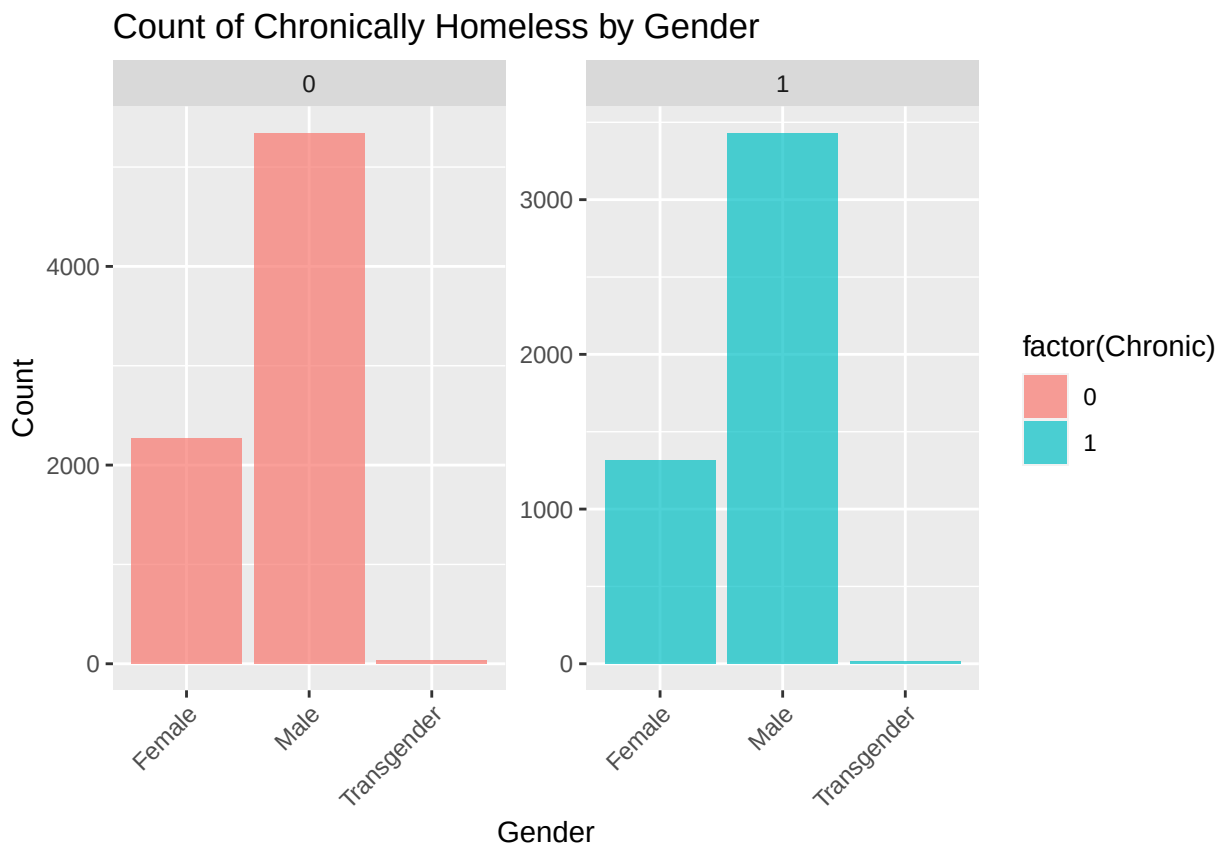
```
# Identifying variables to plot with respect to the binary
# variable chronic homelessness
variables_to_plot <- c("Age", "Gender", "Ethnicity", "Veteran",
    "Times_Homeless_3yrs", "Times_Homeless_Past_Year", "Current_Stint_Duration",
```

```
    "Physical_Sexual_Abuse", "Physical_Disability", "Mental_Illness",
    "Alcohol_Abuse", "Drug_Abuse", "Drug_Alcohol_History", "HIV_Positive",
    "Unemployed_Not_Looking", "Chronic_Time", "Chronic_Condition",
    "Adult_With_Child", "SPA", "Part_Time", "Full_Time")

# List to store ggplots
plots_list <- list()

# For loop will iterate through variables_to_plot and
# create two bar plots (for the binary variable 'Chronic')
# of each variable
for (variable in variables_to_plot) {
    p <- ggplot(homeless_cleaned, aes(x = !!sym(variable), fill = factor(Chronic))) +
        geom_bar(position = "dodge", alpha = 0.7) + labs(title = paste("Count of Chronically Homeless by
        variable), x = variable, y = "Count") + facet_wrap(~Chronic,
        scales = "free_y") + theme(axis.text.x = element_text(angle = 45,
        hjust = 1))
    plots_list[[variable]] <- p

}

# Printing the plots of all variables
print(plots_list)
```

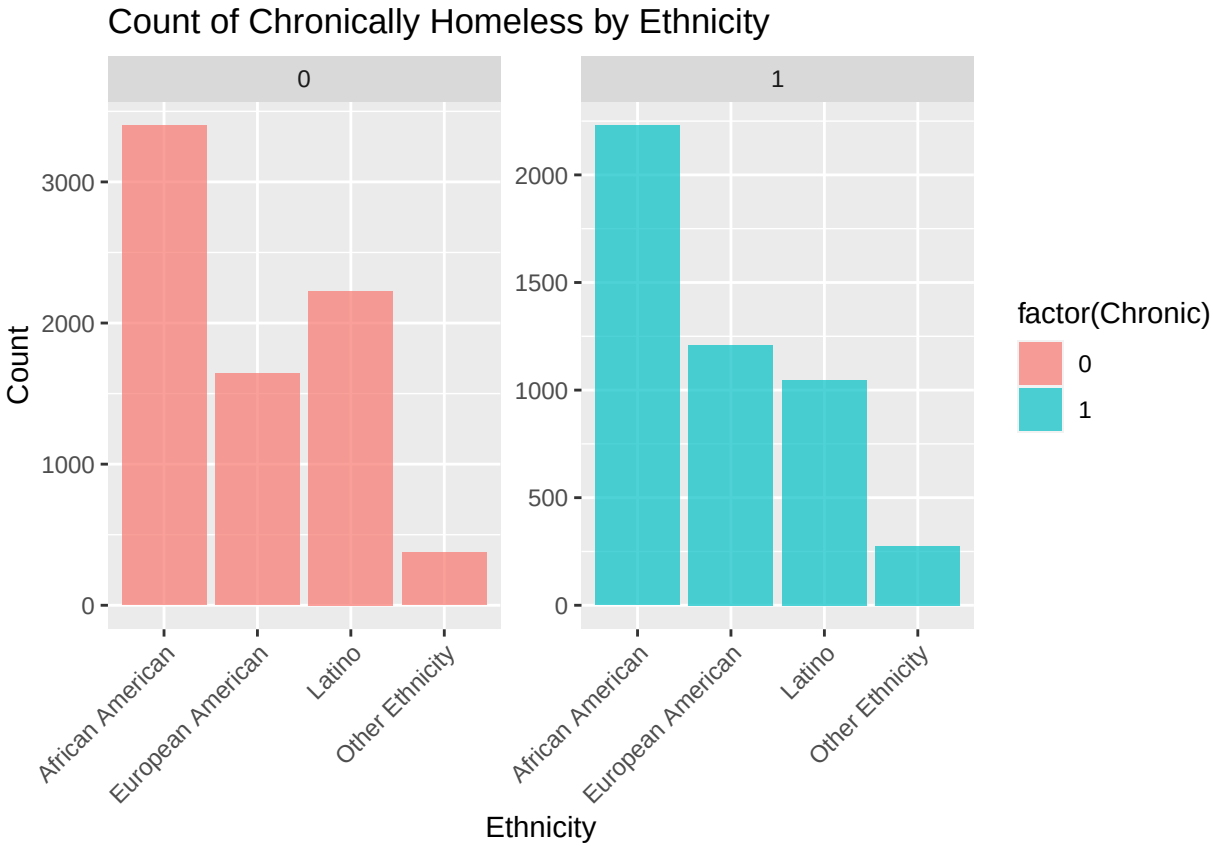
```
## $Age
```



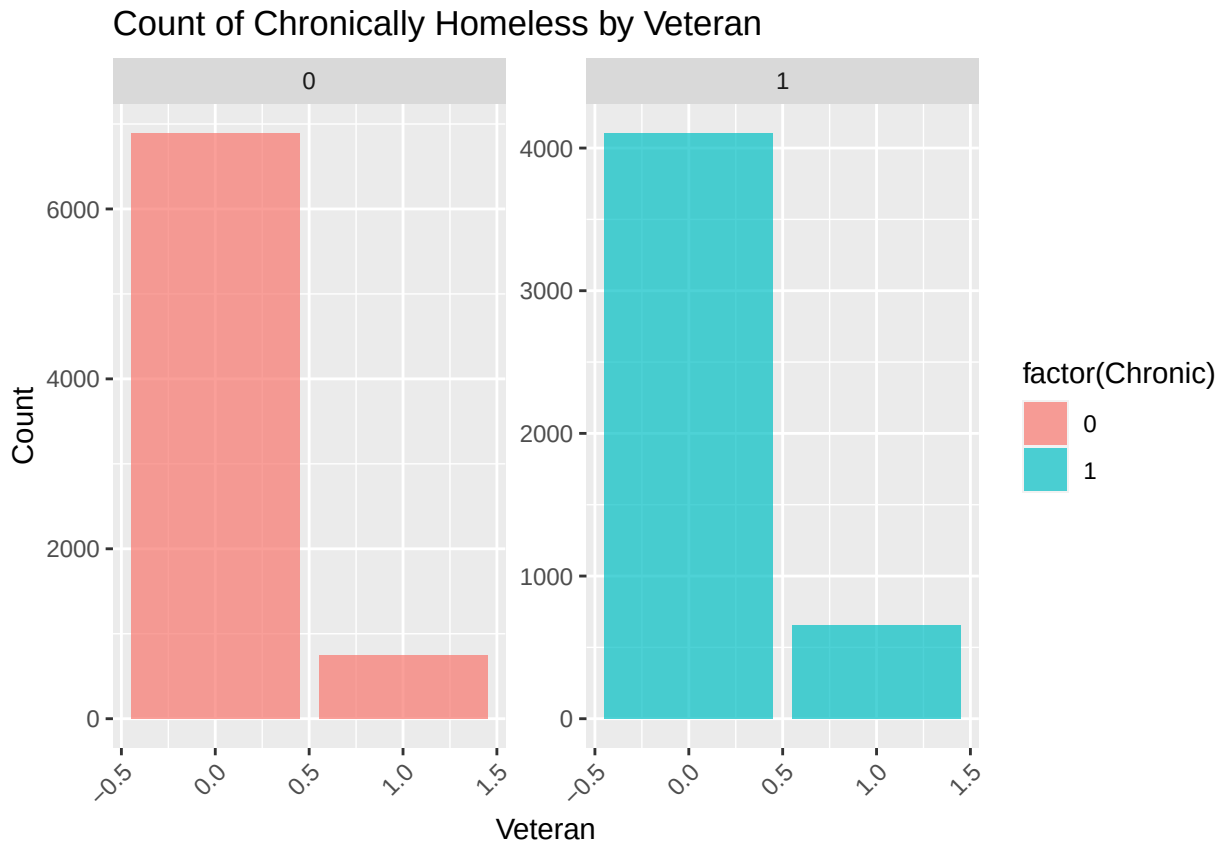
```
##
## $Gender
```



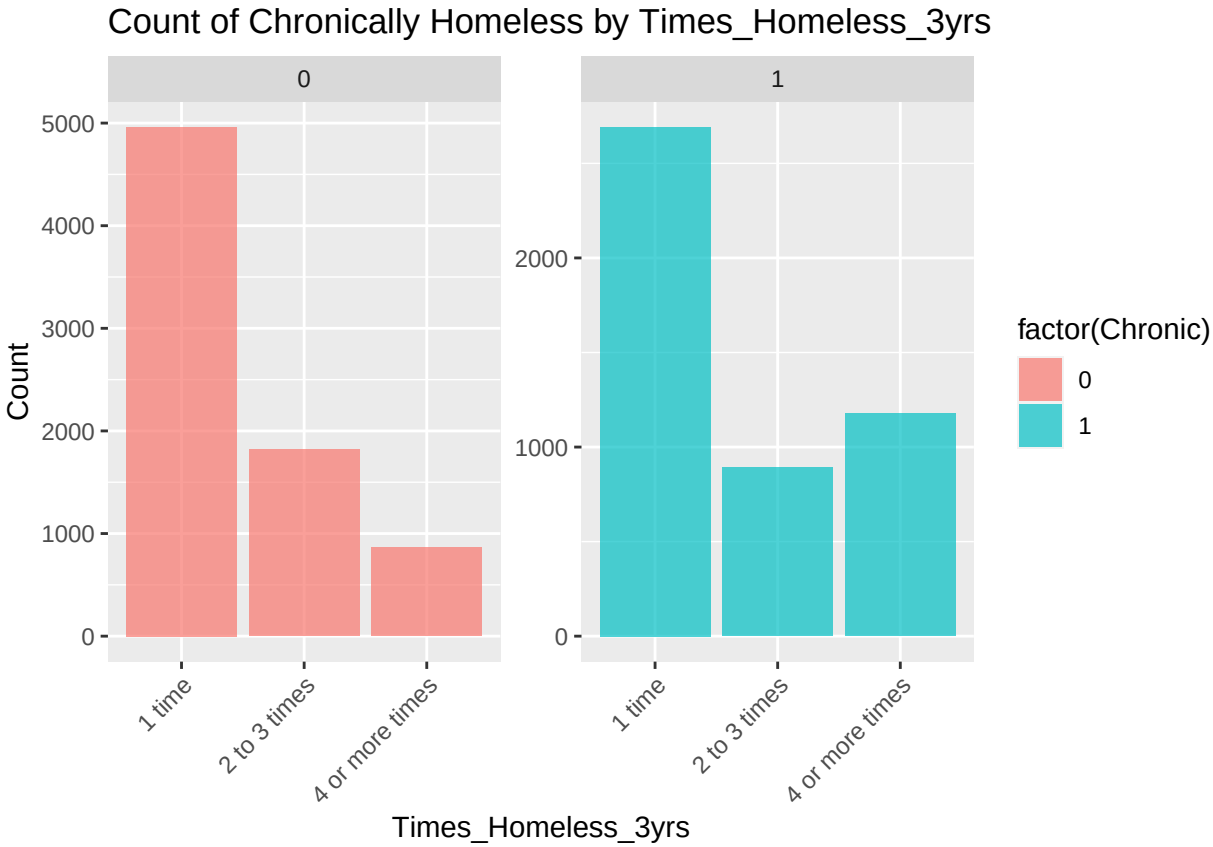

```
##
## $Ethnicity
```


Count of Chronically Homeless by Ethnicity
0
1
3000
2000
1000
0
2000
1500
1000
500
0
Count
African American
European American
Latino
Other Ethnicity
Ethnicity
factor(Chronic)
0
1

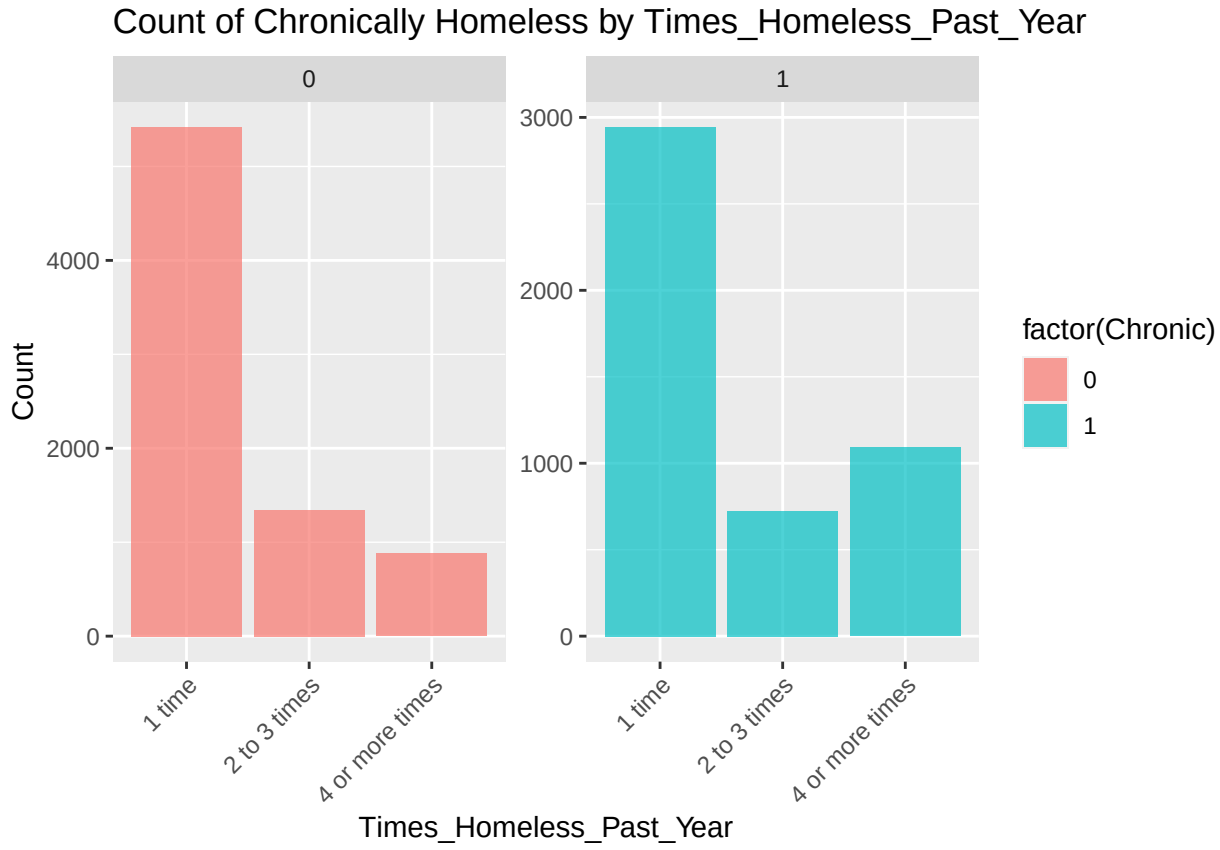
```
##
## $Veteran
```



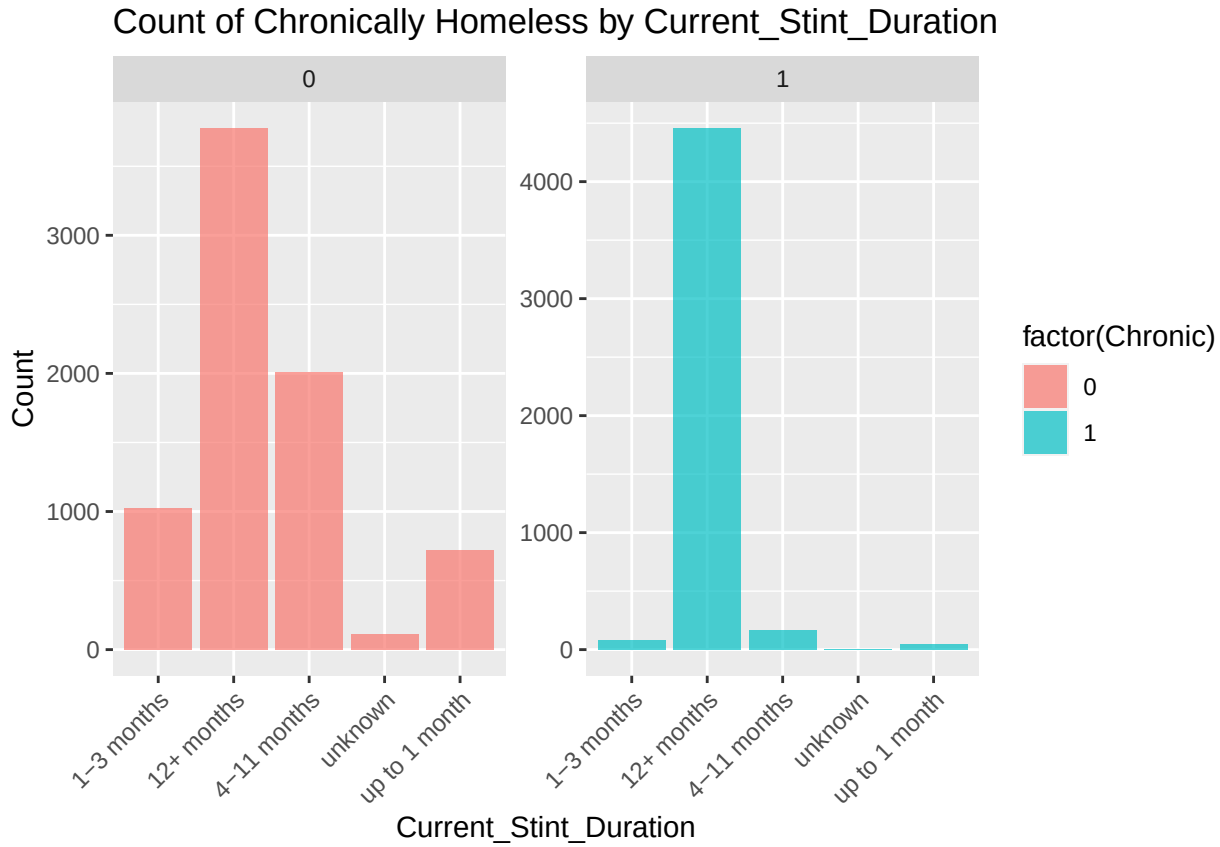

```
##
## $Times_Homeless_3yrs
```



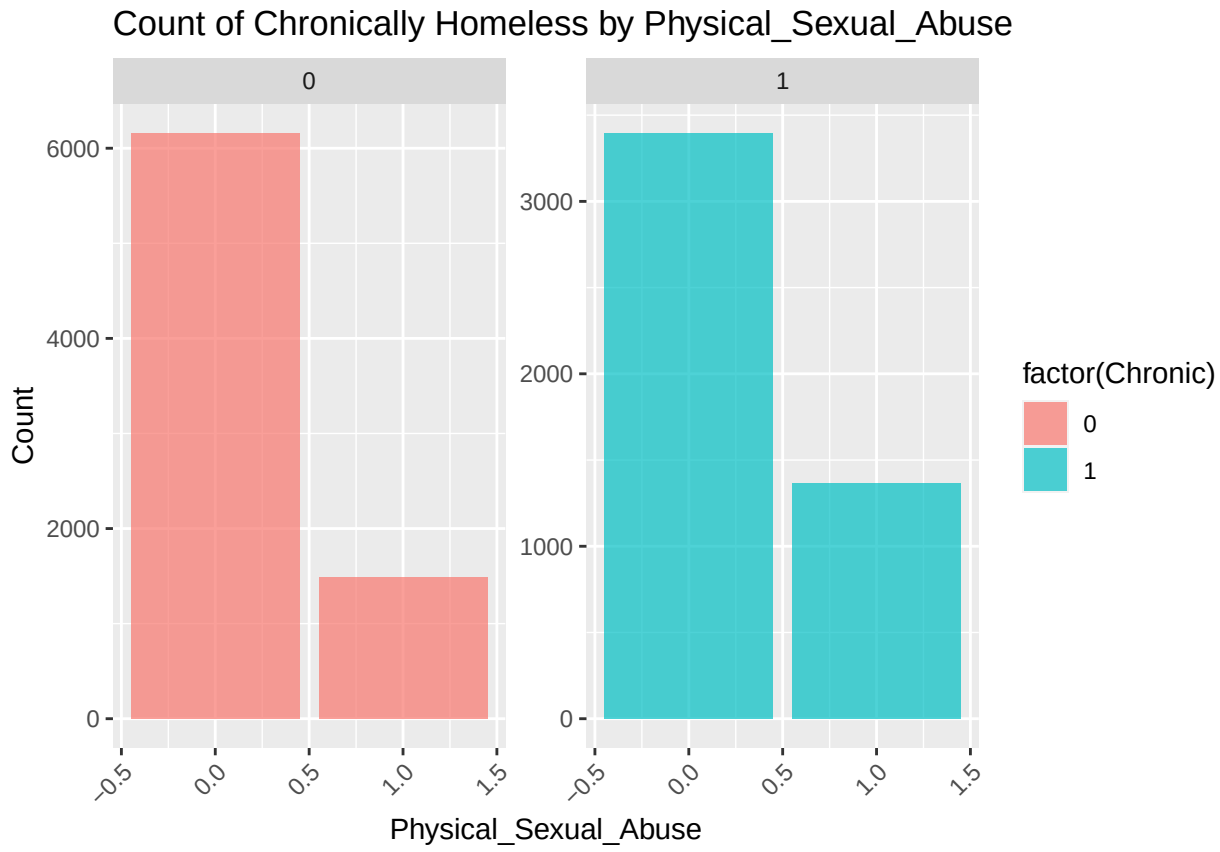

```
##
## $Times_Homeless_Past_Year
```



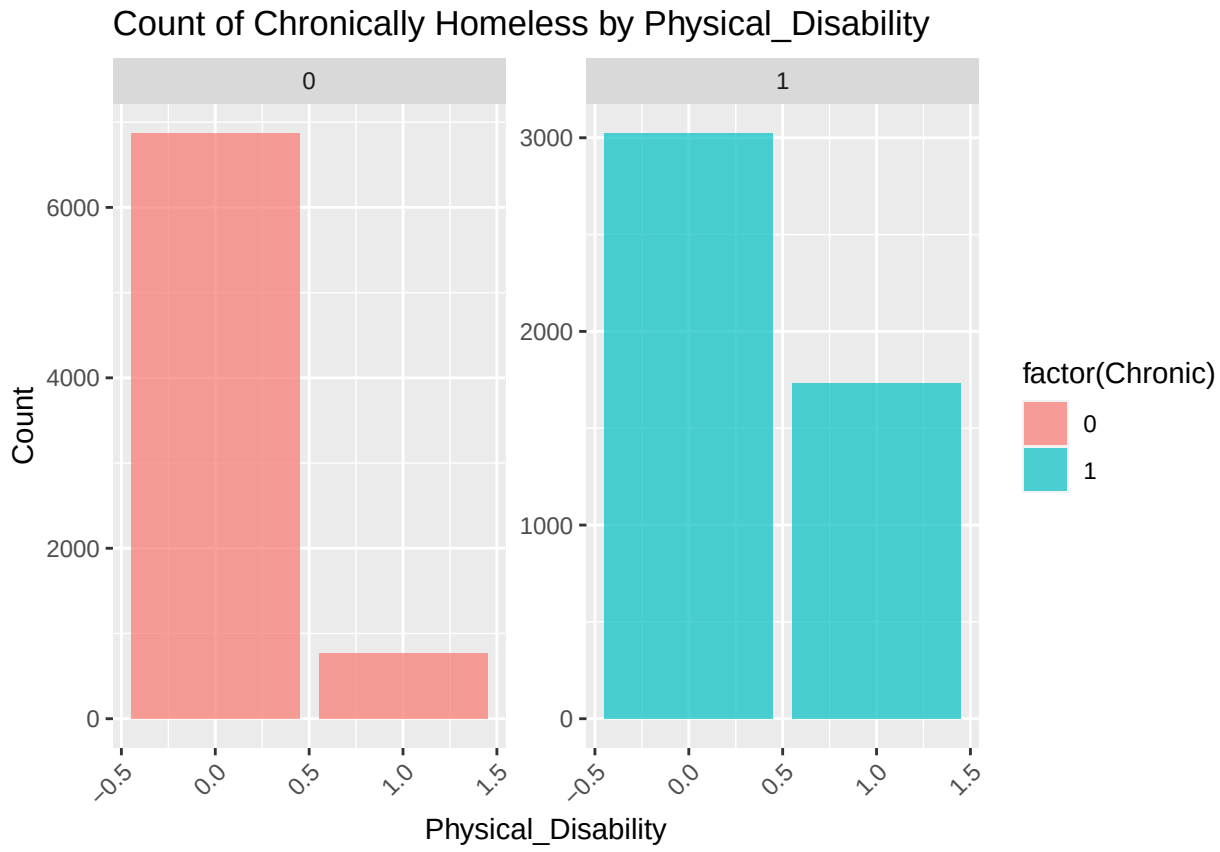

```
##
## $Current_Stint_Duration
```



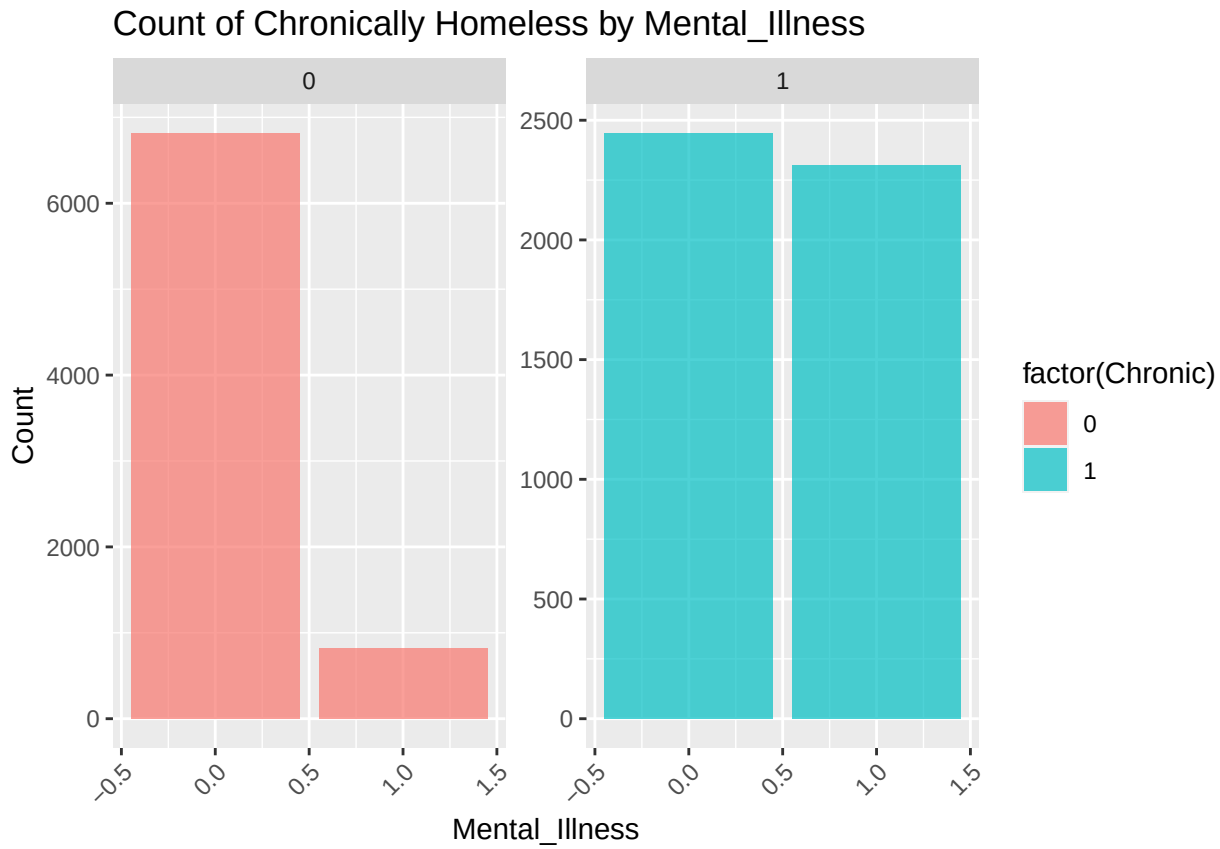

```
##
## $Physical_Sexual_Abuse
```



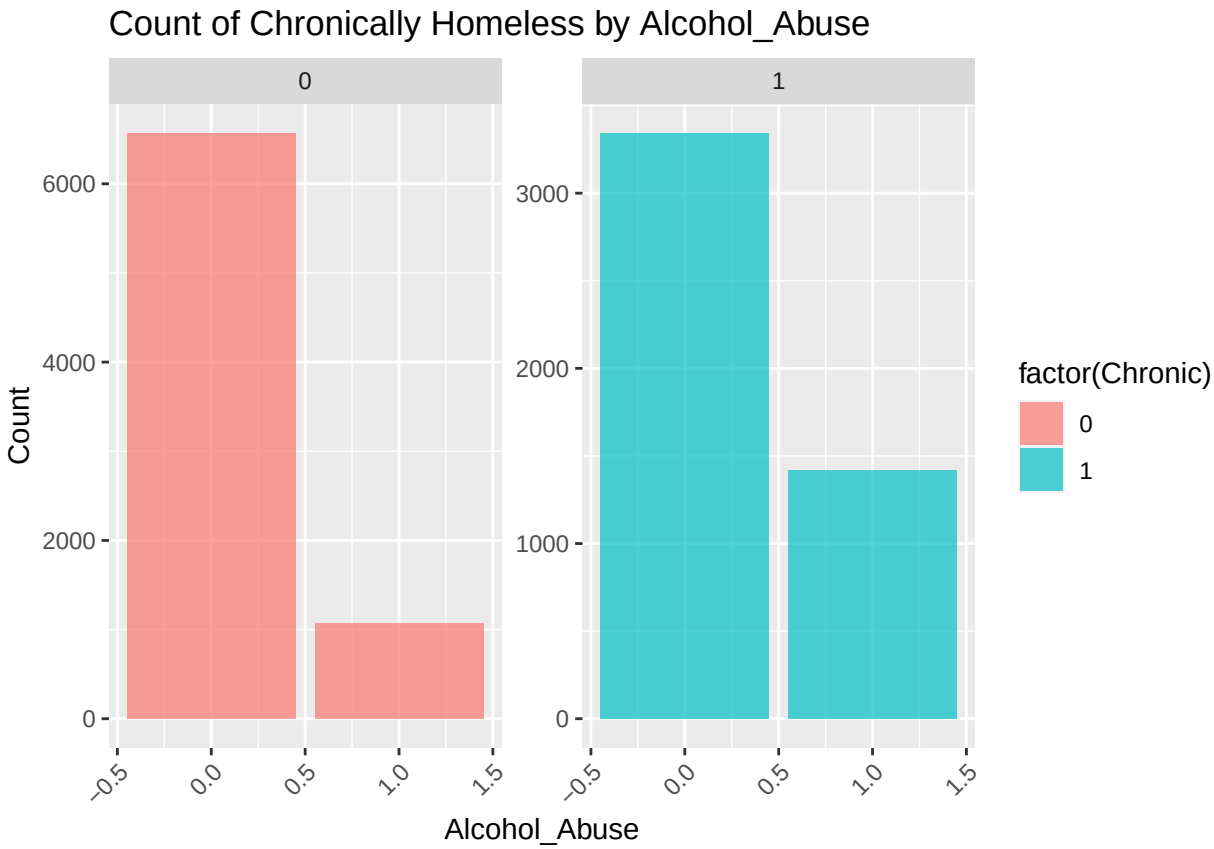

```
##
## $Physical_Disability
```



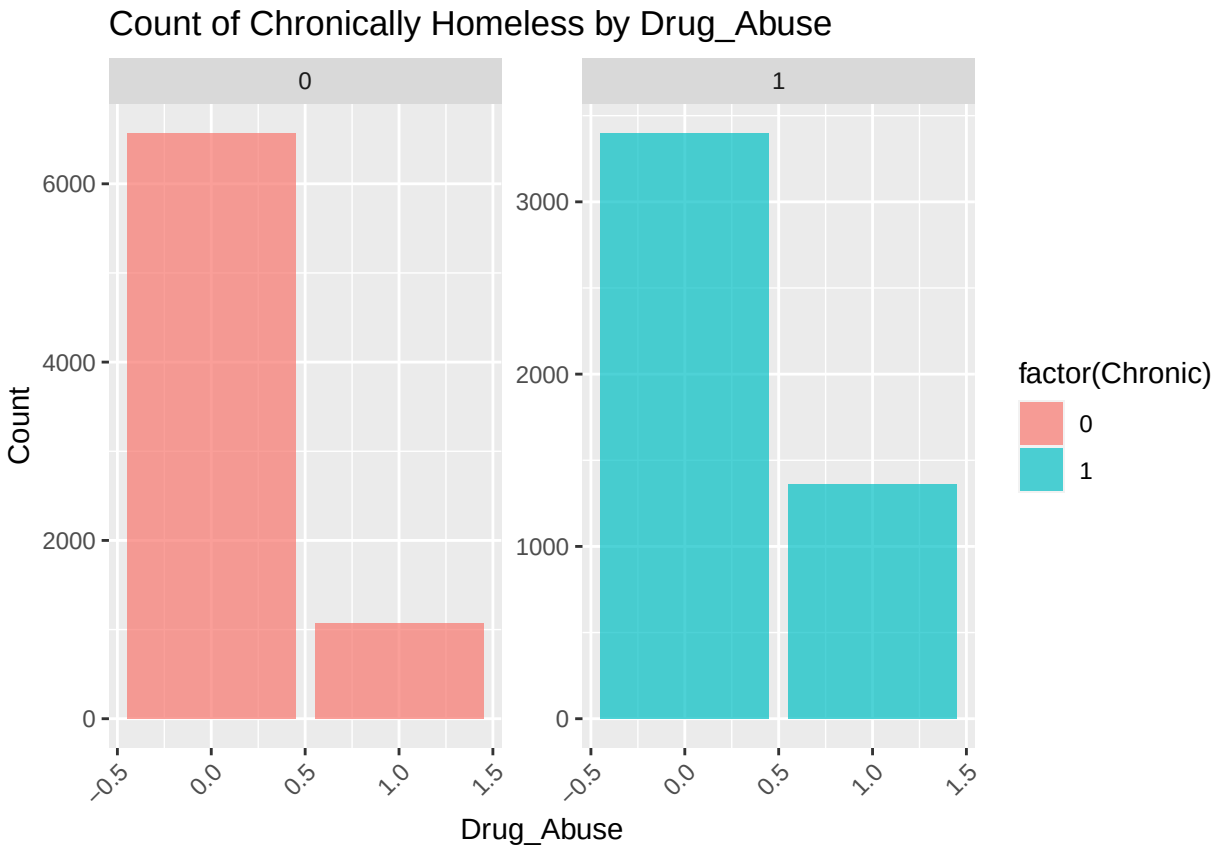

```
##
## $Mental_Illness
```



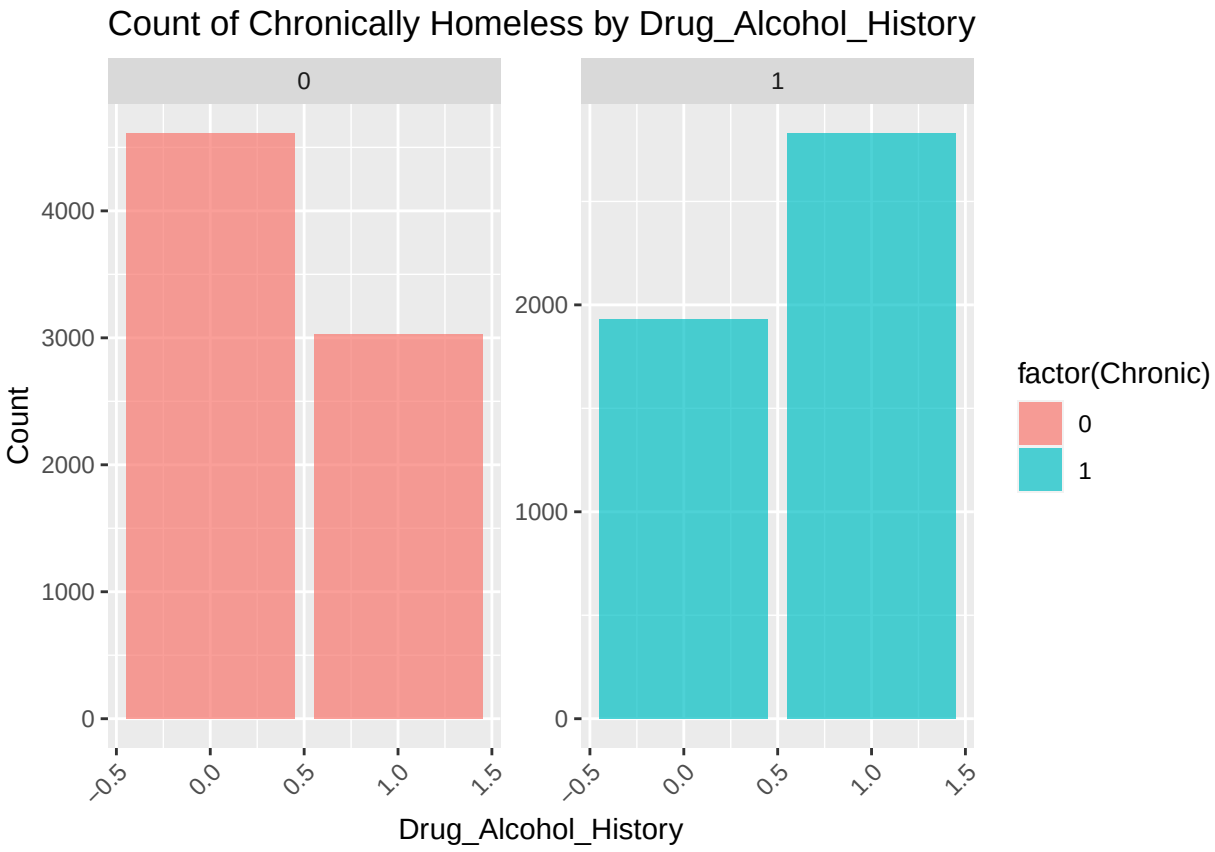

```
##
## $Alcohol_Abuse
```


Count of Chronically Homeless by Alcohol_Abuse
0
1
6000
4000
2000
0
3000
2000
1000
0
Count
-0.5
0.0
0.5
1.0
1.5
Alcohol_Abuse
factor(Chronic)
0
1

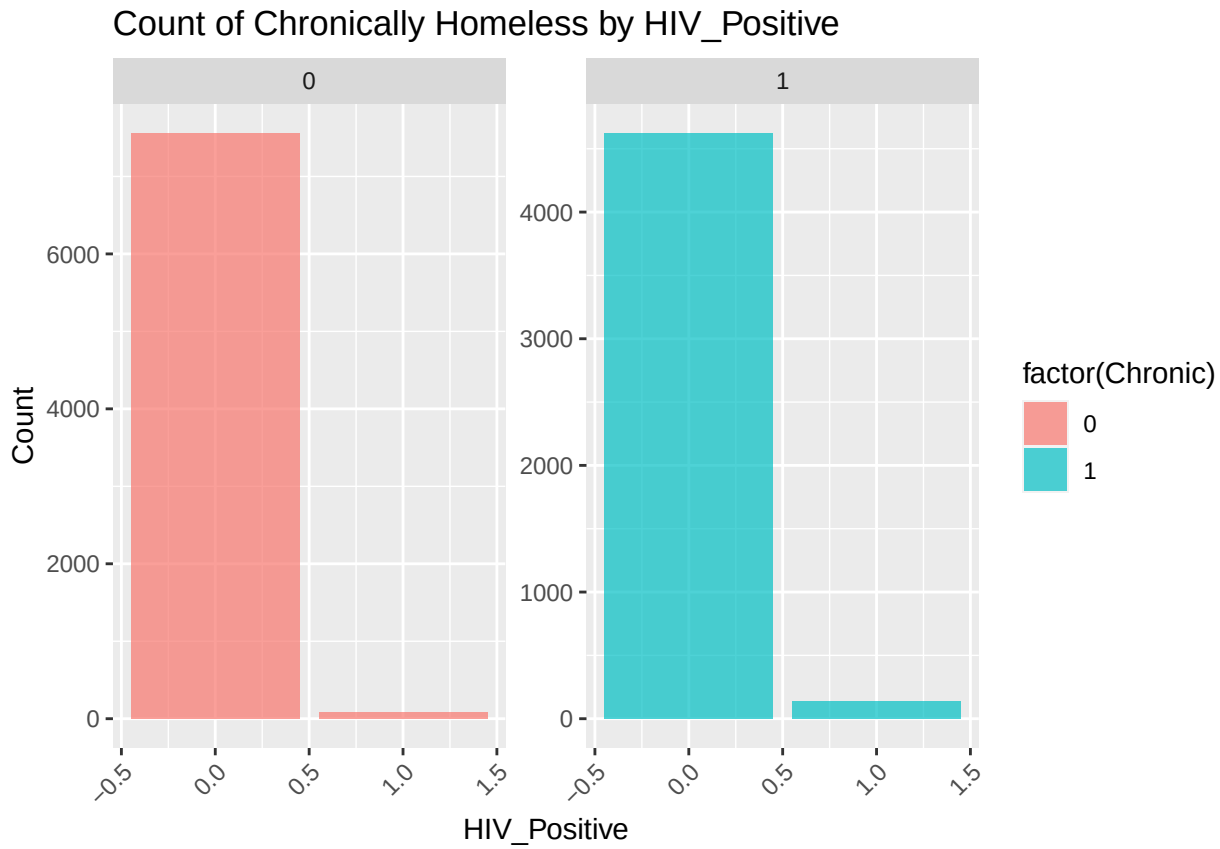
```
##
## $Drug_Abuse
```



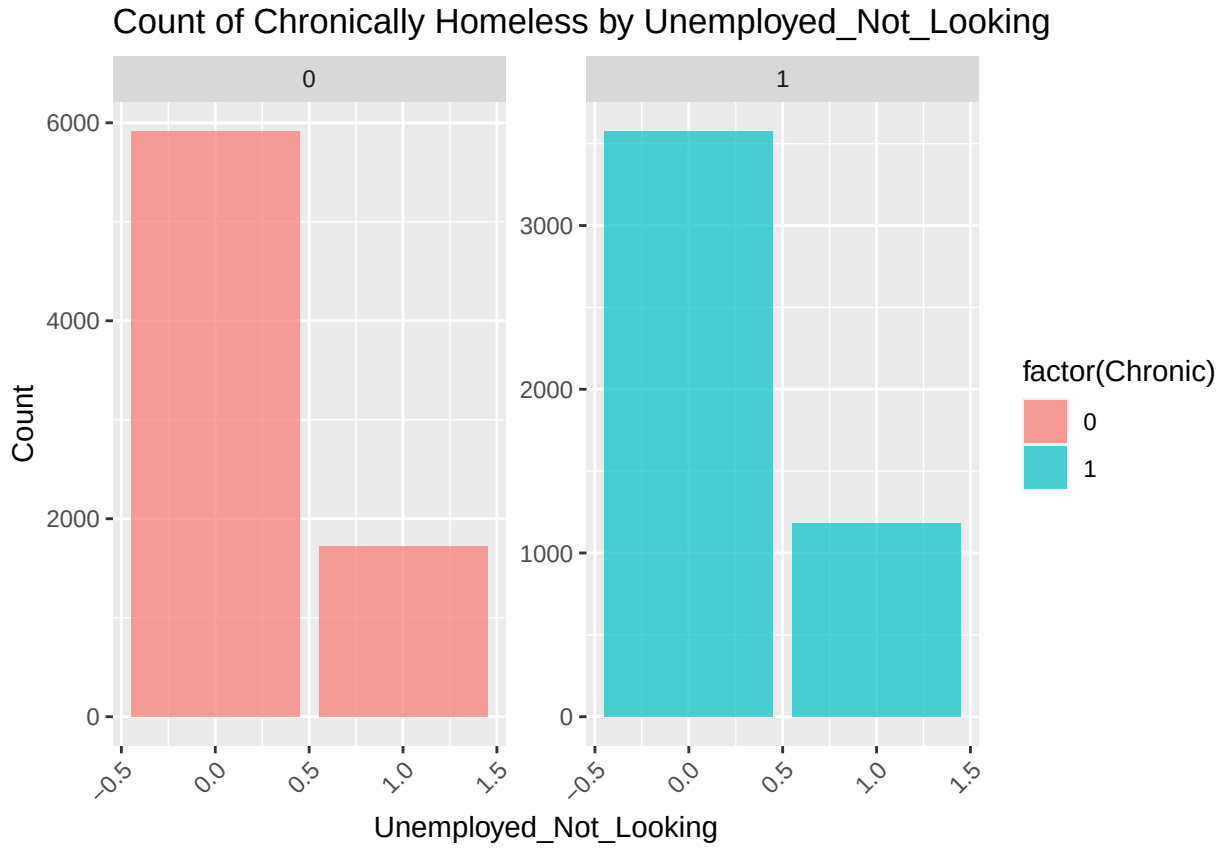

```
##
## $Drug_Alcohol_History
```



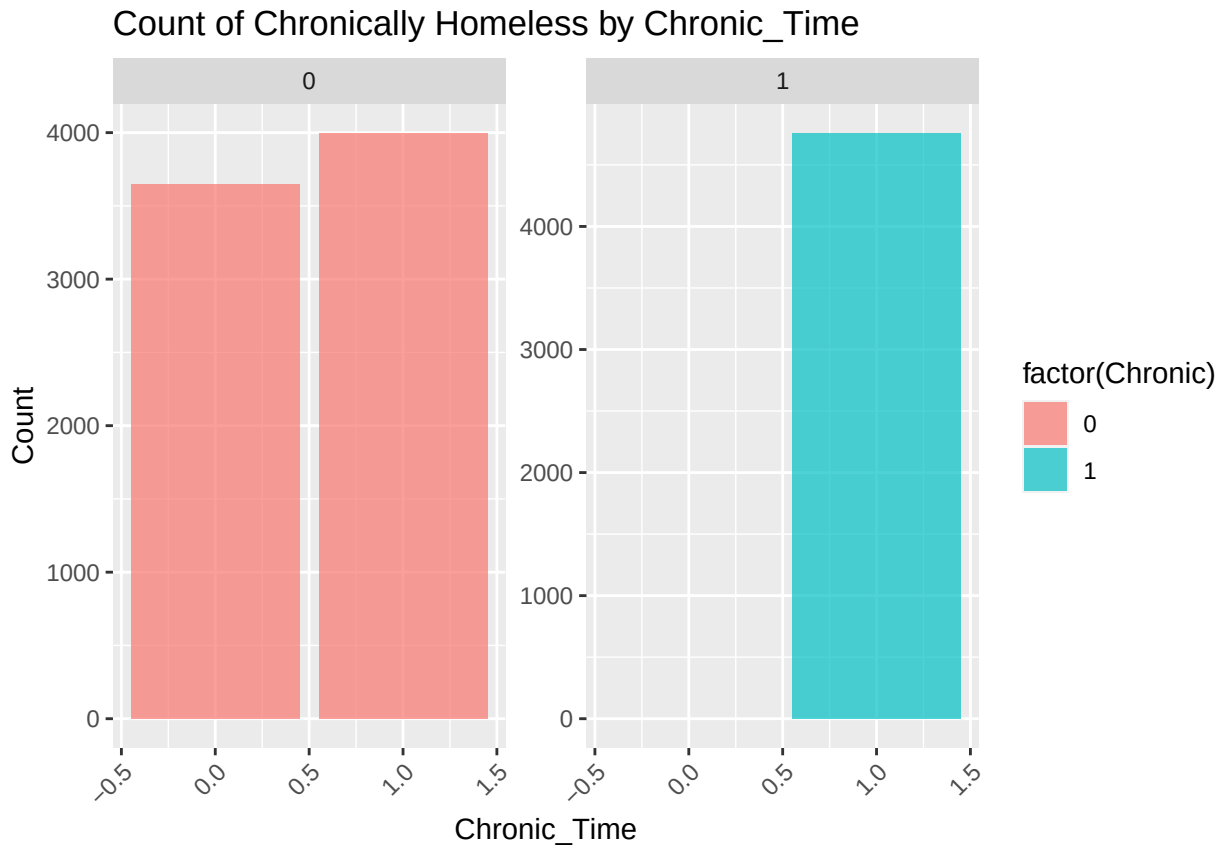

```
##
## $HIV_Positive
```



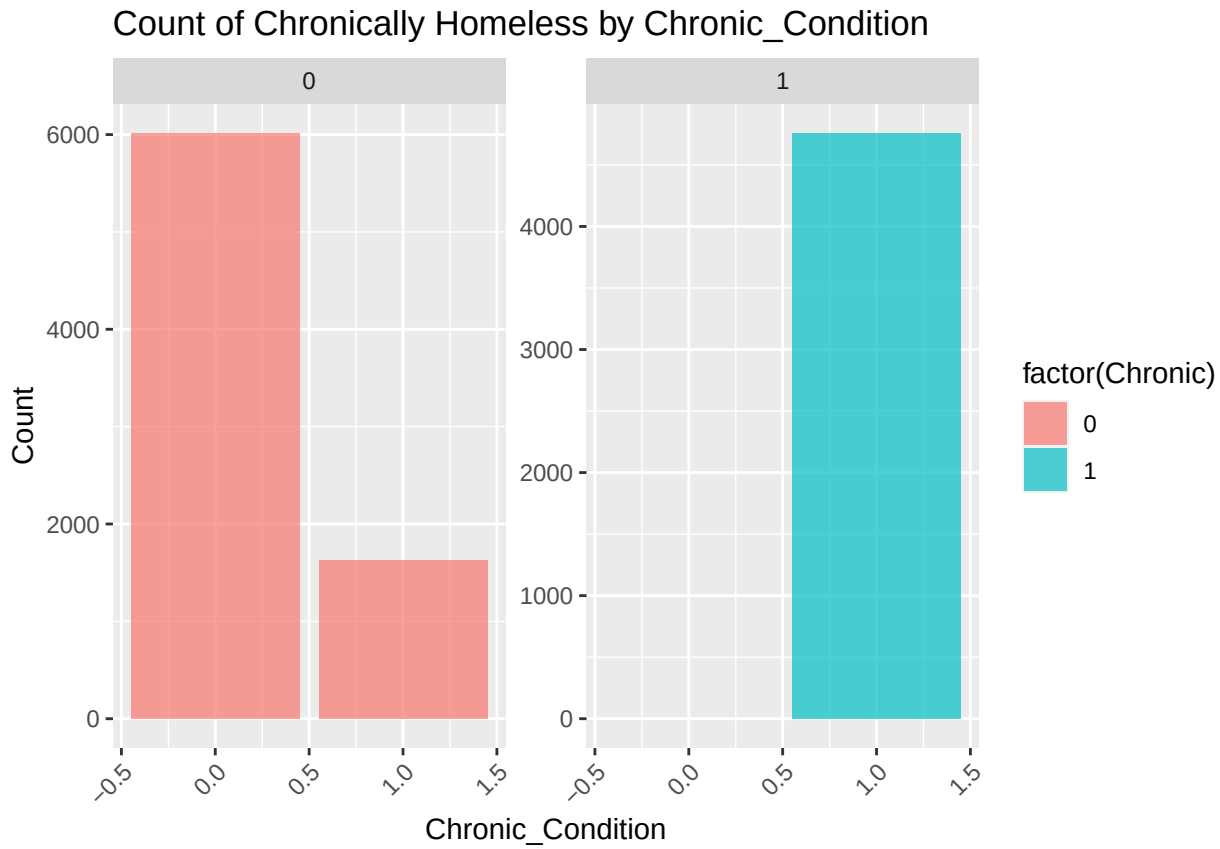

```
##
## $Unemployed_Not_Looking
```



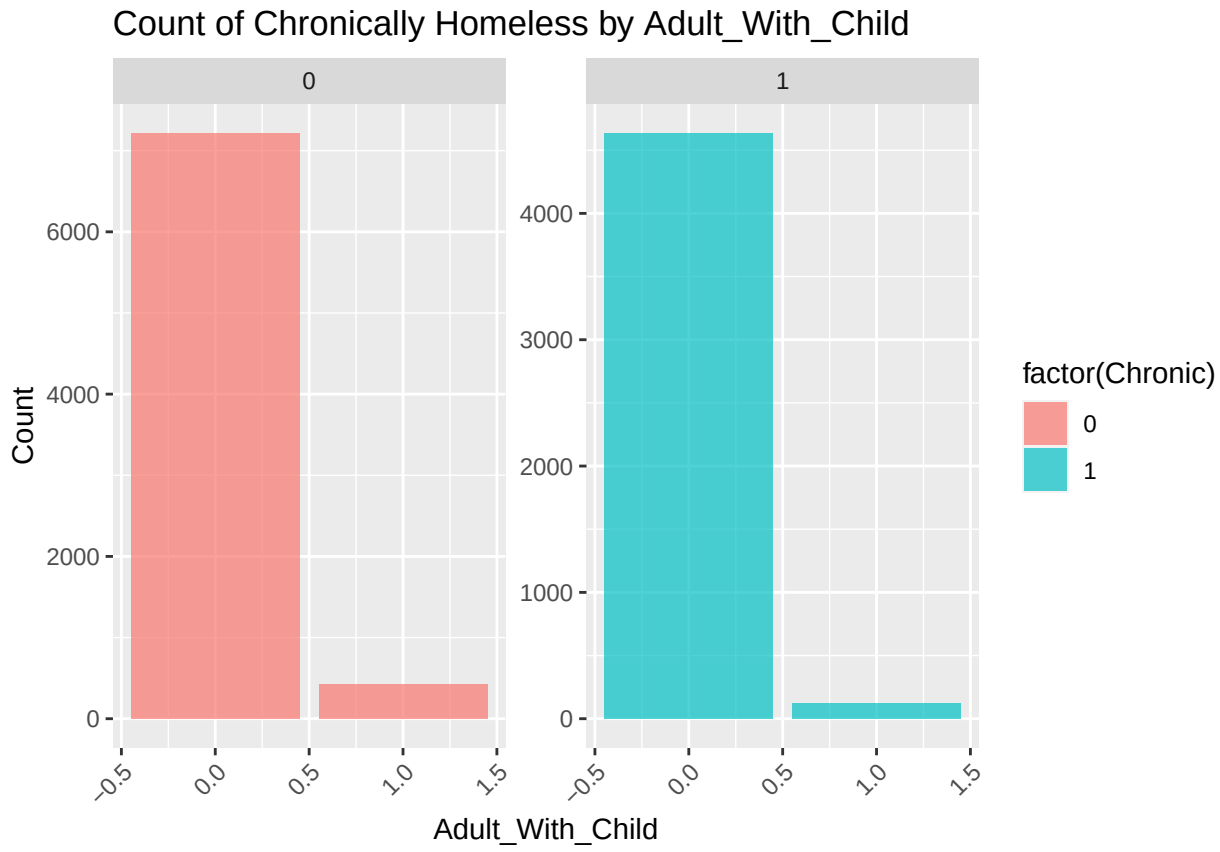

```
##
## $Chronic_Time
```



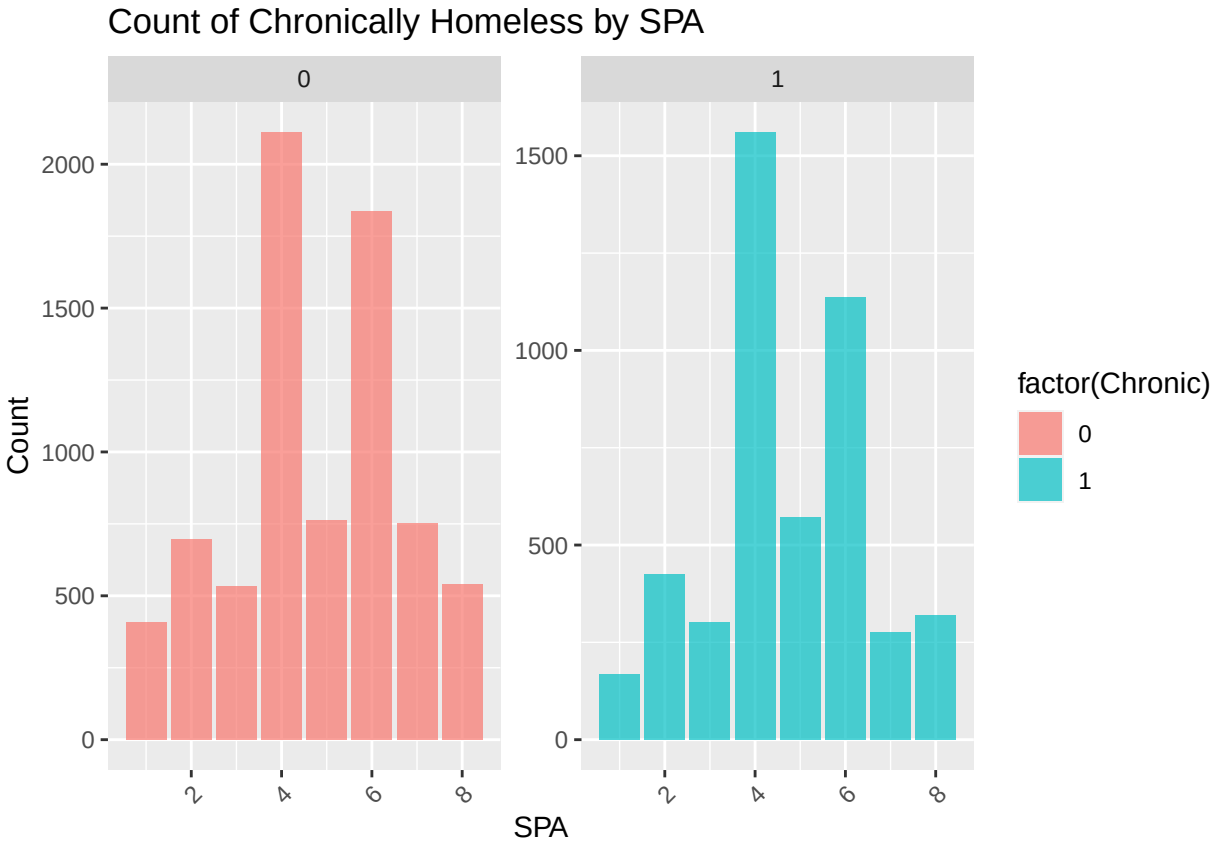

```
##
## $Chronic_Condition
```



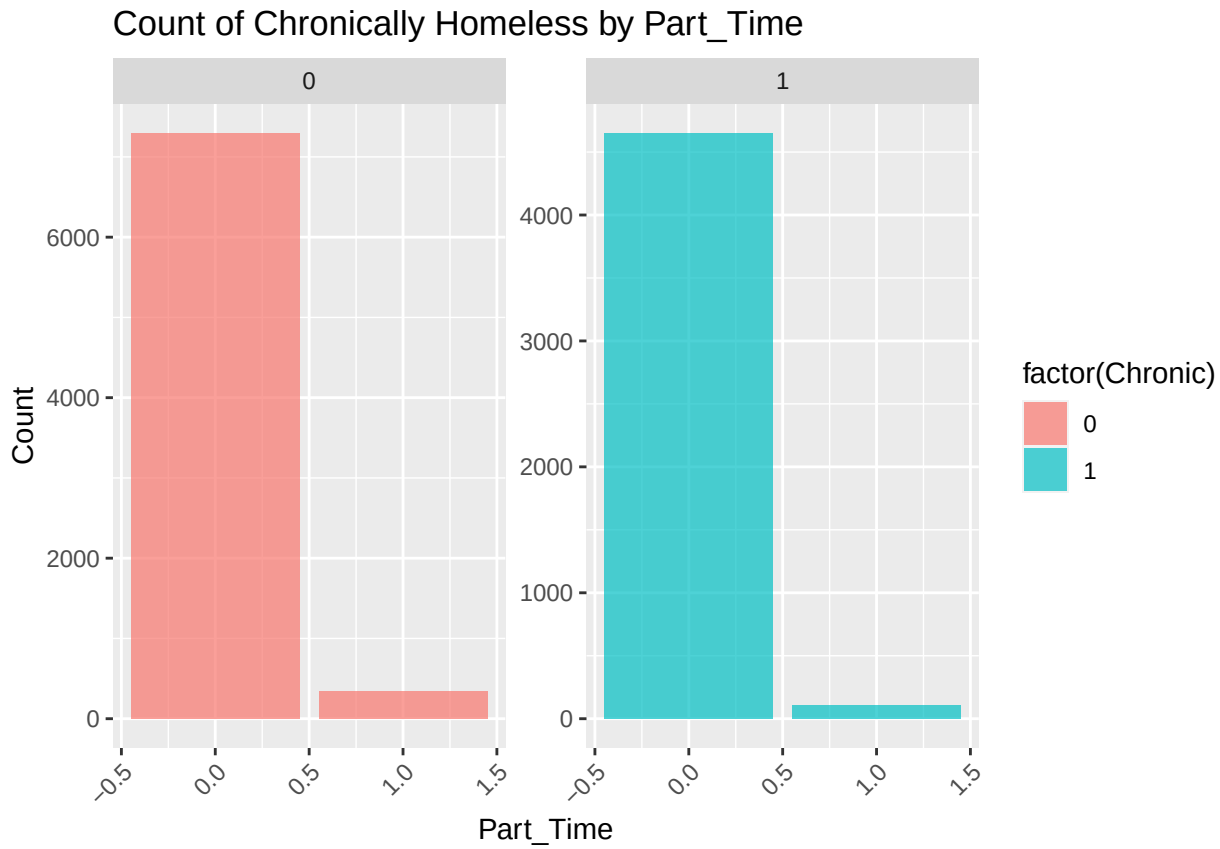

```
##
## $Adult_With_Child
```


Count of Chronically Homeless by Adult_With_Child
0
1
Count
6000
4000
2000
0
4000
3000
2000
1000
0
-0.5
0.0
0.5
1.0
1.5
Adult_With_Child
factor(Chronic)
0
1

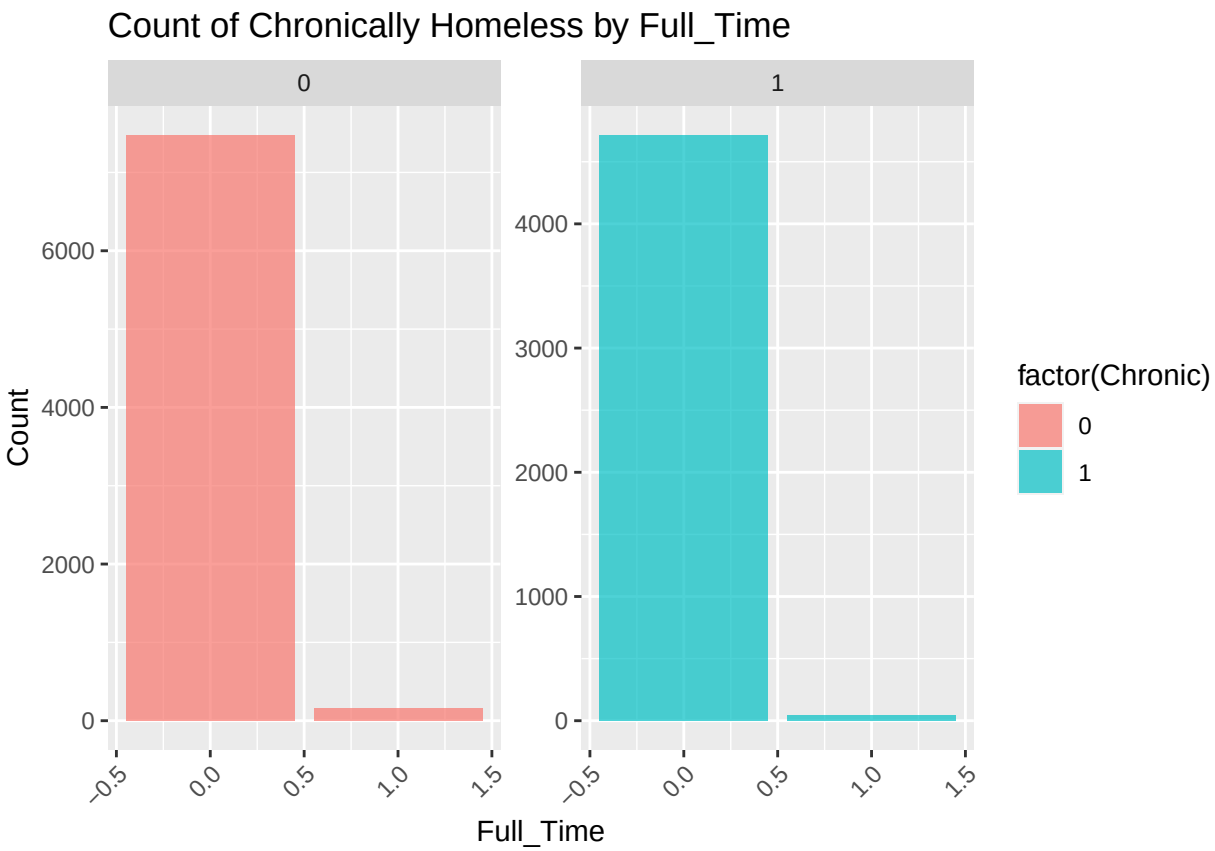
```
##
## $SPA
```


Count of Chronically Homeless by SPA
0
1
2000
1500
1000
500
0
Count
2
4
6
8
SPA
factor(Chronic)
0
1

```
##
## $Part_Time
```


Count of Chronically Homeless by Part_Time
0
1
Count
6000
4000
2000
0
4000
3000
2000
1000
0
-0.5
0.0
0.5
1.0
1.5
Part_Time
factor(Chronic)
0
1

```
##
## $Full_Time
```



Logistic Regression Model Summary

```
summary(lr_fit)
```

```
##
## Call:
## glm(formula = Chronic ~ Age + Gender + Ethnicity + factor(Veteran) +
##     Times_Homeless_3yrs + Times_Homeless_Past_Year + Current_Stint_Duration +
##     factor(Physical_Sexual_Abuse) + factor(Physical_Disability) +
##     factor(Mental_Illness) + factor(Alcohol_Abuse) + factor(Drug_Abuse) +
##     factor(Drug_Alcohol_History) + factor(HIV_Positive) + factor(Unemployed_Not_Looking),
##     family = "binomial", data = homeless_cleaned)
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -5.769483   0.206644 -27.920  < 2e-16
## Age                               0.007551   0.002273   3.322 0.000893
## GenderMale                       -0.055334   0.062922  -0.879 0.379185
## GenderTransgender                -0.956566   0.458469  -2.086 0.036939
## EthnicityEuropean American       -0.091340   0.067015  -1.363 0.172888
## EthnicityLatino                  -0.292722   0.066372  -4.410 1.03e-05
## EthnicityOther Ethnicity         -0.100073   0.124172  -0.806 0.420289
## factor(Veteran)1                  0.239970   0.084139   2.852 0.004343
```

```
## Times_Homeless_3yrs2 to 3 times          0.136488   0.089058   1.533 0.125380
## Times_Homeless_3yrs4 or more times       1.376039   0.109160  12.606  < 2e-16
## Times_Homeless_Past_Year2 to 3 times     0.114221   0.098264   1.162 0.245077
## Times_Homeless_Past_Year4 or more times -0.175577   0.107281  -1.637 0.101711
## Current_Stint_Duration12+ months         4.099045   0.159952  25.627  < 2e-16
## Current_Stint_Duration4-11 months       -0.052256   0.177982  -0.294 0.769062
## Current_Stint_Durationunknown            0.097072   0.521849   0.186 0.852432
## Current_Stint_Durationup to 1 month     -0.064617   0.232834  -0.278 0.781379
## factor(Physical_Sexual_Abuse)1          -0.115631   0.068821  -1.680 0.092926
## factor(Physical_Disability)1             2.142544   0.073294  29.232  < 2e-16
## factor(Mental_Illness)1                  3.003579   0.080256  37.425  < 2e-16
## factor(Alcohol_Abuse)1                   0.751935   0.076359   9.847  < 2e-16
## factor(Drug_Abuse)1                      0.318048   0.076383   4.164 3.13e-05
## factor(Drug_Alcohol_History)1            0.442541   0.066575   6.647 2.99e-11
## factor(HIV_Positive)1                    1.161544   0.215944   5.379 7.49e-08
## factor(Unemployed_Not_Looking)1         -0.262591   0.061459  -4.273 1.93e-05
##
## (Intercept)                             ***
## Age                                     ***
## GenderMale
## GenderTransgender                       *
## EthnicityEuropean American
## EthnicityLatino                         ***
## EthnicityOther Ethnicity
## factor(Veteran)1                        **
## Times_Homeless_3yrs2 to 3 times
## Times_Homeless_3yrs4 or more times      ***
## Times_Homeless_Past_Year2 to 3 times
## Times_Homeless_Past_Year4 or more times
## Current_Stint_Duration12+ months        ***
## Current_Stint_Duration4-11 months
## Current_Stint_Durationunknown
## Current_Stint_Durationup to 1 month
## factor(Physical_Sexual_Abuse)1          .
## factor(Physical_Disability)1            ***
## factor(Mental_Illness)1                 ***
## factor(Alcohol_Abuse)1                  ***
## factor(Drug_Abuse)1                     ***
## factor(Drug_Alcohol_History)1           ***
## factor(HIV_Positive)1                   ***
## factor(Unemployed_Not_Looking)1         ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 16518.9  on 12403  degrees of freedom
## Residual deviance:  9001.6  on 12380  degrees of freedom
## AIC: 9049.6
##
## Number of Fisher Scoring iterations: 6
```

Calculating Difference in Probabilities for Each Predictor

```
age_diff <- homeless_pred(lr_fit, input_data(100, "Female", "European American",
    0, "1 time", "1 time", "up to 1 month", 0, 0, 0, 0, 0, 0,
    0, 0)) - homeless_pred(lr_fit, input_data(18, "Female", "European American",
    0, "1 time", "1 time", "up to 1 month", 0, 0, 0, 0, 0, 0,
    0, 0))
gender_diff <- homeless_pred(lr_fit, input_data(18, "Female",
    "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 0, 0, 0)) - homeless_pred(lr_fit, input_data(18,
    "Transgender", "European American", 0, "1 time", "1 time",
    "up to 1 month", 0, 0, 0, 0, 0, 0, 0, 0))
ethnicity_diff <- homeless_pred(lr_fit, input_data(18, "Female",
    "African American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 0, 0, 0)) - homeless_pred(lr_fit, input_data(18,
    "Female", "Latino", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 0, 0, 0))
veteran_diff <- homeless_pred(lr_fit, input_data(18, "Female",
    "European American", 1, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 0, 0, 0)) - homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 0, 0, 0))
times_homeless3_diff <- homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "4 or more times", "1 time",
    "up to 1 month", 0, 0, 0, 0, 0, 0, 0, 0)) - homeless_pred(lr_fit,
    input_data(18, "Female", "European American", 0, "1 time",
        "1 time", "up to 1 month", 0, 0, 0, 0, 0, 0, 0, 0))
times_homeless_pastyr_diff <- homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "1 time", "4 or more times",
    "up to 1 month", 0, 0, 0, 0, 0, 0, 0, 0)) - homeless_pred(lr_fit,
    input_data(18, "Female", "European American", 0, "1 time",
        "1 time", "up to 1 month", 0, 0, 0, 0, 0, 0, 0, 0))
current_stint_dur_diff <- homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "1 time", "1 time", "12+ months",
    0, 0, 0, 0, 0, 0, 0, 0)) - homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 0, 0, 0))
physical_sexual_abuse_diff <- homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "1 time", "1 time", "up to 1 month",
    1, 0, 0, 0, 0, 0, 0, 0)) - homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 0, 0, 0))
physical_disability_diff <- homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 1, 0, 0, 0, 0, 0, 0)) - homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 0, 0, 0))
mental_illness_diff <- homeless_pred(lr_fit, input_data(18, "Female",
    "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 1, 0, 0, 0, 0, 0)) - homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 0, 0, 0))
alcohol_abuse_diff <- homeless_pred(lr_fit, input_data(18, "Female",
```

```
    "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 1, 0, 0, 0, 0)) - homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 0, 0, 0))
drug_abuse_diff <- homeless_pred(lr_fit, input_data(18, "Female",
    "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 1, 0, 0, 0)) - homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 0, 0, 0))
drug_alcohol_hist_diff <- homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 1, 0, 0)) - homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 0, 0, 0))
hiv_pos_diff <- homeless_pred(lr_fit, input_data(18, "Female",
    "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 0, 1, 0)) - homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 0, 0, 0))
unemployed_not_looking_diff <- homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 0, 0, 0)) - homeless_pred(lr_fit, input_data(18,
    "Female", "European American", 0, "1 time", "1 time", "up to 1 month",
    0, 0, 0, 0, 0, 0, 0, 1))
```