MRSA Forecast

Chris Lee

2024-09-15

Forecasting is beginning to be integrated into decision-making processes not only for business operation but also for infectious disease outbreak response.

The demo uses the data from California Health and Human Services Open Data Portal. All California general acute care hospitals are required to report Methicillin-resistant Staphylococcus aureus (MRSA) bloodstream infection (BSI) cases that occur following hospitalization.

MRSA as its name suggested is caused by strains of Staphylococcus aureus develop antibiotic resistance. The bacteria can develop antibiotic resistance naturally in the environment or through defense mechanisms to block or destroy antibiotic drugs. MRSA is contagious. It can spread through skin- to-skin contact or on surfaces 1.

Goal: In this post, our goal is to predict Standardized Infection Ratio (SIR) for each facility in 2023.

Overview of variables used in the post.

Data Overview

```
```{r}
#| label: Overview data
#| fig-cap: " Distribution of the variables"
#| code-fold: true
#| fig-width: 8
```
```

```
my_glimpse(data = mrsa_combine|> select(-Facility_ID))[-7,]
```

vars	type_range__
Year	continuous, [2014, 2023]
Infections_reported	continuous, [NA, NA]
Patient_Days	continuous, [NA, NA]
SIR	continuous, [NA, NA]
Hospital_Category_RiskAdjustment	categorical, 3values Acute Care Hospital Long-Term Acute Care Hospital Rehabilitation Hospital or Unit Critical Access Hospital NArange5values
Hospital_Type	categorical, 9values Community <125 Beds Major Teaching Community 125-250 Beds Long-Term Acute Care Pediatric Community >250 Beds Free-Standing Rehabilitation Critical Access NArange9values

Distribution of the variables

Data Visualization

Overall the distribution of County, Hospital_Category_RiskAdjustment, and Hospital_Type didn't have major changes from 2014 to 2023. However, two new clinic categories were found in the Hospital_Type after 2016. This could be possible, either there were not many rehabilitation Hospitals or Units or they were not included in the study prior to 2016.

```
```{r}
#| label: Overview categorical data
#| fig-cap: " Distribution of the categorical variables"
#| warning: false
#| code-fold: true
#| fig-width: 8
#| fig-height: 10

freq_data <- function(df, year){
```
```

```
  df <- mrsa_combine |>
       filter(Year == year)|>

       mutate(`Hospital Type`  = case_when(Hospital_Type == "Community, <125 Beds"~ "<125",
                                          Hospital_Type == "Community, 125-250 Beds"~ "125-250",
                                          Hospital_Type == "Community, >250 Beds"~ ">250",
                                          .default = Hospital_Type),
              `Risk Adjustment`   = Hospital_Category_RiskAdjustment)|>
    select(Facility_ID, `Hospital Type`, `Risk Adjustment`, County)

  df <- as.data.frame(df)
  results_list <- vector("list", ncol(df) - 1)

  # Loop through the columns starting from the 2nd column
  for (i in 2:ncol(df)) {
       # Store the result of cal_freq(i) in the list
       results_list[[i - 1]] <- cal_freq(i, df)
      }
  df <- do.call(rbind, results_list)

  # label bar graph
  df$vars <- ifelse(df$Freqx > 0.15, as.character(df$Var1), "")
  return(df)
}

d2014 <- freq_data(df, 2014)
d2014 <- freq_data(df, 2014)
d2016 <- freq_data(df, 2016)
d2019 <- freq_data(df, 2019)
d2023 <- freq_data(df, 2023)
d2014$Year <- 2014
d2016$Year <- 2016
d2019$Year <- 2019
d2023$Year <- 2023

p <-
rbind(d2014,d2016,d2019, d2023)|>
    #filter(Year == 2013)|>
    #select(-Year)|>
    ggplot(aes(x = round(Freqx,2), y = fct_rev(namex), fill = Var1)) +
```

```
    geom_col(position = "fill", color = "white") +
    scale_x_continuous(labels = label_percent()) +
    labs(y = NULL, x = NULL, fill = "var1")+
    guides(fill="none") +
    geom_text(aes(label = vars), position = position_stack(vjust = 0.75), color = "white")+
    theme_void()

p+transition_time(as.integer(Year))+
  enter_fade() +
  exit_fade()+labs(title = "Year: {frame_time}")

#anim_save(filename="animation.gif", mygif)
```

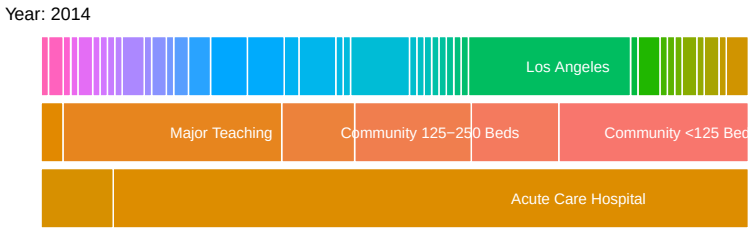


Figure 1: Distribution of the categorical variables

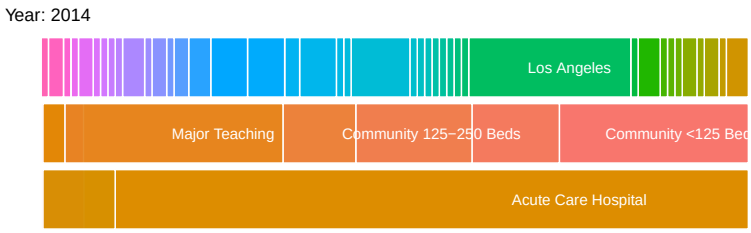


Figure 2: Distribution of the categorical variables

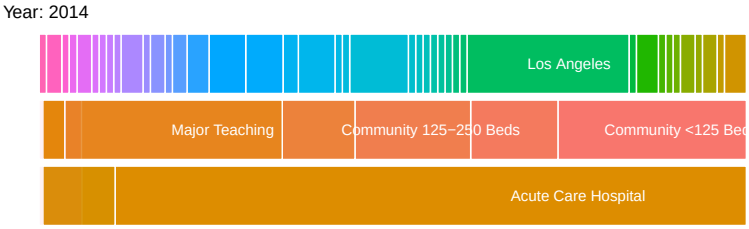


Figure 3: Distribution of the categorical variables

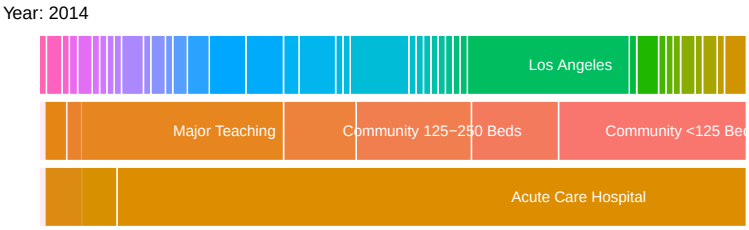


Figure 4: Distribution of the categorical variables

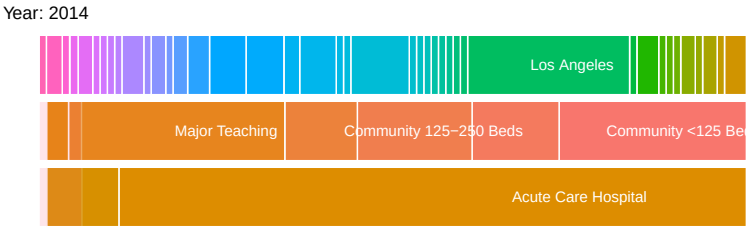


Figure 5: Distribution of the categorical variables

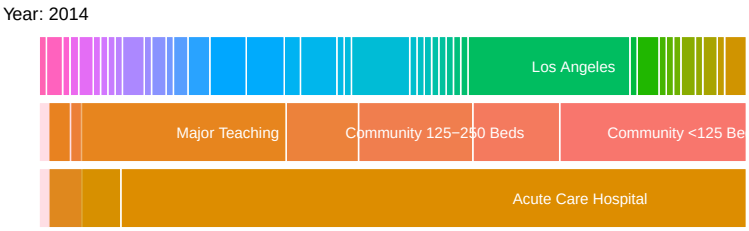


Figure 6: Distribution of the categorical variables

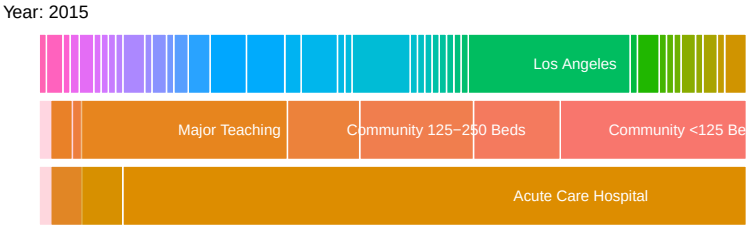


Figure 7: Distribution of the categorical variables

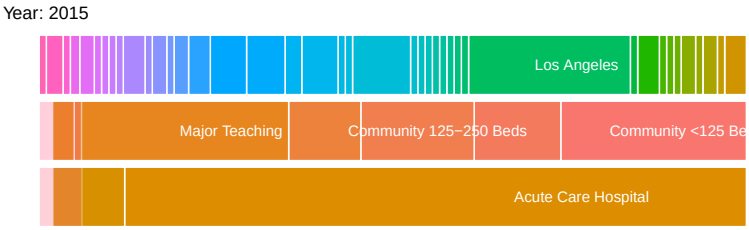


Figure 8: Distribution of the categorical variables

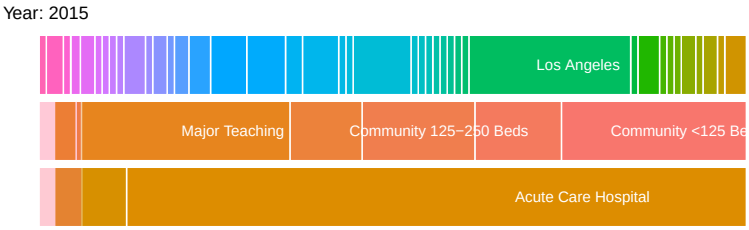


Figure 9: Distribution of the categorical variables

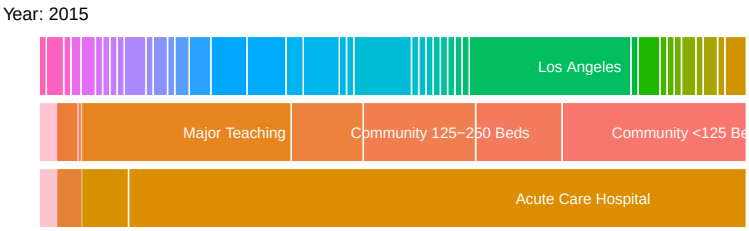


Figure 10: Distribution of the categorical variables

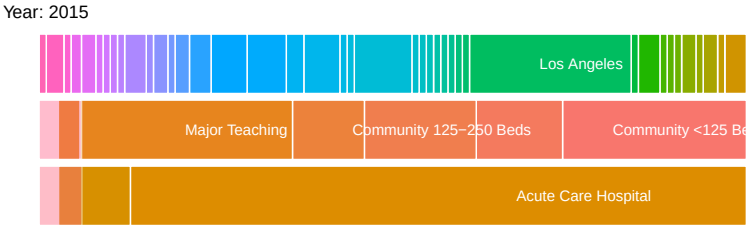


Figure 11: Distribution of the categorical variables

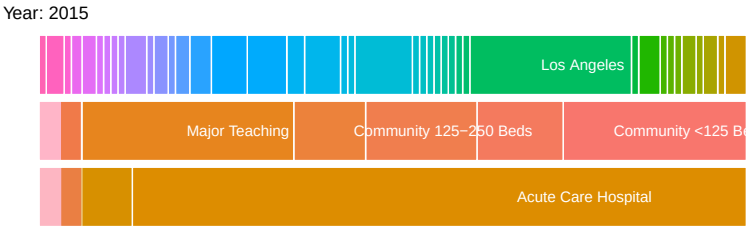


Figure 12: Distribution of the categorical variables

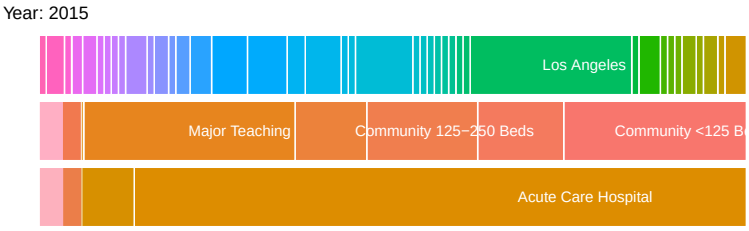


Figure 13: Distribution of the categorical variables

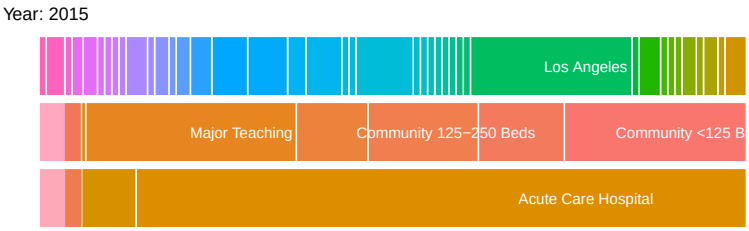


Figure 14: Distribution of the categorical variables

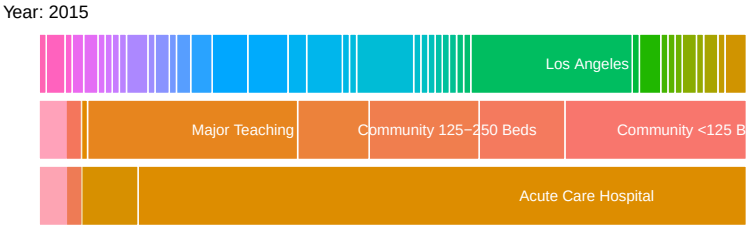


Figure 15: Distribution of the categorical variables

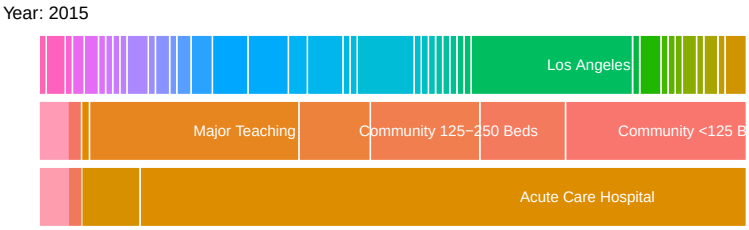


Figure 16: Distribution of the categorical variables

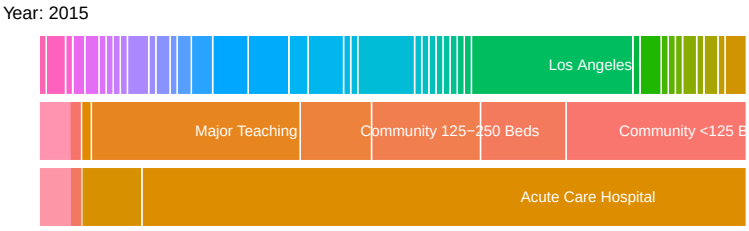


Figure 17: Distribution of the categorical variables

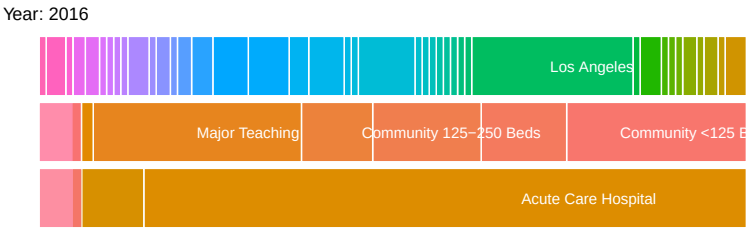


Figure 18: Distribution of the categorical variables

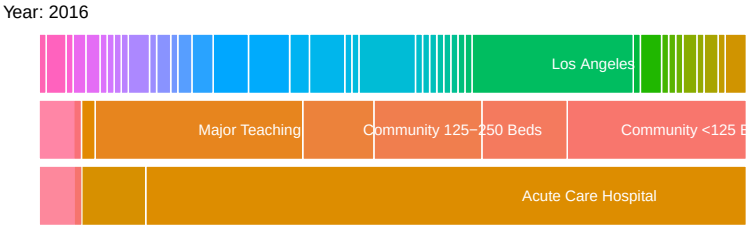


Figure 19: Distribution of the categorical variables

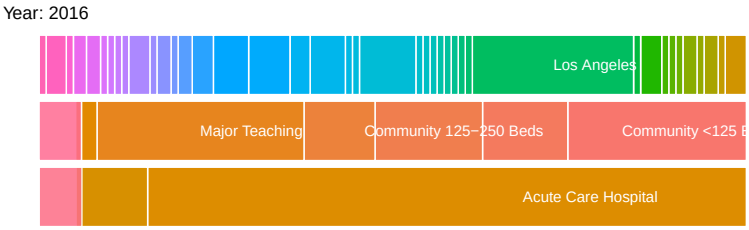


Figure 20: Distribution of the categorical variables

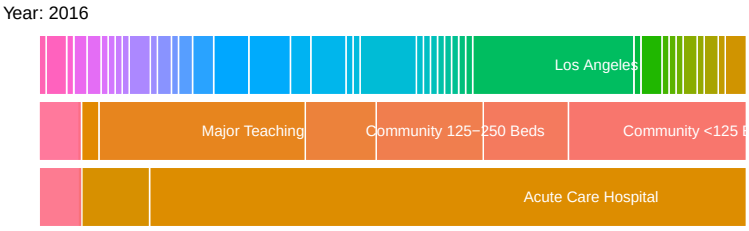


Figure 21: Distribution of the categorical variables

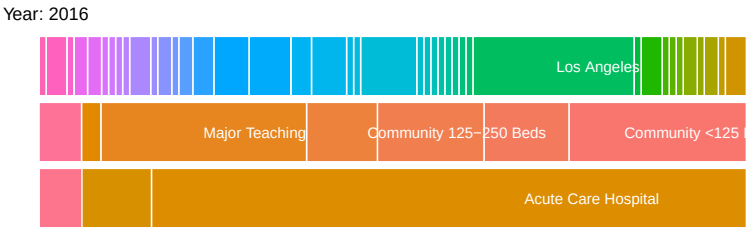


Figure 22: Distribution of the categorical variables

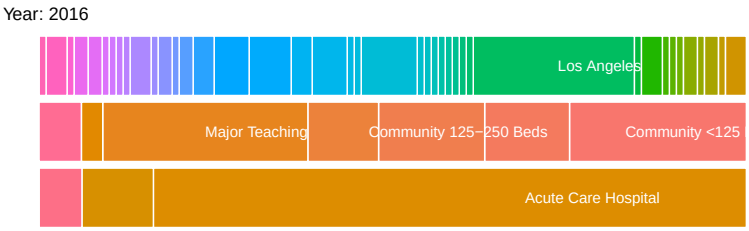


Figure 23: Distribution of the categorical variables

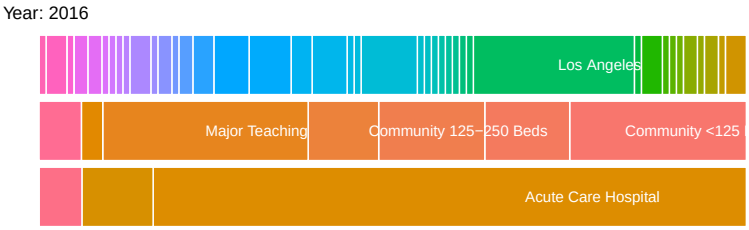


Figure 24: Distribution of the categorical variables

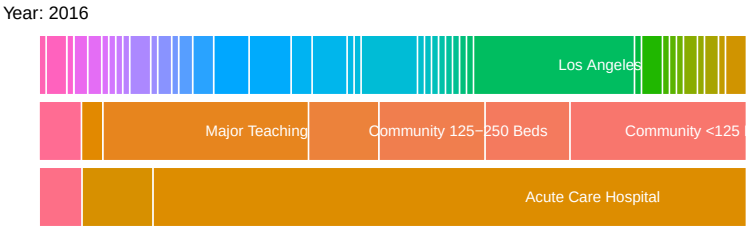


Figure 25: Distribution of the categorical variables

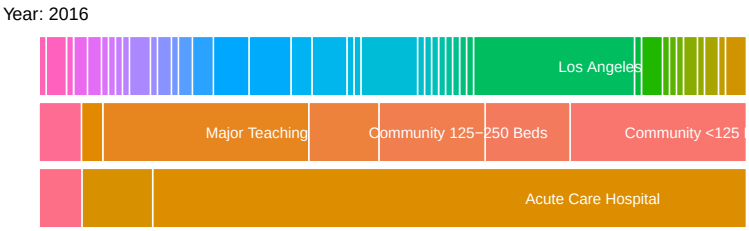


Figure 26: Distribution of the categorical variables

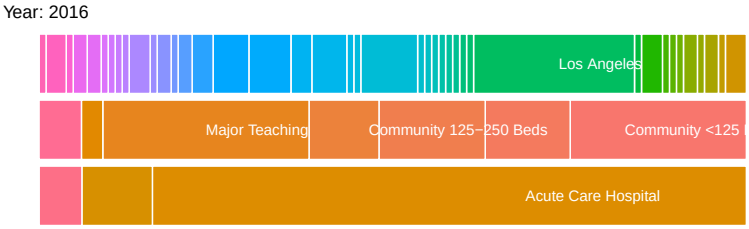


Figure 27: Distribution of the categorical variables

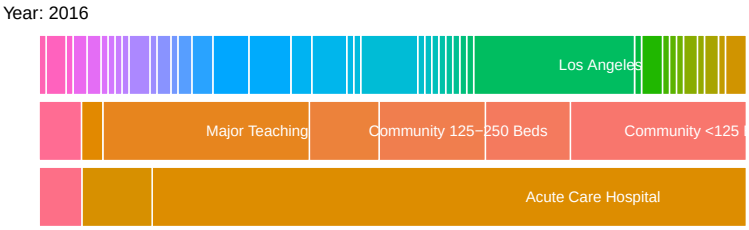


Figure 28: Distribution of the categorical variables

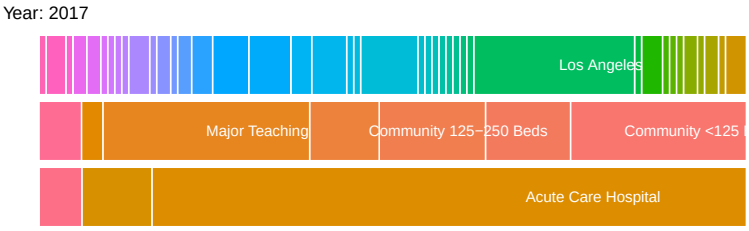


Figure 29: Distribution of the categorical variables

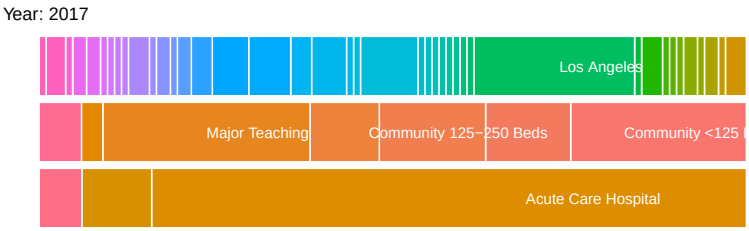


Figure 30: Distribution of the categorical variables

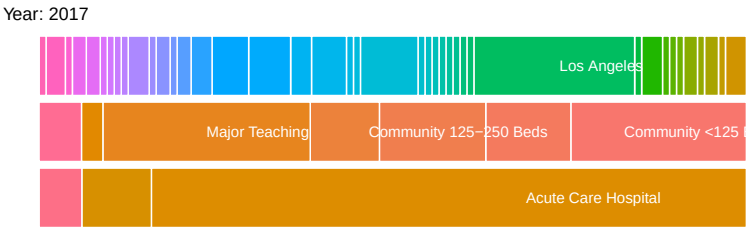


Figure 31: Distribution of the categorical variables

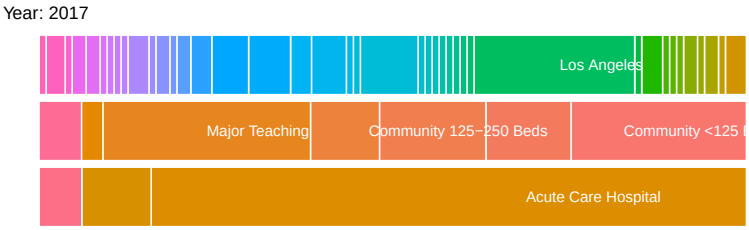


Figure 32: Distribution of the categorical variables

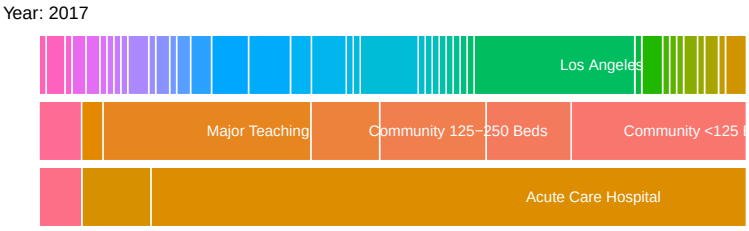


Figure 33: Distribution of the categorical variables

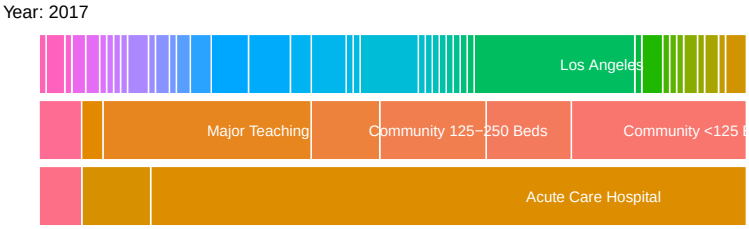


Figure 34: Distribution of the categorical variables

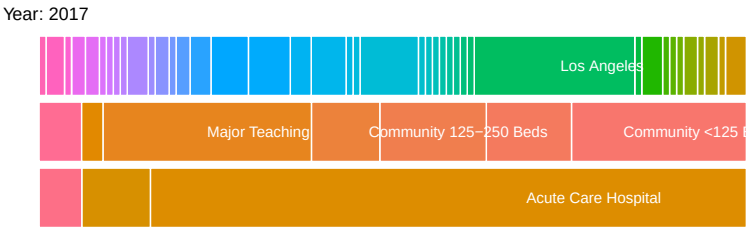


Figure 35: Distribution of the categorical variables

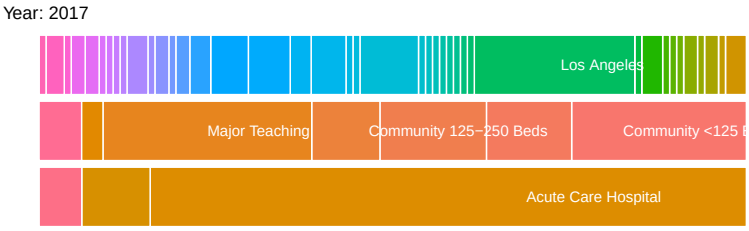


Figure 36: Distribution of the categorical variables

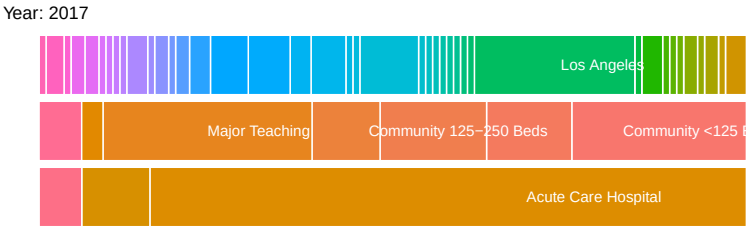


Figure 37: Distribution of the categorical variables

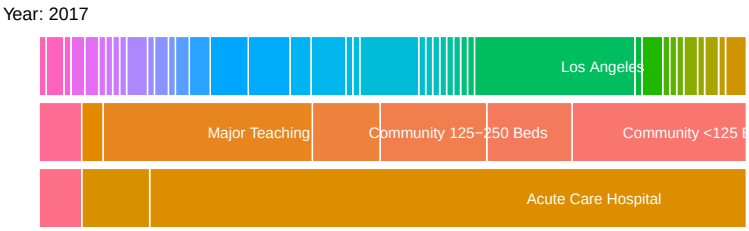


Figure 38: Distribution of the categorical variables

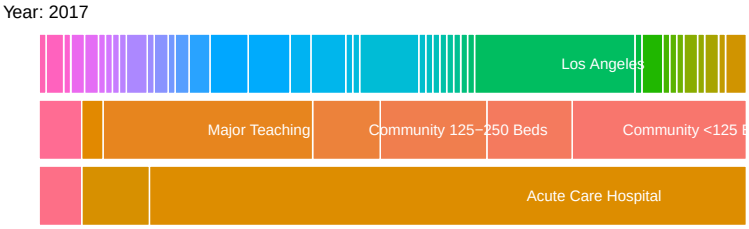


Figure 39: Distribution of the categorical variables

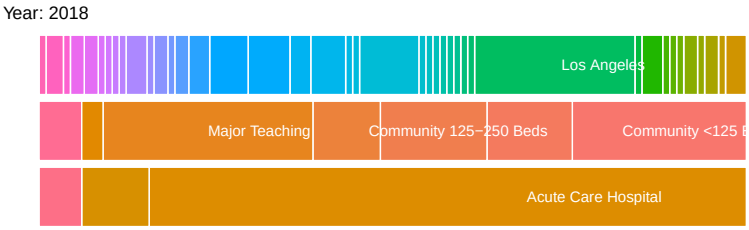


Figure 40: Distribution of the categorical variables

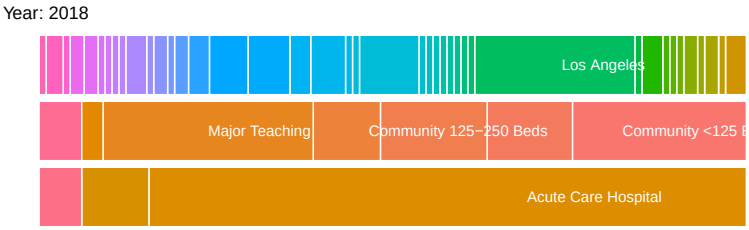


Figure 41: Distribution of the categorical variables

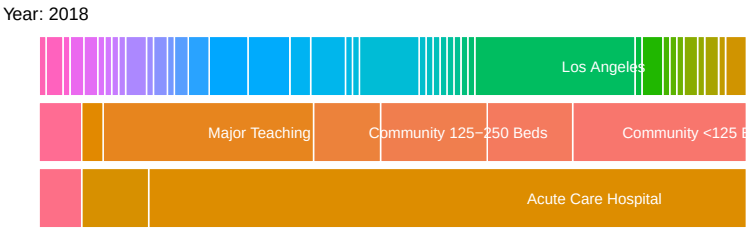


Figure 42: Distribution of the categorical variables

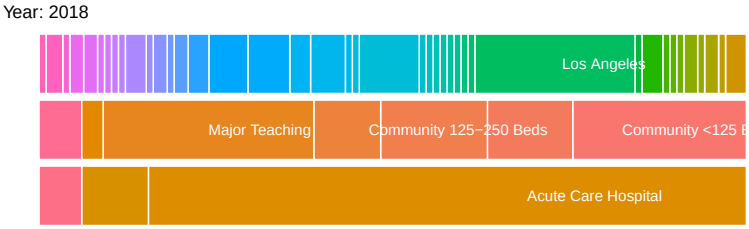


Figure 43: Distribution of the categorical variables

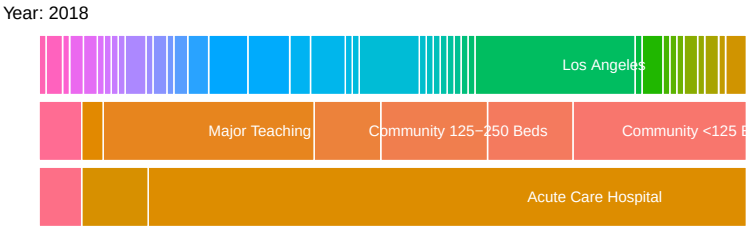


Figure 44: Distribution of the categorical variables

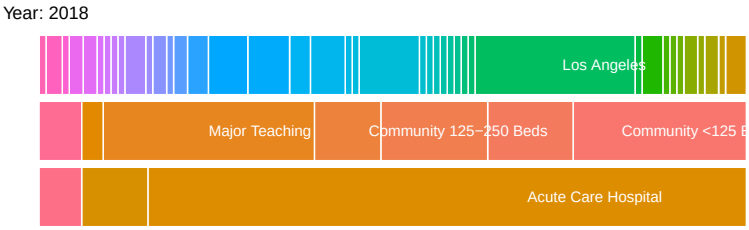


Figure 45: Distribution of the categorical variables

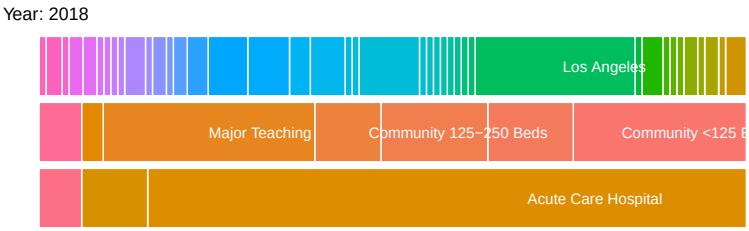


Figure 46: Distribution of the categorical variables

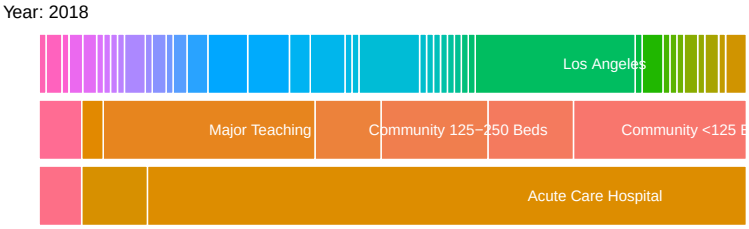


Figure 47: Distribution of the categorical variables

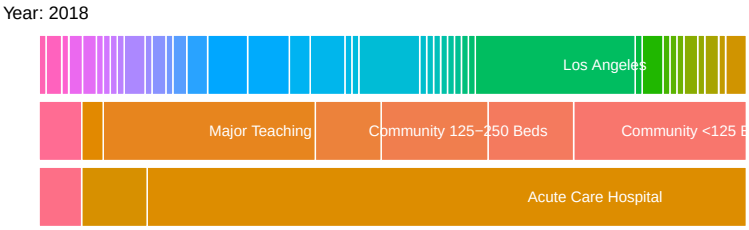


Figure 48: Distribution of the categorical variables

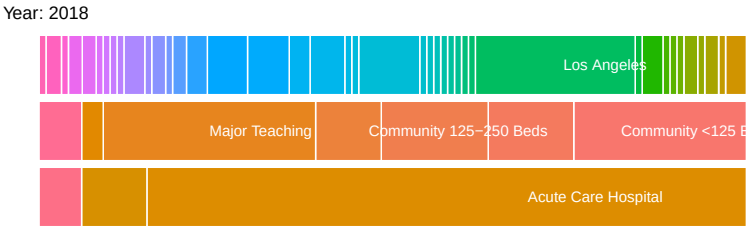


Figure 49: Distribution of the categorical variables

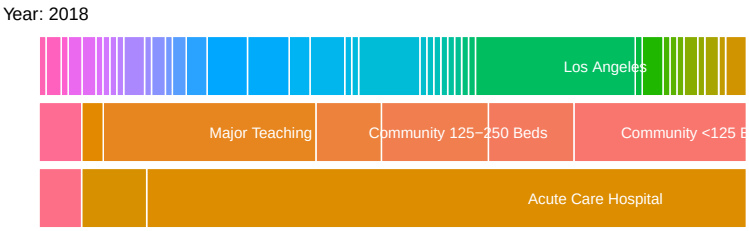


Figure 50: Distribution of the categorical variables

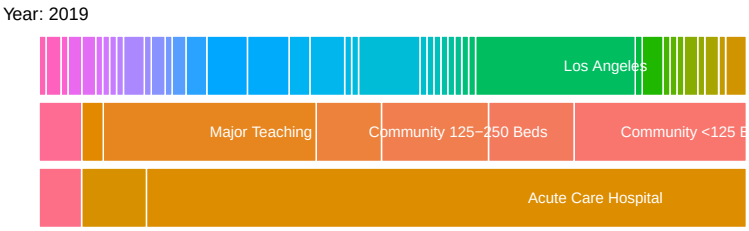


Figure 51: Distribution of the categorical variables

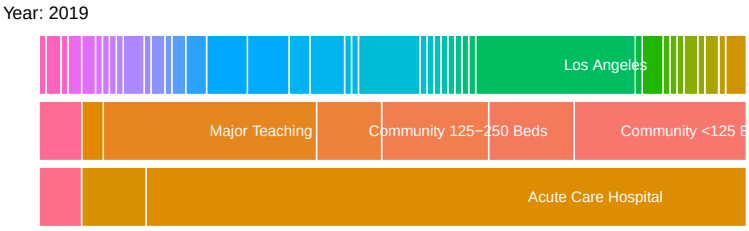


Figure 52: Distribution of the categorical variables

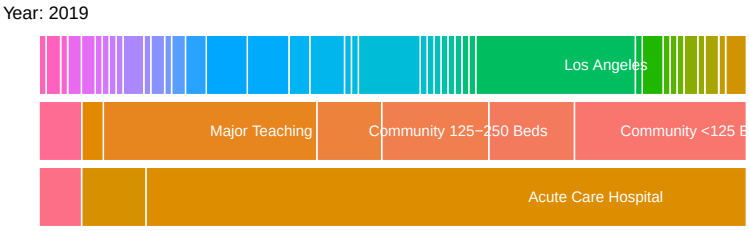


Figure 53: Distribution of the categorical variables

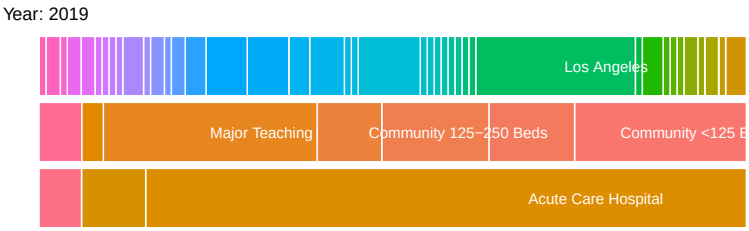


Figure 54: Distribution of the categorical variables

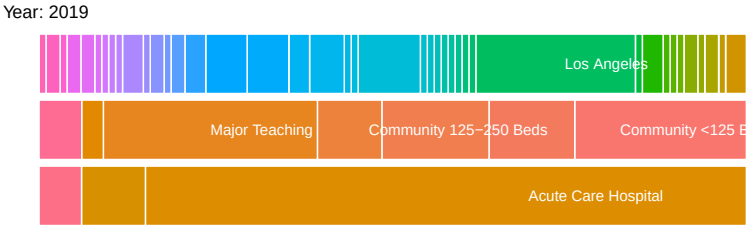


Figure 55: Distribution of the categorical variables

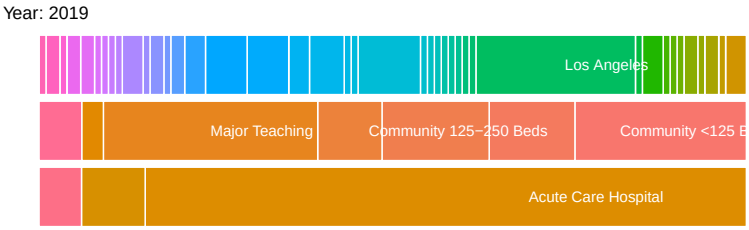


Figure 56: Distribution of the categorical variables

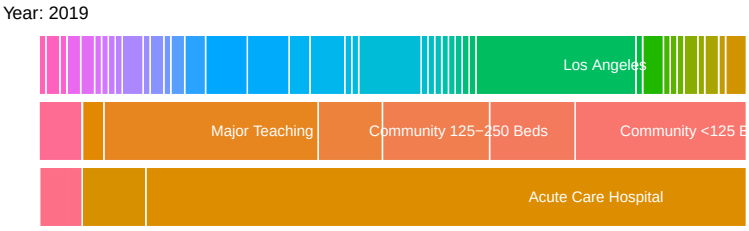


Figure 57: Distribution of the categorical variables

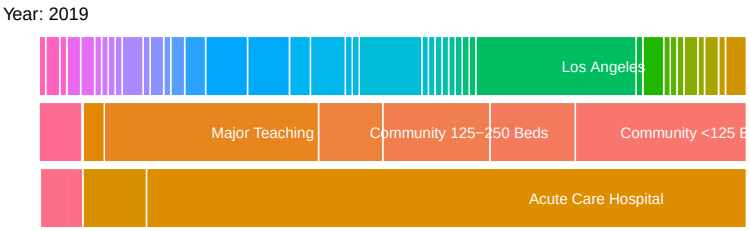


Figure 58: Distribution of the categorical variables

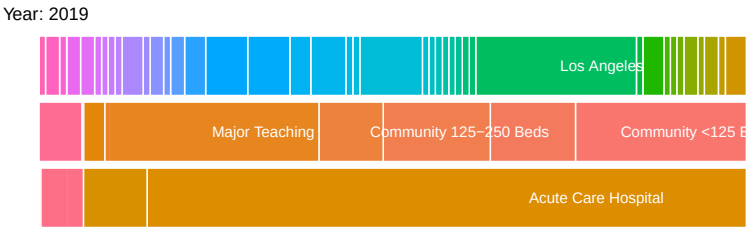


Figure 59: Distribution of the categorical variables

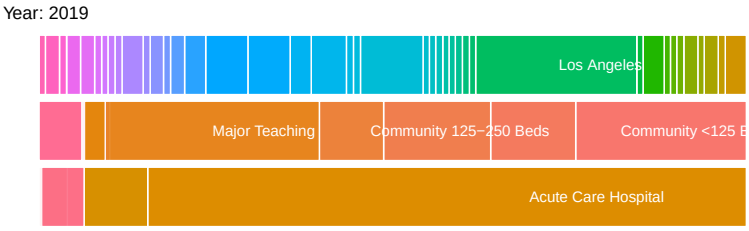


Figure 60: Distribution of the categorical variables

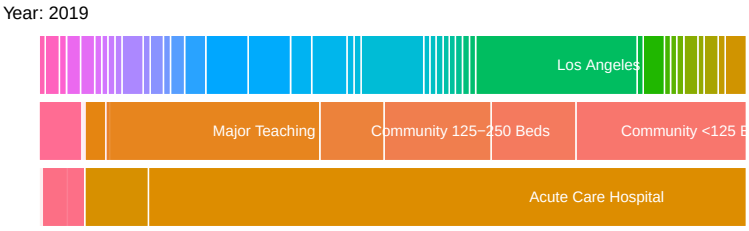


Figure 61: Distribution of the categorical variables

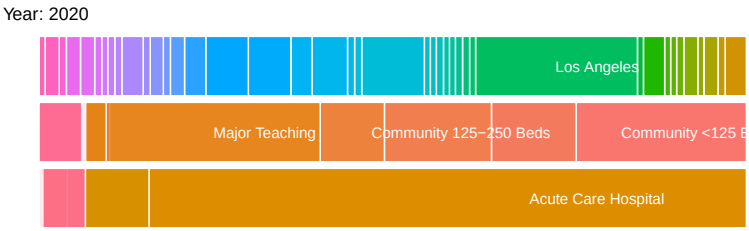


Figure 62: Distribution of the categorical variables

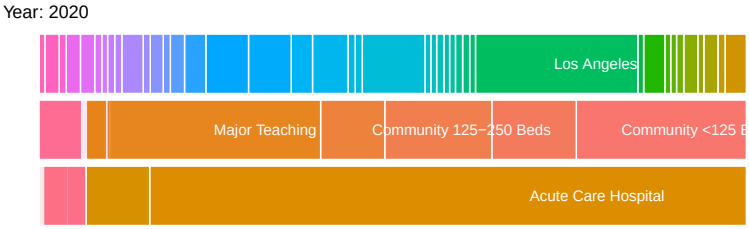


Figure 63: Distribution of the categorical variables

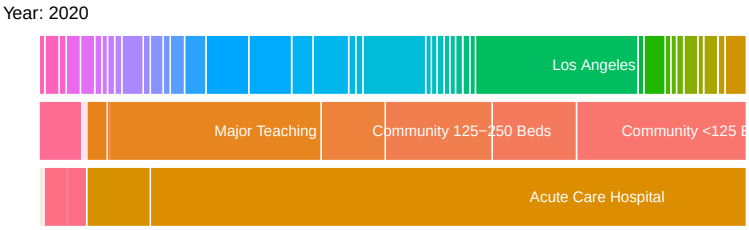


Figure 64: Distribution of the categorical variables

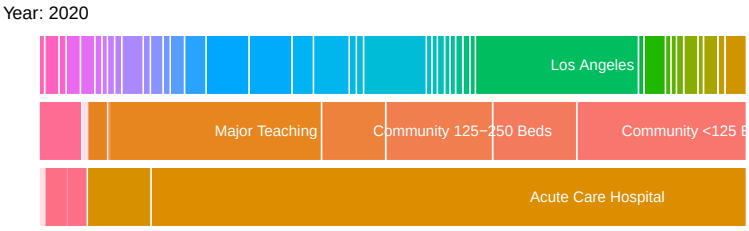


Figure 65: Distribution of the categorical variables

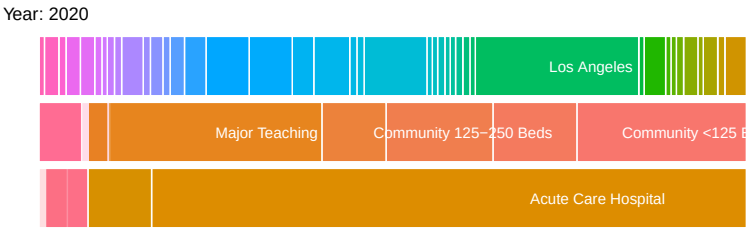


Figure 66: Distribution of the categorical variables

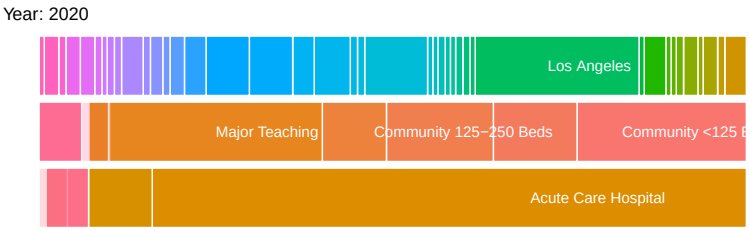


Figure 67: Distribution of the categorical variables

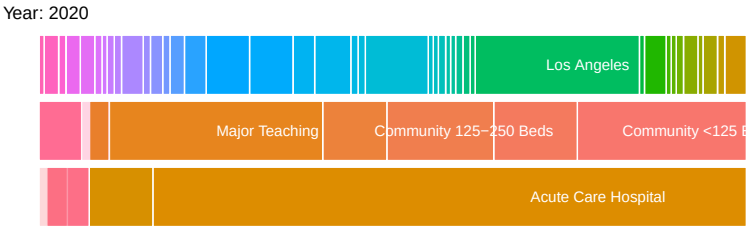


Figure 68: Distribution of the categorical variables

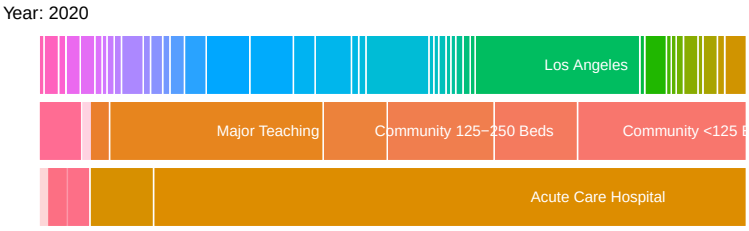


Figure 69: Distribution of the categorical variables

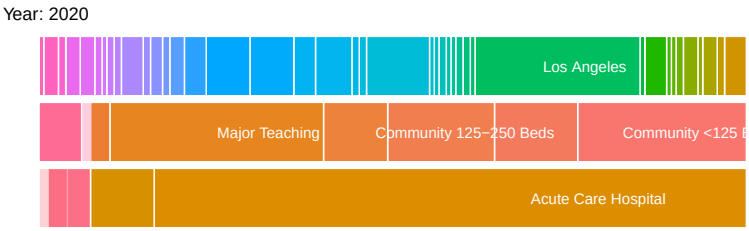


Figure 70: Distribution of the categorical variables

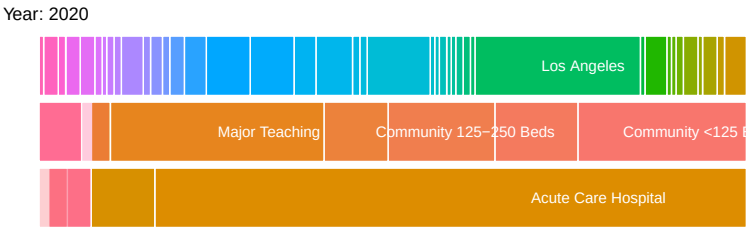


Figure 71: Distribution of the categorical variables

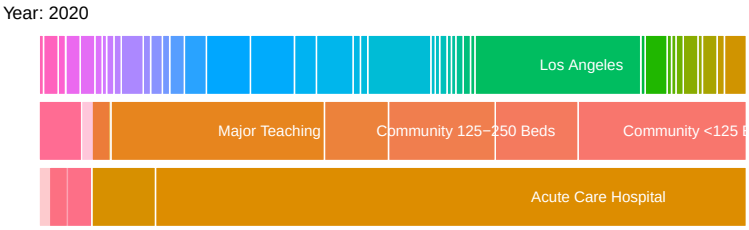


Figure 72: Distribution of the categorical variables

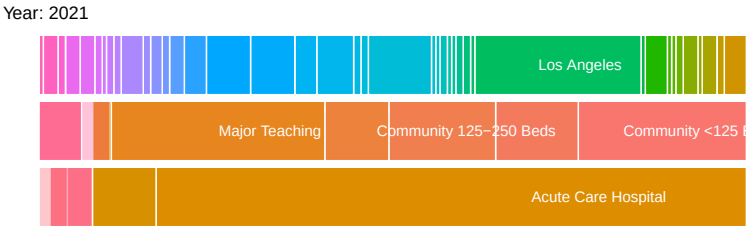


Figure 73: Distribution of the categorical variables

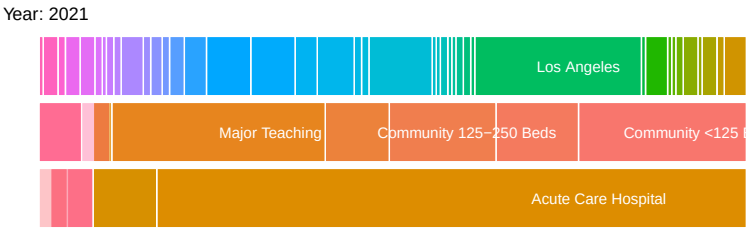


Figure 74: Distribution of the categorical variables

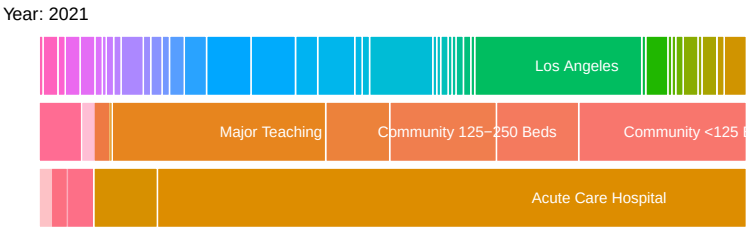


Figure 75: Distribution of the categorical variables

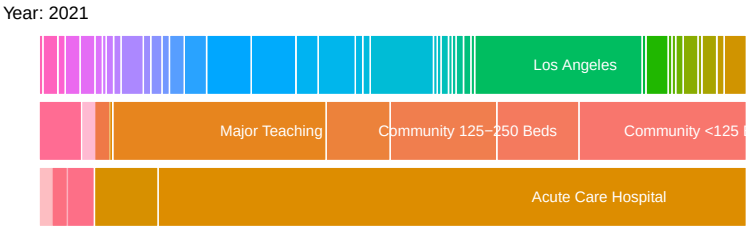


Figure 76: Distribution of the categorical variables

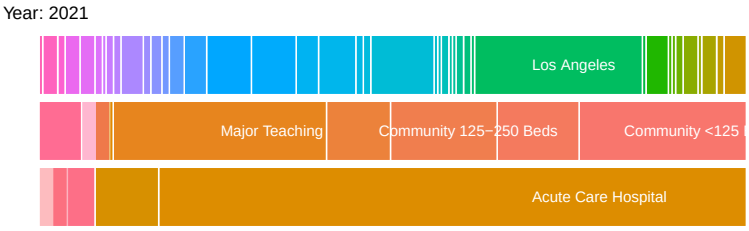


Figure 77: Distribution of the categorical variables

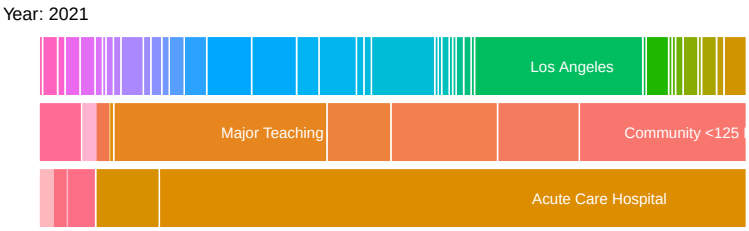


Figure 78: Distribution of the categorical variables

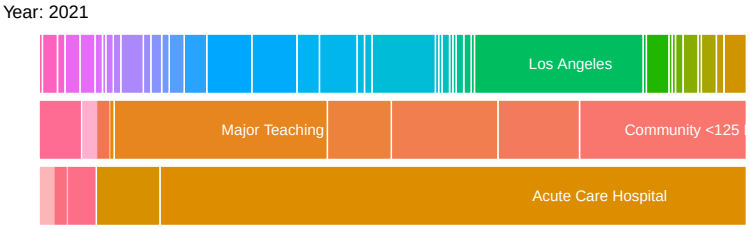


Figure 79: Distribution of the categorical variables

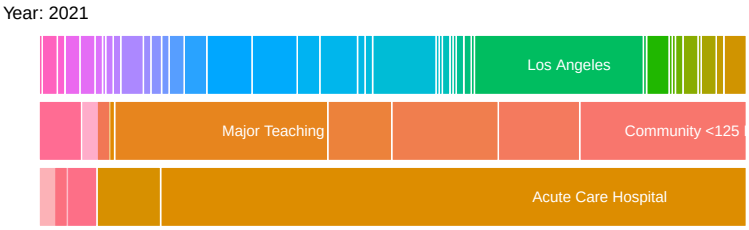


Figure 80: Distribution of the categorical variables

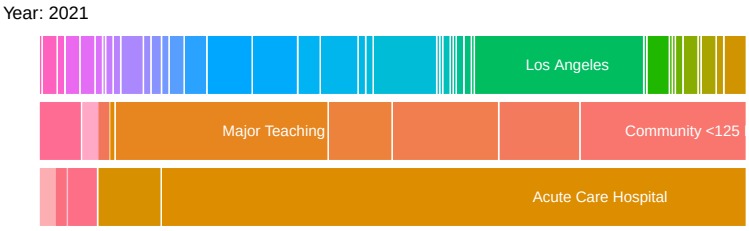


Figure 81: Distribution of the categorical variables

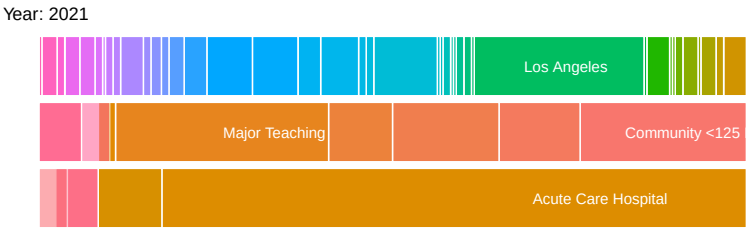


Figure 82: Distribution of the categorical variables

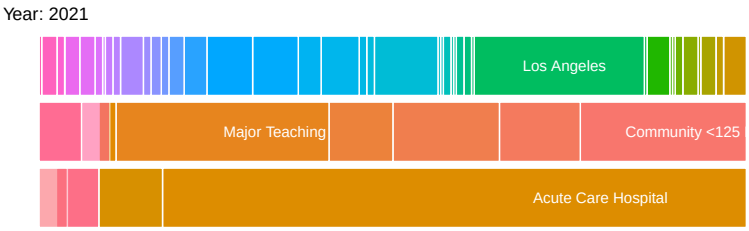


Figure 83: Distribution of the categorical variables

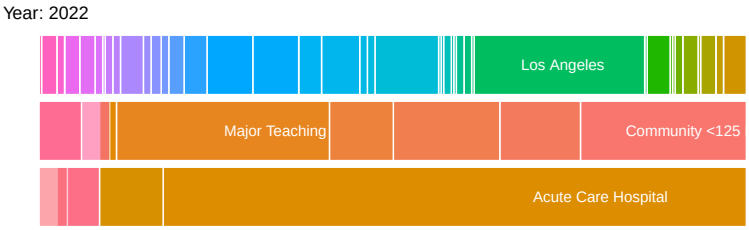


Figure 84: Distribution of the categorical variables

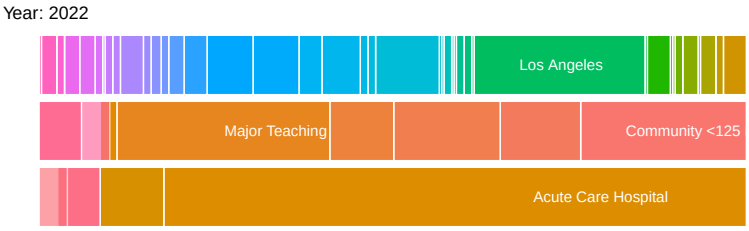


Figure 85: Distribution of the categorical variables

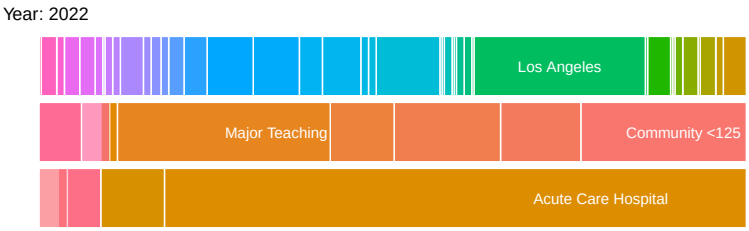


Figure 86: Distribution of the categorical variables

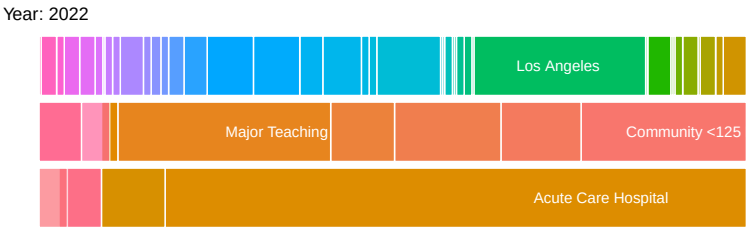


Figure 87: Distribution of the categorical variables

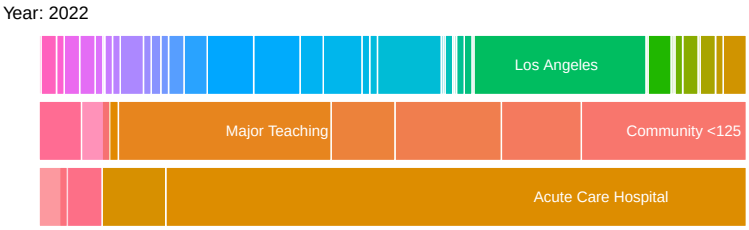


Figure 88: Distribution of the categorical variables

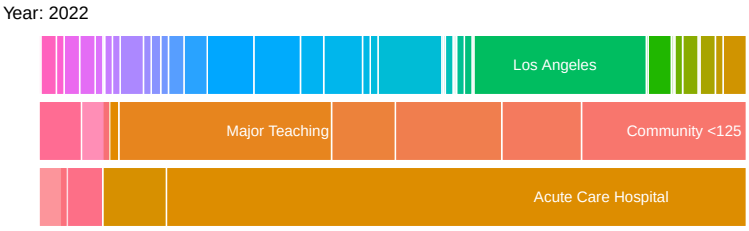


Figure 89: Distribution of the categorical variables

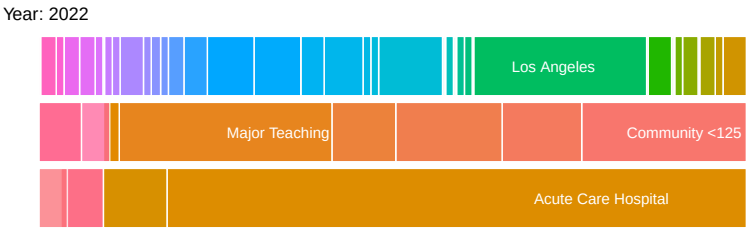


Figure 90: Distribution of the categorical variables

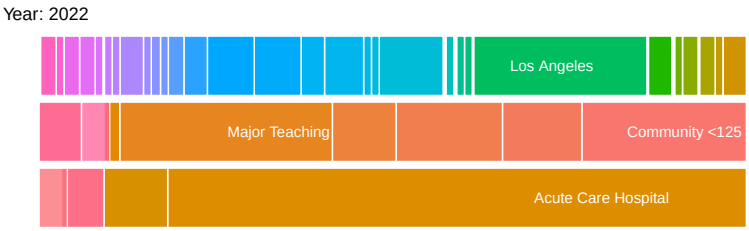


Figure 91: Distribution of the categorical variables

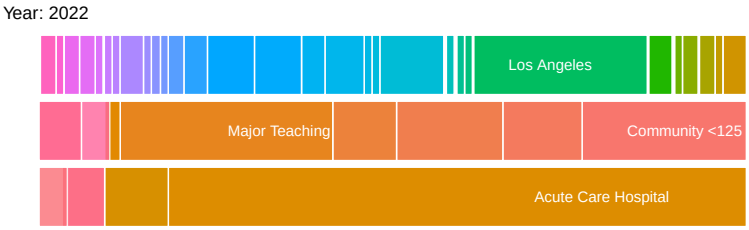


Figure 92: Distribution of the categorical variables

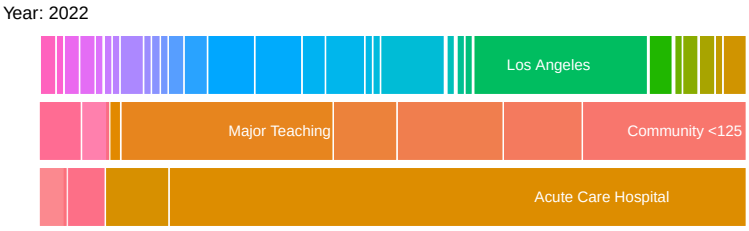


Figure 93: Distribution of the categorical variables

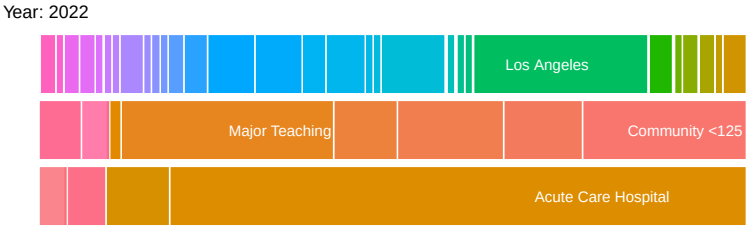


Figure 94: Distribution of the categorical variables

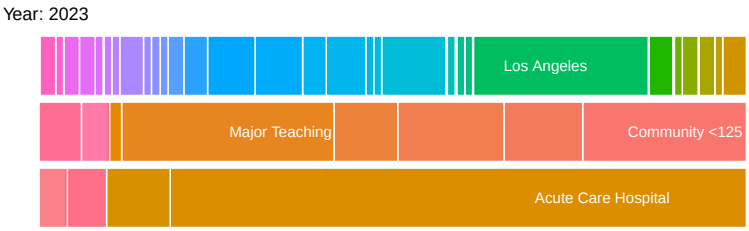


Figure 95: Distribution of the categorical variables

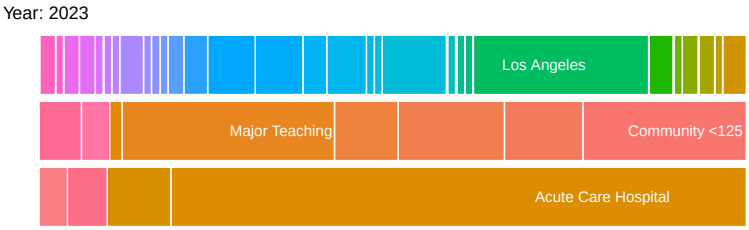


Figure 96: Distribution of the categorical variables

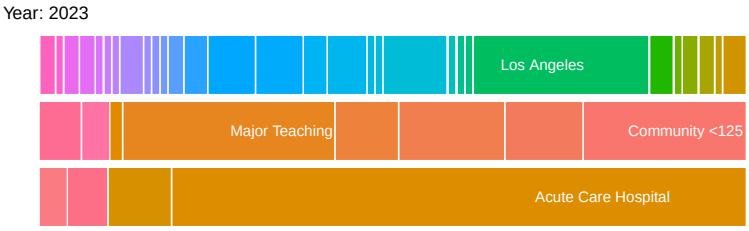


Figure 97: Distribution of the categorical variables

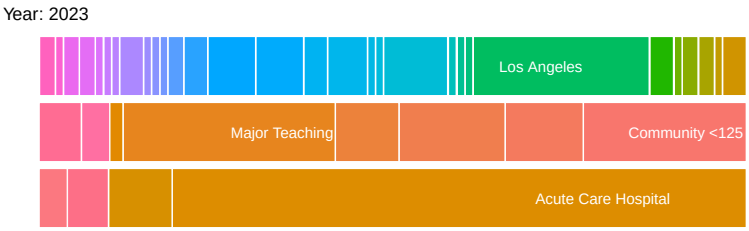


Figure 98: Distribution of the categorical variables

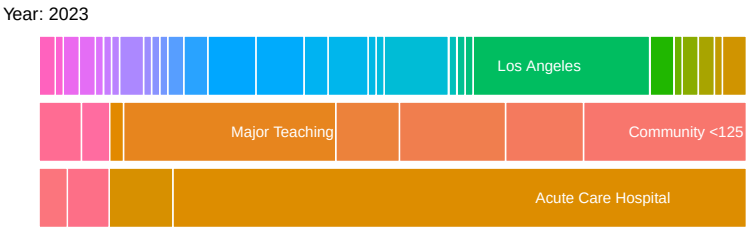


Figure 99: Distribution of the categorical variables

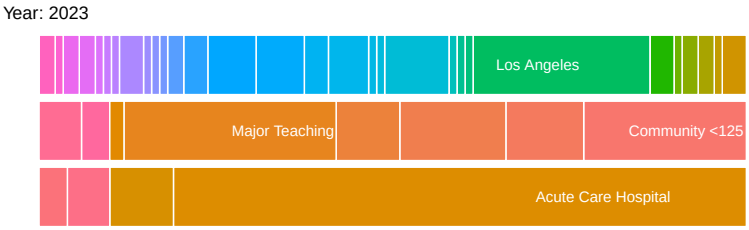


Figure 100: Distribution of the categorical variables

Distribution of SIR values with respect to Hospital_Type, 2014-2023

We can see that the average SIR is about the same in all hospital types except the community hospital <125*. The SIR value in the community hospital <125 ranges from 0 to more than 15, it poses the biggest SIR variation among the hospital types. In addition, We cab use ANOVA analysis to check if the mean SIR was truly the same for the rest of hospital types.

*Critical Access hospital/unit and free-Standing Rehabilitation are not included in the plot due to incomplete SIR data.

```
```{r}
#| label: SIR distribution
#| fig-cap: " Distribution of SIR values with respect to Hospital_Type"
#| code-fold: true
#| fig-width: 8
#| fig-height: 10
#| warning: false
```
```

```
color_palette <- c("#167bb2","#aa85d6","#7A6C5D","#56c552","#d88bb4","#dd847a","#16b29b","red"

mrsa_combine|>
   filter(
          Hospital_Type != "Free-Standing Rehabilitation",
          Hospital_Type != "Critical Access")|>
          mutate(Hospital_Type = case_when(Hospital_Type == "Community <125 Beds"~ "Comm. <125"
                                           Hospital_Type == "Community 125-250 Beds"~ "Comm. 125-250
                                           Hospital_Type == "Community >250 Beds"~ "Comm.>250",
                                           Hospital_Type == "Long-Term Acute Care"~ "Long-Acute",
                                           .default = Hospital_Type))|>

    ggplot(aes(x= Hospital_Type, y= SIR, fill=Hospital_Type))+
     geom_boxplot()+
    scale_fill_manual(values=color_palette)+
       theme(
      legend.position="none",
      plot.title = element_text(size=11)
    ) +
    ggtitle(paste0("Standardized Infection Ratio (SIR) in each hospital type ")) +
    xlab("")+ylab("")
```

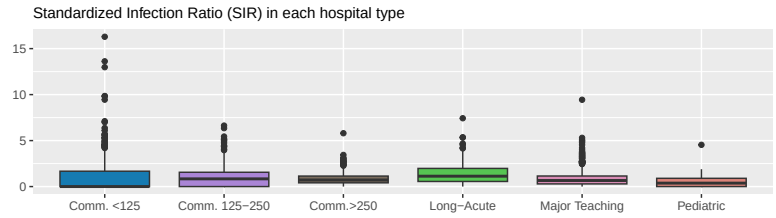


Figure 101: Distribution of SIR values with respect to Hospital_Type

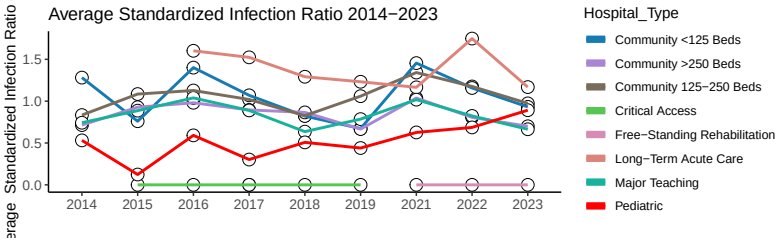
Next, we will be analyzing Hospital_Type, County with respect to a numerical SIR. The SIR ratio was gradually increased in the Pediatric hospitals, unlike in the Long-Term Acute Care hospital, the ratio was slowly decreased except in 2022. In 2021, Community <125 Beds, Community 125-250 Beds, “Community >250 Beds, and Major Teaching, had the highest SIR ra-

tio.

```r
#| title: SIR trend
#| code-fold: true
#| warning: false
#| fig-width: 8
#| fig-height: 8
#| fig-column: page-left

mrsa_combine|>
  group_by(Year, Hospital_Type)|>
  summarise(avg_SIR = mean(na.omit(SIR)))|>
  mutate(Risk_Category = Hospital_Type)|>
  filter(Risk_Category != "NA",
         Year != 2013)|>
  ggplot(aes(x= as.character(Year), y= avg_SIR, color=Hospital_Type))+

    geom_point(shape=21, color="black", fill="white", size=4) +
   # geom_point(size = 4, color = "#0e263560", shape = "circle open") +
  geom_line(aes(group=Risk_Category),linewidth = 1)+
  geom_line(linewidth = 2.5) +
  theme_bw()+
      theme(#axis.text.x=element_blank(),
            axis.text=element_text(size=10),
            legend.position = "right",
            axis.line = element_line(colour = "black"),
            panel.grid.major = element_blank(),
            panel.grid.minor = element_blank(),
            panel.border = element_blank(),
            panel.background = element_blank()
            )+
  scale_color_manual(values=color_palette) +
    labs(y="Average  Standardized Infection Ratio (SIR)", x="",
       title="Average Standardized Infection Ratio 2014-2023")
```



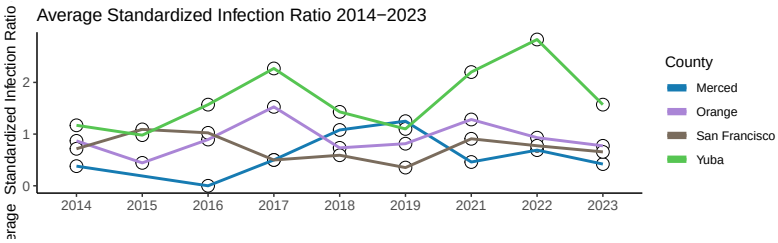
The trends of SIR ratio look similar in the more densely populated counties after 2020.

```{r}
#| title: SIR trend
#| code-fold: true
#| warning: false
#| fig-width: 8
#| fig-height: 8

mrsa_combine|>
  filter(County == "Orange" | County == "San Francisco"| County=="Merced"| County=="Yuba")|>
  group_by(Year, County)|>
  summarise(avg_SIR = mean(na.omit(SIR)))|>
  ggplot(aes(x= as.character(Year), y= avg_SIR, color=County))+

    geom_point(shape=21, color="black", fill="white", size=4) +
   # geom_point(size = 4, color = "#0e263560", shape = "circle open") +
  geom_line(aes(group=County),linewidth = 1)+
  geom_line(linewidth = 2.5) +
  theme_bw()+
      theme(#axis.text.x=element_blank(),
            axis.text=element_text(size=10),
            legend.position = "right",
            axis.line = element_line(colour = "black"),
            panel.grid.major = element_blank(),
            panel.grid.minor = element_blank(),
            panel.border = element_blank(),
            panel.background = element_blank()
            )+
    scale_color_manual(values=color_palette) +
    labs(y="Average  Standardized Infection Ratio (SIR)", x="",
       title="Average Standardized Infection Ratio 2014-2023")
```

```
```



Based on the scatter plots in the previous post, it seems there is a linear pattern between Patient_Days and Infections_Reported variables. We can further compare results between variables using correlation analysis.

```{r}
#| label: Correlation
#| fig-cap: " Correlation of variables"
#| code-fold: true
#| warning: false
#| fig-width: 6
#| fig-height: 6

# spearman correlation

dt <- mrsa_combine|>
      mutate(Hospital_Type= ifelse(is.na(Hospital_Type), "mixed", Hospital_Type),
            RiskAdjustment = ifelse(is.na(Hospital_Category_RiskAdjustment), "mixed",
                                                    Hospital_Category_RiskAdjustment))|>
      mutate(Hospital_Type= as.numeric(as.factor(Hospital_Type)),
             RiskAdjustment = as.numeric(as.factor(RiskAdjustment)),
             County = as.numeric(as.factor(County)))|>
      select(SIR, RiskAdjustment, County,Patient_Days,Infections_Reported)

cor_ <- round(cor(na.omit(dt), method = 'spearman'),2)
cor_matrix <- melt(cor_)
colnames(cor_matrix ) <- c("X1","X2","value")
ggplot(cor_matrix, aes(X1, X2, fill= value)) +
  geom_tile() +
  scale_fill_gradient(low="lightgrey", high="#dd847a")+
```

```
  xlab("")+ylab("")
```

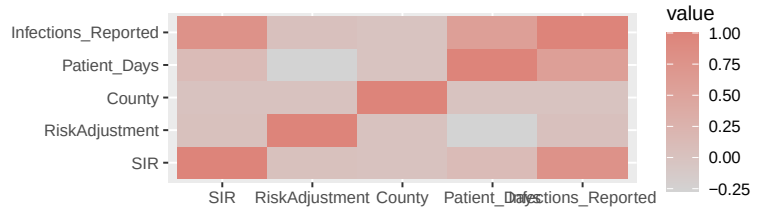


Figure 102: Correlation of variables

The value of the correlation coefficient always ranges between 1 and -1, and we will treat it as a general indicator of the strength of the relationship between variables. Since the data is not normally distributed, we will use Spearman's rho (Spearman's rank correlation coefficient). It's a rank correlation coefficient because it uses the rankings of data from each variable (e.g., from lowest to highest) rather than the raw data itself.

We can see that the SIR has a positive strong correlation with Infections_Reported (correlation coefficients value 0.78), and it has a much weaker correlation with Patient_Days(correlation coefficients value 0.12).The results also show a positive strong correlation between Patient_Days and Infections_Reported (correlation coefficients value 0.58). Weak Correlation is found between Hospital Risk Adjustment and Patient_Days(0.27). No correlations is found between Hospital Risk Adjustment and Infections_Reported(0.04), and between Hospital Risk Adjustment and SIR(0.03). County seems to have no correlation with other variables.

Base forecast model (Xgboost)

```
```{r}
#| title: Forecast base model
#| code-fold: true
#| warning: false
```
```

```
#| fig-width: 8
#| fig-height: 10

# Step 1: Data Preparation
dt_all <- mrsa_combine|>
    mutate(Hospital_Type  = case_when(Hospital_Type == "Community, <125 Beds"~ 1,
                                      Hospital_Type == "Community, 125-250 Beds"~ 2,
                                      Hospital_Type == "Community, >250 Beds"~ 3,
                                      Hospital_Type == "Major Teaching"~ 4,
                                      .default = 0),
          Hospital_Type = as.factor(Hospital_Type),
          Facility_ID = as.factor((Facility_ID)))|>
    select(-c(Hospital_Category_RiskAdjustment,County))

# Step 2: Fill missing values
# Use KNN to fill missing values
dt_no2023 <- kNN(dt_all|>filter(Year != 2023), variable = "SIR", k = 10)
dt_all <-     rbind(dt_no2023, dt_all|> filter(Year == 2023)|> mutate(SIR_imp = FALSE))|>
                group_by(Facility_ID)|>
                arrange(Facility_ID, (as.numeric(Year)))|>
                mutate(lag_IR = dplyr::lag(Infections_Reported, n = 1,default = NA))|>
                mutate(lag_PD = dplyr::lag(Patient_Days, n = 1, default = NA))|>
                mutate(lag_SIR = dplyr::lag(SIR, n = 1, default = NA),
                        impute_SIR = as.factor(ifelse(SIR_imp == TRUE , 1, 0)))

dt <- na.omit(dt_all)|>select(-c(Infections_Reported,Patient_Days,SIR_imp))

# Step 3: Data Splittingv
# Split the data into training and test sets
train_data <- dt|> filter(Year != 2023)
test_data <- dt|> filter(Year == 2023)

# Step 4: Prepare data for XGBoost
train_labels <- train_data$SIR
test_labels <- test_data$SIR

# Convert factors to numeric for xgboost
dtrain <- model.matrix(~. -1, data = train_data |> select(-Year, -SIR))
dtrain <- xgb.DMatrix(data = dtrain, label = train_labels)
dtest <- model.matrix(~. -1, data = test_data |> select(-Year, -SIR))
```

```
dtest <- xgb.DMatrix(data = dtest, label = test_labels)

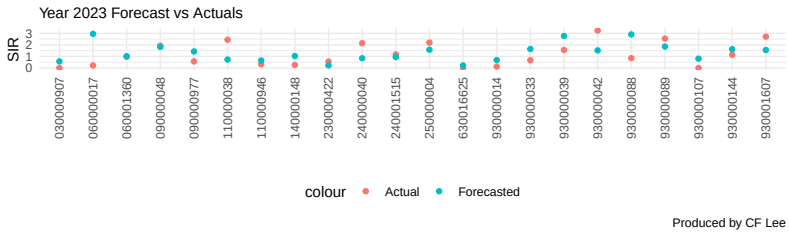
# Step 5: Train XGBoost Model
params <- list(objective = "reg:squarederror", eval_metric = "rmse",booster= "gblinear")
xgb_model <- xgb.train(params = params, data = dtrain, nrounds = 300,  verbose = 0, watchlist =


# Step 6: Forecast and Evaluate
preds <- predict(xgb_model, dtest)

# Evaluate model performance using RMSE
rmse <- sqrt(mean((test_labels - preds)^2))
cat("RMSE:", rmse, "\n")

# Step 7: Plot the Results
test_data$predicted_SIR <- preds
test_data$diff_act_forecast <- abs(test_data$predicted_SIR-test_data$SIR)
test_data <- full_join(test_data|>select(-Hospital_Type), unique(map_risk_list),by= "Facility_I
na.omit(test_data) |>
   filter(Hospital_Type == "Long-Term Acute Care")|>
   #filter(Hospital_Type==4, County == "Los Angeles" )|>
   ggplot(aes(x= Facility_ID))+
   geom_point(aes(y = SIR, color = "Actual"), fill = NA) +
   geom_point(aes(y = predicted_SIR, color = "Forecasted"), fill = NA) +
  theme_minimal()+
  theme(
      legend.position="bottom",
      plot.title = element_text(size=11),
      axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)
    ) +
    labs(title = "Year 2023 Forecast vs Actuals", y = "SIR", x = "", caption ="Produced by CF I
```

```
RMSE: 1.085554
```


Year 2023 Forecast vs Actuals
SIR
3
2
1
0
030000907
060000017
060001360
090000048
090000977
110000038
110000946
140000148
230000422
240000040
240001515
250000004
630016625
930000014
930000033
930000039
930000042
930000088
930000089
930000107
930000144
930001607
colour
Actual
Forecasted
Produced by CF Lee