Covid

Monsuru Moshood

2025-03-24

```r
# Load necessary libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(tidyr)
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```r
library(tseries)

# Load the dataset
covid_data <- read.csv('/Users/moshoodlanre/Desktop/Data Analytic Project/New Covid data result/Covid Da

# View the first few rows of the dataset
head(covid_data)
```

```
##   USMER MEDICAL_UNIT SEX PATIENT_TYPE  DATE_DIED INTUBED PNEUMONIA AGE PREGNANT
## 1     2            1   1            1 03/05/2020      97         1  65        2
## 2     2            1   2            1 03/06/2020      97         1  72       97
## 3     2            1   2            2 09/06/2020       1         2  55       97
## 4     2            1   1            1 12/06/2020      97         2  53        2
## 5     2            1   2            1 21/06/2020      97         2  68       97
## 6     2            1   1            2 9999-99-99       2         1  40        2
##   DIABETES COPD ASTHMA INMSUPR HIPERTENSION OTHER_DISEASE CARDIOVASCULAR
## 1        2    2      2       2            1             2              2
## 2        2    2      2       2            1             2              2
## 3        1    2      2       2            2             2              2
## 4        2    2      2       2            2             2              2
## 5        1    2      2       2            1             2              2
## 6        2    2      2       2            2             2              2
##   OBESITY RENAL_CHRONIC TOBACCO CLASIFFICATION_FINAL ICU
```

```
## 1       2            2       2                   3  97
## 2       1            1       2                   5  97
## 3       2            2       2                   3   2
## 4       2            2       2                   7  97
## 5       2            2       2                   3  97
## 6       2            2       2                   3   2
```

```
# Check for missing values
sum(is.na(covid_data))
```

```
## [1] 0
```

```
# Remove rows with missing values
covid_clean <- covid_data %>% drop_na()

# Check the structure of the cleaned data
str(covid_clean)
```

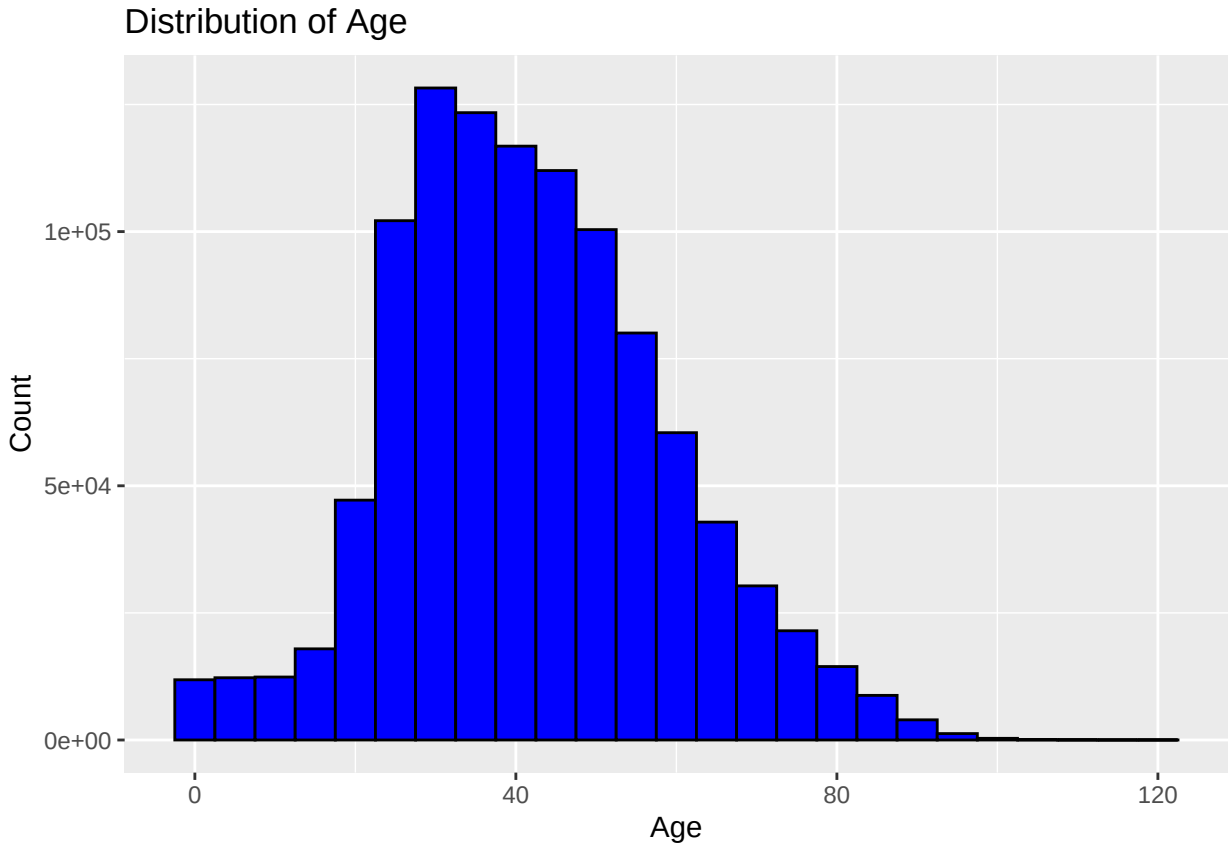
```
## 'data.frame':    1048575 obs. of  21 variables:
##  $ USMER               : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ MEDICAL_UNIT        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ SEX                 : int  1 2 2 1 2 1 1 1 1 1 ...
##  $ PATIENT_TYPE        : int  1 1 2 1 1 2 1 1 2 2 ...
##  $ DATE_DIED           : chr  "03/05/2020" "03/06/2020" "09/06/2020" "12/06/2020" ...
##  $ INTUBED             : int  97 97 1 97 97 2 97 97 2 2 ...
##  $ PNEUMONIA           : int  1 1 2 2 2 1 2 1 2 2 ...
##  $ AGE                 : int  65 72 55 53 68 40 64 64 37 25 ...
##  $ PREGNANT            : int  2 97 97 2 97 2 2 2 2 2 ...
##  $ DIABETES            : int  2 2 1 2 1 2 2 1 1 2 ...
##  $ COPD                : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ ASTHMA              : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ INMSUPR             : int  2 2 2 2 2 2 2 1 2 2 ...
##  $ HIPERTENSION        : int  1 1 2 2 1 2 2 1 1 2 ...
##  $ OTHER_DISEASE       : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ CARDIOVASCULAR      : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ OBESITY             : int  2 1 2 2 2 2 2 2 1 2 ...
##  $ RENAL_CHRONIC       : int  2 1 2 2 2 2 2 1 2 2 ...
##  $ TOBACCO             : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ CLASIFFICATION_FINAL: int  3 5 3 7 3 3 3 3 3 3 ...
##  $ ICU                 : int  97 97 2 97 97 2 97 97 2 2 ...
```

```
# Summary statistics of the cleaned data
summary(covid_clean)
```

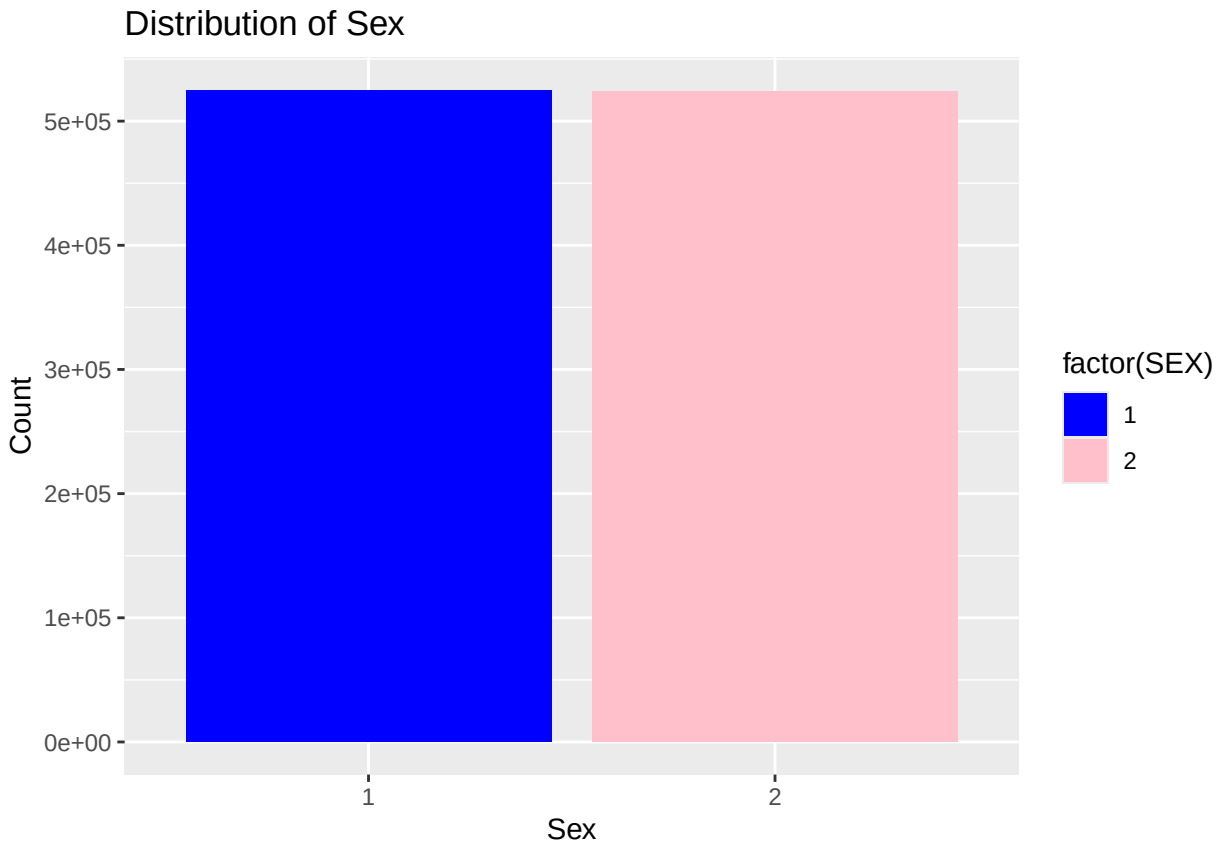
```
##      USMER        MEDICAL_UNIT          SEX          PATIENT_TYPE
##  Min.   :1.000   Min.   : 1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.: 4.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :2.000   Median :12.000   Median :1.000   Median :1.000
##  Mean   :1.632   Mean   : 8.981   Mean   :1.499   Mean   :1.191
##  3rd Qu.:2.000   3rd Qu.:12.000   3rd Qu.:2.000   3rd Qu.:1.000
##  Max.   :2.000   Max.   :13.000   Max.   :2.000   Max.   :2.000
##   DATE_DIED            INTUBED        PNEUMONIA           AGE
##  Length:1048575     Min.   : 1.00   Min.   : 1.000   Min.   :  0.00
##  Class :character   1st Qu.:97.00   1st Qu.: 2.000   1st Qu.: 30.00
##  Mode  :character   Median :97.00   Median : 2.000   Median : 40.00
##                     Mean   :79.52   Mean   : 3.347   Mean   : 41.79
```

```
##                    3rd Qu.:97.00   3rd Qu.: 2.000   3rd Qu.: 53.00
##                    Max.   :99.00   Max.   :99.000   Max.   :121.00
##     PREGNANT        DIABETES           COPD            ASTHMA
##  Min.   : 1.00   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 2.00   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000
##  Median :97.00   Median : 2.000   Median : 2.000   Median : 2.000
##  Mean   :49.77   Mean   : 2.186   Mean   : 2.261   Mean   : 2.243
##  3rd Qu.:97.00   3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000
##  Max.   :98.00   Max.   :98.000   Max.   :98.000   Max.   :98.000
##     INMSUPR        HIPERTENSION    OTHER_DISEASE    CARDIOVASCULAR
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000
##  Median : 2.000   Median : 2.000   Median : 2.000   Median : 2.000
##  Mean   : 2.298   Mean   : 2.129   Mean   : 2.435   Mean   : 2.262
##  3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000
##  Max.   :98.000   Max.   :98.000   Max.   :98.000   Max.   :98.000
##     OBESITY       RENAL_CHRONIC       TOBACCO       CLASIFFICATION_FINAL
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   :1.000
##  1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.:3.000
##  Median : 2.000   Median : 2.000   Median : 2.000   Median :6.000
##  Mean   : 2.125   Mean   : 2.257   Mean   : 2.214   Mean   :5.306
##  3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.:7.000
##  Max.   :98.000   Max.   :98.000   Max.   :98.000   Max.   :7.000
##       ICU
##  Min.   : 1.00
##  1st Qu.:97.00
##  Median :97.00
##  Mean   :79.55
##  3rd Qu.:97.00
##  Max.   :99.00
```

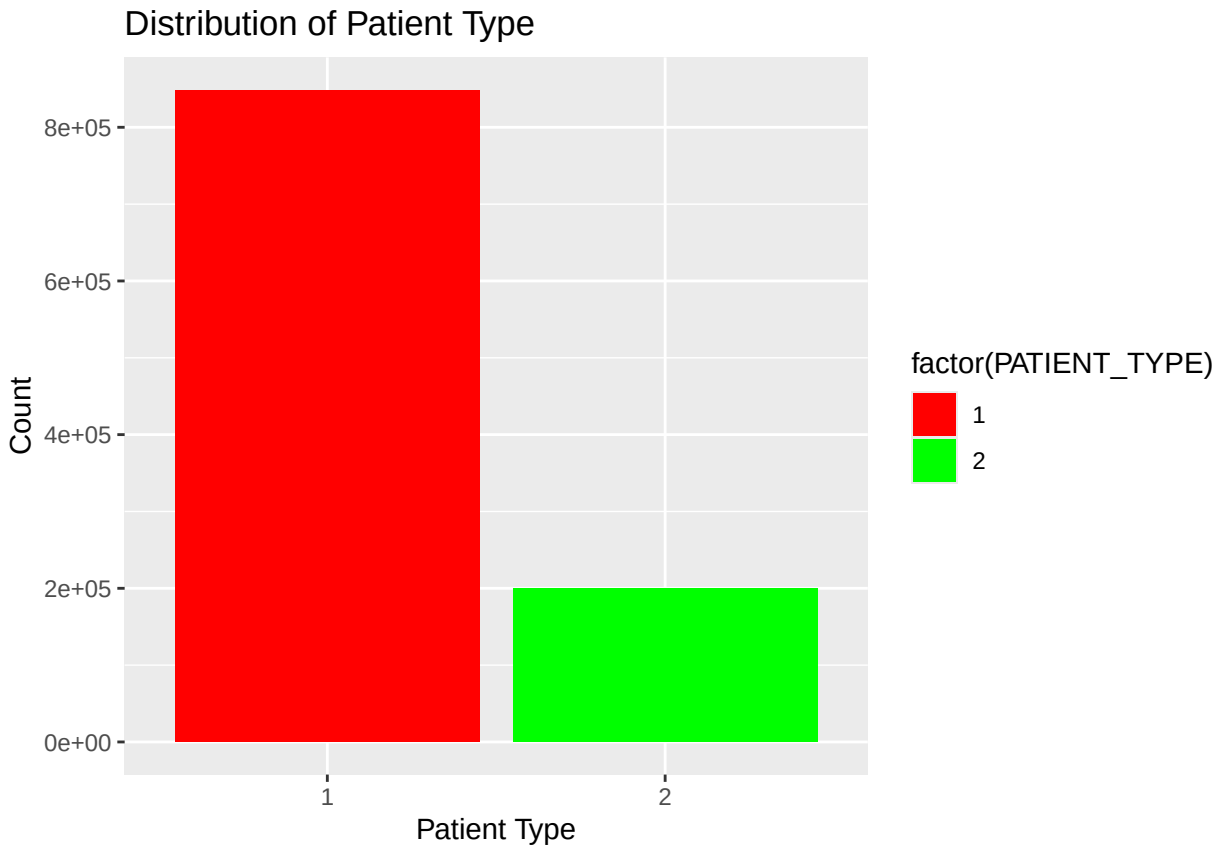
```r
# Exploratory Data Analysis (EDA)

# 1. Distribution of Age
ggplot(covid_clean, aes(x = AGE)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  labs(title = "Distribution of Age", x = "Age", y = "Count")
```



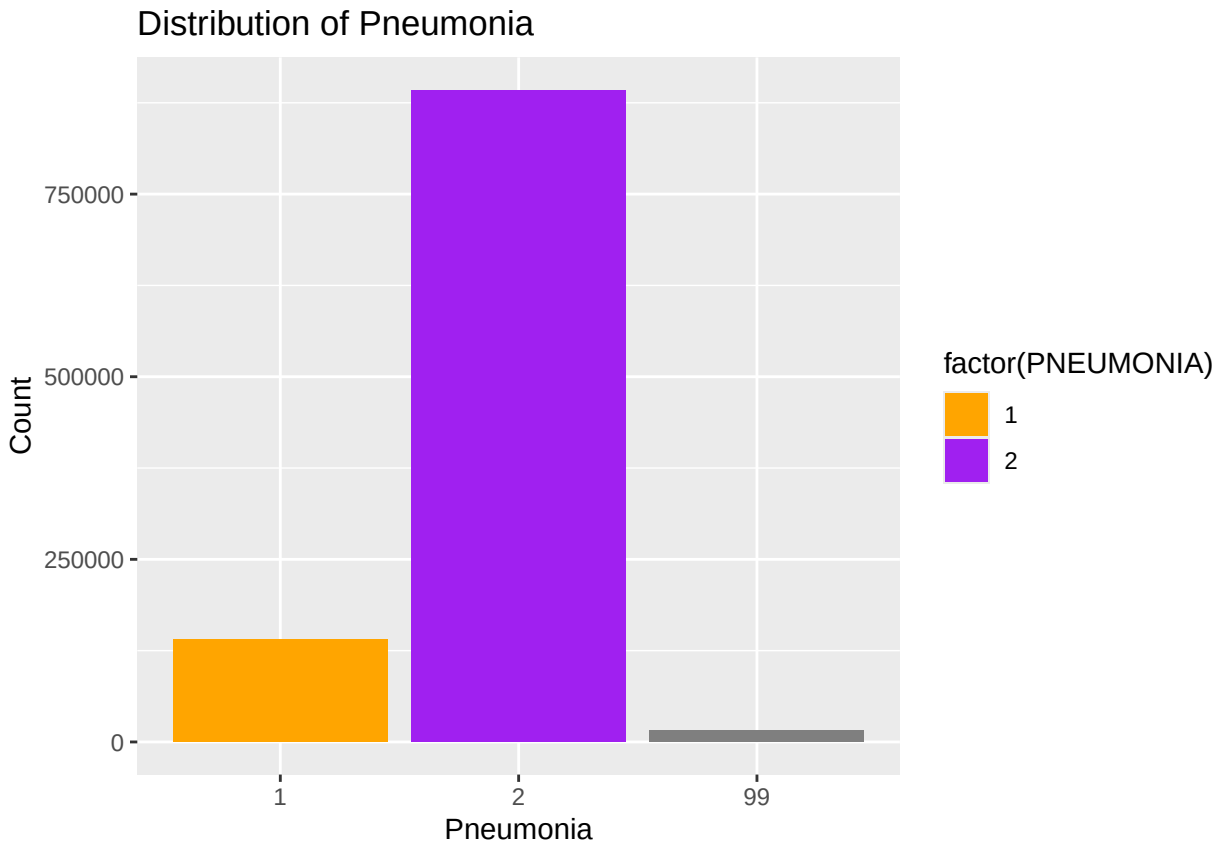

```
# 2. Distribution of Sex
ggplot(covid_clean, aes(x = factor(SEX), fill = factor(SEX))) +
  geom_bar() +
  labs(title = "Distribution of Sex", x = "Sex", y = "Count") +
  scale_fill_manual(values = c("1" = "blue", "2" = "pink"))
```



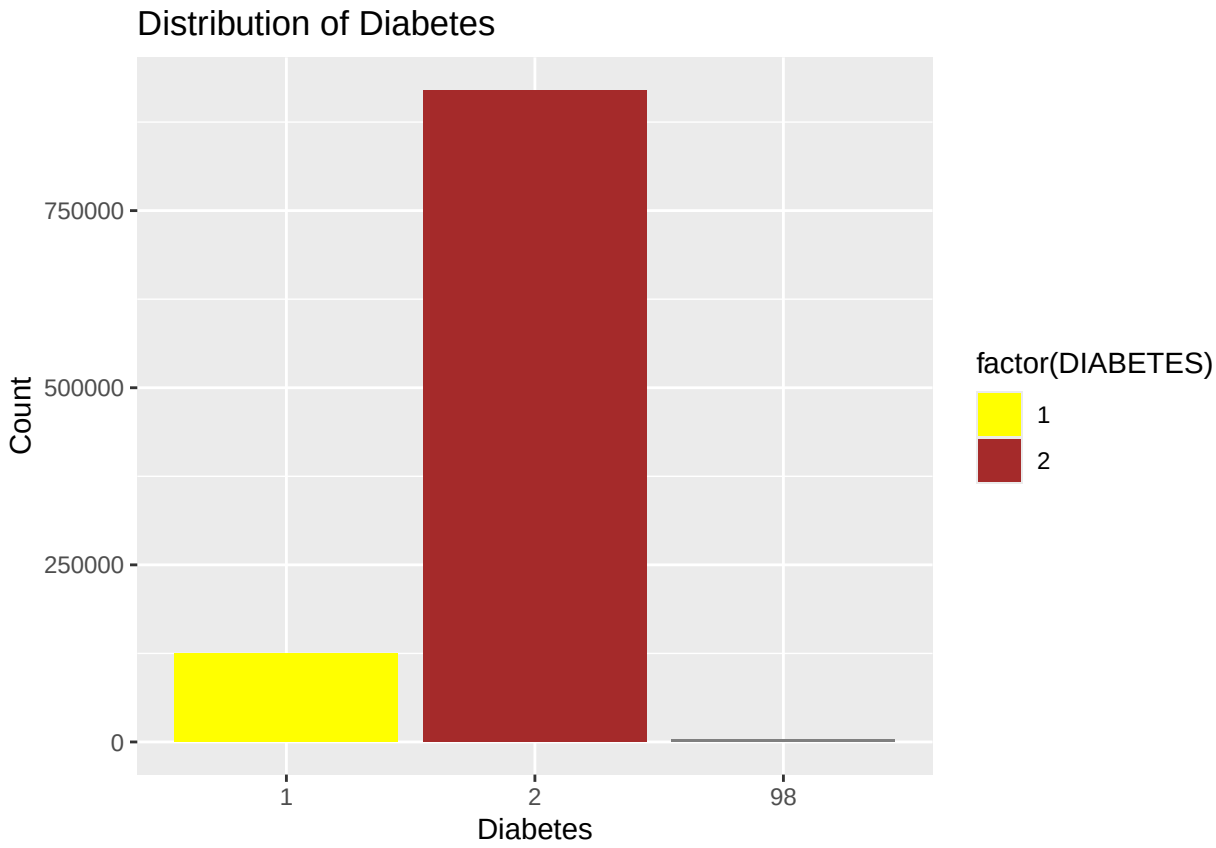

```r
# 3. Distribution of Patient Type (Hospitalized vs. Not Hospitalized)
ggplot(covid_clean, aes(x = factor(PATIENT_TYPE), fill = factor(PATIENT_TYPE))) +
  geom_bar() +
  labs(title = "Distribution of Patient Type", x = "Patient Type", y = "Count") +
  scale_fill_manual(values = c("1" = "red", "2" = "green"))
```



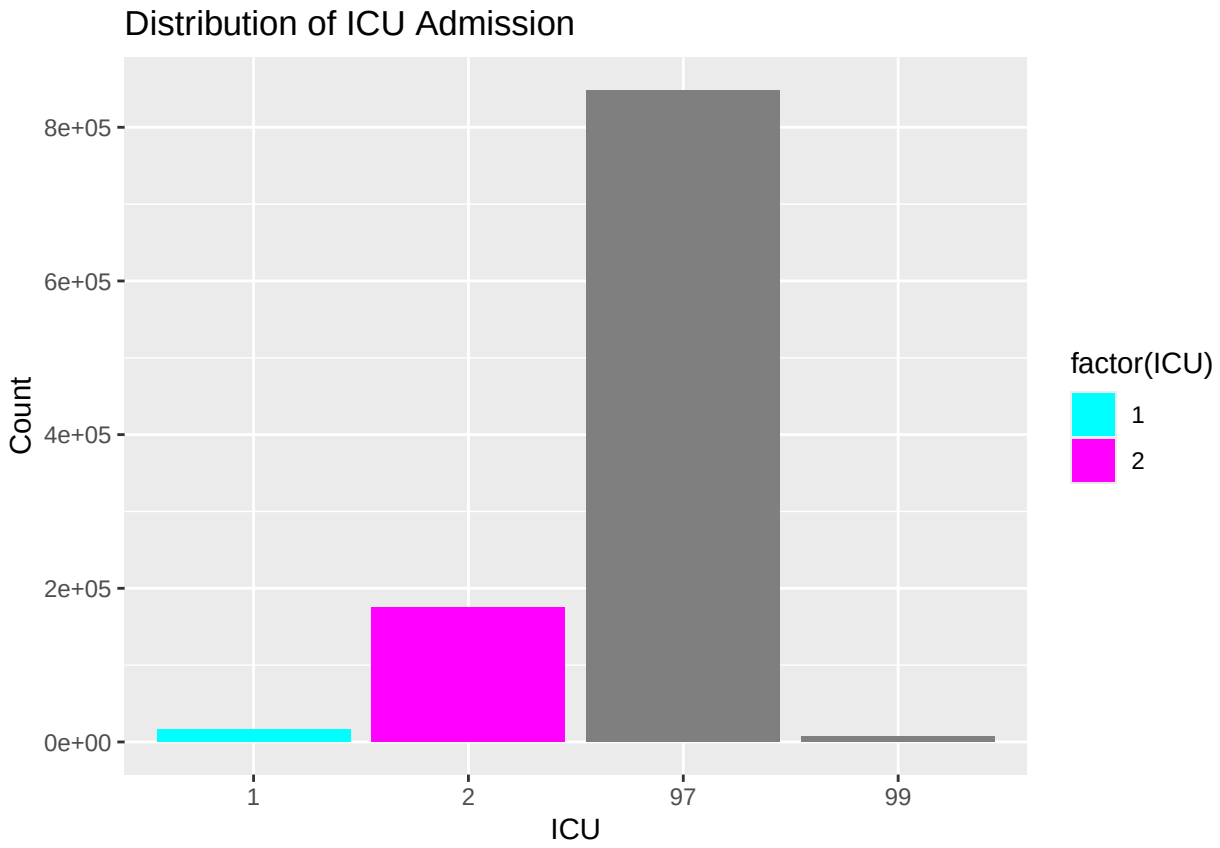

```
# 4. Distribution of Pneumonia
ggplot(covid_clean, aes(x = factor(PNEUMONIA), fill = factor(PNEUMONIA))) +
  geom_bar() +
  labs(title = "Distribution of Pneumonia", x = "Pneumonia", y = "Count") +
  scale_fill_manual(values = c("1" = "orange", "2" = "purple"))
```



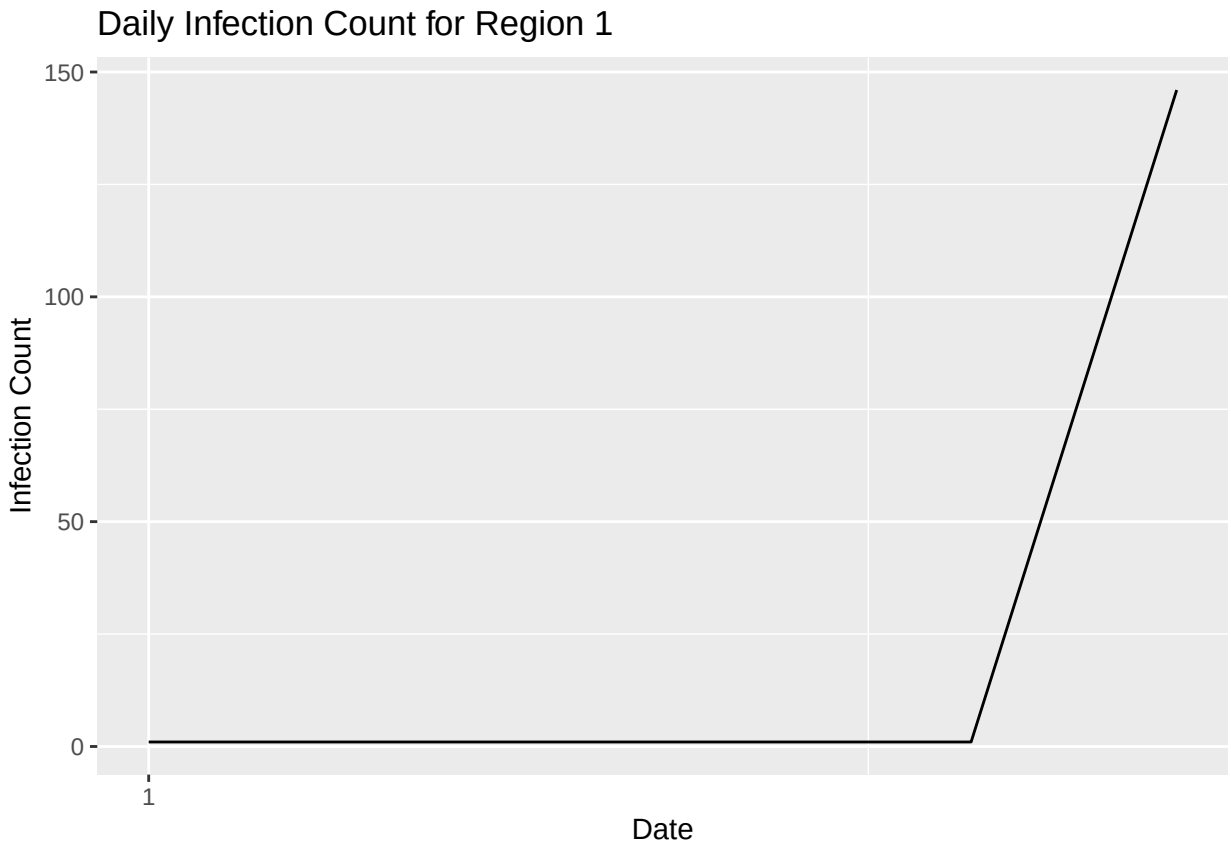

```
# 5. Distribution of Diabetes
ggplot(covid_clean, aes(x = factor(DIABETES), fill = factor(DIABETES))) +
  geom_bar() +
  labs(title = "Distribution of Diabetes", x = "Diabetes", y = "Count") +
  scale_fill_manual(values = c("1" = "yellow", "2" = "brown"))
```




```r
# 6. Distribution of ICU Admission
ggplot(covid_clean, aes(x = factor(ICU), fill = factor(ICU))) +
  geom_bar() +
  labs(title = "Distribution of ICU Admission", x = "ICU", y = "Count") +
  scale_fill_manual(values = c("1" = "cyan", "2" = "magenta"))
```




```r
# Time Series Forecasting with ARIMA

# Convert DATE_DIED to Date format
covid_clean$DATE_DIED <- as.Date(covid_clean$DATE_DIED, format = "%d/%m/%Y")

# Aggregate data by date and region (MEDICAL_UNIT) to get daily infection counts
daily_infections <- covid_clean %>%
  group_by(DATE_DIED, MEDICAL_UNIT) %>%
  summarise(infection_count = n(), .groups = 'drop')

# Plot time series for a specific region (e.g., MEDICAL_UNIT = 1)
region_data <- daily_infections %>% filter(MEDICAL_UNIT == 1)

# Convert to time series object
ts_data <- ts(region_data$infection_count, frequency = 7)  # Assuming weekly seasonality

# Plot the time series
autoplot(ts_data) +
  labs(title = "Daily Infection Count for Region 1", x = "Date", y = "Infection Count")
```




```
# Check for stationarity using Augmented Dickey-Fuller test
adf_test <- adf.test(ts_data)
```

```
## Warning in adf.test(ts_data): p-value greater than printed p-value
```

```
print(adf_test)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = 1.7321, Lag order = 1, p-value = 0.99
## alternative hypothesis: stationary
```

```
# If not stationary, difference the data
if (adf_test$p.value > 0.05) {
  ts_data <- diff(ts_data)
  print("Data was differenced to achieve stationarity.")
}
```

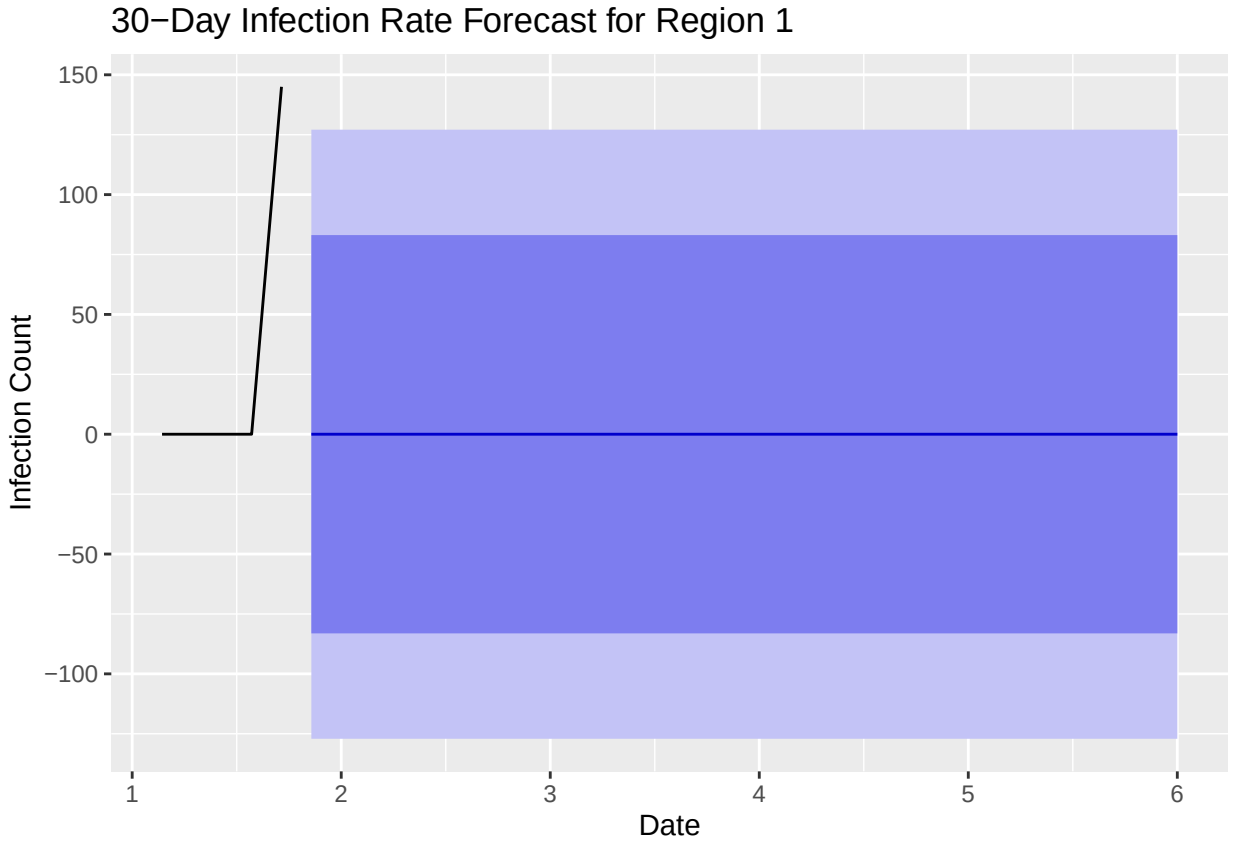
```
## [1] "Data was differenced to achieve stationarity."
```

```
# Fit ARIMA model
arima_model <- auto.arima(ts_data, seasonal = TRUE)
summary(arima_model)
```

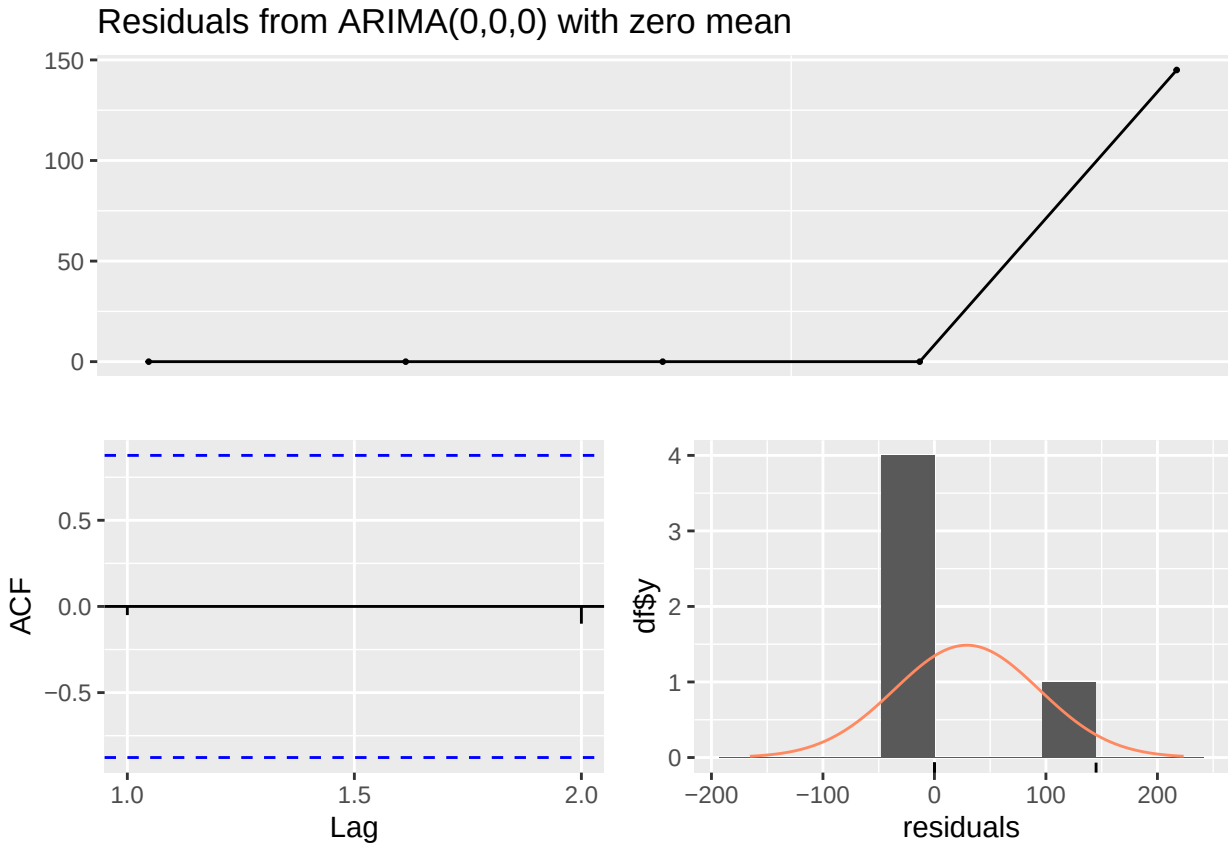
```
## Series: ts_data
## ARIMA(0,0,0) with zero mean
##
## sigma^2 = 4205:  log likelihood = -27.95
```

```
## AIC=57.91   AICc=59.24   BIC=57.52
##
## Training set error measures:
##                ME     RMSE MAE MPE MAPE MASE  ACF1
## Training set 29 64.84597  29 100  100  NaN -0.05
```

```r
# Forecast future infection rates
forecast_result <- forecast(arima_model, h = 30)  # Forecast for the next 30 days

# Plot the forecast
autoplot(forecast_result) +
  labs(title = "30-Day Infection Rate Forecast for Region 1", x = "Date", y = "Infection Count")
```




```r
# Check model residuals
checkresiduals(arima_model)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,0,0) with zero mean
## Q* = 0.53229, df = 3, p-value = 0.9117
##
## Model df: 0.   Total lags used: 3
```

This dataset provides COVID-19 data with 1,048,575 observations and 21 variables. An exploratory data analysis (EDA) was performed on the data, and a time series forecasting model (ARIMA) was applied to the cleaned dataset. The analysis includes a 30-day forecast plot for infection rates, along with diagnostic plots for the ARIMA model residuals.