Data Exploration

Melanie McCord

Introduction

In this part of this study, for initial modeling and analysis, I will be looking at the total number of thefts from January 2001 - March 11, 2023 that were reported. Note that due to the large size of the original dataset (nearly 8 million rows), the raw data is not included in this repository. The raw data can be accessed here.

Data Overview

For this partition of the data, there are two variables: year/month and the number of thefts reported in each month. The full dataset has more variables, which are described below. Each row in the full dataset represents an individual crime that was reported.

Variables

ID	Unique identifier for the record.
Case number	Chicago id for the case number
Date	Date when the incident occurred, this is sometimes a best estimate.
Block	The partially redacted address where the incident occurred.
IUCR	The Illinois Unifrom Crime Reporting Code.
Primary Type	The primary description of the IUCR code.
Description	The secondary description of the IUCR code.
Location description	The primary description of the location where the incident occurred.
Arrest	Whether or not the incident resulted in an arrest.
Domestic	Whether or not the incident was a domestic incident.
Beat	Indicates the beat where the incident occurred.
District	The police district where the incident occurred.
Ward	The city council district where the incident occurred.
Community Area	The community area where the incident occurred.
FBI Code	FBI Code crime classification.
X Coordinate	The X coordinate location where the incident occurred.
Y coordinate	The Y coordinate where the incident occurred.
Year	Year the incident occurred.
Updated on	Date and time the record was last updated.
Latitude	The latitude where the incident occurred.
Longitude	The longitude where the incident occurred.
Location	The location of the incident.

```
chicago_crime <- read.csv("data/thefts_by_month.csv")
chicago_crime <- chicago_crime %>%
  select(-X) %>%
  rename(NumThefts = sum.Count.) %>%
  drop_na(month)
head(chicago_crime)
```

```
##   year month NumThefts
## 1 2022    10      5224
## 2 2015     2      3228
## 3 2019    10      5390
## 4 2001     1      7867
## 5 2017     3      4493
## 6 2008     8      8501
```

```
chicago_crime_monthly <- chicago_crime %>%
 mutate(month = month.name[month]) %>%
  mutate(Month = str_c(year, month, sep = " ")) %>%
  select(Month, NumThefts) %>%
  mutate(Month = yearmonth(Month)) %>%
  filter(year(Month) < 2023) %>%
  as_tsibble(index = Month)
head(chicago_crime_monthly)
```

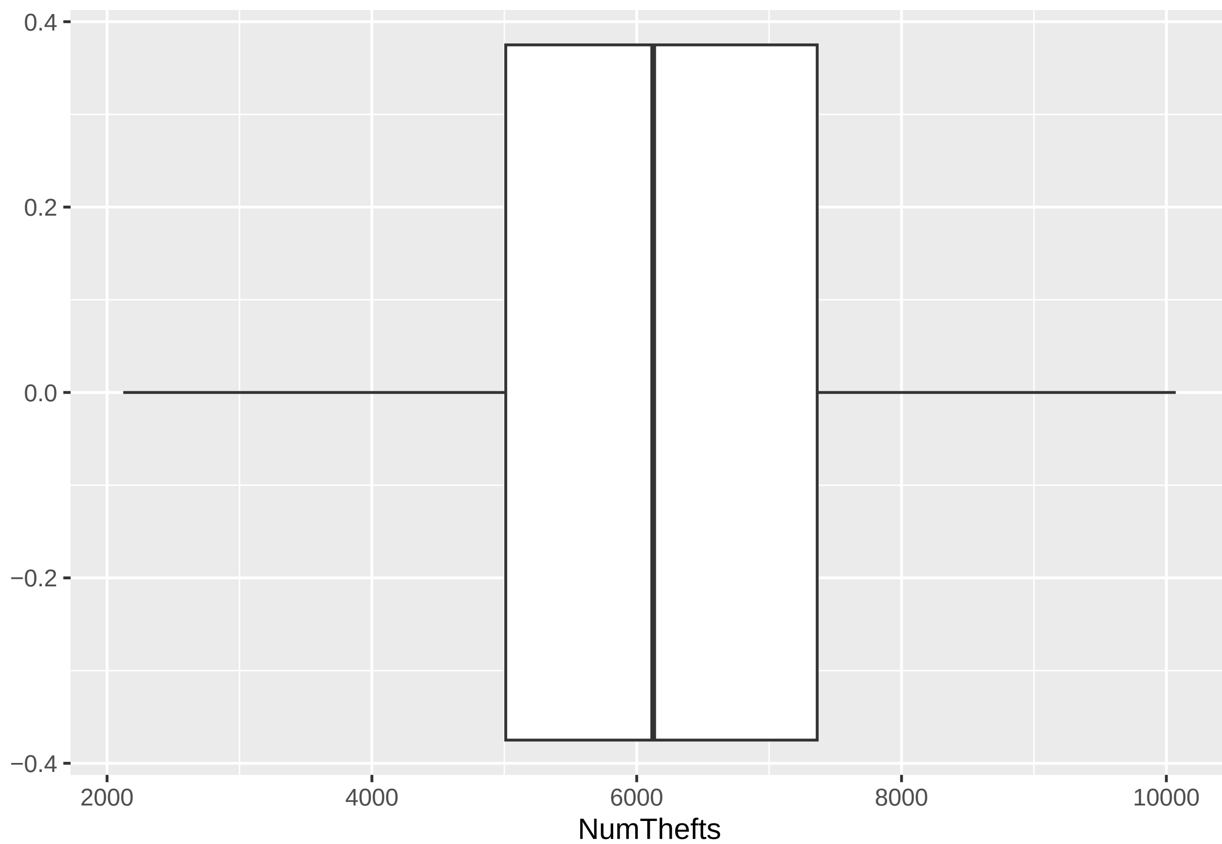
```
## # A tsibble: 6 x 2 [1M]
##      Month NumThefts
##      <mth>     <int>
## 1 2001 Jan      7867
## 2 2001 Feb      6669
## 3 2001 Mar      7765
## 4 2001 Apr      7686
## 5 2001 May      8420
## 6 2001 Jun      8612
```

Data Analysis

Number of Thefts Over Time

Boxplot

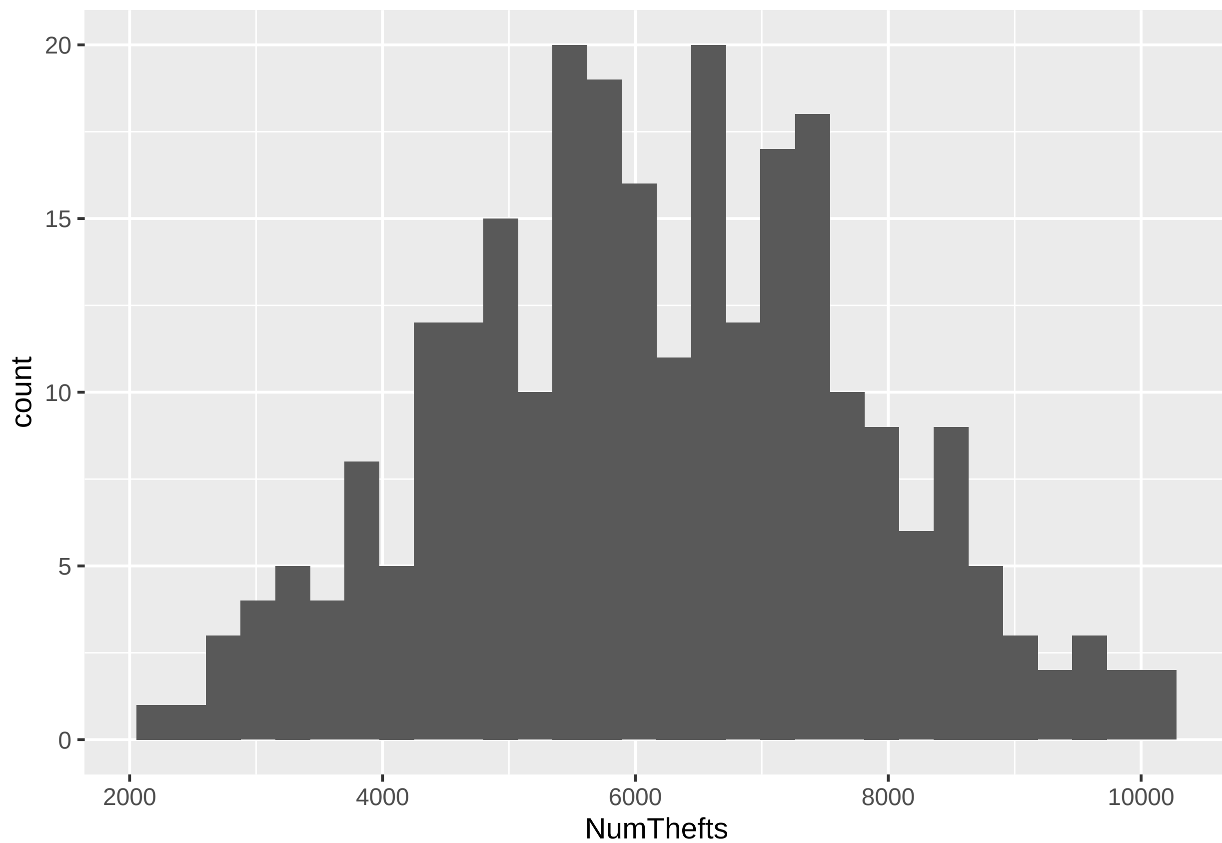
```
ggplot(chicago_crime_monthly, aes(x = NumThefts)) + geom_boxplot()
```



Histogram

```
ggplot(chicago_crime_monthly, aes(x = NumThefts)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Distribution Table

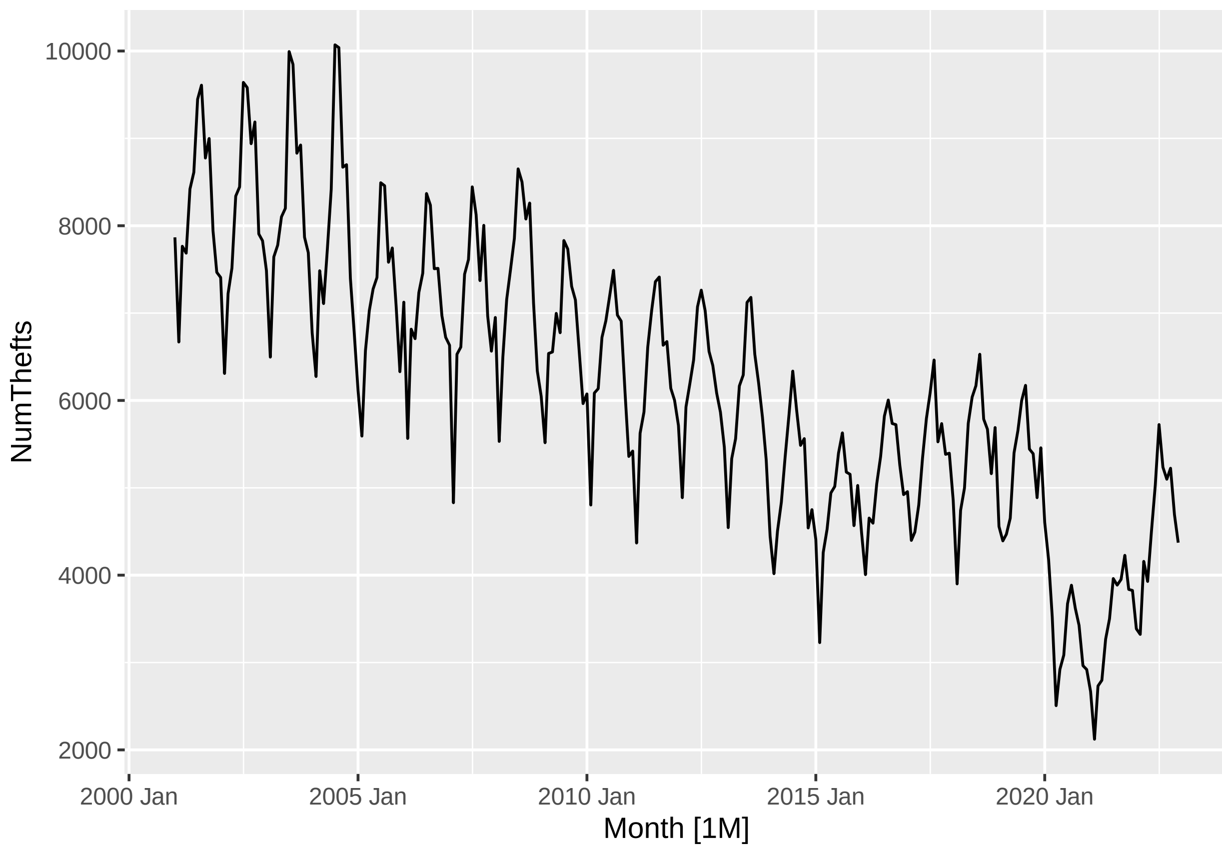
```
summary(chicago_crime_monthly$NumThefts)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2122    5011    6124    6161    7363   10071
```

Number of Thefts Reported in Chicago By Month

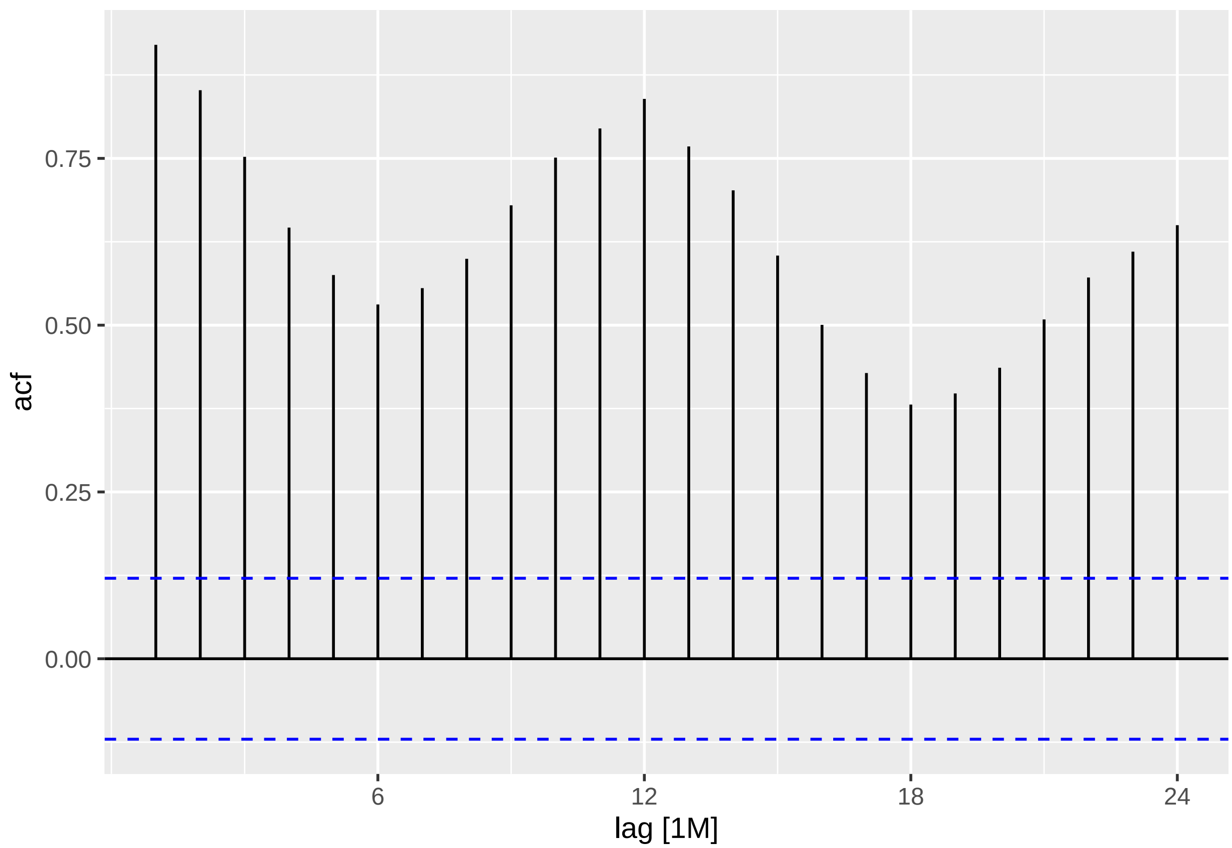
Due to the long time frame of the dataset, it's hard to see exactly where the seasonal pattern is occurring, but there does appear to be a seasonal pattern. There appears to be a general decreasing trend. Initially, I had included the raw data from 2023 as well, but there is a steep drop in March 2023 due to the smaller number of days there, so that is not included here.

```
chicago_crime_monthly %>%
  autoplot(NumThefts)
```



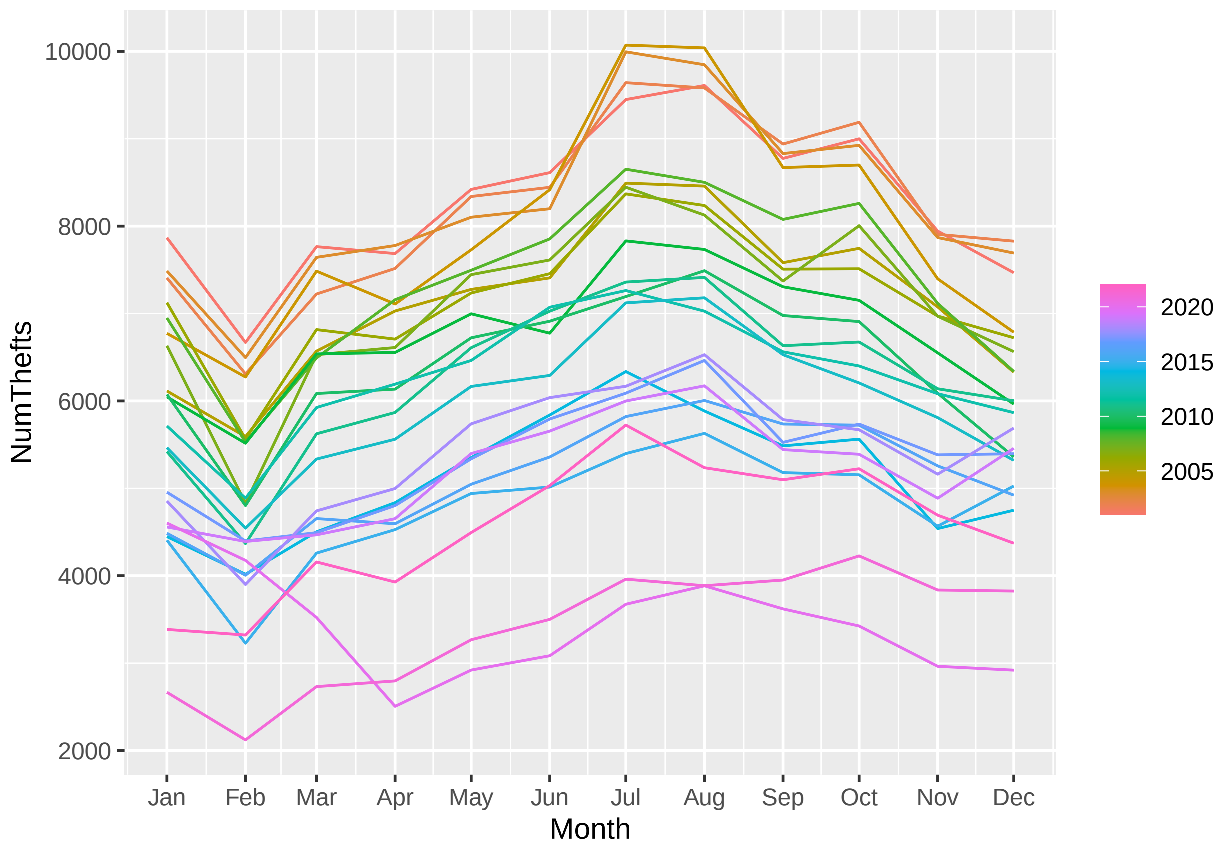
There is really strong positive autocorrelation throughout the monthly data, however it appears to follow a pattern of peaking, sharp decrease, then peaking.

```
chicago_crime_monthly %>%
  ACF(NumThefts) %>%
  autoplot()
```

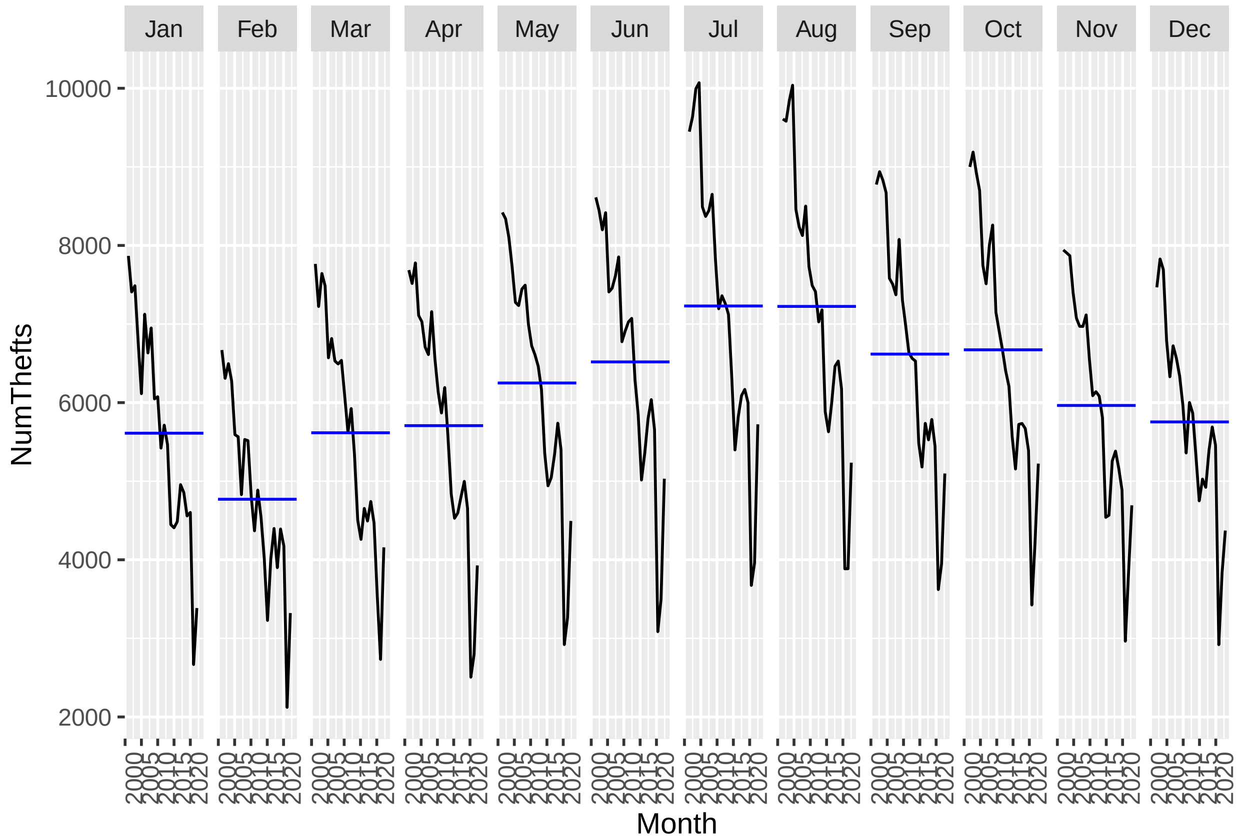


Number of Thefts By Year

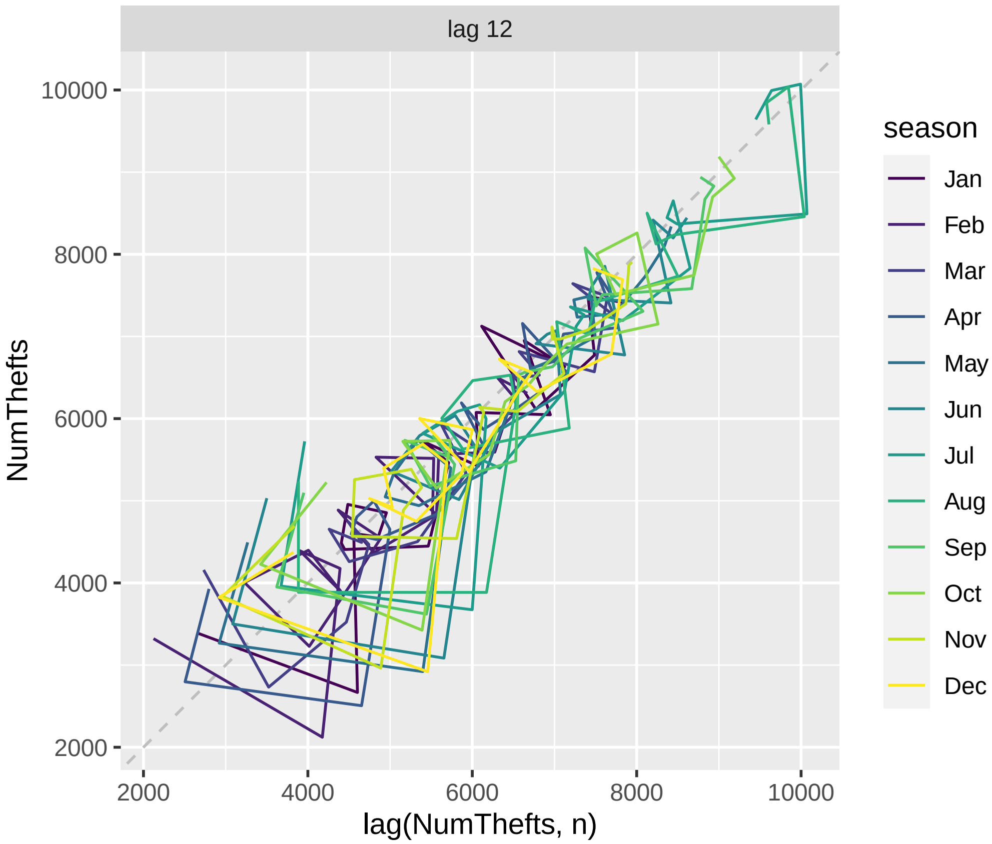
```
chicago_crime_monthly %>%
  gg_season(NumThefts)
```




```
chicago_crime_monthly %>%
  gg_subseries(NumThefts)
```




```r
chicago_crime_monthly %>%
  gg_lag(NumThefts, lags = 12)
```



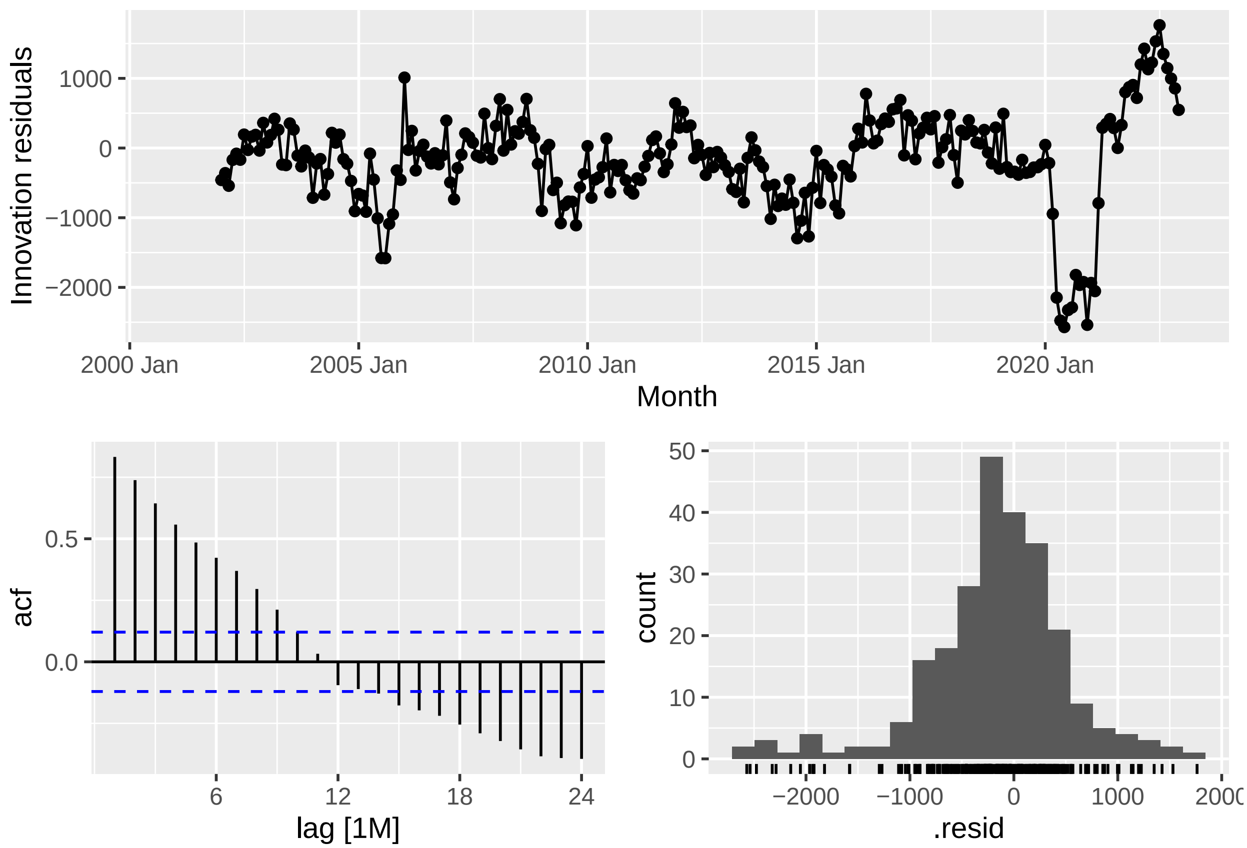
Models

```
chicago_crime.model <- chicago_crime_monthly %>%
  fabletools::model(
    snaive = SNAIVE(NumThefts)
  )
chicago_crime.model %>%
  gg_tsresiduals()
```

```
## Warning: Removed 12 rows containing missing values (`geom_line()`).

## Warning: Removed 12 rows containing missing values (`geom_point()`).

## Warning: Removed 12 rows containing non-finite values (`stat_bin()`).
```

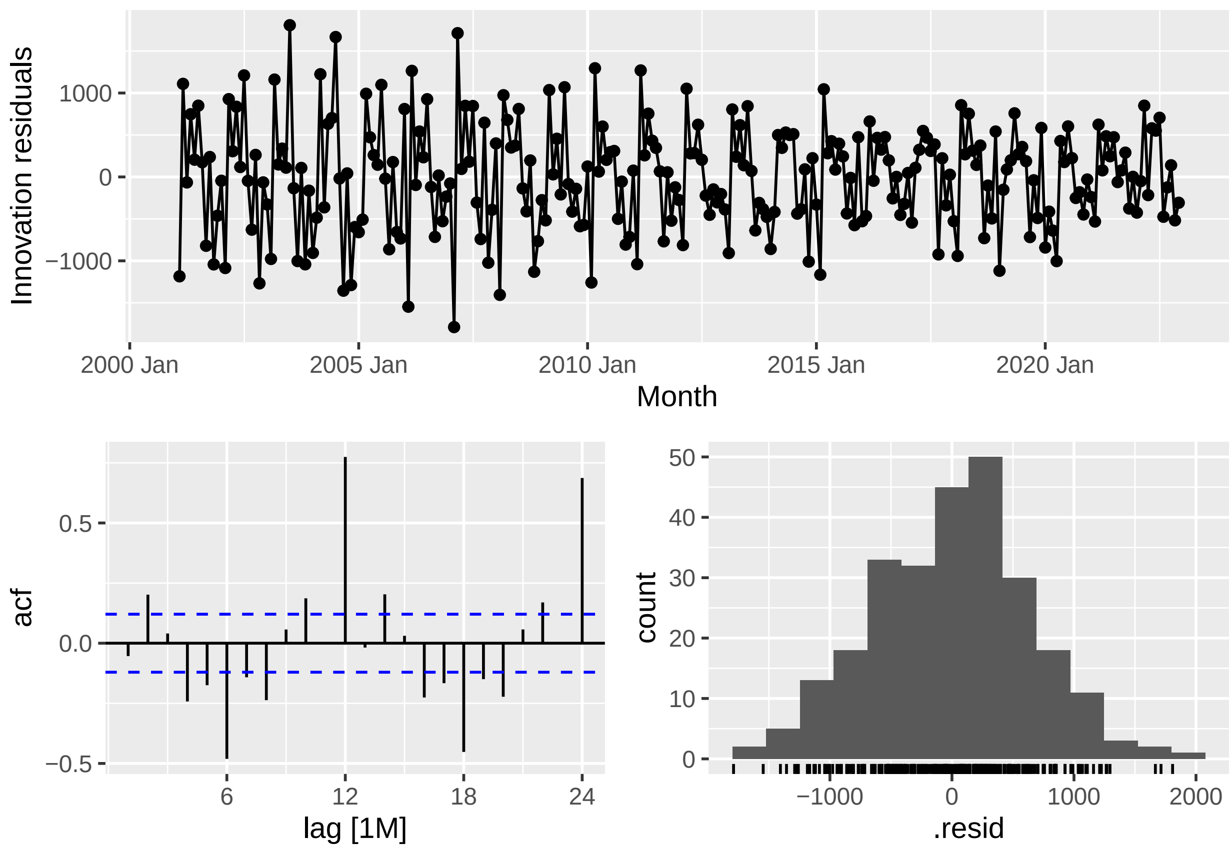



```
chicago_crime.model <- chicago_crime_monthly %>%
  fabletools::model(
    lm = RW(NumThefts ~ drift())
  )
chicago_crime.model %>%
  gg_tsresiduals()
```

```
## Warning: Removed 1 row containing missing values (`geom_line()`).

## Warning: Removed 1 rows containing missing values (`geom_point()`).

## Warning: Removed 1 rows containing non-finite values (`stat_bin()`).
```

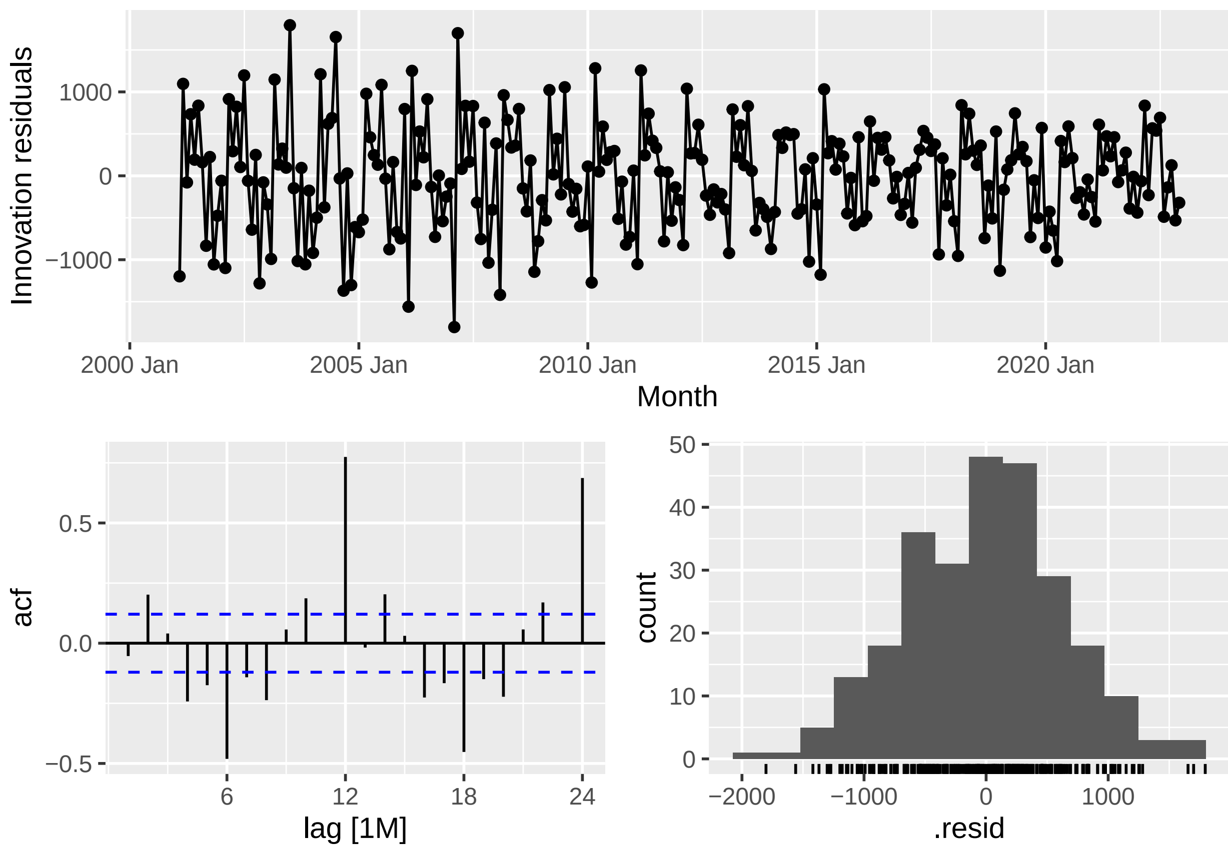



```
chicago_crime.model <- chicago_crime_monthly %>%
  fabletools::model(
    naive = NAIVE(NumThefts)
  )
chicago_crime.model %>%
  gg_tsresiduals()
```

```
## Warning: Removed 1 row containing missing values (`geom_line()`).

## Warning: Removed 1 rows containing missing values (`geom_point()`).

## Warning: Removed 1 rows containing non-finite values (`stat_bin()`).
```



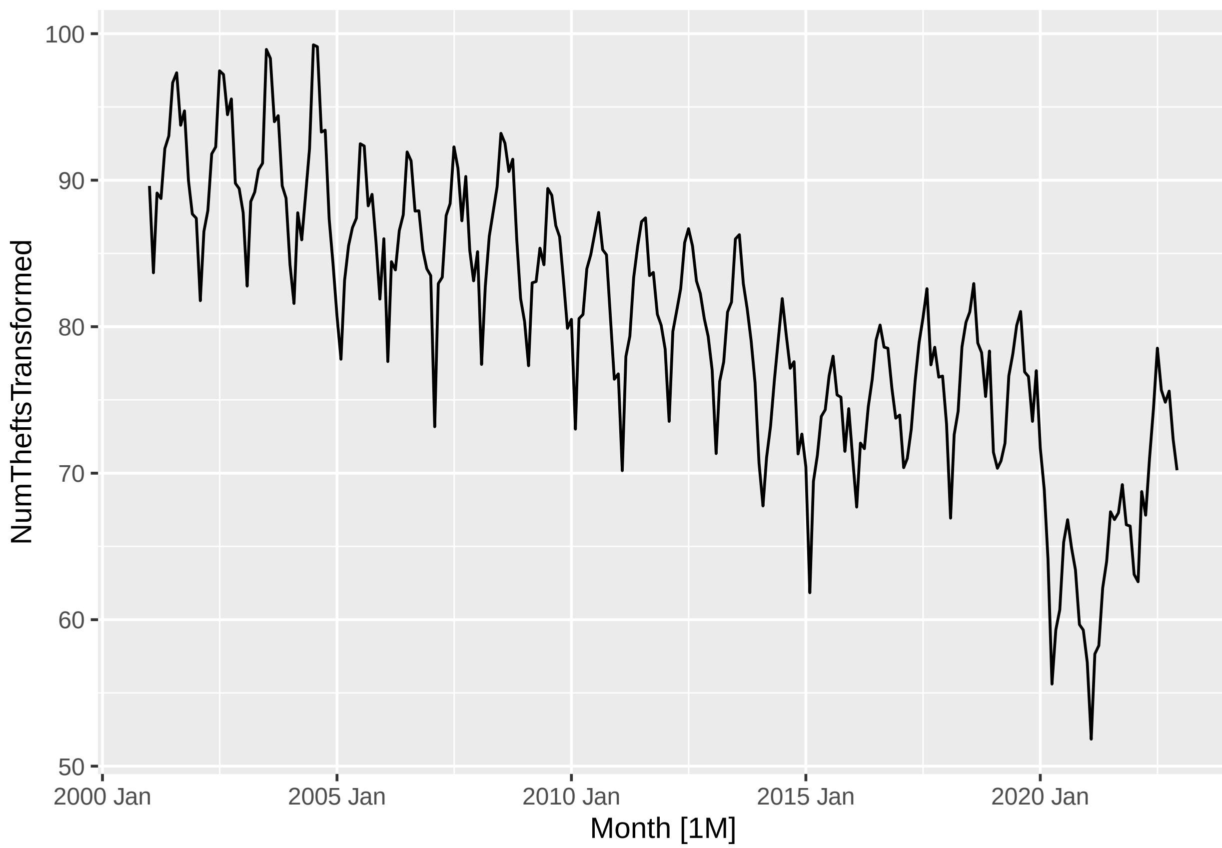

```
all_models <- chicago_crime_monthly %>%
  fabletools::model(
    snaive = SNAIVE(NumThefts),
    lm = RW(NumThefts ~ drift())
  )
all_models %>%
  forecast(h = "5 years") %>%
  autoplot(filter(chicago_crime_monthly, year(Month) > 2010))
```

Transforming the Data

```
lambda <- chicago_crime_monthly |>
  features(NumThefts, features = guerrero) |>
  pull(lambda_guerrero)
lambda
```

```
## [1] 0.4028242
```

```
chicago_crime_monthly <- chicago_crime_monthly %>%
  mutate(NumTheftsTransformed = box_cox(NumThefts, lambda))
chicago_crime_monthly %>%
  autoplot(NumTheftsTransformed)
```

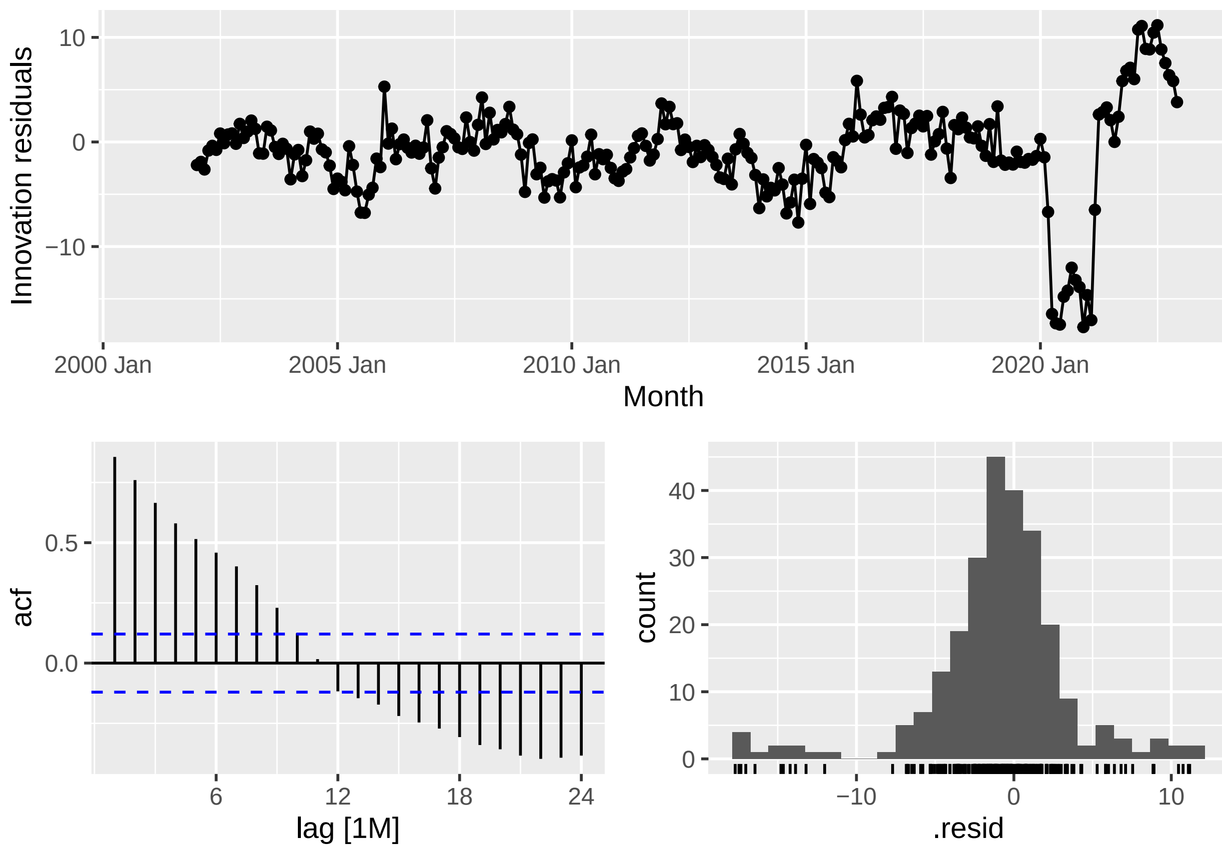



```
snaive.transformed <- chicago_crime_monthly %>%
  fabletools::model(
    snaive = SNAIVE(NumTheftsTransformed)
  )
snaive.transformed %>%
  gg_tsresiduals()
```

```
## Warning: Removed 12 rows containing missing values (`geom_line()`).

## Warning: Removed 12 rows containing missing values (`geom_point()`).

## Warning: Removed 12 rows containing non-finite values (`stat_bin()`).
```

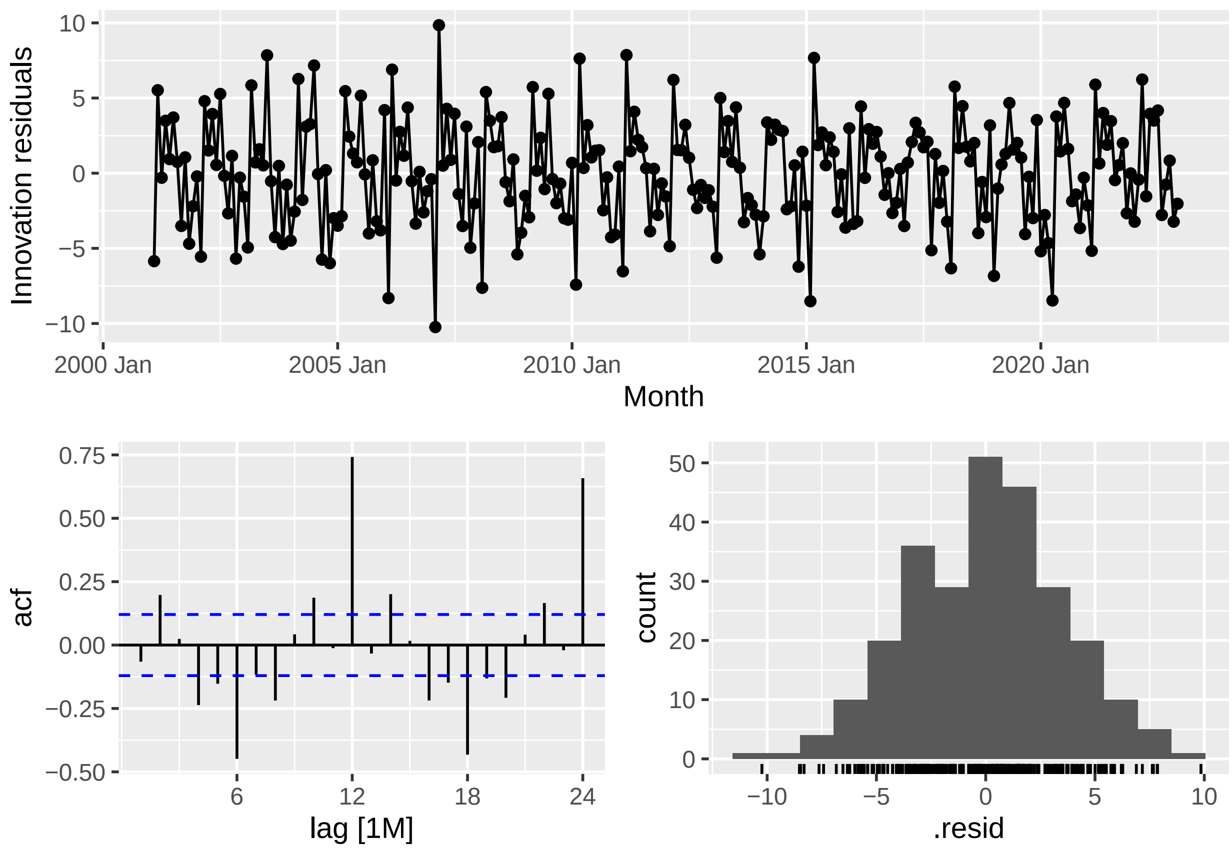



```
lm.transformed <- chicago_crime_monthly %>%
  fabletools::model(
    snaive = RW(NumTheftsTransformed ~ drift())
  )
lm.transformed %>%
  gg_tsresiduals()
```

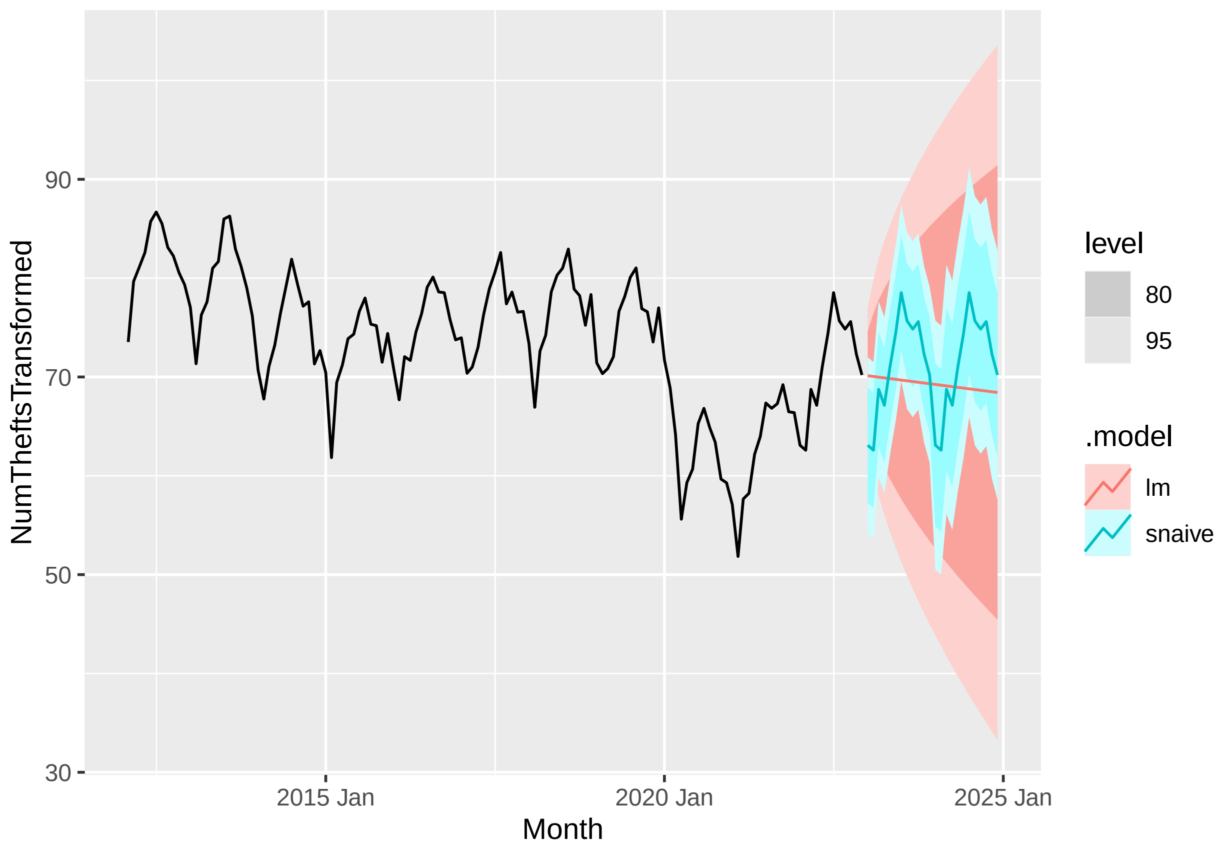
```
## Warning: Removed 1 row containing missing values (`geom_line()`).

## Warning: Removed 1 rows containing missing values (`geom_point()`).

## Warning: Removed 1 rows containing non-finite values (`stat_bin()`).
```




```
all_models.transformed <- chicago_crime_monthly %>%
  fabletools::model(
    snaive = SNAIVE(NumTheftsTransformed),
    lm = RW(NumTheftsTransformed ~ drift())
  )
all_models.transformed %>%
  forecast(h = "2 years") %>%
  autoplot(filter(chicago_crime_monthly, Month > yearmonth("Jan 2012")))
```



Arrests Over Time

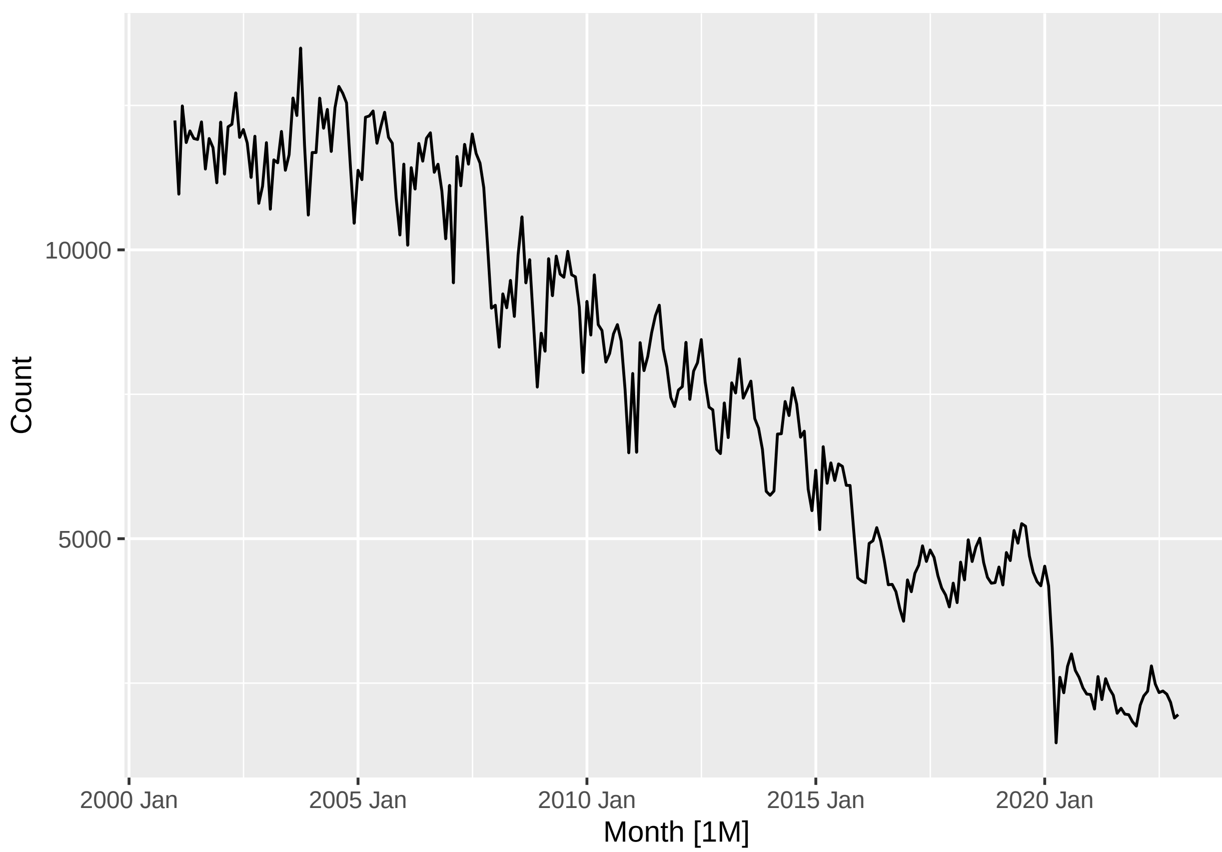
```
arrests = read.csv("data/arrests_by_month.csv")
arrests <- arrests %>%
  select(-X) %>%
  rename(Count = sum.Count.) %>%
  drop_na(month)
head(arrests)
```

```
##   Arrest year month Count
## 1  false 2012     1 18749
## 2   true 2019     3  4761
## 3  false 2005     2 20774
## 4   true 2014     9  6757
## 5   true 2015     5  6313
## 6  false 2009     9 24304
```

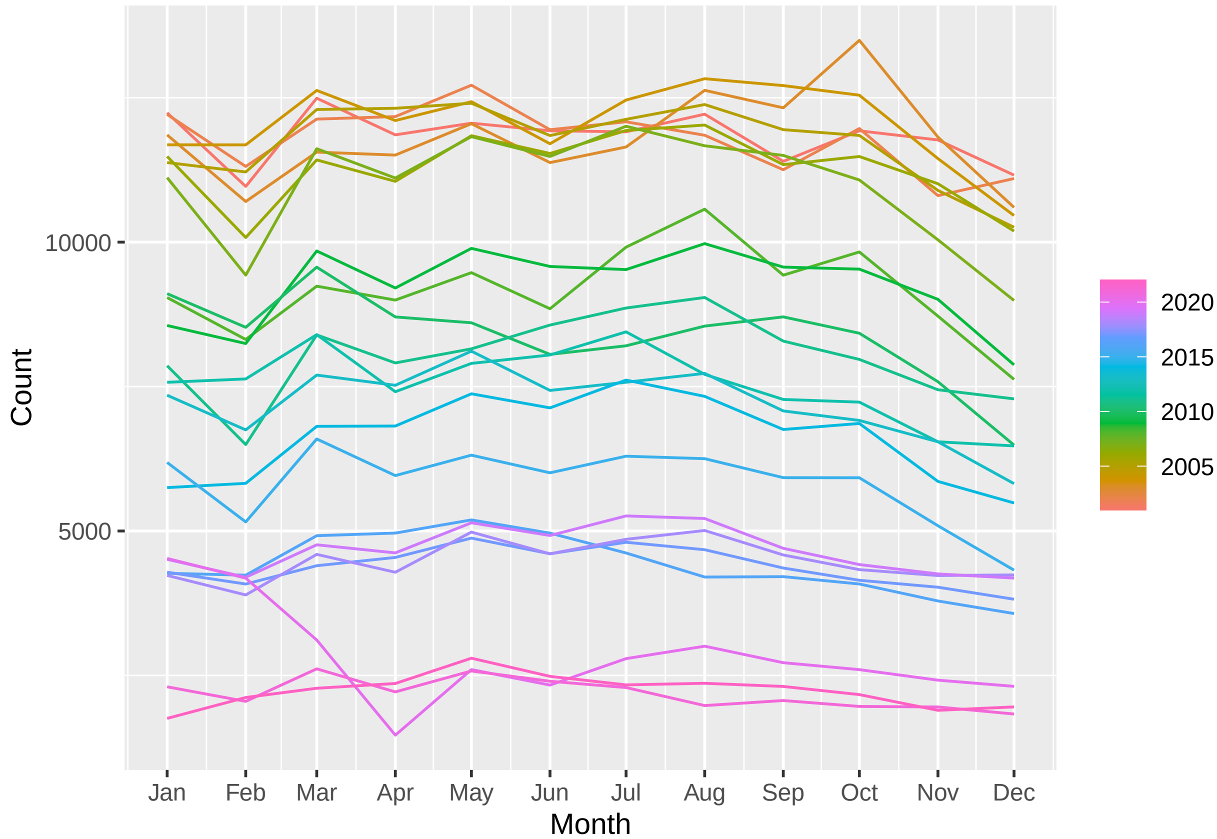
```
arrests_ts <- arrests %>%
 mutate(month = month.name[month]) %>%
  mutate(Month = str_c(year, month, sep = " ")) %>%
  mutate(Month = yearmonth(Month)) %>%
  filter(year(Month) < 2023) %>%
  select(-c(year, month)) %>%
  filter(Arrest == "true") %>%
  as_tsibble(key = Arrest, index = Month)
head(arrests_ts)
```

```
## # A tsibble: 6 x 3 [1M]
## # Key:       Arrest [1]
##   Arrest Count    Month
##   <chr>  <int>    <mth>
## 1 true   12239 2001 Jan
## 2 true   10964 2001 Feb
## 3 true   12491 2001 Mar
## 4 true   11857 2001 Apr
## 5 true   12059 2001 May
## 6 true   11927 2001 Jun
```

```
arrests_ts %>%
  autoplot(Count)
```



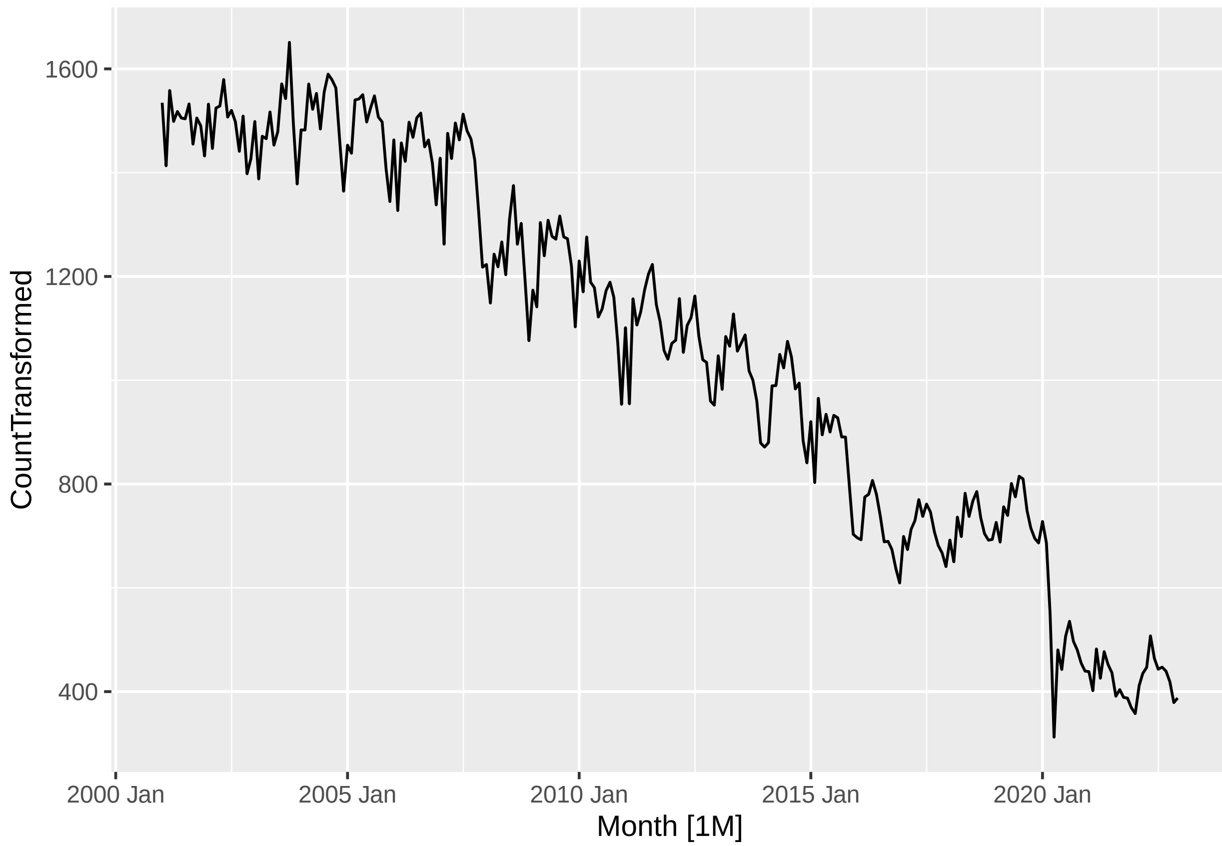

```
arrests_ts %>%
  gg_season(Count)
```




```
lambda <- arrests_ts |>
  features(Count, features = guerrero) |>
  pull(lambda_guerrero)
lambda
```

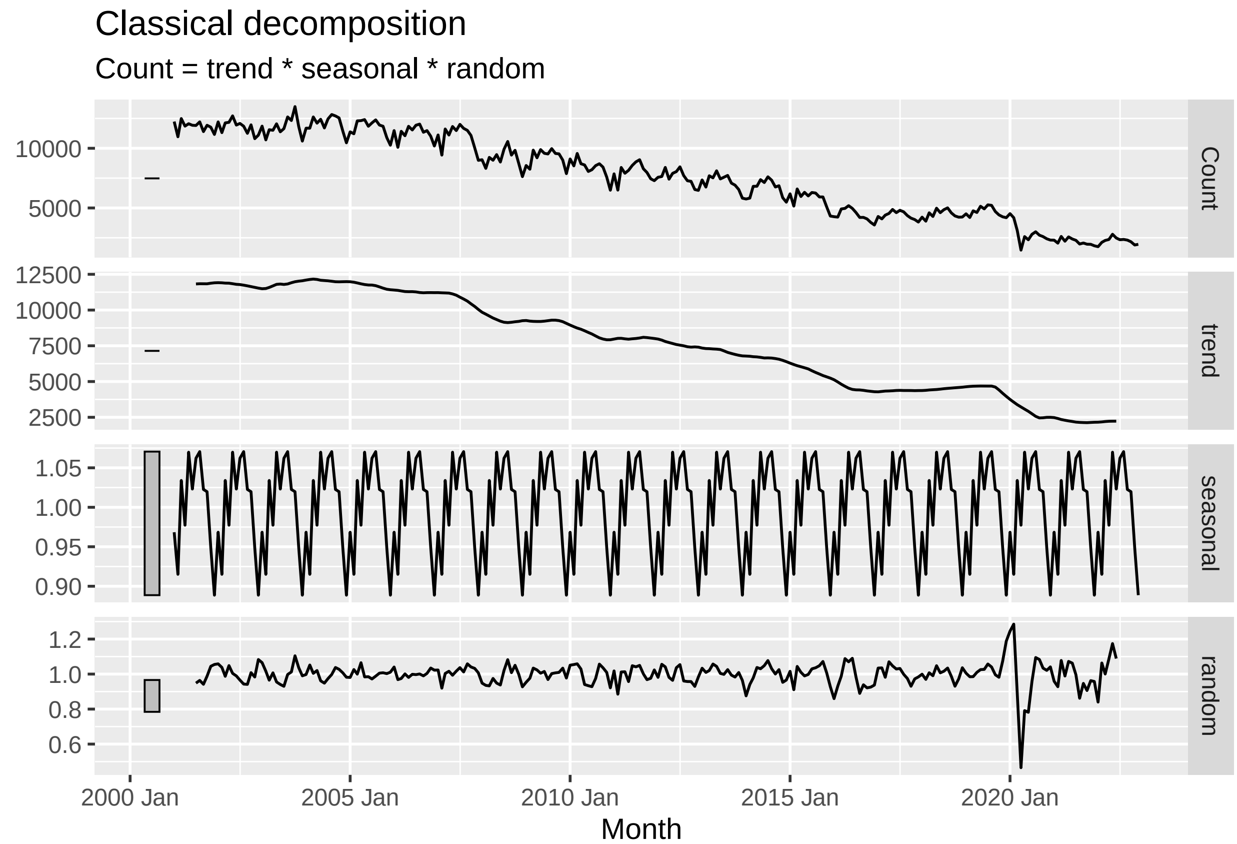
```
## [1] 0.7487626
```

```
arrests_ts <- arrests_ts %>%
  mutate(CountTransformed = box_cox(Count, lambda))
arrests_ts %>%
  autoplot(CountTransformed)
```

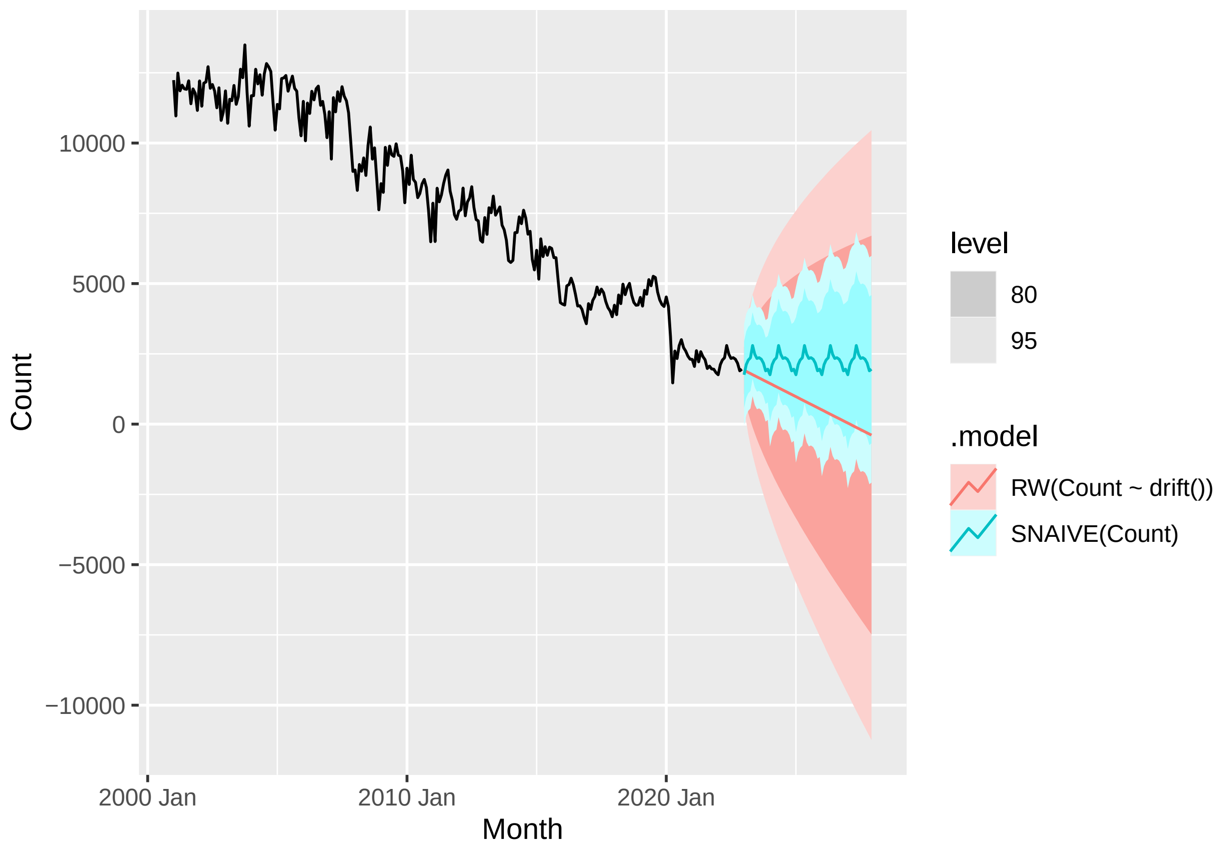



```
arrests_ts %>%
  model(
    classical_decomposition(Count, type = "multiplicative")
  ) %>%
  components() %>%
  autoplot()
```

```
## Warning: Removed 6 rows containing missing values (`geom_line()`).
```




```
arrests_ts %>%
  model(
    RW(Count ~ drift()),
    SNAIVE(Count)
  ) %>%
  forecast(h = "5 years") %>%
  autoplot(arrests_ts)
```



Count
10000
5000
0
−5000
−10000
2000 Jan
2010 Jan
2020 Jan
Month
level
80
95
.model
RW(Count ~ drift())
SNAIVE(Count)