Bikes Sold Data Forecast

Javern Wilson

2024-08-12

About Dataset

The Bikes Sold dataset represents time series with daily entries. The main variables are the bike prices and the number of bikes sold.

Date	Bike Price	Number of Bikes Sold
8/7/2024	$107.82	398880731
8/6/2024	$103.84	409012100
8/5/2024	$92.06	552842400
8/2/2024	$103.76	482027500
8/1/2024	$117.53	523462300
7/31/2024	$112.90	473174200

Summary of Bikes Sold Dataset

```
##       Date              bike_price       num_bikes_sold
##  Min.   :1999-01-22   Min.   :  0.030   Min.   :1.968e+07
##  1st Qu.:2005-06-13   1st Qu.:  0.280   1st Qu.:3.460e+08
##  Median :2011-10-27   Median :  0.460   Median :5.069e+08
##  Mean   :2011-10-30   Mean   :  6.302   Mean   :6.060e+08
##  3rd Qu.:2018-03-19   3rd Qu.:  4.200   3rd Qu.:7.359e+08
##  Max.   :2024-08-07   Max.   :139.800   Max.   :9.231e+09
```

Aggregate Dataset from Daily to Monthly

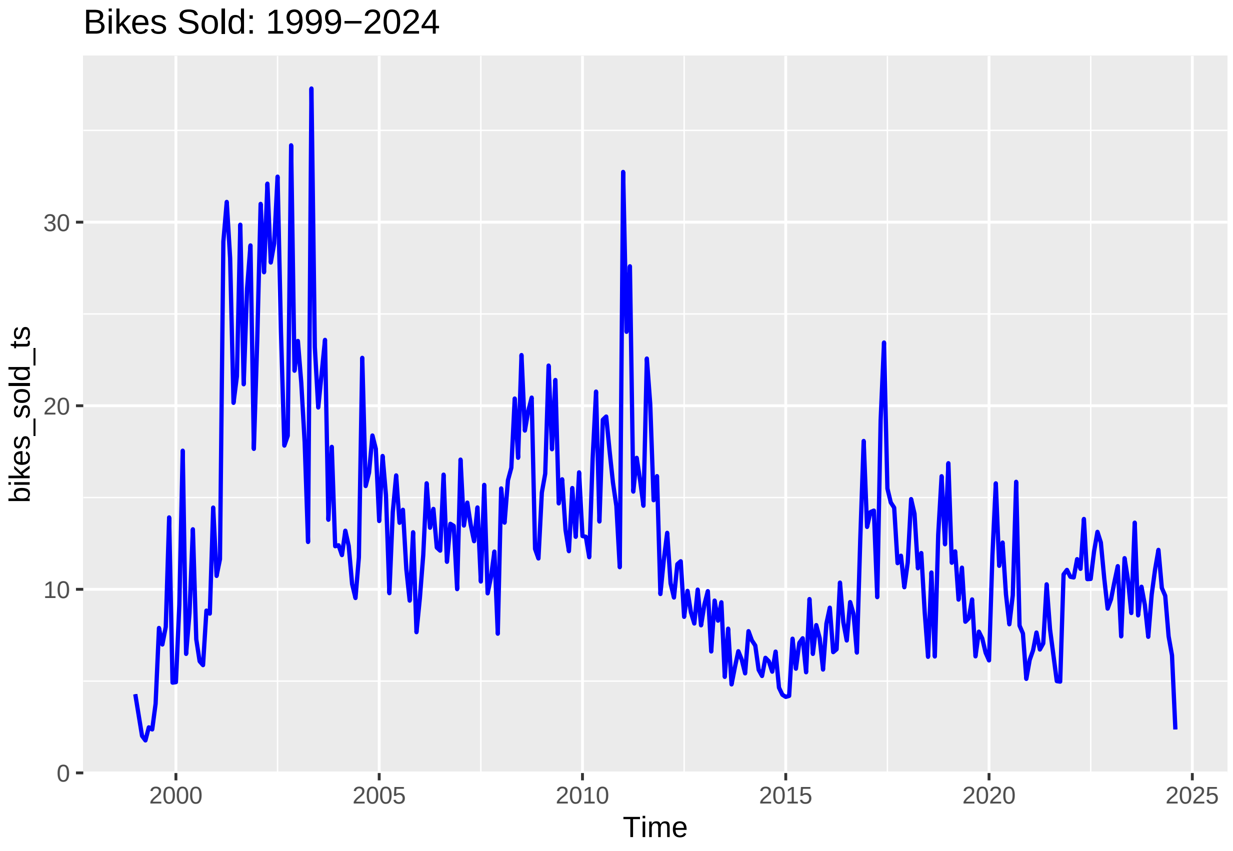
For forecasting and analysis, the data was aggregated monthly due to its length. A `Total_Sales = (Avg_Bike_Price * Total_Bikes_Sold)` column was added just for visibility.

YearMonth	Avg_Bike_Price	Total_Bikes_Sold	Total_Sales
1999-01-01	0.0416667	4.284288	0.1785120
1999-02-01	0.0415789	3.146592	0.1308320
1999-03-01	0.0421739	2.014512	0.0849599
1999-04-01	0.0395238	1.767024	0.0698395
1999-05-01	0.0400000	2.482512	0.0993005

YearMonth	Avg_Bike_Price	Total_Bikes_Sold	Total_Sales
2024-04-01	86.09273	10.074181	867.3137
2024-05-01	95.89591	9.647971	925.2009
2024-06-01	124.54632	7.442539	926.9408
2024-07-01	122.80500	6.405439	786.6199
2024-08-01	105.00200	2.366225	248.4584

Analysis/Observation

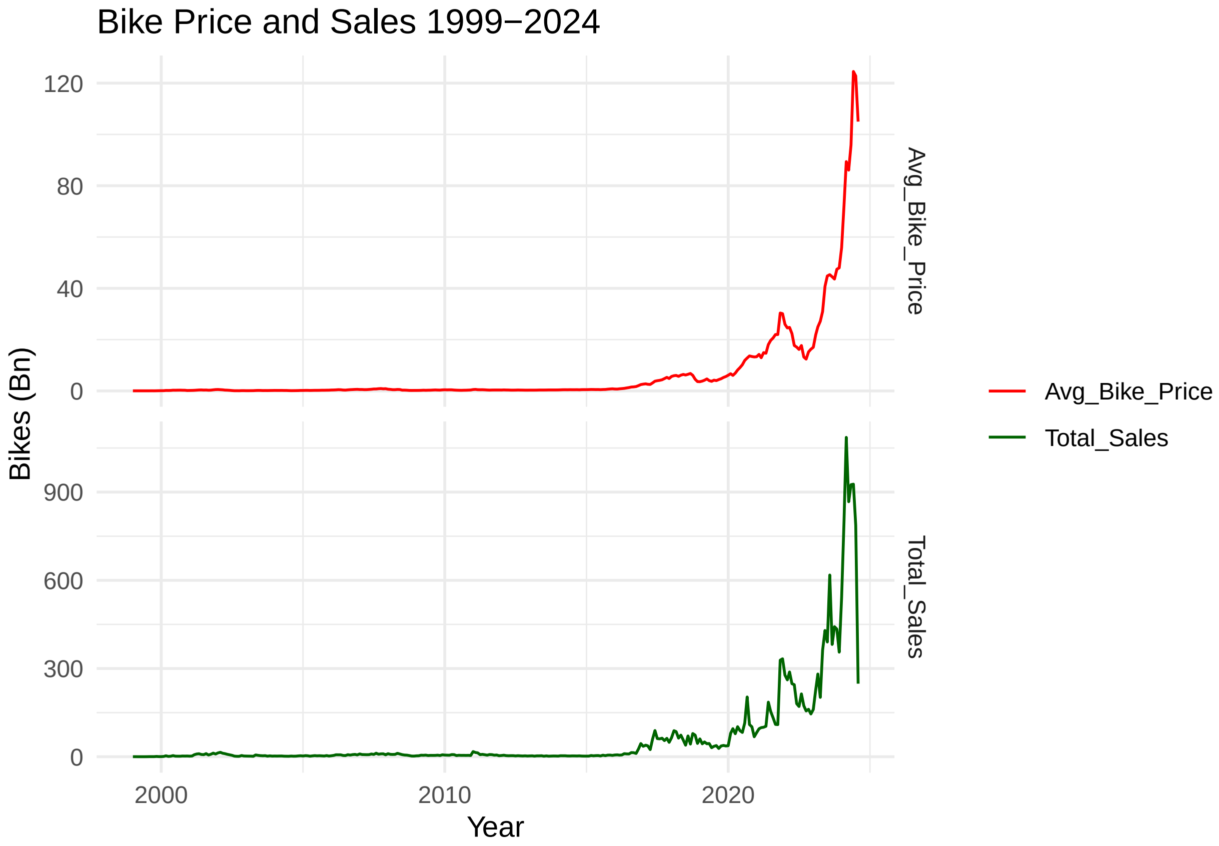
Illustration Time Series Graphs



Key Observations:

- **Early Growth (Late 1990s - Early 2000s):**
  - The series begins with a period of rapid growth, reaching a peak around the early 2000s. This indicates a sharp increase in bike sales during this time.
- **Fluctuations and Decline (2000s):**
  - After the peak, there is a significant decline, followed by several fluctuations, with smaller peaks and troughs.
  - The overall trend during the mid-to-late 2000s is downward, possibly indicating decreasing sales with occasional short-lived increases.

- **Stabilization and Further Decline (2010s):**
  - After 2010, the series shows a general stabilization with some volatility, but the overall trend is still downward, though less steep than in the previous decade.
- **Recent Years:**
  - There is another notable decline, with sales dropping to levels comparable to those seen at the beginning series.
  - Suggests that bike sales are continuing to decline, possibly due to market saturation, changing consumer preferences, or other external factors affecting demand.



Key Observations:

- **Low Activity (Late 1990s - 2010s):**
  - For the majority of the timeline both plots, from the late 1990s through the 2010s, bike sales remained relatively flat with minimal fluctuations. The sales figures during this period were consistently low, showing little to no significant growth.
- **Sudden and Exponential Growth (Late 2010s - Early 2020s):**
  - Starting around the late 2010s, there is a dramatic and exponential increase. This could be attributed to various factors such as:
    - ∗ More demand for new features in bike technology
    - ∗ Shift in consumer behavior (e.g., the COVID-19 pandemic)
    - ∗ Sharp rise in bike prices
    - ∗ Shift towards higher-value products.

- **Sharp Decline:**
  - After reaching the peak, there is a noticeable and steep decline in sales. However, even after this drop, the sales remain substantially higher than the earlier years of the time series.
  - The sharp decline after the peak could indicate a market correction, a decrease in demand after a surge, or market saturation.

What story does the graphs tell?

Higher Number of Bikes Sold but Lower Revenue

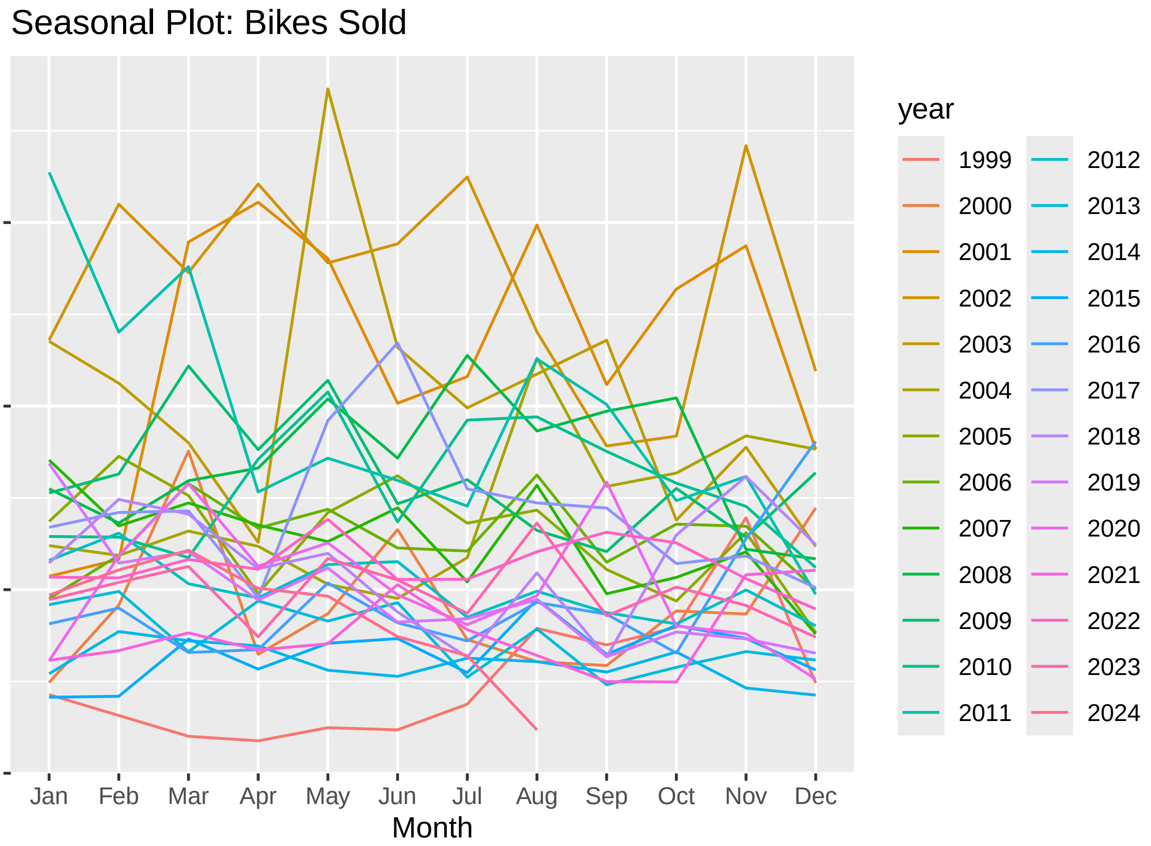
- Lower-Priced Bikes: If the number of bikes sold is higher, but the revenue is low, it could suggest that the bikes being sold were of lower value or priced more afford-ably. During the early 2000s, there may have been high-volume sales of lower-cost bikes.
- Market Saturation: A large number of bikes sold in earlier periods might reflect market saturation, where the demand was high, but the price point was kept low to maintain or increase sales volume.

Sudden Increase in Revenue with Fewer Units Sold

- Higher-Priced Bikes: The exponential increase in revenue in the later years, despite a potentially lower number of bikes sold, suggests a shift towards selling more expensive, premium bikes. This could be due to advancements in technology, electric bikes, or a focus on high-end models that command a higher price.
- Pandemic Influence: The spike in revenue might also correspond with the COVID-19 pandemic, where demand for bikes surged, and consumers were willing to pay more, either due to increased interest in outdoor activities or supply chain disruptions that drove prices up.

Going forward, we will focus on the `Total_Bikes_Sold` series which is used to help forecast bikes to be sold in the future months

Seaonsality



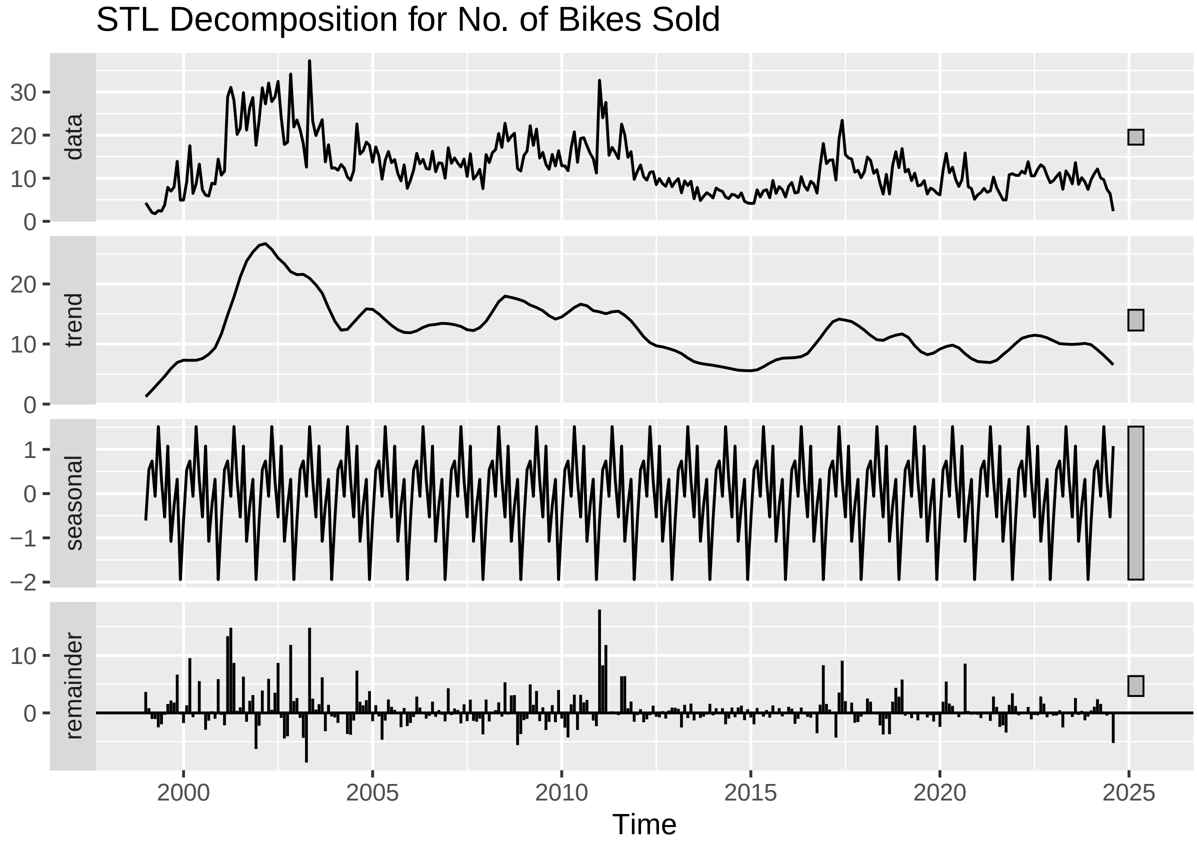
Peaks and Troughs by Month: The seasonal peaks and troughs are relatively consistent across many years, indicating that bike sales follow a similar pattern annually. This recurring pattern is a hallmark of seasonality, where sales increase and decrease at predictable times of the year.

April and June: There are noticeable peaks in bike sales around April and June across many years. This suggests that bike sales tend to increase during the spring and early summer months. The increase may be driven by factors such as warmer weather, outdoor activities, and perhaps holiday periods.

September and November: Another set of peaks is observed around September and November, indicating a possible secondary increase in sales. This could be due to end-of-summer activities, back-to-school periods or clearance sales when bike sales might spike again.

December, January and February: There is a drop in sales in these months across all years, which might reflect a seasonal decline during winter months.

Decomposition with STL (Seasonal and Trend decomposition using Loess)



To confirm the details aforementioned:

- The trend component shows the overall direction of bike sales over time (downwards)
- The seasonal component highlights the repetitive patterns or cycles within each year. Indicates consistent seasonal effects
- As for the remainder component, it captures noise or irregular effects not explained by trend or seasonality

Modeling

Create Train and Test sets

Train

```
##           Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
## 1999 4.284288 3.146592 2.014512 1.767024 2.482512 2.366160 3.764064 7.894800
##           Sep      Oct
## 1999 6.992640 7.953504
```

Test

```
##            Jan       Feb       Mar       Apr May Jun       Jul       Aug
## 2019                                                  8.415984  9.450328
## 2020  6.125412 11.848652 15.773952 11.278304
##            Sep       Oct       Nov       Dec
## 2019  6.347768  7.696172  7.321800  6.543532
## 2020
```

Build and Train Models

Models Applied:

- ARIMA
- Seasonal ARIMA (SARIMA)
- Exponential Smoothing (ETS)
- Holt-Winters
- Linear Regression
- Random Forest

Evaluate and Compare Models

Below are the accuracy scores (MAE and RMSE) for the different models trained. These metrics indicate how well each model performed on unseen data after training.

```
## [1] "Model Comparison:"
```

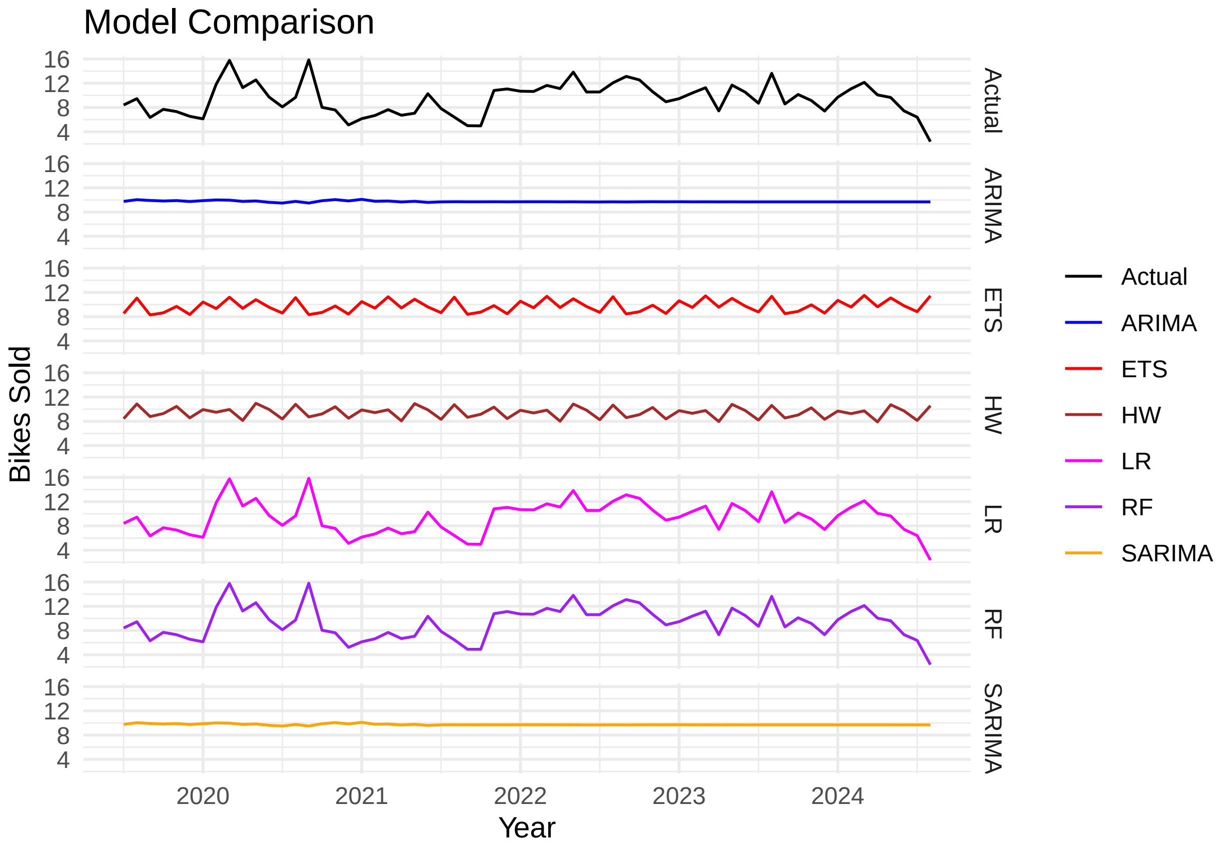
```
##               Model      MAE     RMSE
## 1             ARIMA 2.165712 2.684964
## 2               ETS 1.989633 2.641244
## 3            SARIMA 2.165712 2.684964
## 4      Holt-Winters 2.068275 2.654406
## 5     Random Forest 2.441044 3.186866
## 6 Linear Regression 2.437078 3.185128
```

Interpretation:

ETS (Exponential Smoothing State Space Model) achieved the lowest Mean Absolute Error (MAE) of 1.9896 and the lowest Root Mean Squared Error (RMSE) of 2.6412, indicating the best overall performance in terms of accuracy.

Holt-Winters model came close, with an MAE of 2.0683 and an RMSE of 2.6544. It performed slightly worse than ETS but still showed strong results.

Random Forest and Linear Regression were the least accurate among the models tested.



You may notice that the Random Forest (RF) and Linear Regression (LR) models have the highest Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) but fits closest to the actual data.

Here's why this might happen:

Overfitting: Both models captures not only the true patterns but also the noise within the training data which lead to to higher error rates when applied to the test data.

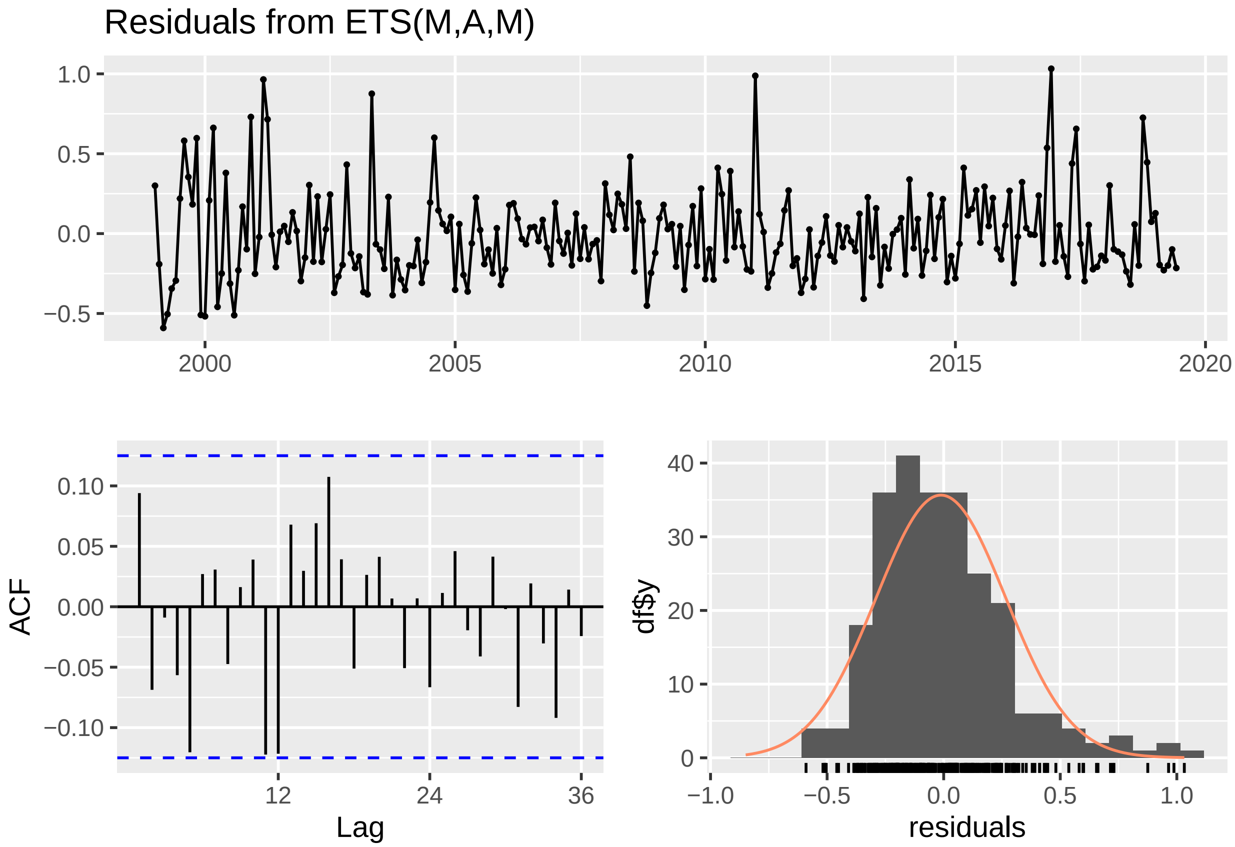
Model Complexity: Both models compared to ARIMA and ETS are complex. While this complexity allows them to capture more intricate patterns, it can also lead to higher errors because it may be sensitive to variations and noise in the data that other models might smooth over or ignore.

Variance in Data: If the data has a lot of variability, models like ARIMA and ETS may smooth out these fluctuations, leading to lower MAE and RMSE but less detailed fitting. RF and LR, on the other hand, might capture these fluctuations, leading to a better visual fit but higher error metrics due to the variability it introduces.

Closer Look at Chosen Models

ETS - Exponential Smoothing (ETS) Model

Residuals from ETS (M,A,M) Model

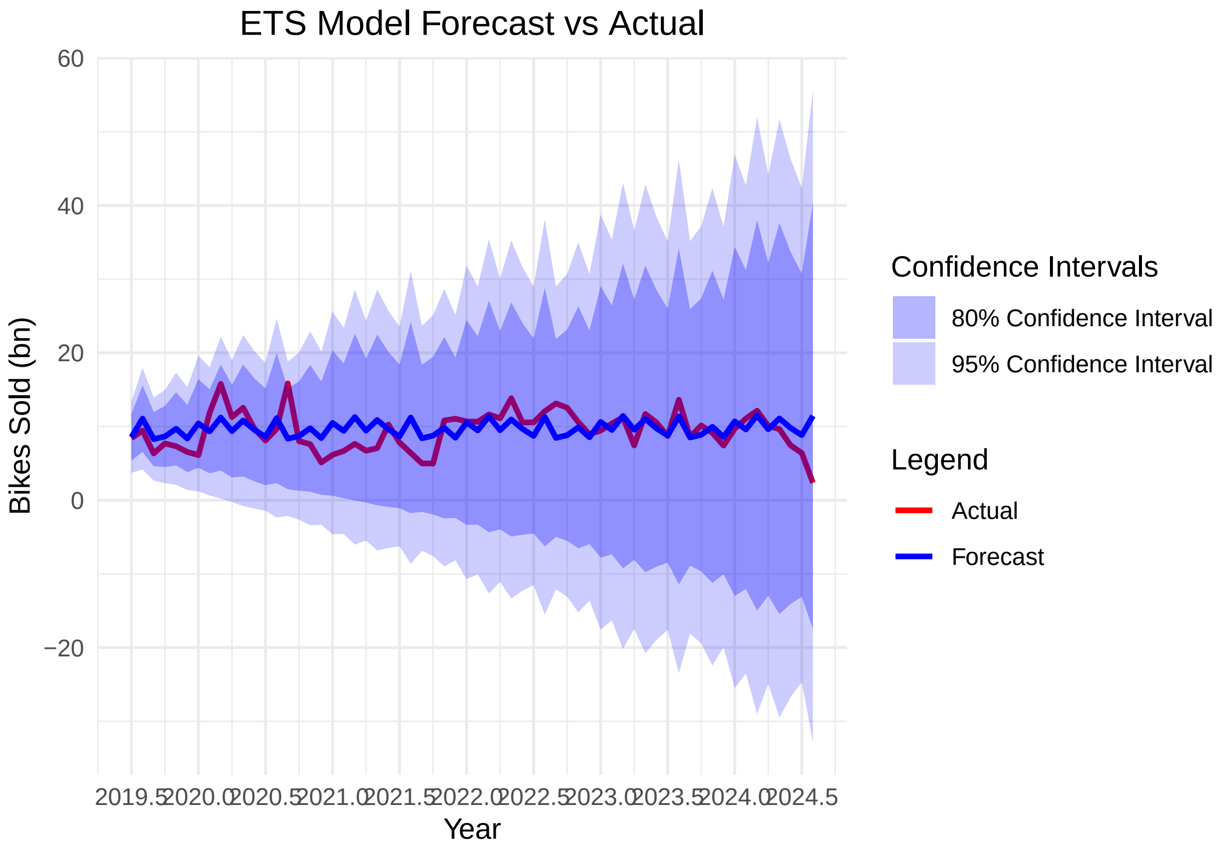


```
##
##  Ljung-Box test
##
## data:  Residuals from ETS(M,A,M)
## Q* = 26.55, df = 24, p-value = 0.3258
##
## Model df: 0.   Total lags used: 24
```

Residuals from ETS (M,A,M) Model

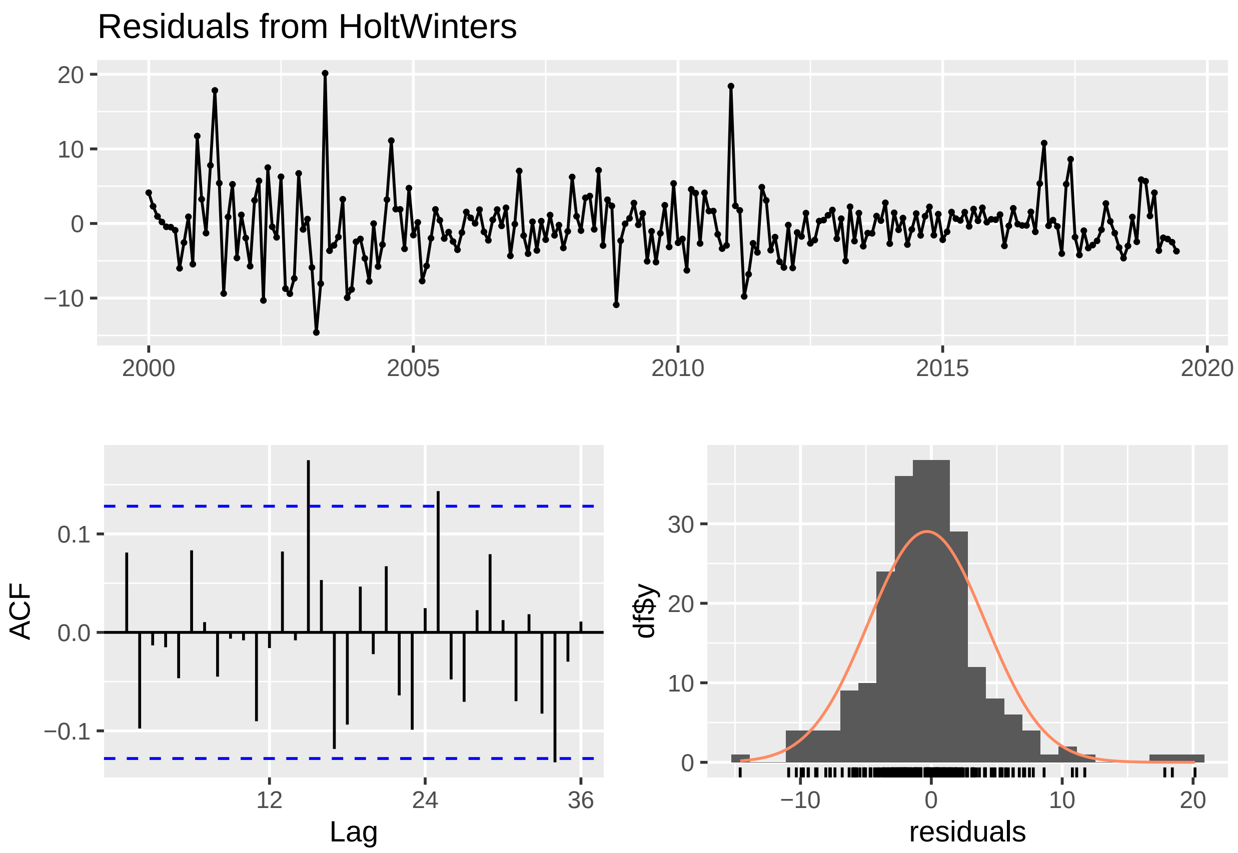
- The residuals fluctuate around zero, with some variability especially around 2000–2005.
- The ACF plot shows very little auto-correlation, with most of the auto-correlations staying within the significance bounds (blue dashed lines). This indicates that the residuals are fairly uncorrelated, which is a good sign, but there are still some small auto-correlations at certain lags.
- The histogram of the residuals is approximately normal but slightly skewed to the right. The residuals are mostly centered around zero, but there are some deviations from normality, which may indicate some non-randomness in the residuals.

ETS on the Test set



Holt-Winters Model

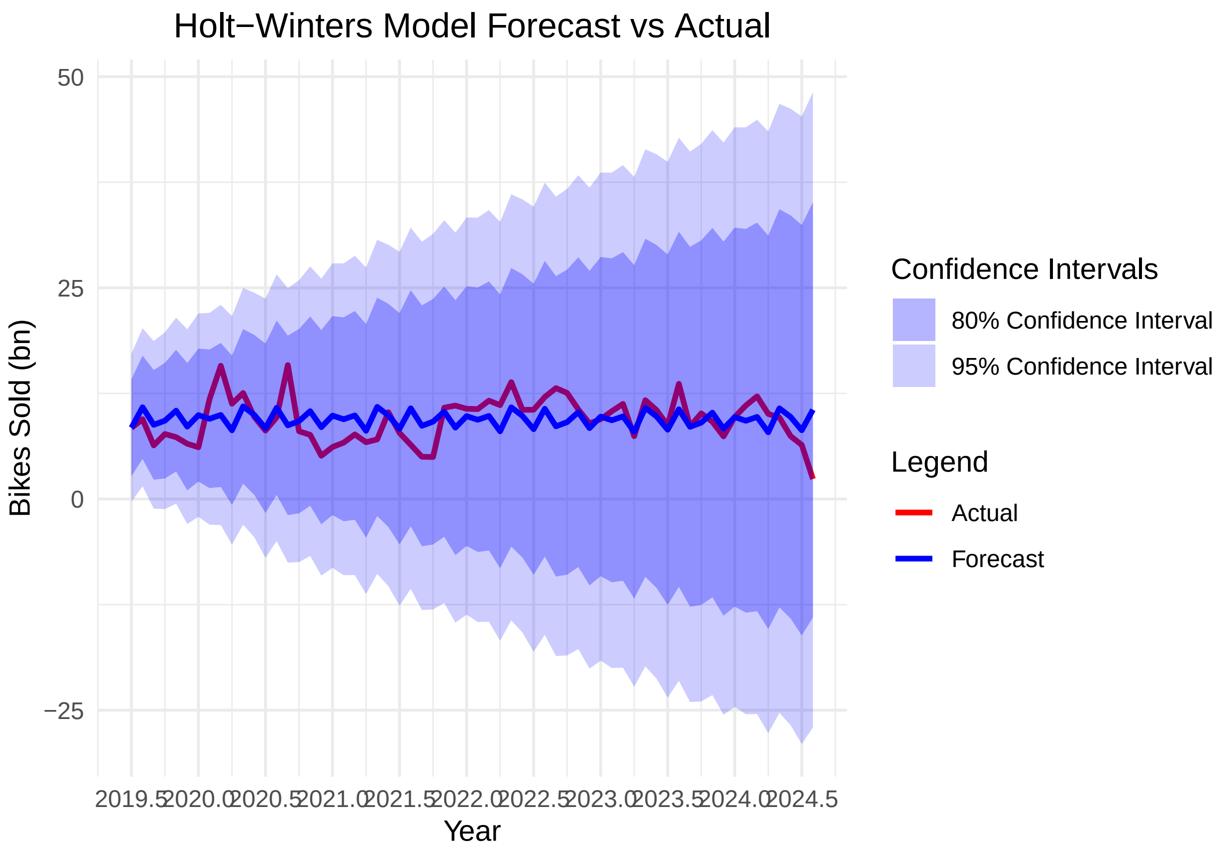
Residuals from Holt-Winters Model



```
##
##  Ljung-Box test
##
## data:  Residuals from HoltWinters
## Q* = 30.331, df = 24, p-value = 0.174
##
## Model df: 0.   Total lags used: 24
```

- The residuals also fluctuate around zero but exhibit a more pronounced pattern compared to the ETS model, especially around the early 2000s and around 2010. This suggests that the Holt-Winters model might not fully capture all the dynamics of the data.
- The ACF plot shows significant auto-correlations at certain lags. This indicates that the residuals are not completely random and that the Holt-Winters model might be missing some periodic components or trends in the data.
- The histogram is roughly normal, but there is a slight skew to the right and a few outliers. The distribution is more spread out compared to the ETS model, suggesting that the residuals have larger variability.

Holt-Winters on the test set



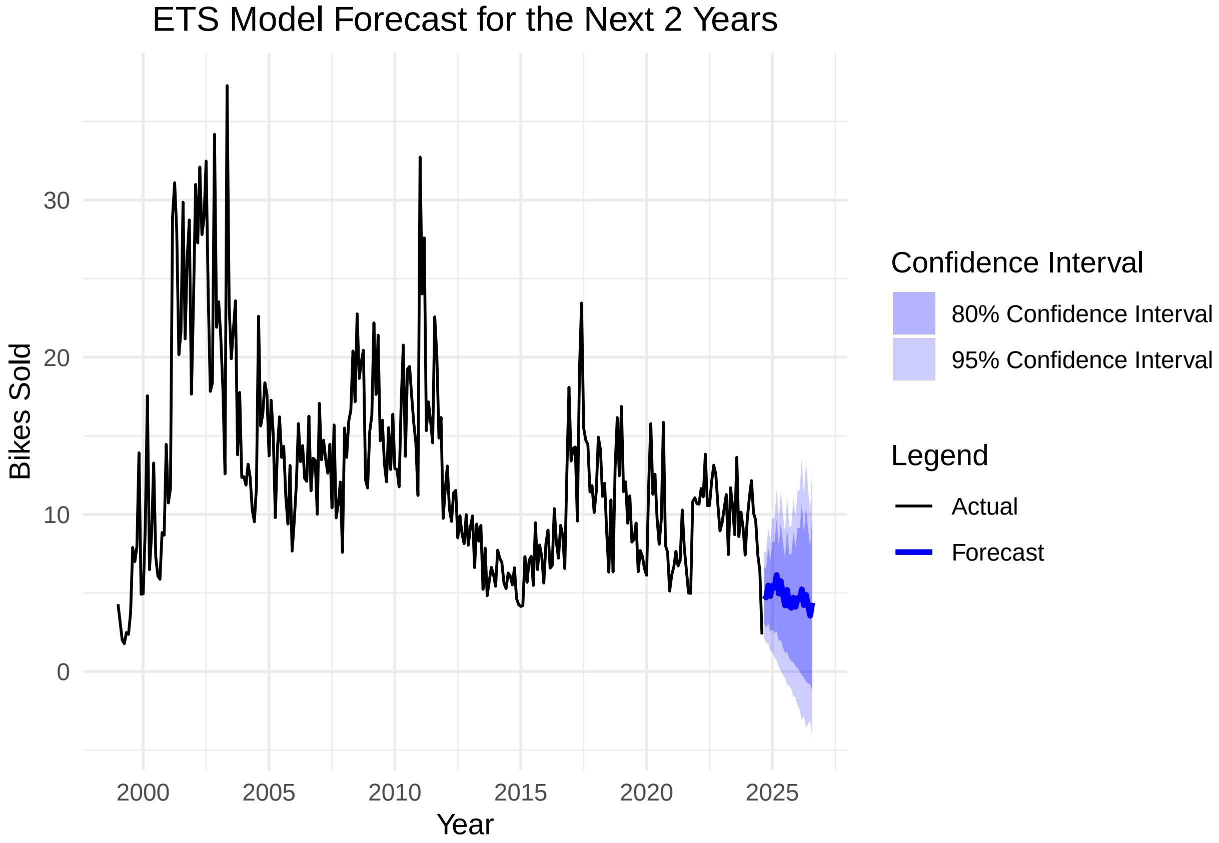
Overall

- The ETS model appears to perform better overall, as indicated by less autocorrelation in the residuals, a tighter distribution around zero, and fewer pronounced patterns in the time series plot.
- The Holt-Winters model shows more significant residual patterns and auto-correlations, suggesting it may not be capturing all the important components of the data. Nonetheless, this model performed better than the others (ARIMA, Linear Regression and Random Forest).

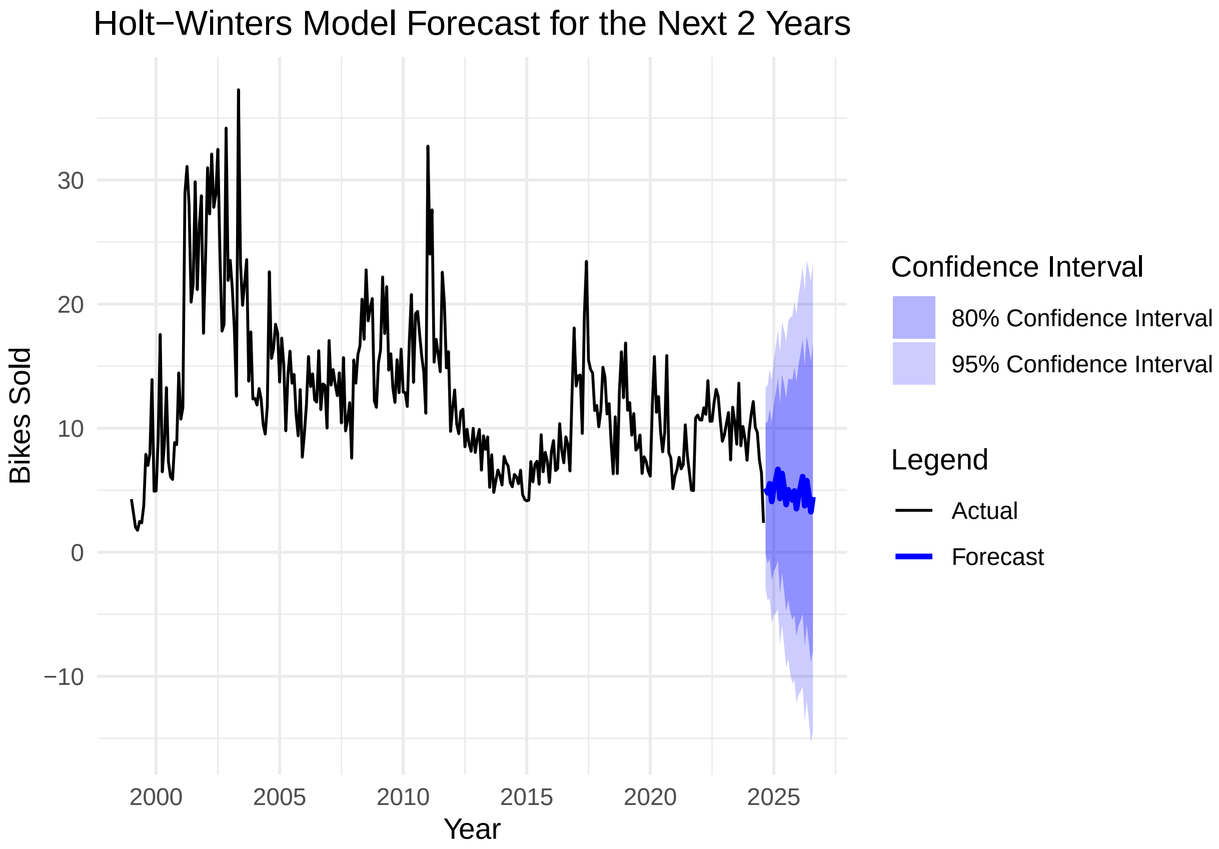
Forecasting

Forecasting for the next 2 years

ETS - Exponential Smoothing (ETS) Model



Holt-Winters Model



Both forecasts predict a continued decline in bike sales over the next two years.