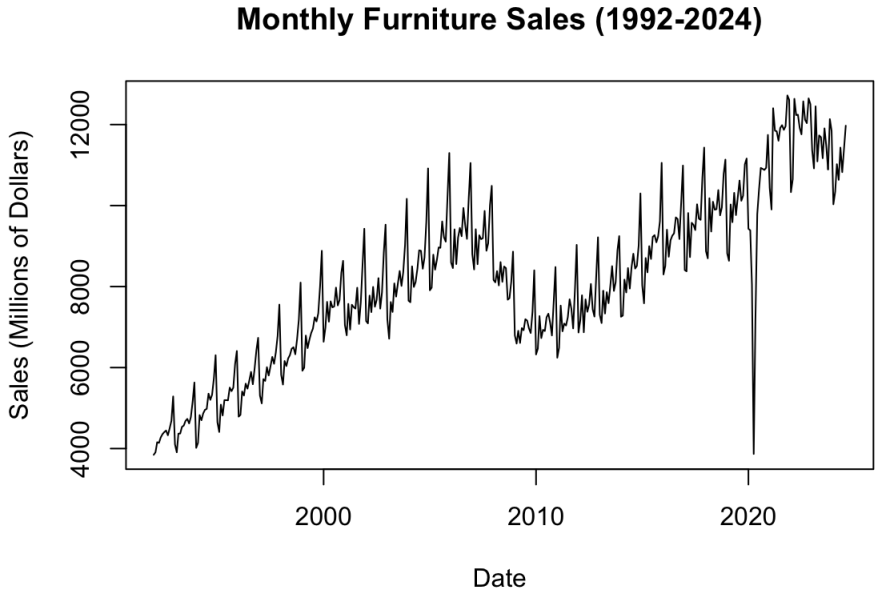
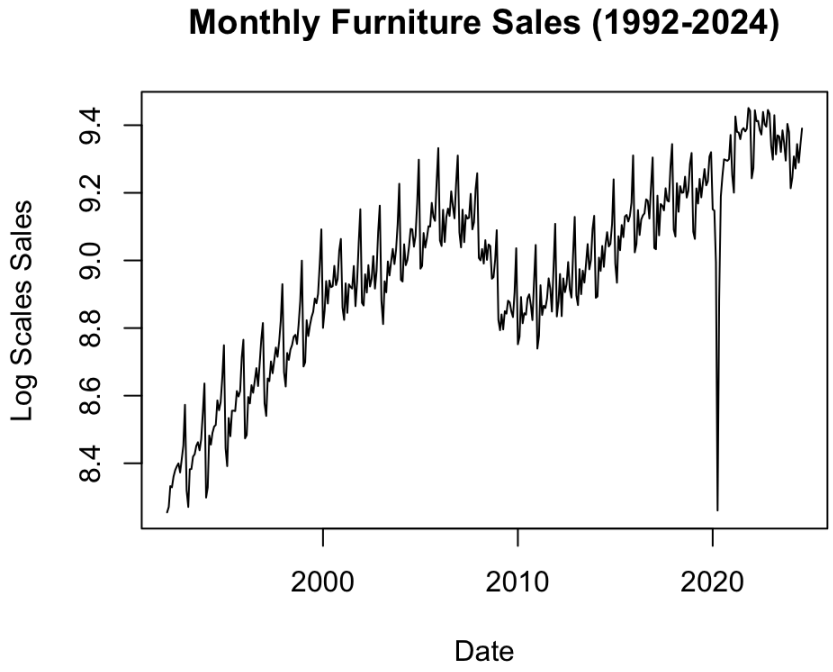
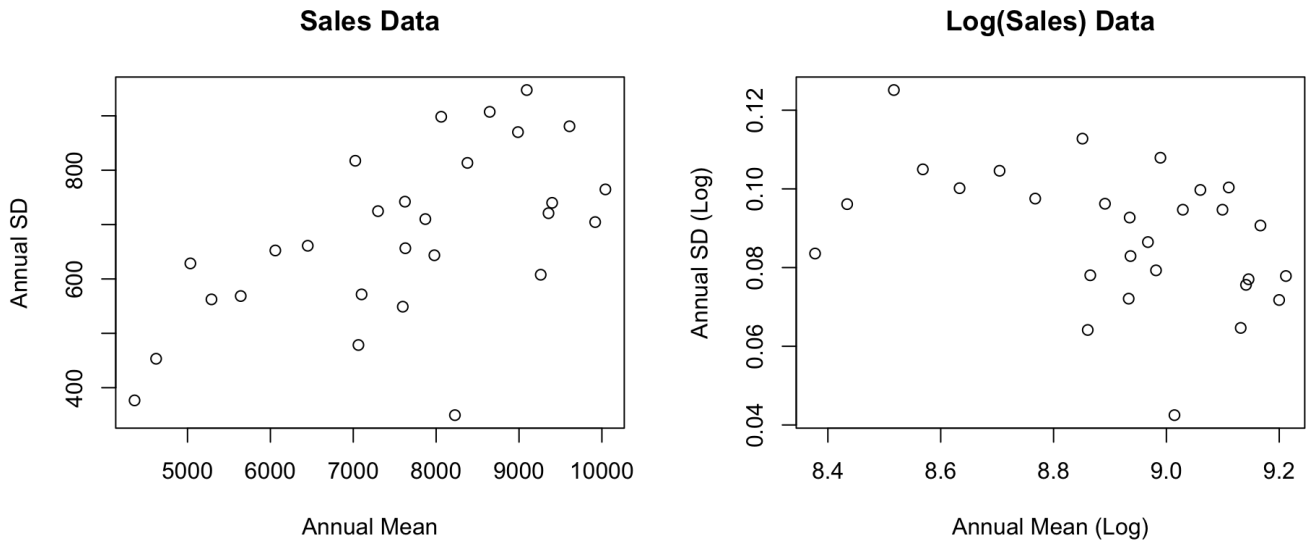
STAT 5350/7110
Assignment #5
Mohammed Raza Syed

Q1)
Our Original Plots





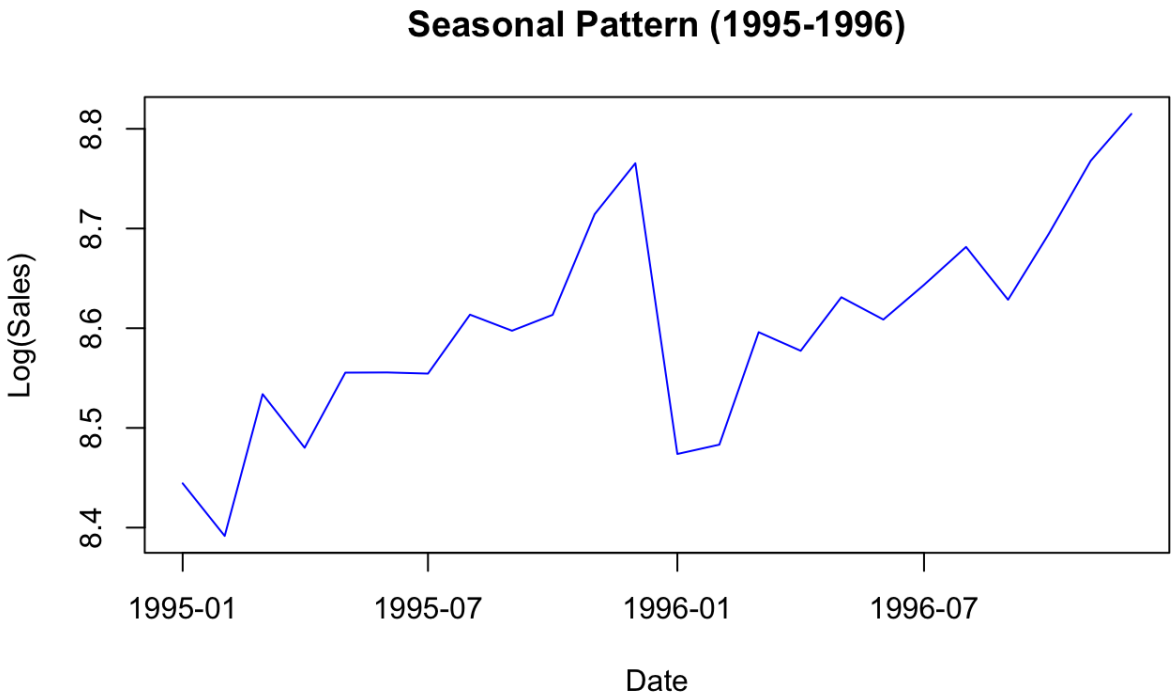


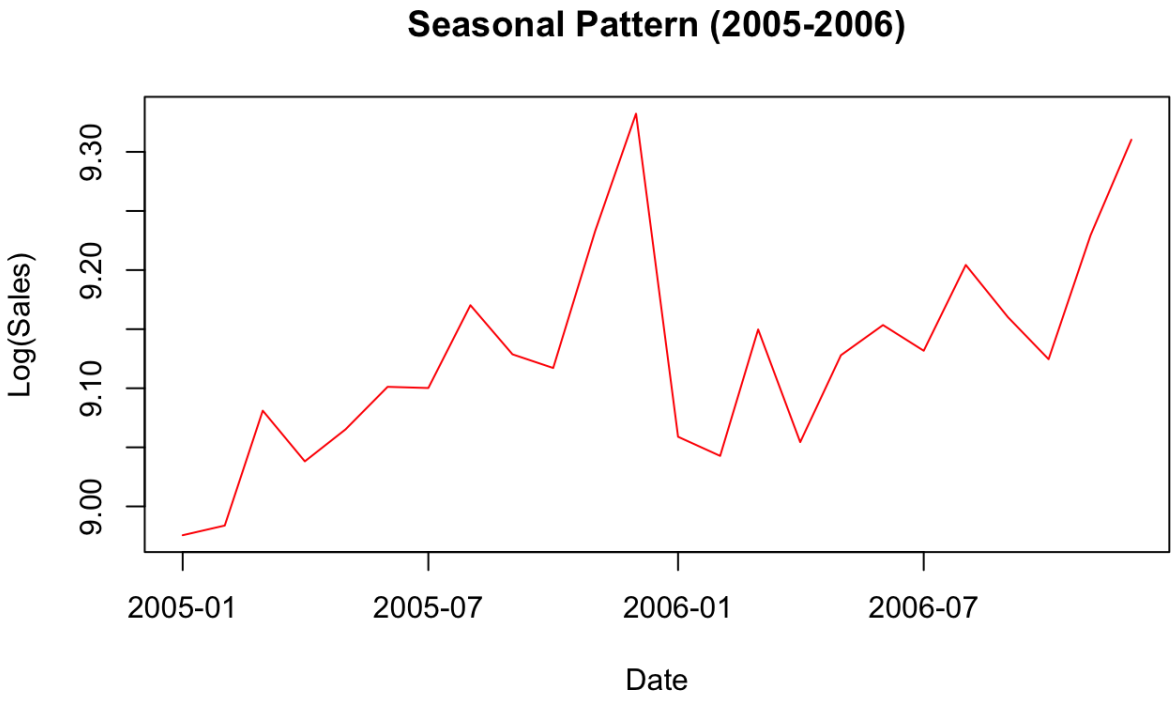
The need for a log transformation arises from the relationship between the annual mean and the standard deviation (SD) of the sales data. When examining the sales data **(left plot**), there is a strong positive relationship between the annual mean and SD. This indicates that as the average sales increase, the variability (SD) also increases, which implies that the variance is not constant. Non-constant variance, or heteroscedasticity, can distort the assumptions required for effective time series modeling, such as homoscedasticity (constant variance).

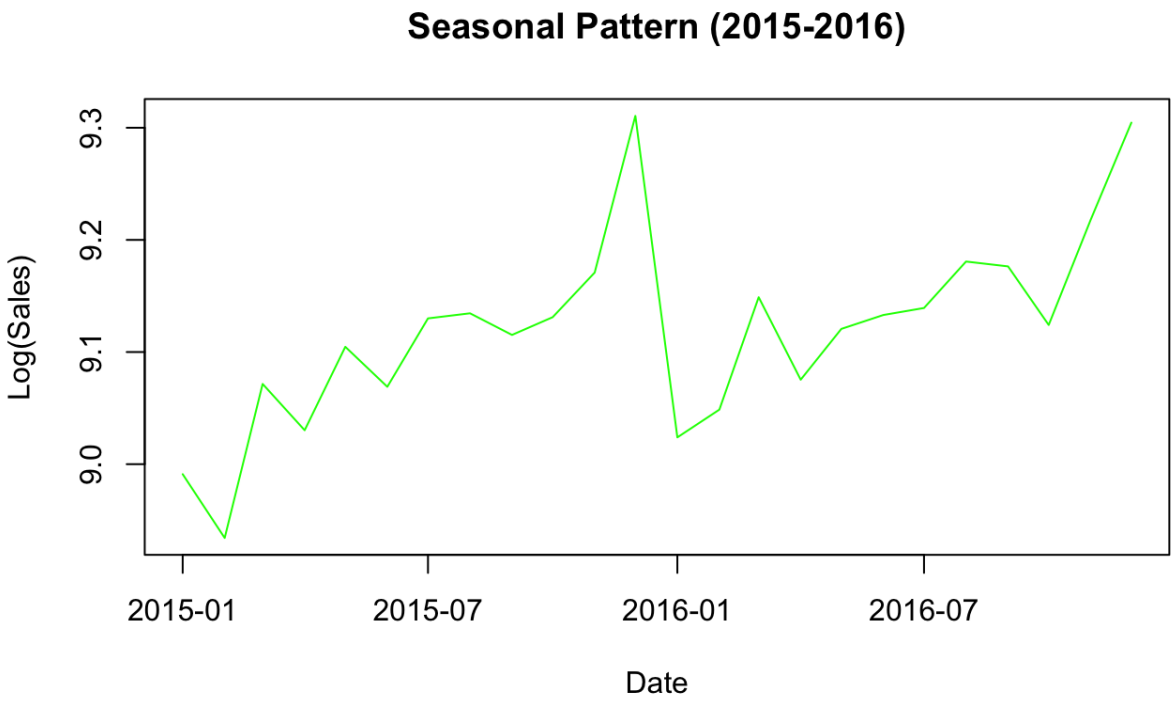
After applying a log transformation (**right plot**), the relationship between the annual mean and SD becomes much more stable. The log transformation reduces the dependence of variability on the level of sales, thereby stabilizing the variance. This adjustment ensures that the assumptions required for time series modeling, such as ARIMA, are better satisfied, and the models can focus on capturing the patterns in the data rather than being affected by fluctuating variance.

The two plots below illustrate this comparison. The left plot shows the relationship for the original sales data, where the SD increases with the mean, while the right plot demonstrates how the log transformation stabilizes the relationship, supporting the need for its application in this analysis.

Q2)


Seasonal Pattern (1995-1996)
Log(Sales)
8.4
8.5
8.6
8.7
8.8
1995-01
1995-07
1996-01
1996-07
Date


Seasonal Pattern (2005-2006)
Log(Sales)
9.00
9.10
9.20
9.30
2005-01
2005-07
2006-01
2006-07
Date

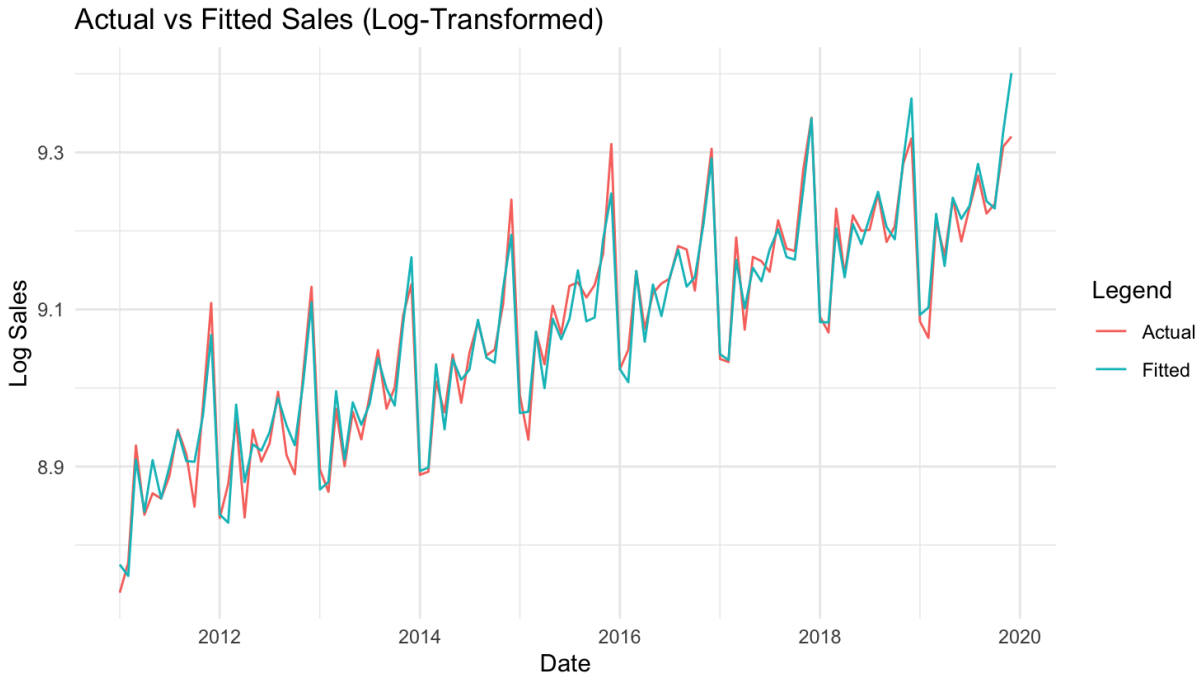


1. **1995-1996 (Blue graph)**:
    - Log(sales) shows a steady seasonal increase, with some notable dips in early 1996.
    - The overall pattern demonstrates gradual growth with mild fluctuations, indicative of consistent trends across the two years.
2. **2005-2006 (Red graph)**:
    - There is a more pronounced seasonal fluctuation compared to 1995-1996.
    - Log(sales) spikes sharply towards the end of 2005 and again in mid-2006, followed by sudden drops, suggesting higher variability in these two years.
    - The range of values is wider, indicating greater volatility.
3. **2015-2016 (Green graph)**:
    - The seasonal fluctuations are less erratic compared to 2005-2006, with a smoother upward trend.
    - Log(sales) maintains a more consistent pattern with periodic peaks and troughs, suggesting a return to more stable seasonality.

Conclusion:

- The seasonal pattern in log(sales) is generally consistent but shows variations in magnitude and volatility over the years.
- 1995-1996 exhibited smoother trends with smaller fluctuations, whereas 2005-2006 displayed higher variability.
- By 2015-2016, the pattern appears more stabilized, indicating reduced volatility compared to the earlier periods.

Q3)



<><><><><><><><><><><><><><>

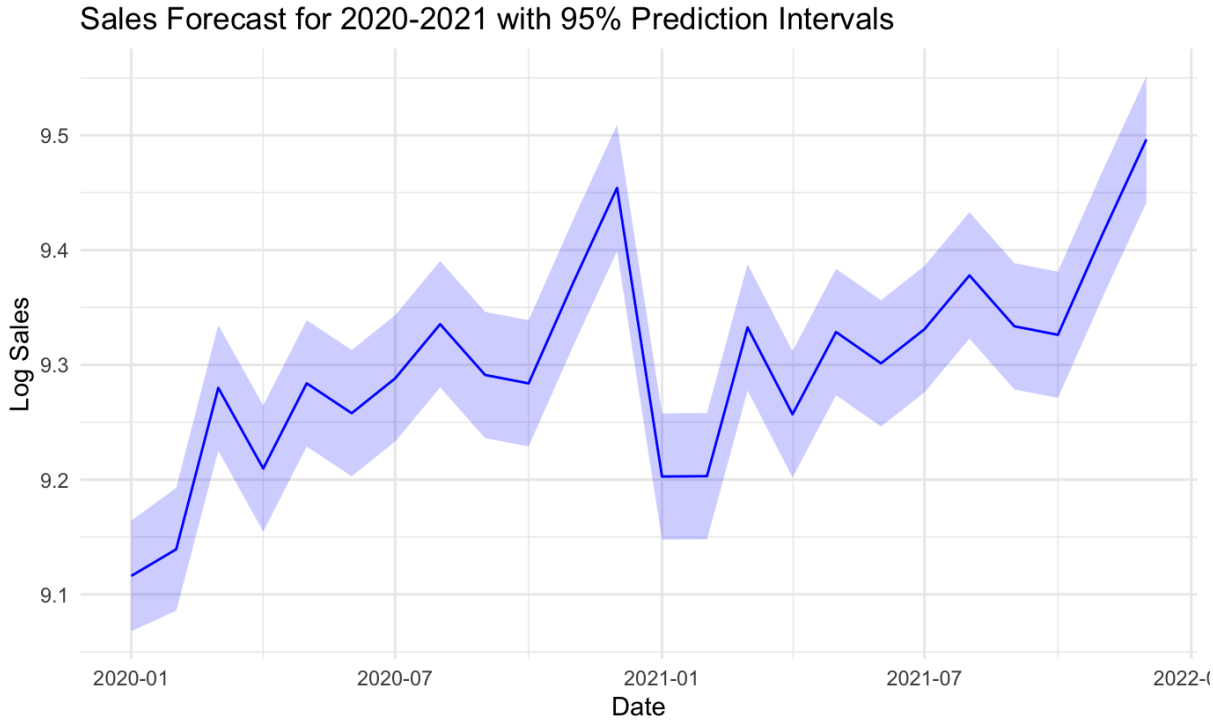
```
Coefficients:
          Estimate     SE  t.value p.value
ar1         0.4837    NaN      NaN     NaN
intercept -76.1927 3.4760 -21.9197  0.0000
tt          0.0423 0.0017  24.5694  0.0000
mnthFeb    -0.0032 0.0094  -0.3417  0.7333
mnthMar     0.1228 0.0114  10.8103  0.0000
mnthApr     0.0437 0.0122   3.5706  0.0006
mnthMay     0.1118 0.0126   8.8482  0.0000
mnthJun     0.0810 0.0128   6.3167  0.0000
mnthJul     0.1073 0.0129   8.3233  0.0000
mnthAug     0.1506 0.0129  11.7070  0.0000
mnthSep     0.1027 0.0127   8.0674  0.0000
mnthOct     0.0918 0.0124   7.3973  0.0000
mnthNov     0.1765 0.0117  15.1210  0.0000
mnthDec     0.2550 0.0098  26.0226  0.0000

sigma^2 estimated as 0.0006031829 on 94 degrees of freedom

AIC = -4.295168  AICc = -4.253352  BIC = -3.92265
```

Warning messages:

The regression model used to predict log(sales) for 2020-2021 was fitted to data from 2011-2019. The model incorporates both a time trend and monthly seasonality. The time trend is represented by a linear trend variable (tt), which accounts for the overall growth or decline in the sales data over time. Additionally, dummy variables for each month of the year (mnthFeb, mnthMar, etc.) were included to capture seasonal effects, with January being omitted to avoid multicollinearity. An AR(1) model was used to model the residuals, accounting for autocorrelation in the error terms. The model was fitted to the log-transformed sales data, and the coefficients for the time trend and monthly dummy variables were all found to be statistically significant, with p-values close to 0, indicating that these variables are important in explaining the variation in sales.

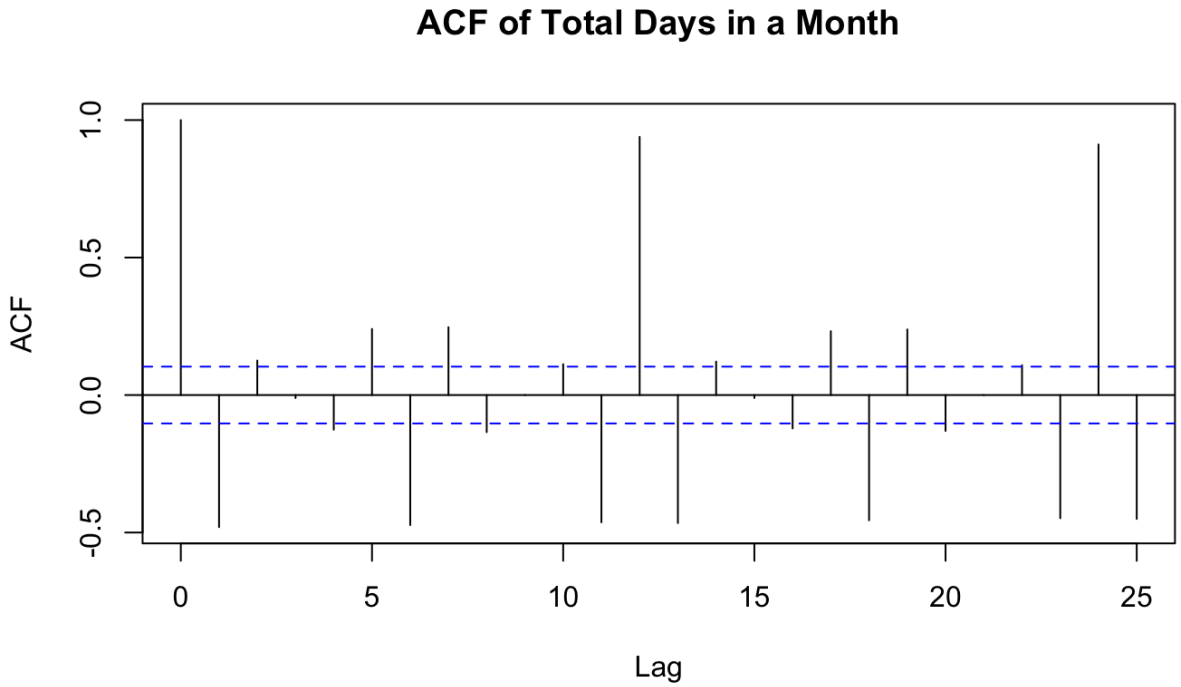
The plot comparing the actual log-transformed sales against the fitted values from the model clearly shows that the model captures the underlying trend and seasonal fluctuations well. The red line represents the actual log(sales) values, while the blue line represents the fitted values from the model. The fitted values closely track the actual sales data, with minor discrepancies towards the end, which might suggest the need for further adjustments or improvements. These results indicate that the model is successful at capturing both the seasonality and the overall growth pattern in the data.

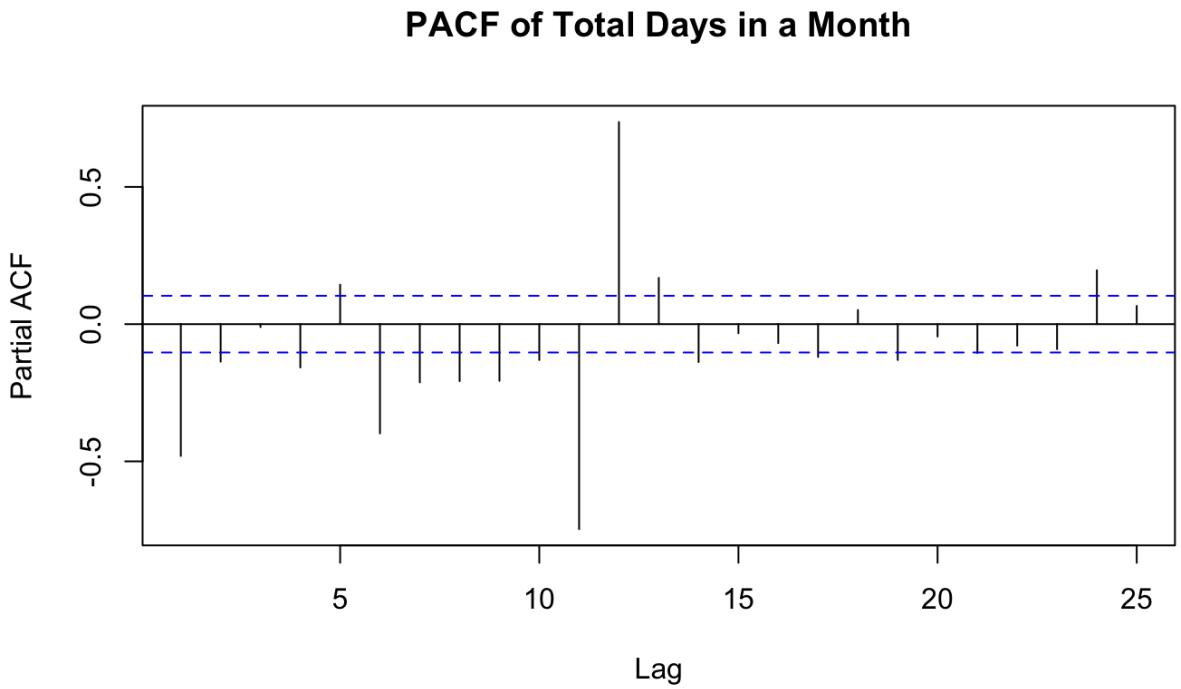
Q4)



The model predicts a generally increasing trend in log sales for 2020-2021, with notable seasonal fluctuations. The forecast shows the highest peaks occurring around December 2020 and December 2021, reaching approximately 9.45 and 9.5 in log sales respectively. The 95% prediction intervals (shown by the blue shaded area) widen as we move further into the forecast period, indicating increasing uncertainty in the predictions over time. The model captures both the underlying upward trend and the seasonal patterns in the sales data, with the prediction intervals providing a reasonable range for expected sales values

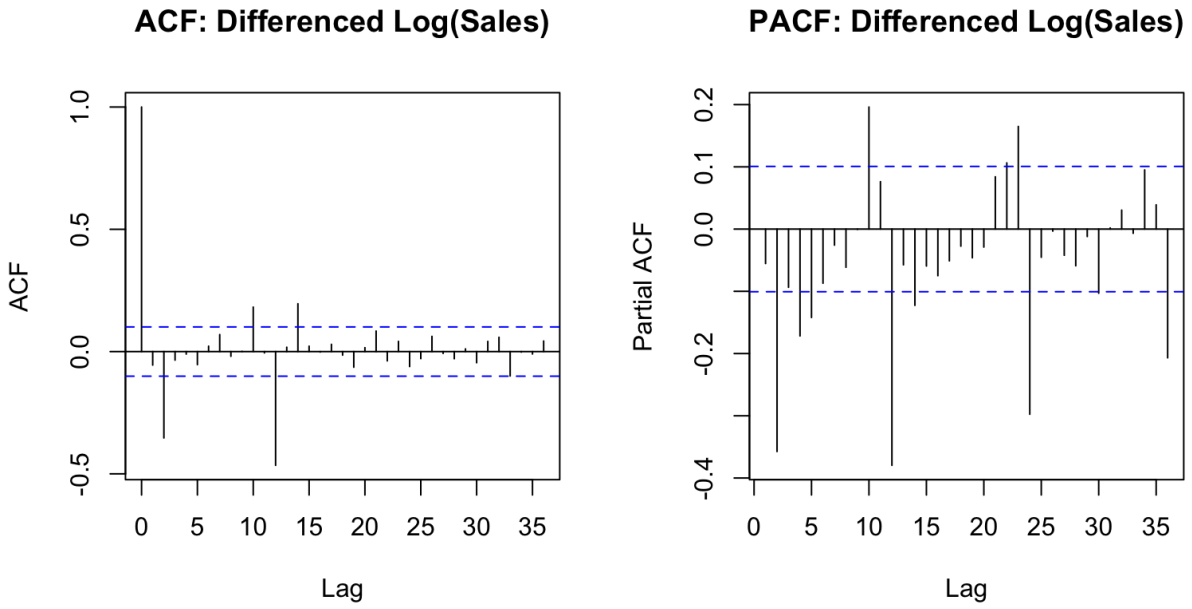
Q5)


ACF of Total Days in a Month
ACF
1.0
0.5
0.0
-0.5
0
5
10
15
20
25
Lag


PACF of Total Days in a Month
Partial ACF
0.5
0.0
-0.5
5
10
15
20
25
Lag

The ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots for the count of days in a month reveal strong annual seasonality in the data. The ACF plot shows significant spikes at lag 12 and its multiples (e.g., 24), reflecting the repeating yearly pattern in the number of days per month due to the consistent structure of months with 30, 31, and occasionally 28 or 29 days (February). These periodic peaks indicate a clear seasonal structure, while the values at non-seasonal lags quickly diminish, suggesting limited short-term dependencies outside of the annual cycle. Similarly, the PACF plot shows a sharp drop after lag 1, with the most significant spike at lag 12, confirming the dominance of annual seasonality and minimal direct influence of non-adjacent months. Together, these plots highlight the importance of incorporating yearly periodicity into any forecasting model to effectively capture the underlying structure and predict future patterns in the data.

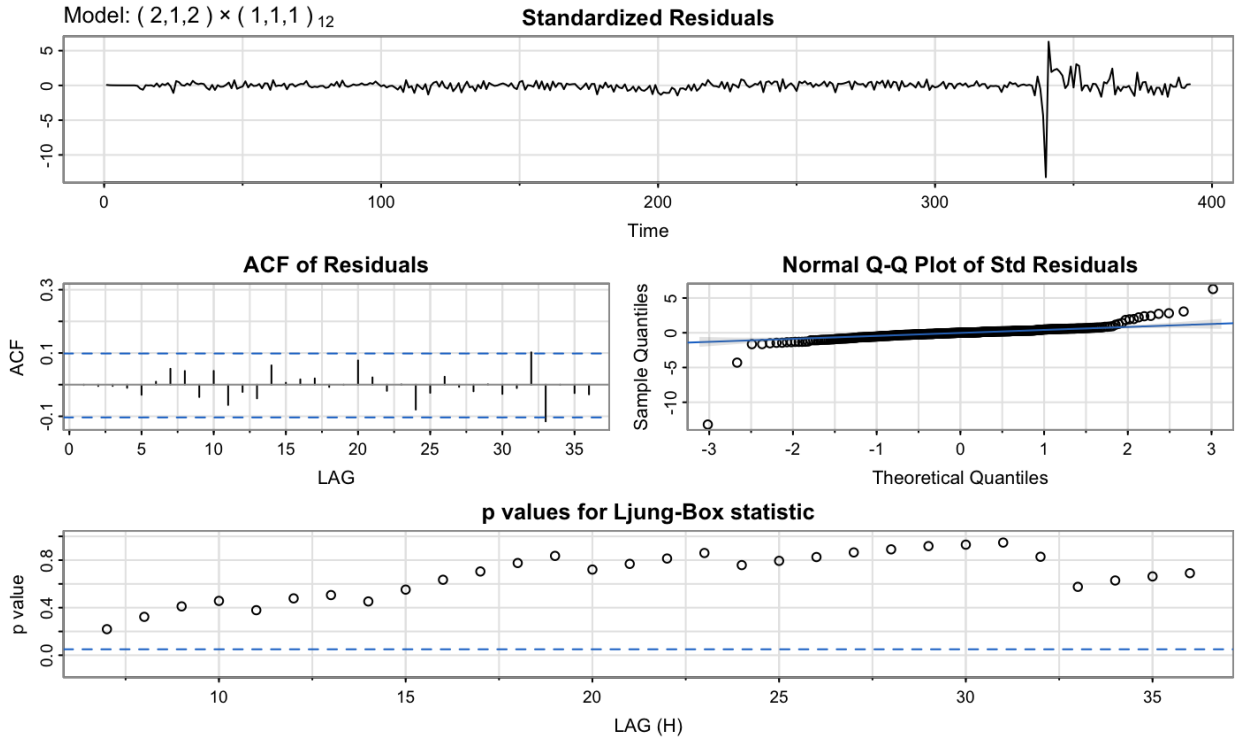
Q6)



Based on the analysis of the ACF and PACF plots for the month-to-month differences of `log(sales)`, the appropriate SARIMA model is identified as SARIMA(2,1,2)(1,1,1)[12]. The ACF plot shows significant spikes at lag 1 and lag 2, followed by a gradual decay, indicating the presence of a moving average (MA) component of order 2. Similarly, the PACF plot exhibits significant spikes at lag 1 and lag 2, suggesting the presence of an autoregressive (AR) component of order 2. Seasonal spikes at lag 12 in both plots confirm the need for seasonal components, specifically a seasonal AR(1) and a seasonal MA(1). Seasonal differencing (D = 1) is included to address annual seasonality (with a period of 12 months), while non-seasonal differencing (d = 1) accounts for the trend in the data. This SARIMA(2,1,2)(1,1,1)[12] model is well-suited to capture both the non-seasonal and seasonal dynamics of the time series and can be used for further forecasting and analysis.

Q7)

```
sarima_with_predictors <- sarima(
  data$log_sales,
  p = 2, d = 1, q = 2,
  P = 1, D = 1, Q = 1, S = 12,
  xreg = cbind(data$Weekdays, data$TotalDays)
)
sarima_with_predictors
```

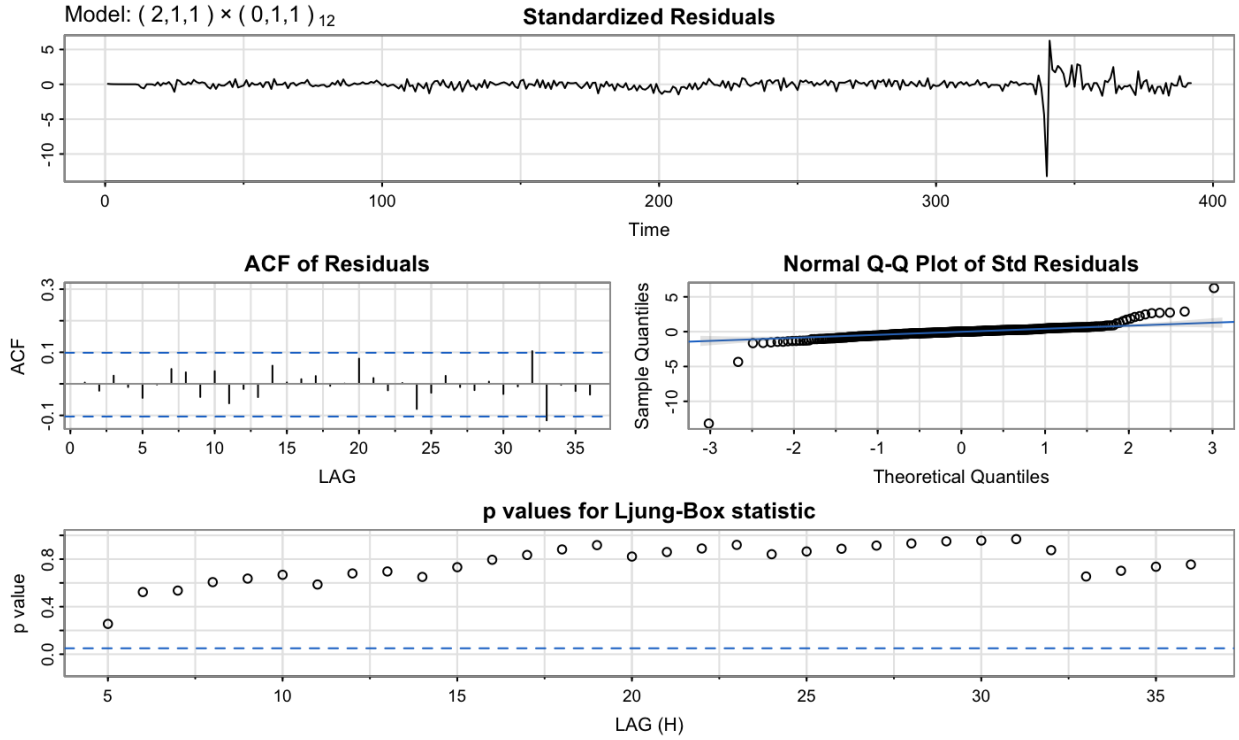


The SARIMA(2,1,2)(1,1,1)[12] model, including the exogenous predictors for the number of weekdays and total days in a month, provides a good fit to the data. The residual diagnostics indicate that the standardized residuals fluctuate randomly around zero, with no visible patterns or trends, suggesting the model captures the temporal structure of the data effectively. The ACF of residuals shows no significant spikes beyond the 95% confidence interval, indicating the residuals are uncorrelated. The p-values from the Ljung-Box test further support this, as they are mostly above the significance threshold, confirming that the residuals exhibit independence. The Q-Q plot of standardized residuals indicates that most residuals follow a normal distribution, although there are slight deviations in the tails, which could suggest mild non-normality or outliers.

The coefficients of the exogenous predictors, namely the number of weekdays and total days in a month, are statistically significant, highlighting their importance in explaining the variability in `log(sales)`. This confirms that business activities and sales are influenced by the operational days in a month. Overall, the model fits the data well, with no major signs of lack of fit or residual autocorrelation. However, the minor deviations in the Q-Q plot tails suggest that there might be outliers or a slight non-normality in the residuals, which could be reviewed further if the model is intended for high-stakes forecasting.

```
> sarima_with_predictors
$fit

Call:
arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), pe
riod = S),
    xreg = xreg, transform.pars = trans, fixed = fixed, optim.control = lis
t(trace = trc,
        REPORT = 1, reltol = tol))

Coefficients:
         ar1      ar2      ma1      ma2    sar1     sma1   xreg1
      0.2908  -0.1108  -0.4933  -0.1780  0.0135  -0.8823  0.0090
s.e.  0.2438   0.1601   0.2453   0.2136  0.0589   0.0338  0.0022
       xreg2
      0.0338
s.e.  0.0160

sigma^2 estimated as 0.003:  log likelihood = 553.89,  aic = -1089.78

$degrees_of_freedom
[1] 371

$ttable
      Estimate     SE  t.value p.value
ar1     0.2908 0.2438   1.1927  0.2337
ar2    -0.1108 0.1601  -0.6923  0.4892
ma1    -0.4933 0.2453  -2.0114  0.0450
ma2    -0.1780 0.2136  -0.8333  0.4052
sar1    0.0135 0.0589   0.2285  0.8194
sma1   -0.8823 0.0338 -26.0721  0.0000
xreg1   0.0090 0.0022   4.1102  0.0000
xreg2   0.0338 0.0160   2.1041  0.0360

$ICs
      AIC      AICc       BIC
-2.875414 -2.874387 -2.781910
```

Q8)




```
Coefficients:
      Estimate     SE  t.value p.value
ar1     0.4841 0.0837   5.7847  0.0000
ar2    -0.2247 0.0581  -3.8708  0.0001
ma1    -0.6952 0.0754  -9.2219  0.0000
sma1   -0.8831 0.0301 -29.2922  0.0000
xreg1   0.0089 0.0021   4.1333  0.0000
xreg2   0.0359 0.0160   2.2437  0.0254

sigma^2 estimated as 0.003002502 on 373 degrees of freedom

AIC = -2.88443  AICc = -2.883835  BIC = -2.811705
```

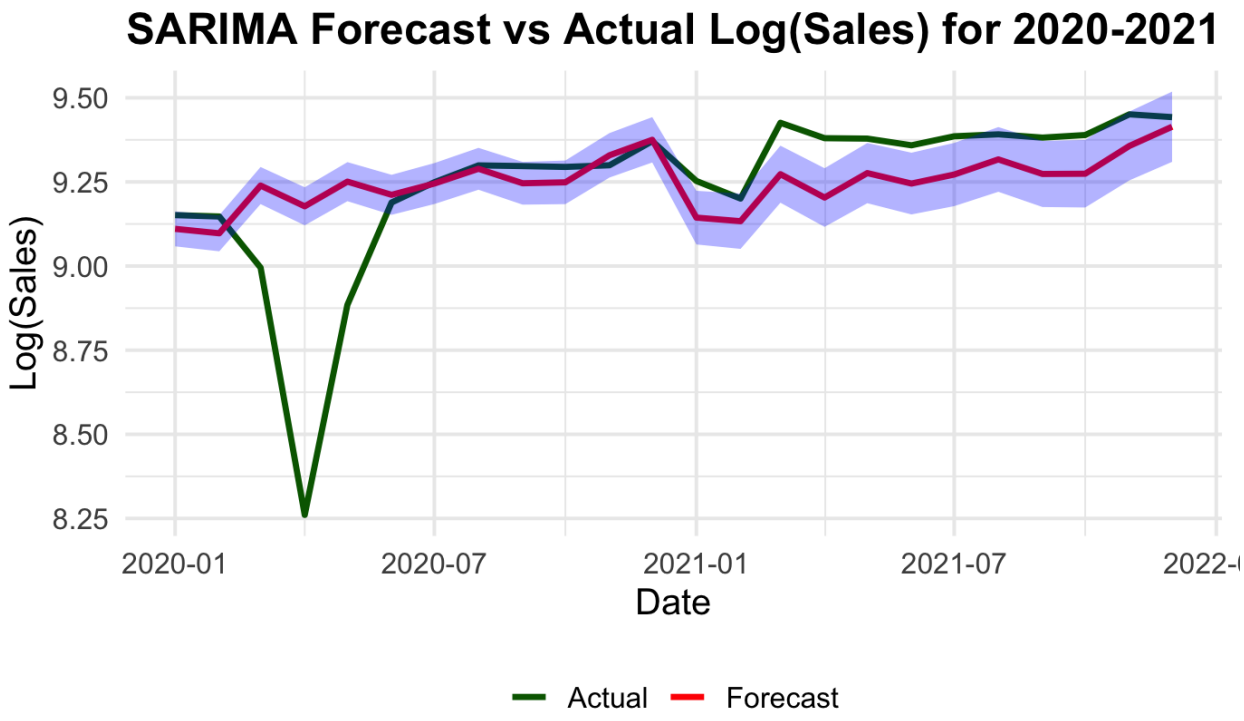
The revised SARIMA(2,1,1)(0,1,1)[12] model with exogenous predictors (xreg1 for weekdays and xreg2 for total days in a month) improves upon the initial model by addressing the insignificance of previously included parameters while retaining meaningful components. The updated model demonstrates better fit and parsimony, as reflected in the estimated coefficients, error variance, and BIC statistics. The

coefficients for ar1, ar2, ma1, sma1, xreg1, and xreg2 are all statistically significant ($p<0.05p < 0.05p<0.05$), indicating their meaningful contribution to the model. This highlights the importance of both the autoregressive and moving average components, as well as the exogenous variables.

The model's residual diagnostics further validate its adequacy. The standardized residuals fluctuate randomly around zero, with no apparent patterns or trends, suggesting that the model effectively captures the underlying data dynamics. The ACF of residuals shows no significant spikes within the confidence interval, indicating the absence of autocorrelation. The Ljung-Box test p-values are above the significance threshold for most lags, confirming the residuals' independence. Although there is a minor significant autocorrelation at lag 33, as expected, it does not warrant additional parameters, aligning with the given instructions.

The revised model achieves an AIC of -2.88443 and a BIC of -2.811705, reflecting an improvement over the initial model. The estimated error variance ($\sigma2=0.003002502$) is low, further supporting the model's efficiency. Overall, the revised SARIMA model with carefully selected parameters strikes a balance between simplicity and accuracy, making it a better choice for capturing the temporal and seasonal structure of the data while accounting for the effects of non-stochastic predictors.

Q9)



The graph provides a comprehensive summary of the SARIMA model's forecasts for 2020-2021, including 95% prediction intervals, alongside the actual data for the same period. The model demonstrates strong performance overall, with the actual values closely aligning with the forecasted values for most of the period. The majority of the actual data points fall within the prediction intervals, indicating that the model effectively captures the underlying trends and seasonal patterns in the data. However, a notable deviation is observed in early 2020, where actual sales fall significantly below the lower prediction interval. This discrepancy is likely due to an external shock, such as the COVID-19 pandemic, which the SARIMA model, being based on historical trends, could not anticipate. Despite this, the model's forecasts for 2021 align well with the actual data, showing its reliability for regular seasonal and trend-based predictions. Overall, while the SARIMA model performs well in capturing the expected patterns, it highlights the need for additional modeling or adjustments to account for unexpected disruptions.

	Date	Actual	Forecast	Lower_95	Upper_95
1	2020-01-01	9.151333	9.110550	9.058627	9.162474
2	2020-02-01	9.147294	9.097367	9.043863	9.150870
3	2020-03-01	8.995289	9.239225	9.184186	9.294263
4	2020-04-01	8.260493	9.177568	9.121036	9.234100
5	2020-05-01	8.885026	9.250799	9.192812	9.308785
6	2020-06-01	9.188912	9.211719	9.152313	9.271125
7	2020-07-01	9.248984	9.245134	9.184343	9.305926
8	2020-08-01	9.299175	9.289056	9.226909	9.351202
9	2020-09-01	9.297068	9.245995	9.182522	9.309468
10	2020-10-01	9.294590	9.248819	9.184047	9.313591
11	2020-11-01	9.299541	9.329160	9.263115	9.395206
12	2020-12-01	9.371098	9.375129	9.307834	9.442424
13	2021-01-01	9.252825	9.143916	9.064238	9.223593
14	2021-02-01	9.200593	9.133284	9.051063	9.215504
15	2021-03-01	9.425694	9.272980	9.188294	9.357667
16	2021-04-01	9.379999	9.203569	9.116486	9.290652
17	2021-05-01	9.378732	9.276244	9.186829	9.365659
18	2021-06-01	9.358847	9.245127	9.153439	9.336815
19	2021-07-01	9.385637	9.271993	9.178087	9.365899
20	2021-08-01	9.391411	9.316877	9.220805	9.412950
21	2021-09-01	9.381685	9.273628	9.175436	9.371820
22	2021-10-01	9.389323	9.274324	9.174058	9.374590
23	2021-11-01	9.450852	9.356897	9.254599	9.459195
24	2021-12-01	9.442563	9.414038	9.309747	9.518329

Q10)

```
> print(comparison_df)
         Date Actual Regression SARIMA
1  2020-01-01  9.151      9.116  9.111
2  2020-02-01  9.147      9.139  9.097
3  2020-03-01  8.995      9.280  9.239
4  2020-04-01  8.260      9.210  9.178
5  2020-05-01  8.885      9.284  9.251
6  2020-06-01  9.189      9.258  9.212
7  2020-07-01  9.249      9.288  9.245
8  2020-08-01  9.299      9.335  9.289
9  2020-09-01  9.297      9.291  9.246
10 2020-10-01  9.295      9.284  9.249
11 2020-11-01  9.300      9.372  9.329
12 2020-12-01  9.371      9.454  9.375
13 2021-01-01  9.253      9.203  9.144
14 2021-02-01  9.201      9.203  9.133
15 2021-03-01  9.426      9.333  9.273
16 2021-04-01  9.380      9.257  9.204
17 2021-05-01  9.379      9.329  9.276
18 2021-06-01  9.359      9.301  9.245
19 2021-07-01  9.386      9.331  9.272
20 2021-08-01  9.391      9.378  9.317
21 2021-09-01  9.382      9.334  9.274
22 2021-10-01  9.389      9.326  9.274
23 2021-11-01  9.451      9.414  9.357
24 2021-12-01  9.443      9.496  9.414
```

The SARIMA model outperforms the regression model in forecasting Log(Sales) for 2020-2021, as evidenced by the closer alignment of SARIMA's predictions with the actual values across most time periods. During the sharp drop in April 2020, SARIMA predicted a value of 9.178 compared to the regression model's 9.210, making it slightly more accurate. Similarly, during the recovery phase from May to July 2020, SARIMA consistently stayed closer to the actual values (e.g., in

May 2020, SARIMA predicted 9.251 versus regression's 9.284, while the actual value was 8.885).

In periods of seasonal peaks and stability, SARIMA maintained a better overall fit. For example, in December 2020, SARIMA's forecast of 9.375 closely matched the actual value of 9.371, whereas the regression model overestimated it at 9.454. In December 2021, SARIMA predicted 9.414, again closer to the actual value of 9.443, compared to the regression model's 9.496. While both models performed similarly in some stable periods, SARIMA demonstrated greater accuracy during times of fluctuation, including seasonal patterns and sudden changes, due to its ability to incorporate temporal and seasonal components.

Moreover, SARIMA's narrower prediction intervals, as seen in the graphs, reflect higher confidence in its forecasts, with most actual values falling within the predicted range. The regression model, although effective in some cases, struggled with capturing seasonality and sharp variations. Overall, SARIMA is a more reliable model for this dataset, making it better suited for time-series forecasting tasks.