**Assignment 3: MGSC 661**

**Gina Yassin (261261130)**

**Othmane Zizi (261255341)**

**Helia Mahmood-Zadeh (261224416)**

**Submission Instruction:** Please submit

1. A completed word document with all your text-based written responses and results from R (screenshot/typed answer). In your submitted word file, you only need to replace the red parts with your answers.
2. The completed R script.

***Example 1: Boston Bikesharing***

For this example, you will apply what you learn to a data set of bike-sharing system. The data set contains the number of daily trips from a station near MIT, time range is Jan 1st 2015 to Nov 30th 2018.

A building with a triangular roof

AI-generated content may be incorrect.

**Q1:** Read bikesharing.csv into R and store it in bikedata. How many data points do we have?

***Answer:***

**There are 1424 observations in the data.**

**Q2** Create a time series object for the number of daily trips (i.e., bikedata$trip) using command tsbike = ts(…,…).

**no need to include any answer for Q2**

**Q3** Plot tsbike. Label the y-axis “Trips” and title the plot “Daily Bluebike Trips”. Briefly describe the patterns that you see in the plot.

***Answer:***

Our plot is:



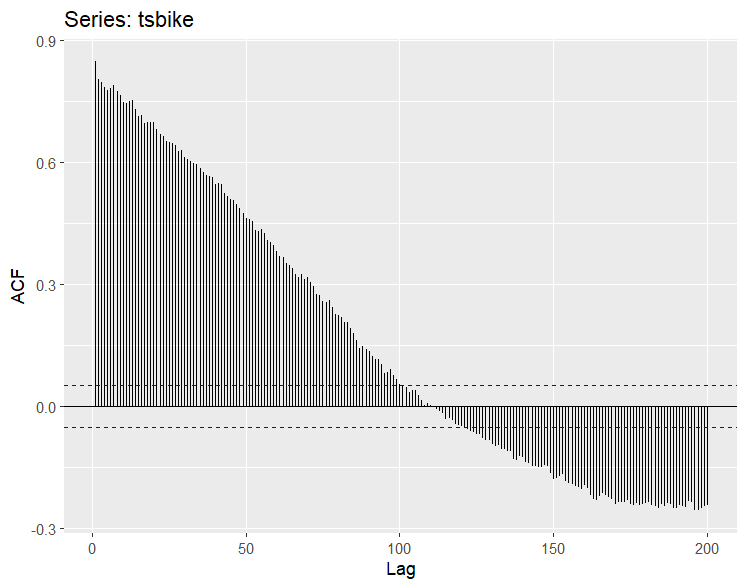
Our observations are

**The series shows a clear upward trend in daily Bluebike trips, with early values much lower than the peaks seen in later years. It exhibits a repeating yearly pattern with regular peaks and troughs occurring at consistent intervals, indicating strong seasonality. These recurring peaks likely correspond to warmer months when ridership is typically higher, while the troughs align with colder periods when usage tends to drop. Day-to-day values are highly variable, with sharp rises and drops that reflect the influence of short-term factors. The size of the seasonal swings increases over time, suggesting that variation grows alongside the overall trend. The series also includes several days with near-zero trips, representing isolated extreme low-usage events.**

**Q4** Plot the ACF for the first 200 lags. Briefly describe what you see and what it means?

***Answer:***

Our plot is



Our observations are

**The ACF starts very high at lag 1 (around 0.85–0.90) and decreases slowly and smoothly as the lag increases. The autocorrelations remain above the significance bands for many lags before gradually tapering toward zero, and the function eventually crosses zero and becomes slightly negative around lags 100–150. This slow, persistent decay indicates strong dependence in the series, meaning past values have a long-lasting influence on future values. Overall, the pattern suggests non-stationarity with trend and clear underlying structure, so the data are far from random and must be modeled accordingly.**

**Q5** Now, plot the ACF for the first 800 lags. Briefly describe what you see.

***Answer:***

Our plot is

A graph with a black line

AI-generated content may be incorrect.

Our observations are

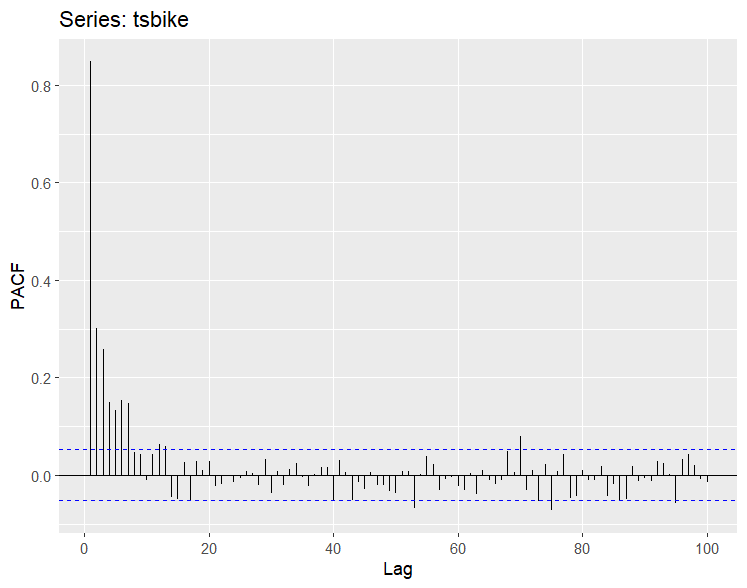
**The ACF begins very high (close to +1.0) at Lag 1 and decays very slowly, remaining statistically significant for hundreds of lags. This slow, smooth decline indicates strong persistence and non-stationarity, driven by a long-term upward trend in bike trips. The autocorrelation crosses the significance bands around Lag 100 and then gradually becomes negative before rising again, consistent with long-range dependence. This pattern implies that differencing (d = 1) will be necessary to stabilize the mean before fitting an ARIMA model.**

**Unlike the shorter 0–200 lag view, the extended 0–800 ACF plot reveals large, repeating wave patterns. These cycles peak around Lag 365 and Lag 730, corresponding to one and two years of daily data. This confirms strong annual seasonality in the Bluebike trips. Because these seasonal peaks sit atop a slowly decaying baseline, their magnitude is amplified by the underlying trend. As a result, the true strength of seasonality will become clearer only after the series is differenced to remove the trend component.**

**Q6** Plot the PACF for the first 100 lags. Briefly describe what you see and what it means?

***Answer:***

Our plot is



Our insights are:

**The PACF shows one dominant spike at lag 1 (around 0.85), indicating that yesterday’s ridership is the strongest predictor of today’s trips once all other lags are controlled for. Several smaller but still significant spikes appear within the first 10 lags, reflecting short-term dependence over the next few days. After lag 10, the partial autocorrelations fall close to zero and remain inside the significance bands, showing that the series does not have meaningful long-run partial dependence.**

**When compared to the ACF, which decays slowly and shows strong seasonal waves, the PACF clarifies that the long-horizon patterns in the ACF are driven by trend and yearly seasonality rather than by true long-term autoregressive effects. In other words, the series has strong short-run structure but its long-run persistence comes from non-stationarity, not from high-order AR behavior.**

**This pattern suggests that the stationary component of the series is well described by a low-order AR process (likely AR(1)), while the overall series requires differencing and possibly seasonal modeling. As a result, an ARIMA or seasonal ARIMA model would be more appropriate for forecasting than simple naïve methods.**

**Q7** We are now going to test the accuracy of Naive method:

* 1. Use the first 1400 data points as training data (call it train\_data) and the remaining data points as test data (call it test\_data). Create both train\_data and test\_data

**no need to include any answer for this part in this submitted word doc**

* 1. Run Naive method with h = 24 (i.e., we want to forecast what happen in the next 24 days) and store it in bike\_naive. Show the point forecasts together with their 80% and 99% prediction intervals.

***Answer:***

The forecast for the h=24 is:

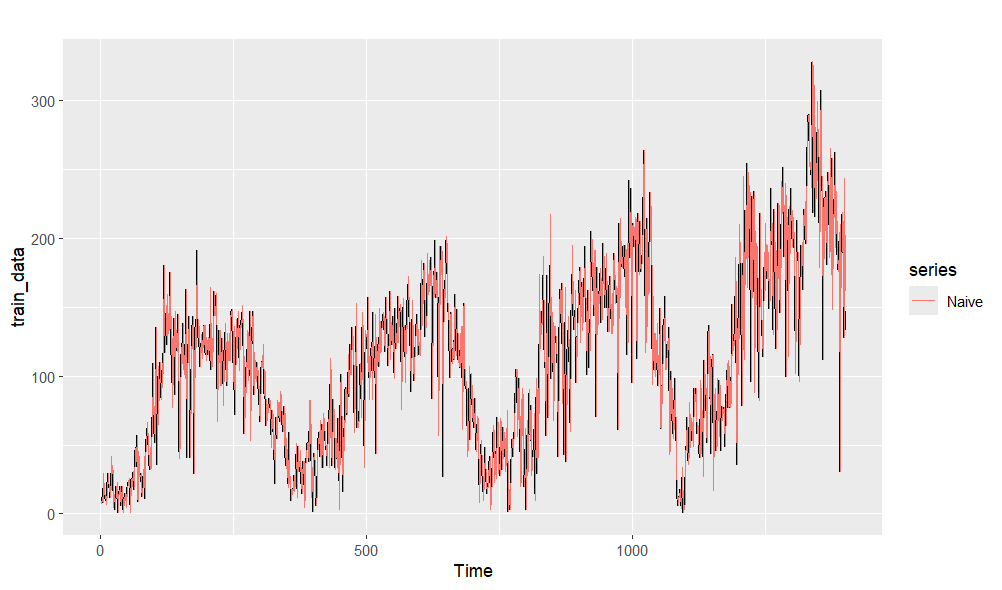
A screenshot of a computer

AI-generated content may be incorrect.

* 1. Plot train\_data and the fitted values under Naive method in the same graph.

***Answer:***

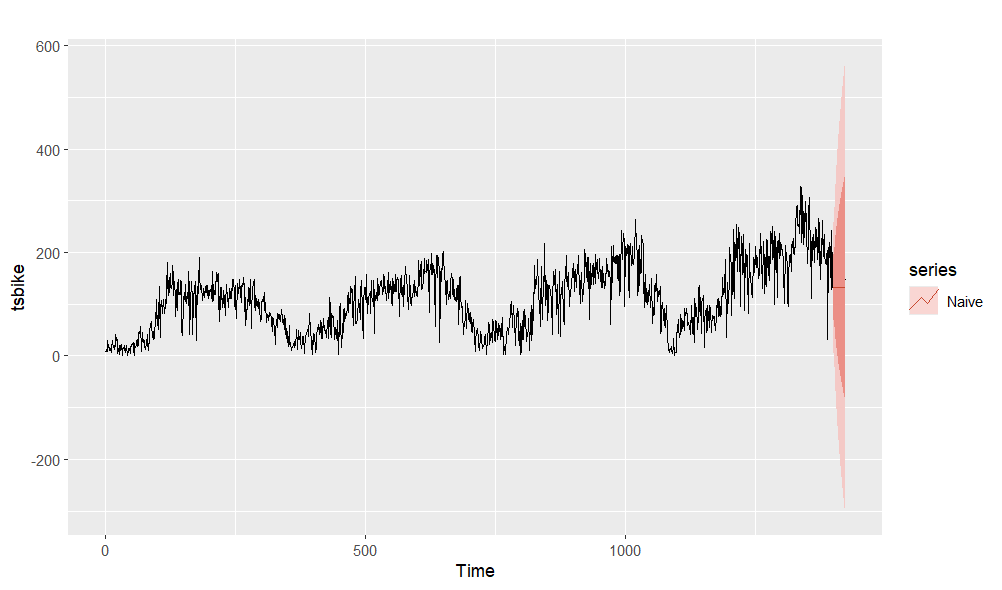
The graph is:



* 1. Plot tsbike together with the point forecasts under Naive method in the same graph.

***Answer:***

The graph is

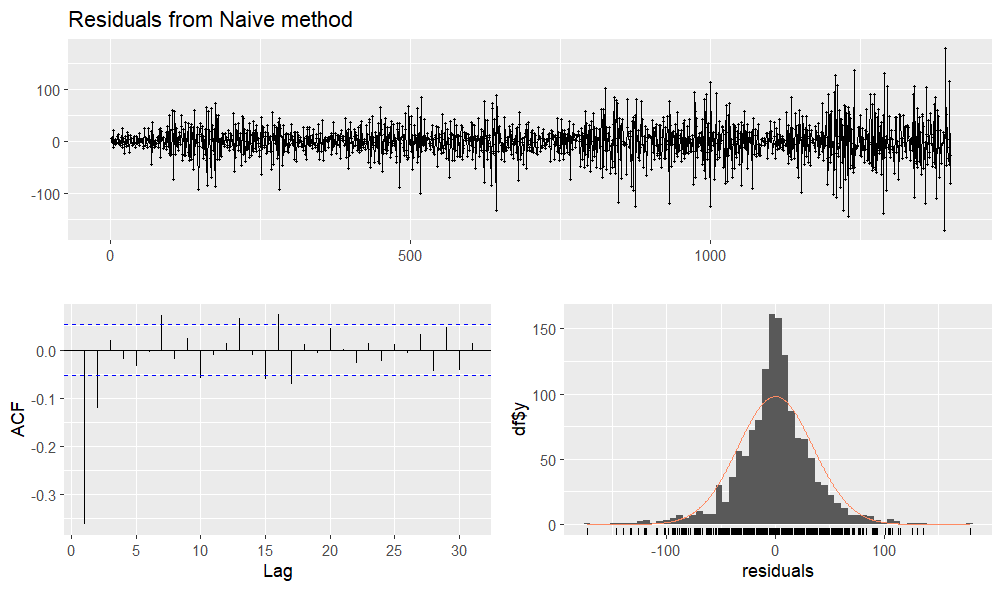


**Since the confidence fans seem to obstruct the actual data, we’ve added it in dashed lines so it can be compared with the Naive forecast in the following chart:** A graph showing a graph of a graph

AI-generated content may be incorrect.

* 1. Apply checkresiduals to the residuals of Naive method. Are the residuals random (i.e., are they white noise series)? Provide brief explanations.

***Answer:***



**The residuals from the Naive method do not behave like white noise. In the time plot, the residuals show visible clustering, where consecutive positive or negative errors persist over stretches of time, indicating that meaningful structure remains in the data.**

**The ACF plot also shows a large, significant negative spike at lag 1 and several additional lags that fall outside the significance bands. If the residuals were truly random, all autocorrelations would lie within those bands; instead, the presence of significant autocorrelation confirms that the model has not captured all temporal patterns.**

**The histogram of residuals is roughly centered at zero but shows mild asymmetry and heavier tails, which further suggests deviation from the properties of white noise. Overall, the residual diagnostics show that the Naive method leaves systematic structure in the errors, meaning the residuals are not random and the model is not fully adequate for this time series.**

* 1. What is the MAPE of Naive method? Briefly describe what it means. Do you think Naive is a good method to forecast what happen in the next 24 days?

A blue background with white text

AI-generated content may be incorrect.

***Answer:***

The MAPE is **58.46889**

It means that **on average, the Naive forecast is off by about 58% compared to the actual values in the test period. This indicates relatively large forecasting errors.**

My thought on Naïve method is **given the high MAPE and the volatility in the bike trip data, the Naive method does not perform well for forecasting the next 24 days. The series has trend and strong seasonal patterns that Naive cannot capture, leading to poor accuracy.**

***Example 2: Australian liquor sales***

In this example, you will apply what you learn to Australian liquor sales data set

**Instructions:** We are going to read xlsx type of file (not csv) into R. To do this, please install package readxl and load it. After that, run the following codes:

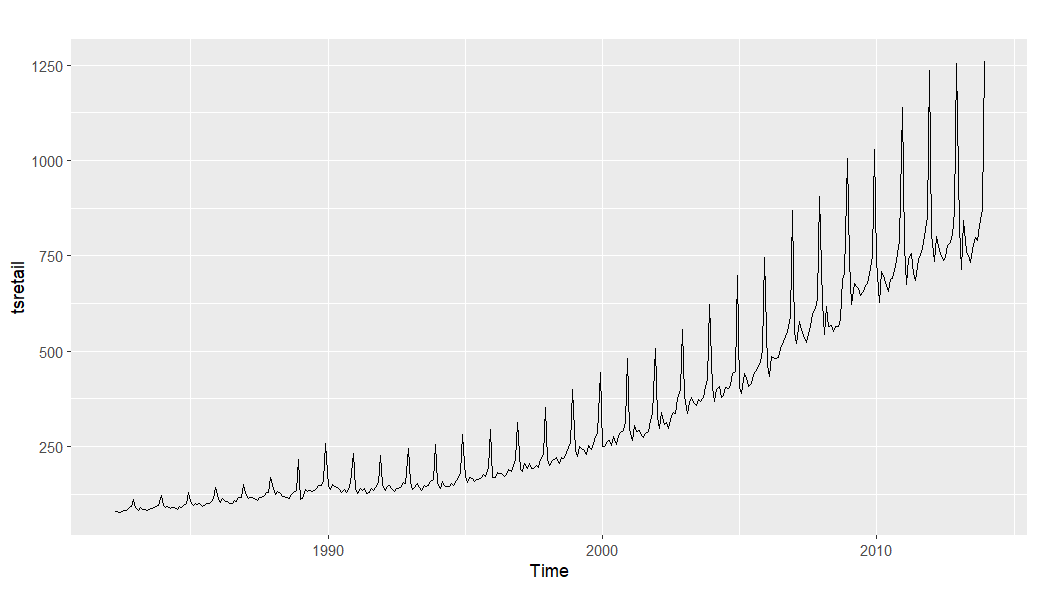
retaildata = read\_excel("retail.xlsx", skip=1)

tsretail = ts(retaildata$A3349462J, start = c(1982,4), freq=12)

**Q1:** Plot tsretail using autoplot.

***Answer:***

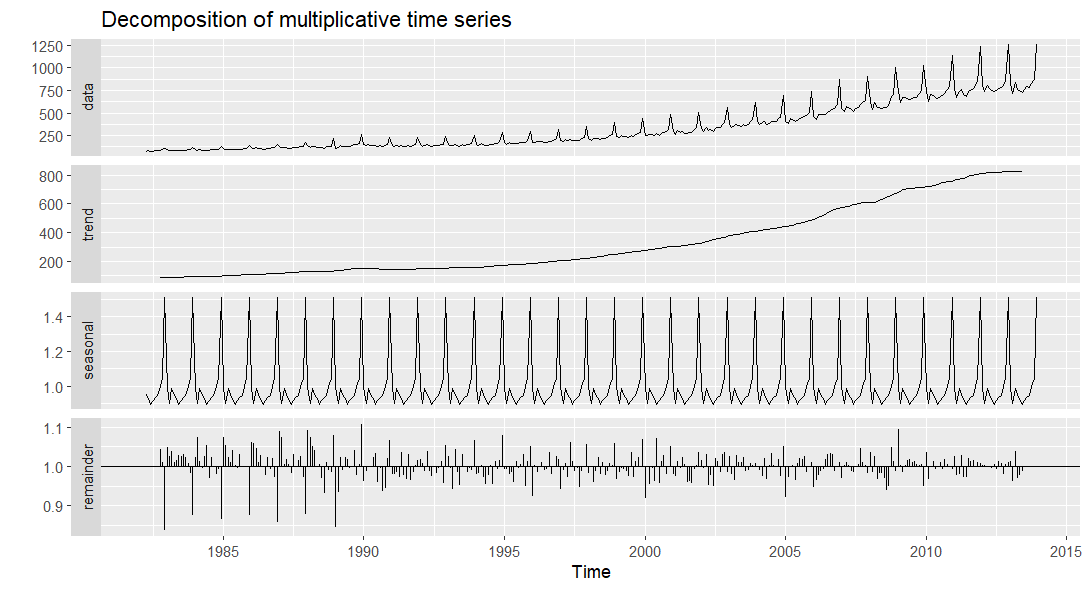
The graph is



**Q2:** Run multiplicative decomposition using decompose and plot it.

***Answer:***

The graph is



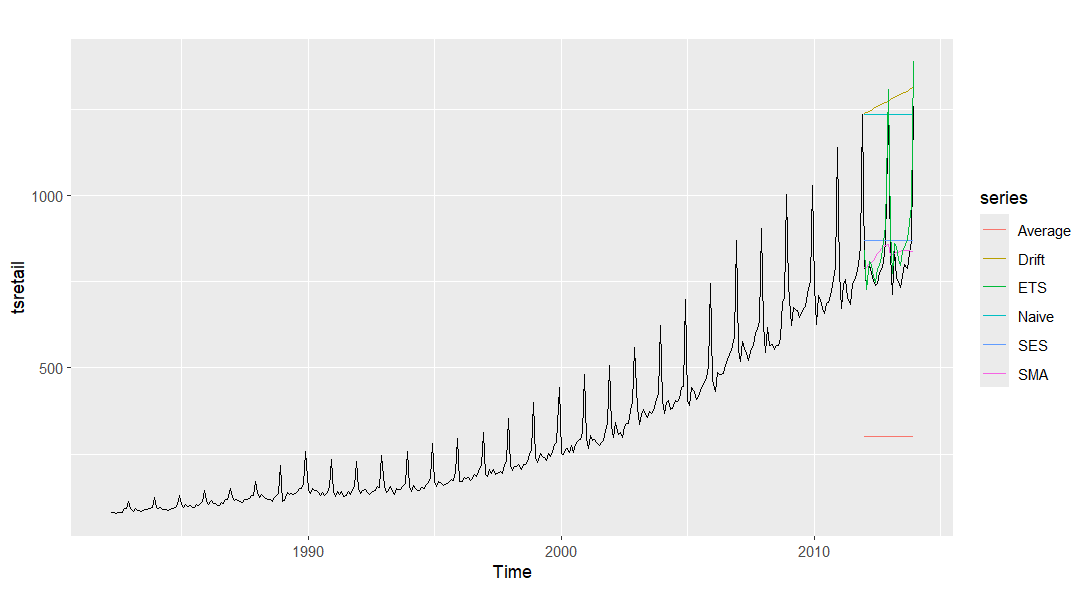
**Q3:** We will use all data points until the end of 2011 as training data set and the remaining data points as test data set. Create train\_data and test\_data.

**no need to include any answer Q3 in this submitted word doc**

**Q4:** The test data set created above has 24 data points. Use train\_data and h = 24, forecast liquor sales for the next 24 months using Naive, Average, Drift, SMA, SES, and ETS. Plot tsretail together with the forecasts from all methods in the same graph. And report the MAPE for each method.

***Answer:***

The graph is



The MAPEs are:

**Naive: 53.13**

**Average: 62.68**

**Drift: 57.76**

**SMA: 8.76**

**SES: 12.87**

**ETS: 6.18**

**\*\*Please note due to different R versions, some values may be slightly different (approximately 8.5 and 8.7 or 53 and 55). For example: One team member got the first set of results shown below, meanwhile the other team member got the second set of results using the same exact code. We have included BOTH screenshots below.\*\***

A computer screen shot of a program

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

***Example 3: Uschange***

In this example, you will apply what you learn to USchange data.csv. The data set contains information on the percentage changes in quarterly personal comsumption expecidture, personal disposable income, production, savings and the unemployment rate in the US from 1970 to 2016. For this HW, we will only focus on the variables Consumption and Income.

**Q1:** Read the data into R and name it usdata

**Q2:** Create a ts object using the variable ‘Consumption’, call it tsdata. Remember that this a quarterly data and the data set starts from quarter 1 of year 1970. Use the right frequency.

**Q3:** Use all the data points up to the end of year 2012 as training data set and the remaining data points starting from year 2013 as test data set. Call them train\_data and test\_data.

**no need to include any answer for Q1-Q3 in this submitted word doc**

**Q4:** Forecast Consumption using Naive, Average, Drift, SMA, SES, ETS, ARMA, ARIMA, methods. Report the MAPE of each method.

***Answer:***

The MAPEs are:

**\*\*Please note due to different R versions, some values may be slightly different (approximately 8.5 and 8.7 or like 55 and 57 for example).**

**For example: One team member got the first results below, meanwhile the other team member got the second set of results using the same exact code. We have included BOTH screenshots below.\*\***

**Naive: 55.19426**

**Average (meanf): 48.33**

**Drift: 55.19426**

**SMA: 53.06192**

**SES: 50.64277**

**ETS: 50.64277**

**ARMA: 37.96675**

**ARIMA: 37.96675**

**SARIMA: 38.74213**

A computer screen with white text

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.