Laboratory 1 Written Report

ANA 535 Forecasting

Cleaning, Sorting and Filtering Time Series Data

Name KOHEI NISHITANI Student ID# 1122867

Date April 6, 2025

# Table of Contents

* Introduction
* Background, Data, Methods and Procedures
* Results
* Conclusions
* References
* Attachment or Appendices

# Introduction

The first laboratory session concentrates on developing practical abilities to analyze time series data by cleaning data and sorting as well as filtering and conducting initial exploratory data analysis (EDA). This lab requires study of Amtrak data which includes a twelve-year record of passenger numbers and distance information between 1991 and 2024. The long duration of the data requires special attention towards handling dates while we must identify problematic entries before exploring natural patterns that will affect further forecasting operations.

This exercise aims to teach data conversion techniques that transform raw information into a format which R programming language uses for smooth analysis. The transformation and visualization process of the Amtrak dataset uses packages which include dplyr, lubridate, tsibble, and ggplot2. The analysis improves when researchers filter their data to a specific year then order it by descending values to examine date understandings along with the impact of minor data adjustments on analysis outputs. The effects of missing data elements along with atypical values on future predictive modeling and statistical conclusions also receive investigate here.

**Background**

I spent time reviewing various resources before this laboratory to build additional skills in handling time series within R. For example, I read the main materials of “R for Data Science” Chapter 5 by Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemunc to enrich the concept of Tidyverse.

The external online research supplementing my studies served as my additional reading sources. The Domino Data Lab site featured a brief blog entry titled “Time Series with R” that illustrated how to operate on date-time data and find outliers and made simple visualizations in R. This collection of materials identified optimal techniques for data importation into Amtrak while revealing how to manipulate the data effectively and generate preliminary visuals to detect unusual patterns.

**Data**

The primary dataset is an Amtrak time series from January 1991 to June 2024, containing around 400 monthly observations. It includes four columns: Date, Ridership, PassengerMiles, and RidersReported. Because RidersReported has early inconsistencies, this analysis focuses on Ridership and PassengerMiles. Each row corresponds to one month’s data. We simply format the Date column for time-based operations, verify no missing values, and finalize the dataset for subsequent exploration and analysis.

**Methods and Procedures**

All data wrangling and visualization tasks were performed in **R** (using RStudio). First, I loaded the relevant libraries—such as **dplyr**, **tidyverse**, **lubridate**, **tsibble**, **fable**, and **feasts**—which offer functions for time series handling, plotting, and data transformation.

In the initial phase, the Amtrak dataset was imported and prepared for analysis, with columns renamed for clarity and the date field converted into a recognized date format. This ensured that each observation’s monthly timestamp was properly identified. Next, the dataset was examined for missing values, revealing none, and summarized to confirm that all variables conformed to expectations. The data was then reformatted as a tidy structure, making it easier to perform downstream analyses.

Subsequently, a monthly time series representation was created for both ridership and passenger miles, starting from 1991 and extending into mid-2024. By plotting these variables, it became clear that notable fluctuations occurred during the late 1990s, in 2008, and again around 2020, suggesting external economic or global events influenced Amtrak traffic. To investigate subsections of the data more closely, filtering techniques were applied to focus on particular years, and the observations were sorted in reverse chronological order to illustrate how date handling simplifies comparisons.

A polynomial trend (up to cubic) was then applied to highlight broad patterns in passenger miles, revealing declines in the mid-1990s and later unexpected drops in 2020. While these polynomial curves provided a general sense of overall movement, they did not capture abrupt or drastic changes, implying that external shocks like lockdowns require more specialized modeling. Finally, a modern approach using time-series tibbles and extended plotting methods demonstrated that passenger miles is potentially a more comprehensive measure than simple passenger counts, especially when evaluating long-distance travel variations.

**Results**

This laboratory produced multiple time-series plots to explore **PassengerMiles** and **Ridership** from January 1991 to June 2024. Figure 1 shows the **PassengerMiles** series, revealing a slow decline through the mid-1990s, a partial recovery around the early 2000s, and a sharp drop after 2020. The latter dip corresponds to COVID-19 shutdowns, when overall travel was significantly reduced. According to the course guidance, PassengerMiles is more appropriate than simple ridership counts, because passengers can travel vastly different distances, and thus total mileage may better capture real demand.

Next, we fitted a **polynomial trend** (up to cubic) to illustrate broad patterns in the data (Figure 2). While a cubic model did highlight the 1990s downturn and post-2010 growth, it failed to capture the abrupt pandemic-related plunge in 2020. This discrepancy suggests that external shocks (like lockdowns) are not easily accounted for with a purely polynomial fit. Additionally, the data likely exhibit seasonality or more complex cycles, which simple trend lines cannot fully represent.

Lastly, for Ridership (Figure 3), we saw an analogous pattern, but the raw passenger count alone may understate or overstate the true impact of extended travel distances. Indeed, the professor’s note indicates that focusing on PassengerMiles helps reflect the intensity of ridership across different trip lengths. Overall, these time plots underscore the importance of distinguishing between ridership quantity and total travel distances—especially in analyzing historical events like the 2008 recession or 2020 pandemic, both of which markedly affected travel behavior.

**Conclusions**

Overall, the Amtrak data from 1991 to mid-2024 underscores the importance of carefully selecting the metric used to describe passenger activity. While raw ridership totals highlight general fluctuations, passenger miles provide a more nuanced view of overall demand, especially when travelers take longer trips. No missing values were detected, and polynomial trend lines captured some of the dataset’s broad contours—such as mid-1990s declines and post-2010 improvements—yet they fell short of explaining sudden shifts tied to major external factors, including lockdowns. This discrepancy indicates that future models must account for events that cause abrupt, time-specific changes rather than simply smoothing them. Furthermore, the mismatch between “RidersReported” and “Ridership” at the beginning of the dataset confirms that even apparently minor discrepancies can affect interpretation. By converting the data into monthly time series objects and tibbles, the analysis has laid a foundation for more advanced forecasting methods. Going forward, incorporating seasonality, handling non-stationary patterns, and potentially using more sophisticated approaches (such as ARIMA variants or other techniques) will be crucial to fully understanding the dynamics of Amtrak’s passenger volumes.

**References**

Domino Data Lab. (2022). Time series with R. [Blog post]. Retrieved September 24, 2025, from <https://domino.ai/blog/time-series-with-r>

**Wickham, H., & Grolemund, G.** (2017). R for data science. O’Reilly. Retrieved September 25, 2025, from <https://r4ds.hadley.nz/>