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

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 learn 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. Furthermore, this exercise also touches the concept of Auto correlation and stationary which are crucial statistic concept in the time-series analysis. Finally, this exercise also covers data transformation from wide to long by using additional data set.

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. Also Rpubs Stationary Testing features the basic concept of Autocorrelation function and Ljung-Box test, Augmented Dickey-Fuller t-statistic test and Kwiatkowski-Phillips-Schmidt-Shin(KPSS).

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. We also use additional hierarchical sales dataset which contains 237 variables of daily sales data. Given this dataset, we learn how to use pivot_longer() for data transformation.

Methods and Procedures

The tasks for data cleaning and chart creation occurred exclusively through RStudio by using the R programming environment. I initiated the process by loading necessary libraries such as dplyr, tidyverse, lubridate and tsibble, fable alongside feasts because they provide functionalities for time series analysis and data processing and visualization.

The first step involved importing the Amtrak dataset followed by an analysis preparation process which included column renaming for better understanding and conversion of the date field to an

acceptable date format. Every observation received proper identification of its date value for monthly time tracking. All variables in the dataset showed expected values after assessment for missing data which revealed none. The data needed a tidy format transformation which simplified operations for further analysis procedures.

A monthly time series format was created to examine both ridership and passenger miles from 1991 up to the middle of 2024. The visual presentation of these two variables demonstrated significant changes throughout the 1990s followed by major shifts in 2008 and another substantial alteration in 2020 which indicated external economic or worldwide events affected Amtrak traffic statistics. Date handling became easier to analyze different data sections by using filtering methods combined with reverse chronological sorting of observations.

The application of polynomial trends from cubic to lower order demonstrated an overall decrease in passenger miles which intensified during two historical periods before 2020 and in the mid-1990s and then again in 2020. The general trends from polynomial curves did not show abrupt changes because lockdown events demand specialized modeling techniques. Modifying plotting approaches with time-series tibbles revealed passenger miles as a better measurement than passenger count because they offer enhanced assessment of distance travel fluctuations.

Results

This laboratory produced multiple time-series plots to explore PassengerMiles and Ridership from January 1991 to June 2024. Figure 1 shows the difference of PassengerMiles series and Ridership, revealing a passenger mile 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 in both plots,

when overall travel was significantly reduced. Behind this trend, PassengerMiles is more appropriate than simple ridership counts, because passengers can travel vastly different distances, and thus total mileage may better capture real demand. By using filtering, we visualize the annual trend as reference(Figure 2).

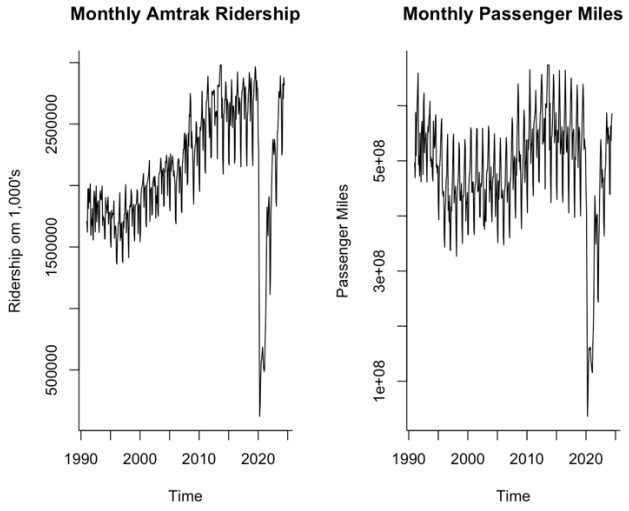

Figure 1. Amtrak Ridership and Passenger Miles

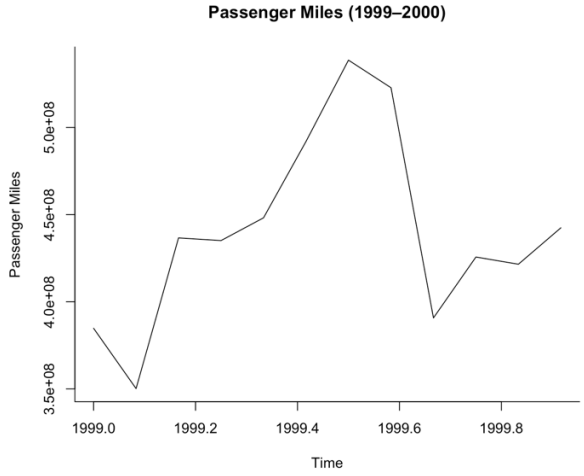

Figure 2. Amtrak Passenger Miles in 1999

Next, we visualized autocorrelation and see if data is stationary yet. From a stationarity standpoint, a key rule of thumb is that the ACF of a stationary series should drop off relatively quickly toward zero, reflecting only short‐range dependence. In this ACF plot(Figure 3.), however, the bars remain strong and persist far into higher lags, which signals ongoing correlation rather than a rapid decay. Then, we fitted a polynomial trend (up to cubic) to illustrate broad patterns and trend in the data (Figure 4). 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.

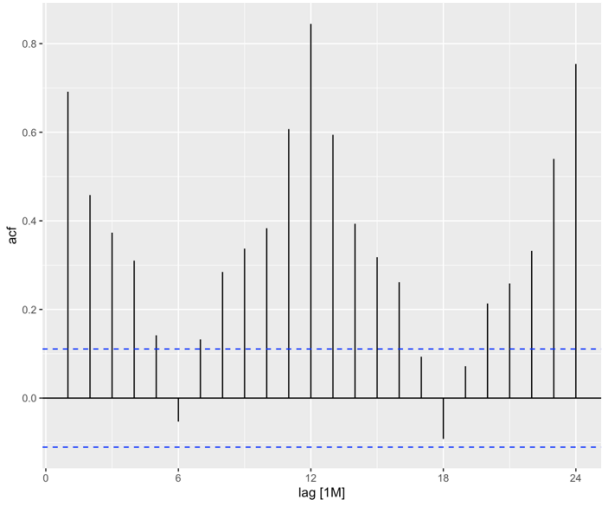

Figure 3. ACF plot

Lastly, for Ridership (Figure 4), 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.

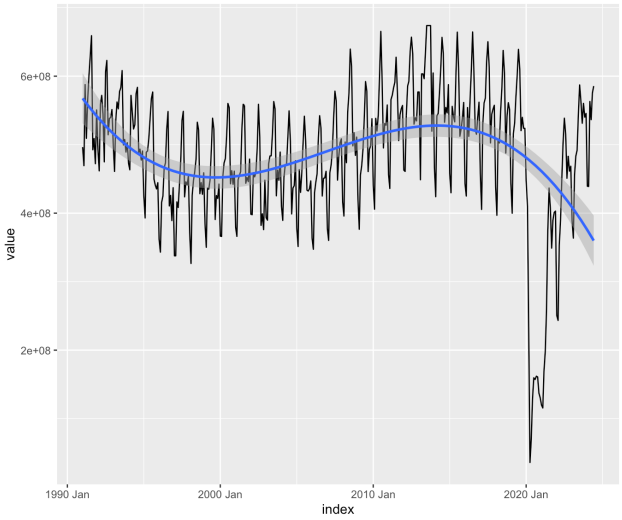

Figure4. Trend

Besides of Amtrak dataset, additional sales data shows noisy multiple product trend (Figure 5). This requires us to select product and narrow and filter the period of data for further deep dive.

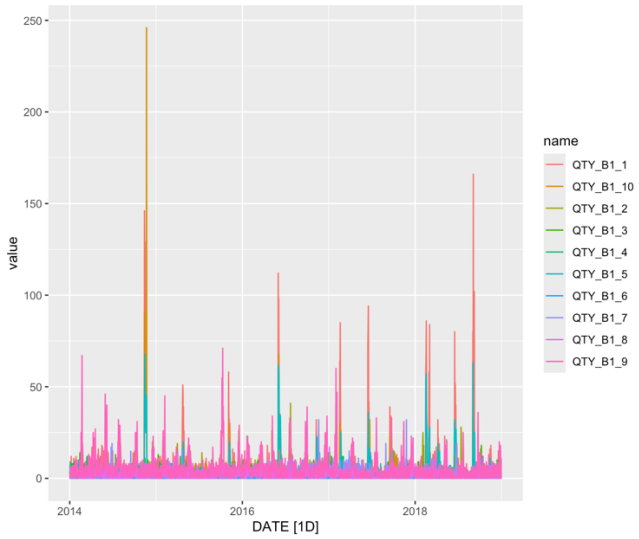

Figure 5.

Conclusions

Researching Amtrak passenger data from 1991 to mid-2024 demonstrates how crucial it is to select the most suitable measurement variable which describes transportation activity. Overall demand manifests best through passenger miles because they accurately represent how much travelers actually use the railroad system rather than simpler ridership numbers. Analysis of the dataset revealed that no values are missing and polynomial trend lines revealed some dataset patterns including mid-1990s declines and post-2010 trends but failed to explain abrupt point changes caused by major events like Covid-lockdowns and financial crisis etc. Data interpretation may be affected by seemingly insignificant differences between two variables like "RidersReported" and "Ridership" which become apparent at the beginning of the dataset. The analysis establishes future forecasting potential by transforming the data into time series tibbles and monthly objects. The future analysis of Amtrak's passenger volumes needs sophisticated methods such as ARIMA variants or other techniques to handle non-stationary patterns while integrating seasonal effects.

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/

Kyle T. Rich (2017). RPubs, from

https://rpubs.com/richkt/269797