A Time Series Analysis of Winning Berlin Marathon Times

BIOS 611 Project Fall 2024 UNC-Chapel Hill

Thomas Joyce

2024-12-07

Introduction

Thousands of athletes from around the world traverse a relatively flat yet still challenging 26.2 miles at the Berlin Marathon in late September every year. This race is one of the seven Abbott World Marathon Majors, along with Tokyo, Boston, London, Sydney, Chicago, and New York. As one might expect, the Berlin Marathon not only has a highly competitive professional field, but also a rapidly expanding amateur contingent. While elite runners come to Berlin to pursue world records (Scheer at al., 2021), most people take four hours or longer to complete the race.

Published publicly on Kaggle, the Berlin Marathon dataset contains finishing times for 882,539 runners (158,404 females and 724,135 males) and weather conditions between 1974 and 2019. Demographic variables in the Berlin Marathon dataset include age, gender, and country. Weather variables encompass precipitation (mm), hours of sunshine, hours of clouds, atmospheric pressure (mbar), average temperature (Celsius), minimum temperature, and maximum temperature.

While numerous studies have used the Berlin Marathon dataset to investigate the impact of weather conditions on past running performances (Knechtle et al., 2021; Scheer et al., 2021; Weiss et al., 2024), I was particularly interested in using using previous finishing times and weather features to forecast future winning times. After conducting exploratory data analysis, I created four ARIMA time series models to forecast winning male and female times from 2021 to 2024 and compared the forecasts to the actual results. The time series models usually predicted winning times to within 2-3 minutes of the actual value. Adding weather features to the models did not improve the predictive performance.

Data Preprocessing

The Berlin Marathon dataset consists of two files: a runner csv file and a weather csv file. The original runner csv file had 884,944 observations. However, 2405 observations had no finishing time recorded, so these observations were removed from the dataset. The runner csv file did not have any data (all male and female observations were missing) for 1978 and 1980; female observations were also missing in 1976, 1994, and 2019. While the complete timeline of the Berlin Marathon dataset spans from 1974 to 2019, it should be noted that runner data was missing for the years specified above. In the runner csv file, observations with a numerical value for age were mapped into one of the following age categories: <20, 20-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80+. However, runners with an arbitrary letter value for age were placed into an age category of "Unknown." Only male observations in 2019 had values for Country, so this feature was not used in any analyses.

The weather csv file had 46 rows, each representing the weather conditions on the day of the Berlin Marathon from 1974 to 2019. There was no missing data in the weather csv file. Temperatures were transformed from degrees Celsius to degrees Fahrenheit.

Exploratory Data Analysis

Before conducting time series analysis, it was essential to obtain an understanding of how Berlin Marathon participation counts and finishing times have evolved over the years. I also wanted to determine which features might be useful predictors of finishing time. A variety of data visualizations were constructed to meet these objectives.

The number of runners in the Berlin Marathon steadily increased between 1974 to 2019, with over 40,000 runners in 2018 (Figure 1). This rise in participants is likely due to the increasing popularity in marathons among the general population over the past half century, which coincides with progressive trends in cardiovascular health and fitness.

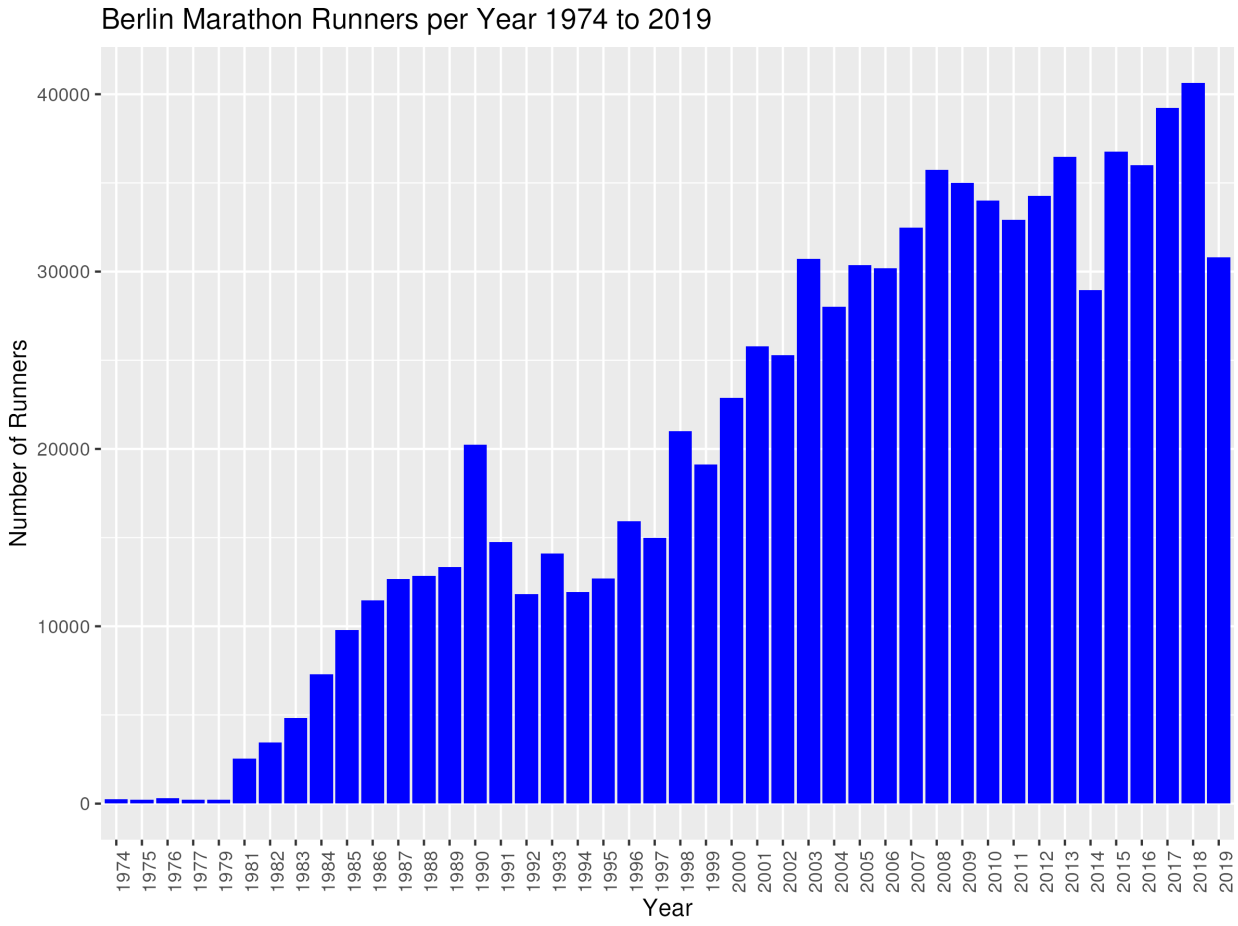

Figure 1: Berlin Marathon runners per year (Only male observations were recorded in 1976, 1994, and 2019)

While the number of runners at the Berlin Marathon has steadily increased, average times have slowed (Figure 2). In the late 1980's, the average finishing time at the Berlin Marathon was around 3 hours and 35 minutes. However, the average finishing times were close to 4 hours and 10 minutes in the 2010's. It seems reasonable to hypothesize that the observed increase in runners stems from the fact that more ordinary people (i.e., not lifelong runners or professional athletes) are beginning to run marathons. If this is the case, it makes sense that average times have slowed because new runners typically won't be as fast as seasoned runners.

A related potential reason for the trend in slowing average finishing times is that the proportion of middle aged runners relative to younger runners in the Berlin Marathon has increased over the years (Figure 3). The 20-29 year old age group had the highest proportion of runners in the 1980's, but towards the 2000's,

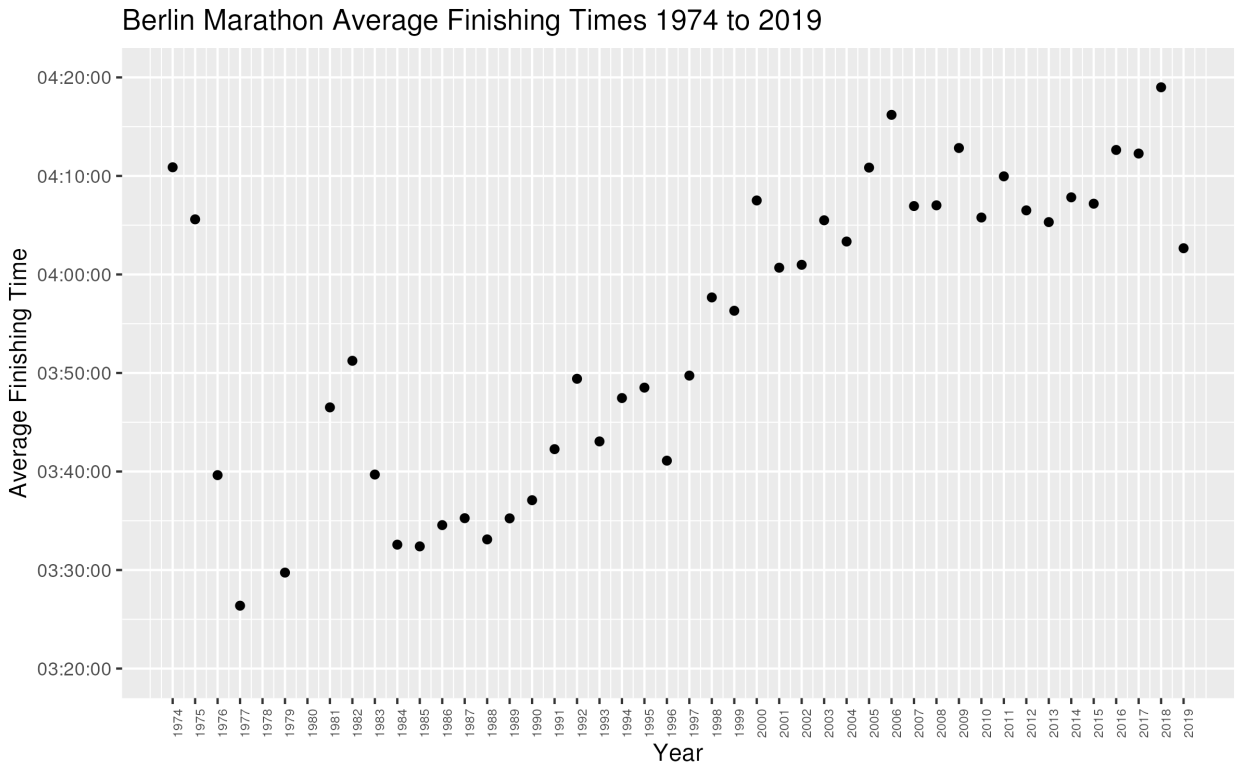

Figure 2: Berlin Marathon average finishing times (Only male times were recorded in 1976, 1994, and 2019)

the proportions of runners aged 40-64 increased substantially.

Figure 4 shows the average Berlin Marathon finishing times by age between 1974 and 2019. We can see that older age groups generally have slower average finishing times compared to younger age groups. Hence, more older runners entering the race relative to younger runners would make the overall average finishing time slower.

Figure 5 shows the distributions of marathon times by gender aggregated across all years. We can see that there have been a lot more male finishers than female finishers. Both time distributions are bell-shaped and can be approximated by a smoothed kernel density function. The average finishing time for males is about 3 hours and 56 minutes, and the average finishing time for females is about 4 hours and 26 minutes.

Figure 6 shows the winning Berlin Marathon times by gender between 1974 and 2019. Unlike the average finishing times, the winning times have become faster over the years. There was a significant improvement in winning times for both males and females in the early 1980's, and since then winning times have continued to speed up gradually. I was particularly interested in using the Berlin Marathon dataset to forecast winning male and female times in 2021 to 2024 (there was no race in 2020 due to COVID).

Moving on to the weather visualizations, Figure 7 shows the precipitation on the day of the Berlin Marathon each year. Most years had little to no rain on race day. However, 2010 was an outlier with about 30 mm of rain.

Figure 8 shows the average, minimum, and maximum temperatures on the day of the Berlin Marathon each year. Average race day temperatures were usually between 50-60 degrees Fahrenheit, which is quite tolerable for most runners. In recent years, the maximum temperatures have exceeded 70 degrees, which is a bit hot. Since the race begins around 9 or 10 am in the morning, faster runners may not experience the maximum temperatures until near the end of the race. However, runners taking over 4 hours to finish definitely had to deal with the heat in some years.

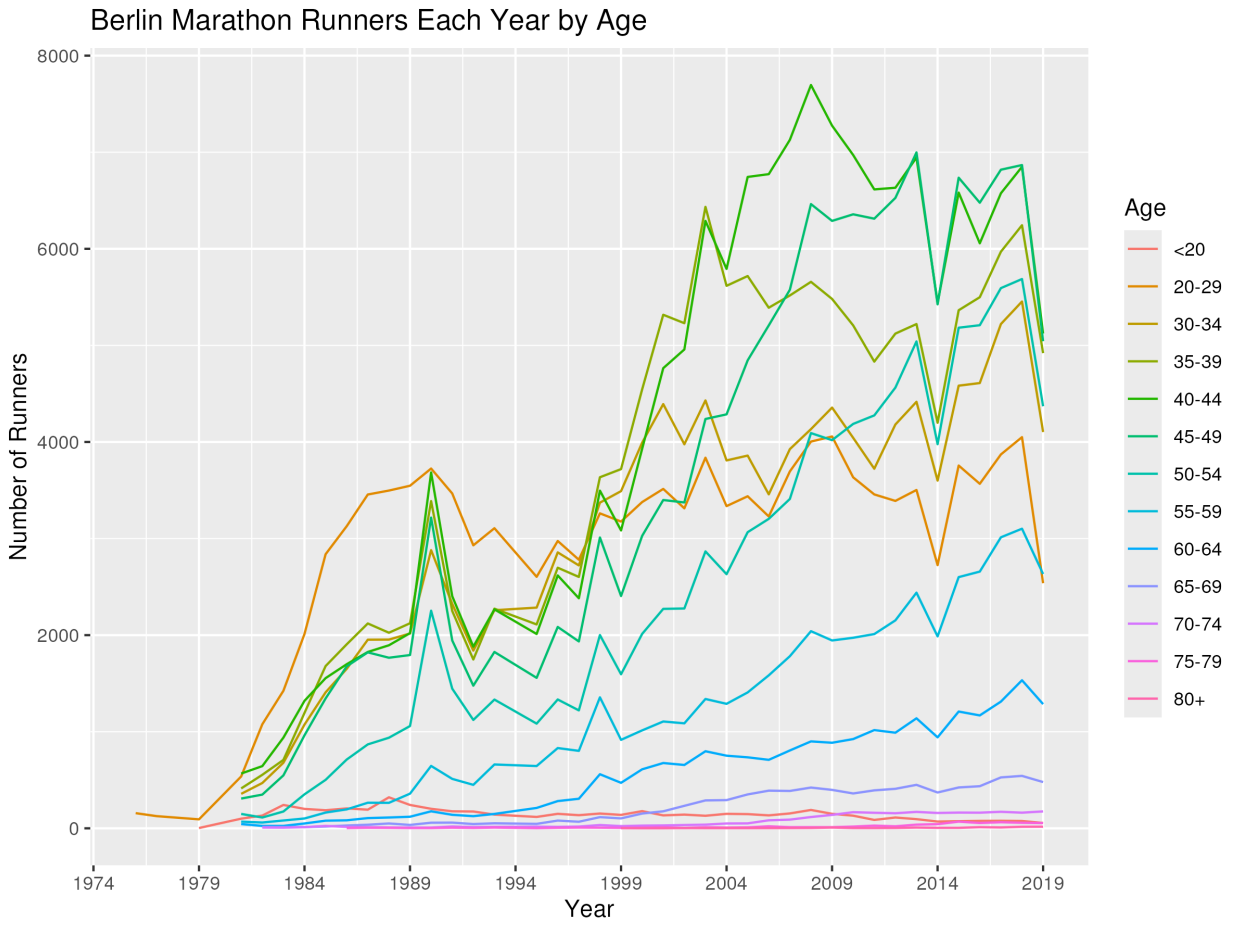

Figure 3: Annual runners by age

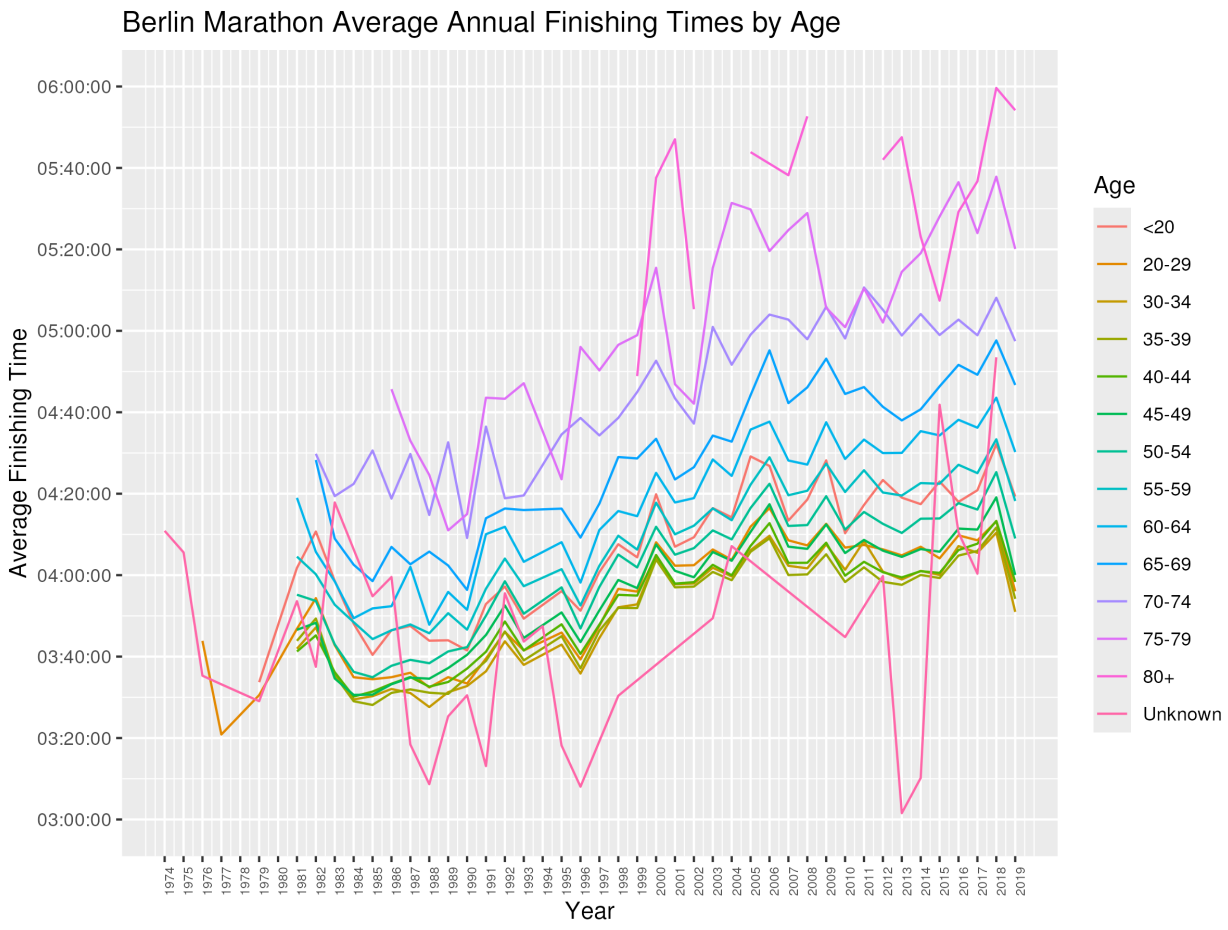

Figure 4: Berlin Marathon average finishing times by age

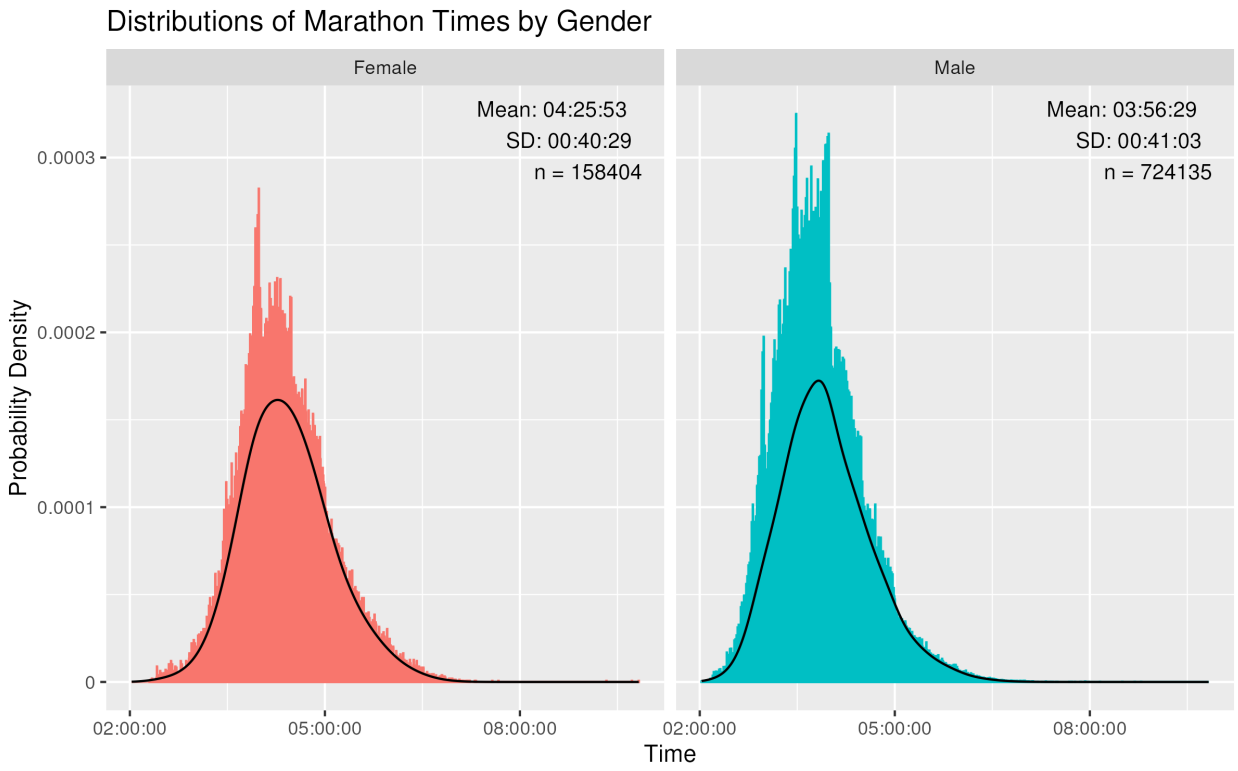

Figure 5: Berlin Marathon time distributions by gender

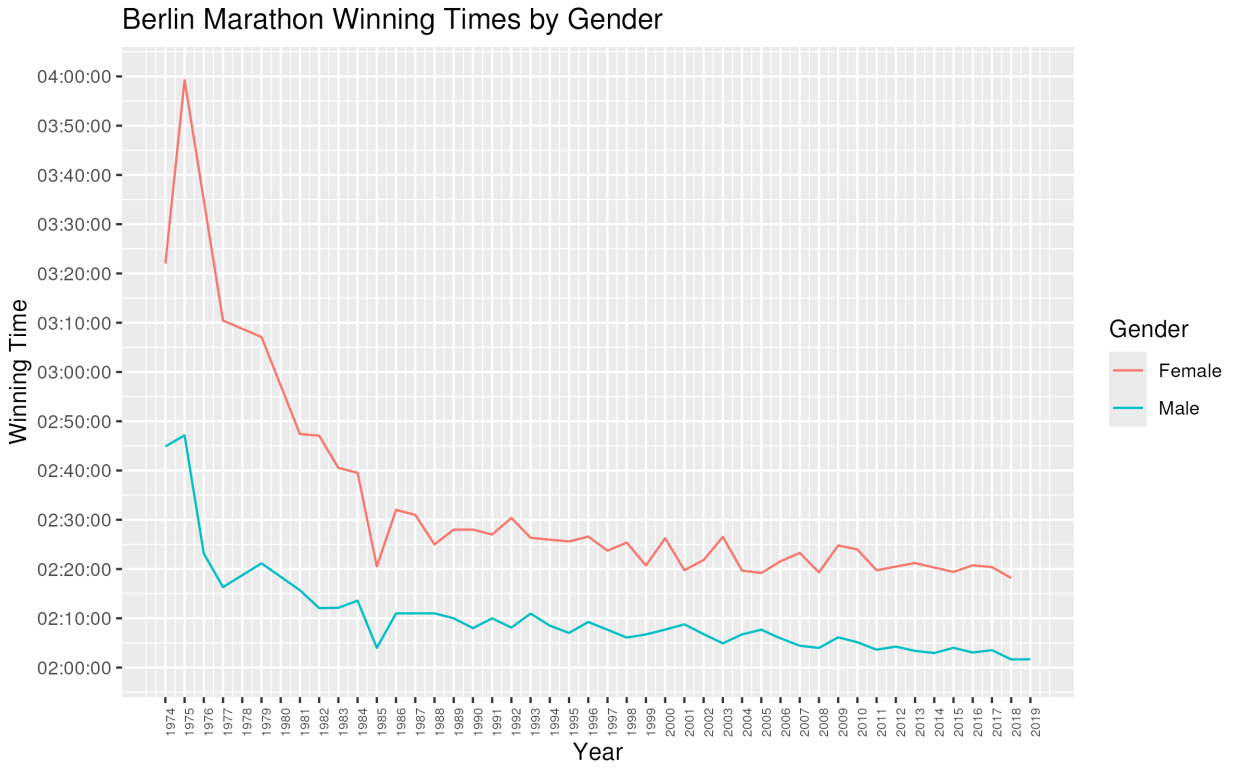

Figure 6: Berlin Marathon winning times by gender

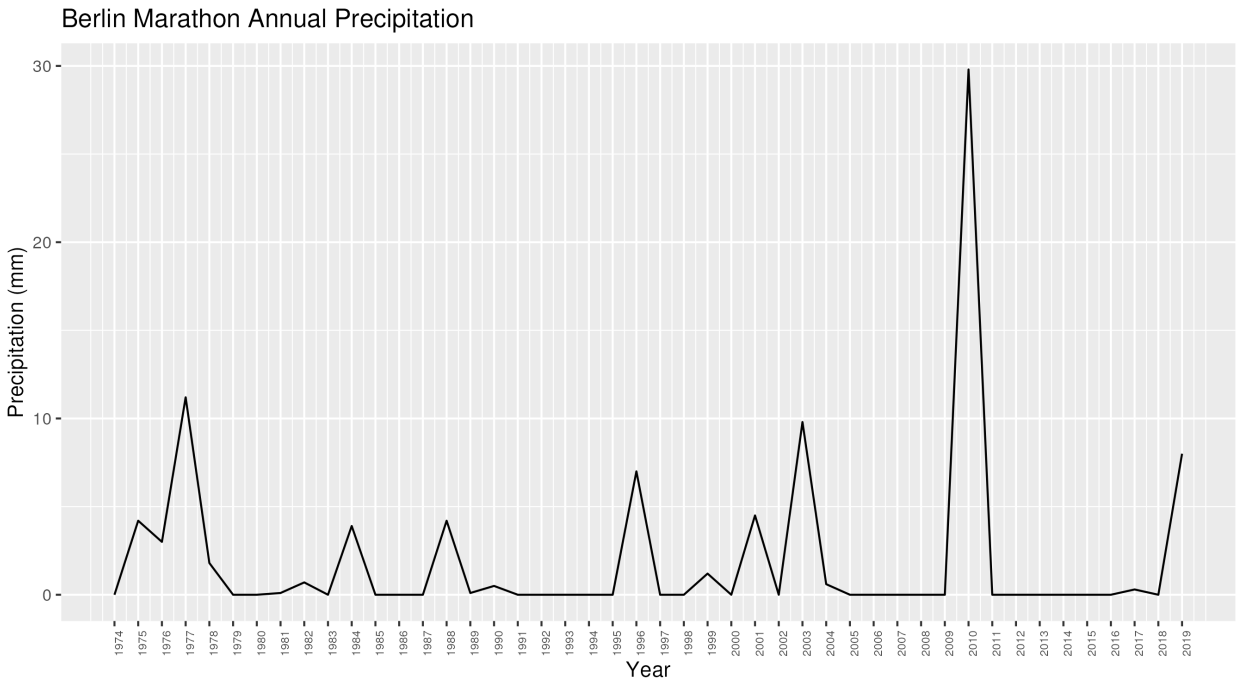

Figure 7: Berlin Marathon annual precipitation trends

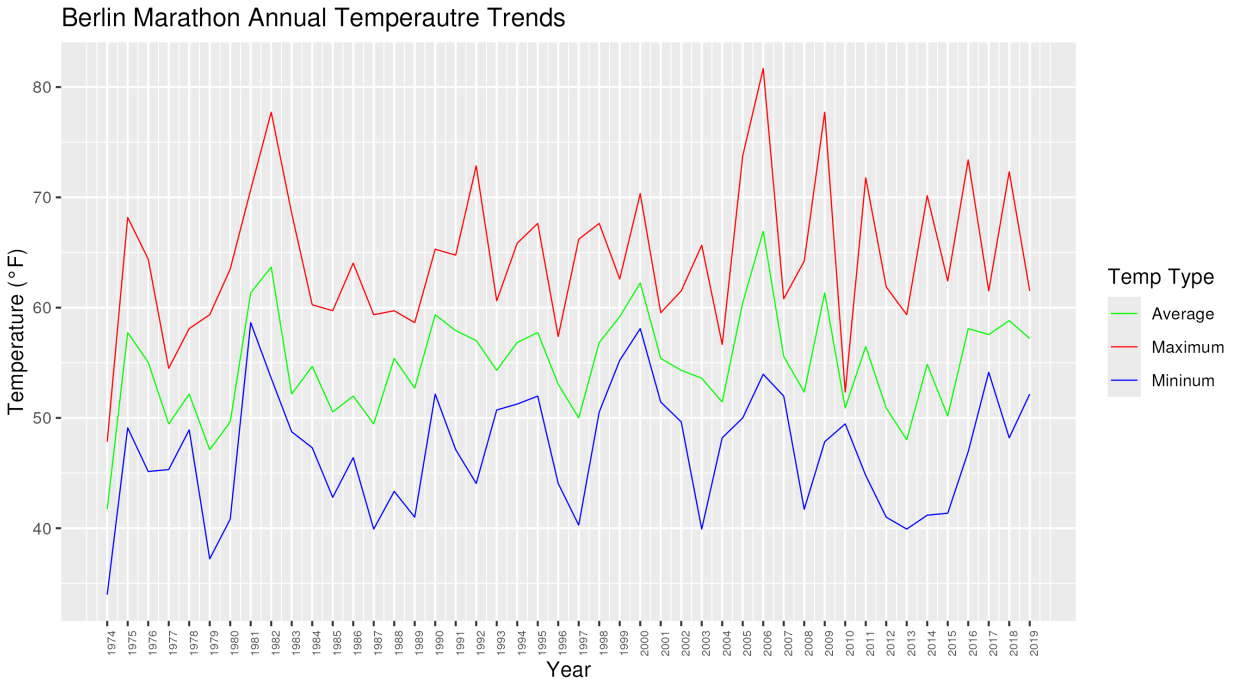

Figure 8: Berlin Marathon annual temperature trends

To obtain a deeper understanding of how the weather and time variables in the Berlin Marathon dataset relate to one another, I created a heatmap of feature correlations (Figure 9). The Pearson correlations among weather variables were mostly what I expected. For example, cloud hours and sunshine hours had a strong negative correlation of -0.88, while maximum temperature and sunshine hours had a moderate positive correlation of 0.66. However, most of the weather variables had weak to no correlation with finishing time (in seconds). There was a slight positive correlation of 0.25 between finishing time and year. This makes sense because the average finishing times have slowed over the years.

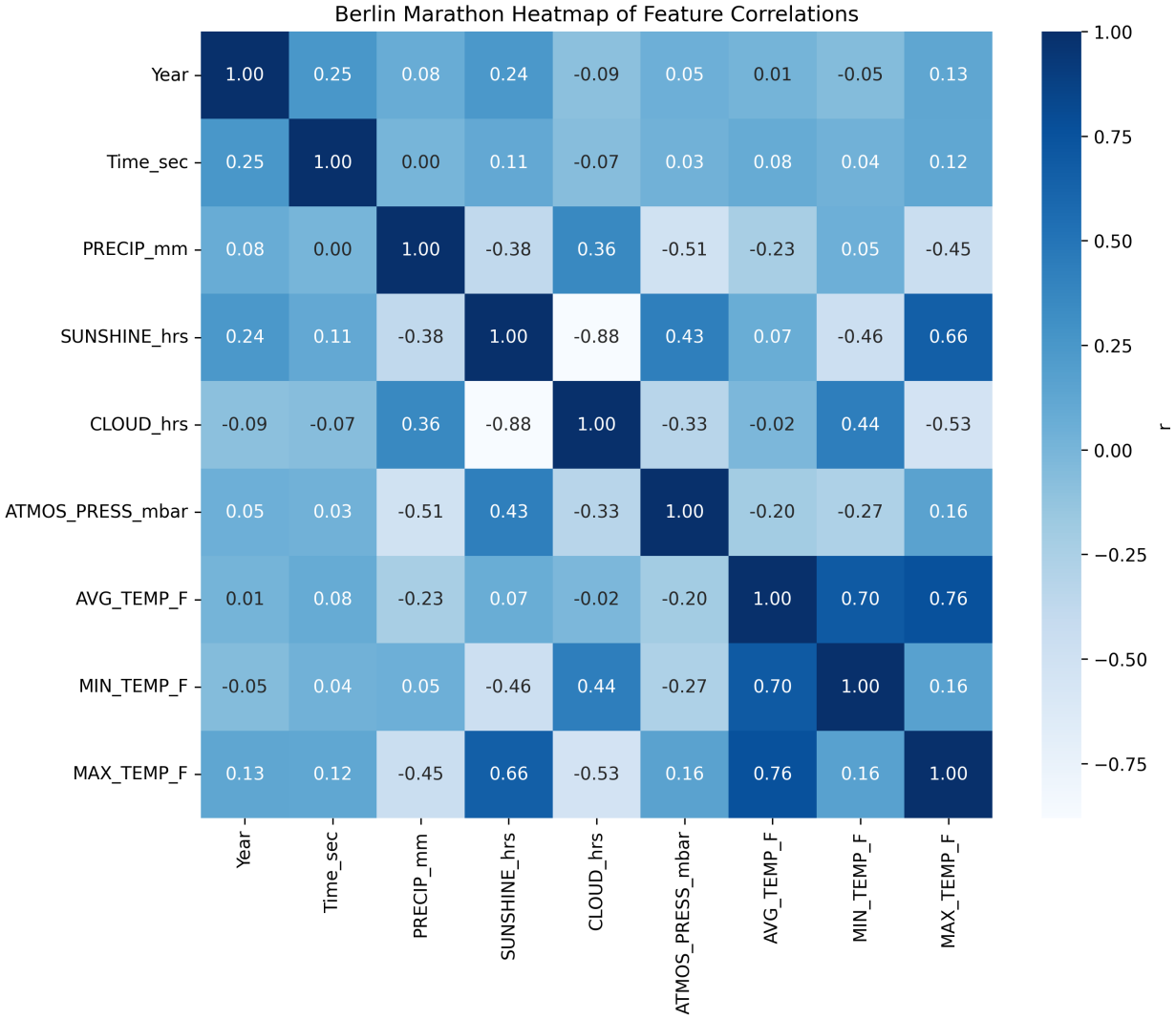

Figure 9: Heatmap of feature correlations for the Berlin Marathon dataset

Time Series Analysis for Forecasting Future Winning Times

After exploring time and weather trends in the Berlin Marathon dataset, I wanted to develop a model to predict future male and female winning times using past winning times and weather variables. Time series was a natural approach to this problem since the winning Berlin Marathon times and weather data have been recorded annually from 1974 to 2019. I decided to use the Autoregressive Integrated Moving Average (ARIMA) time series method from the forecast package in R due to its flexibility in model selection and potential to incorporate external regressors (Hyndman et al., 2024). ARIMA(p,d,q) includes p autoregressors to forecast future values using a linear combination of previously observed values, d differencing components

to make the time series stationary, and q moving average components that leverage past errors to forecast future values. I used the auto.arima function in the forecast package to automatically select the values of p, d and q that yield the best model on the training data. When training the models, a random seed value was set to make the results reproducible.

Four total ARIMA time series models were created to forecast winning Berlin Marathon times for the next five years. However, I was only able to compare the model predictions to the actual results for four years (2021 to 2024) since the Berlin Marathon was canceled in 2020 due to COVID. The first model used previous male winning times to forecast winning male times for the next five years. The second model used previous male winning times and additional weather covariates to forecast winning male times for the next five years. The third model used previous previous female winning times to forecast female winning times for the next five years. The fourth model used previous male winning times and additional weather covariates to forecast female winning times for the next five years. The following weather features were included as covariates in the second and fourth models: precipitation (mm), hours of sunshine, hours of clouds, and maximum temperature (Fahrenheit). Since the ARIMA model is scale invariant, the covariates were not standardized or normalized.

Figures 10-13 display the forecasts for each time series model; the black line represents previous winning times (1974 to 2019), the light blue line indicates forecasted winning times for the next five years (2020 to 2024), the dark gray bands represent 95% confidence limits, and the light gray bands represent 80% confidence limits.

The first time series model forecasted male winning times without using any weather covariates (Figure 10). This was an ARIMA(0,1,0) model, so the next value in the series only depended on the previous value plus a random error term. The forecasted winning times for the next five years were all 02:41:41 (Table 1). Despite its lack of variation, this model offered forecasts that were generally pretty close to the true winning times, with a mean absolute error of just 1 minute, 48 seconds.

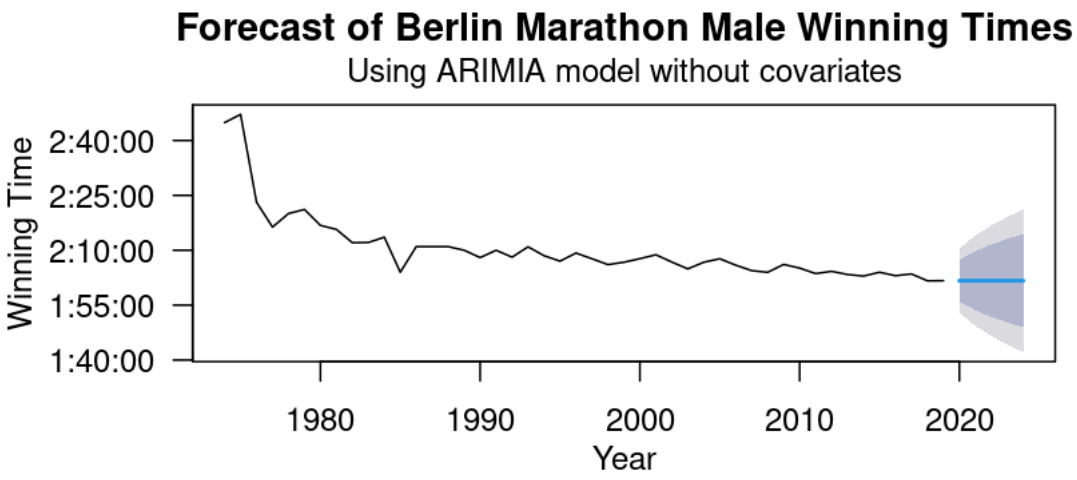

Figure 10: ARIMA(0,1,0) model for male winning times without covariates

Table 1: Comparison of actual vs. forecasted male winning times for the ARIMA model without covariates. (MAE = 00:01:48.25)

Year	Actual	Forecast	Error
2020	NA	02:01:41	NA
2021	02:05:45	02:01:41	00:04:04
2022	02:01:09	02:01:41	-00:00:32
2023	02:02:42	02:01:41	00:01:01
2024	02:03:17	02:01:41	00:01:36

The second model forecasted male winning times using weather covariates (Figure 11). This was an ARIMA(0,1,0) model; weather variables included as additional additive terms. The forecasted times had some variance when weather covariates were included (Table 2). Overall, the second model's predictive performance was acceptable, but slightly worse than the male winning time model without covariates.

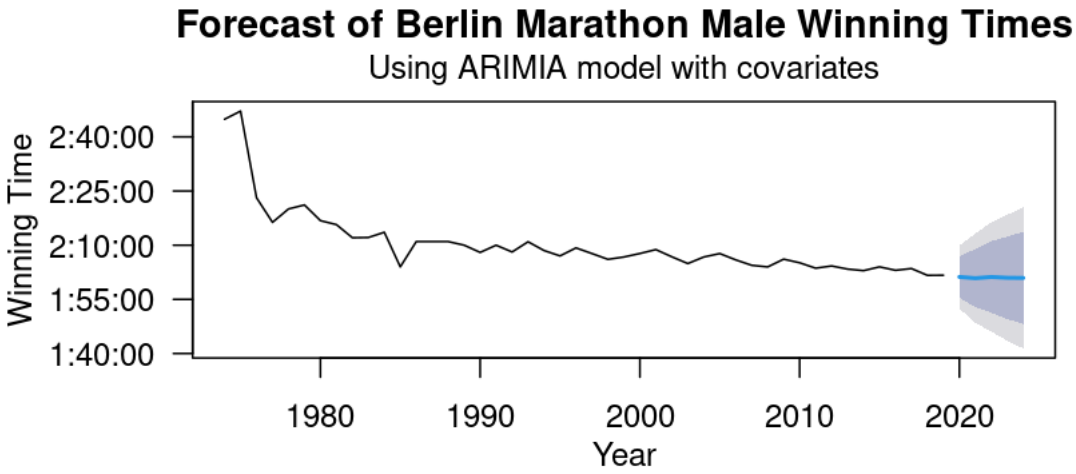

Figure 11: ARIMA(0,1,0) model for male winning times with covariates

Table 2: Comparison of actual vs. forecasted male winning times for the ARIMA model with covariates. (MAE = 00:02:15.25)

Year	Actual	Forecast	Error
2020	NA	02:01:13	NA
2021	02:05:45	02:00:52	00:04:53
2022	02:01:09	02:01:12	-00:00:03
2023	02:02:42	02:00:58	00:01:44
2024	02:03:17	02:00:56	00:02:21

The third model forecasted female winning times without using any weather covariates (Figure 12). This was an ARIMA(2,1,2) model, meaning that it had 2 autoregressors, 1 differencing component, and 2 moving average components. For this model, the forecasted values for the next five years were all between 2 hours, 17 minutes and 2 hours, 19 minutes (Table 3). Female winning times had a lot more variation than the

male winning times between 2021 and 2024, and the model did not do a great job capturing that variation. However, the overall performance was decent, with a mean absolute error of 3 minutes and 8 seconds.

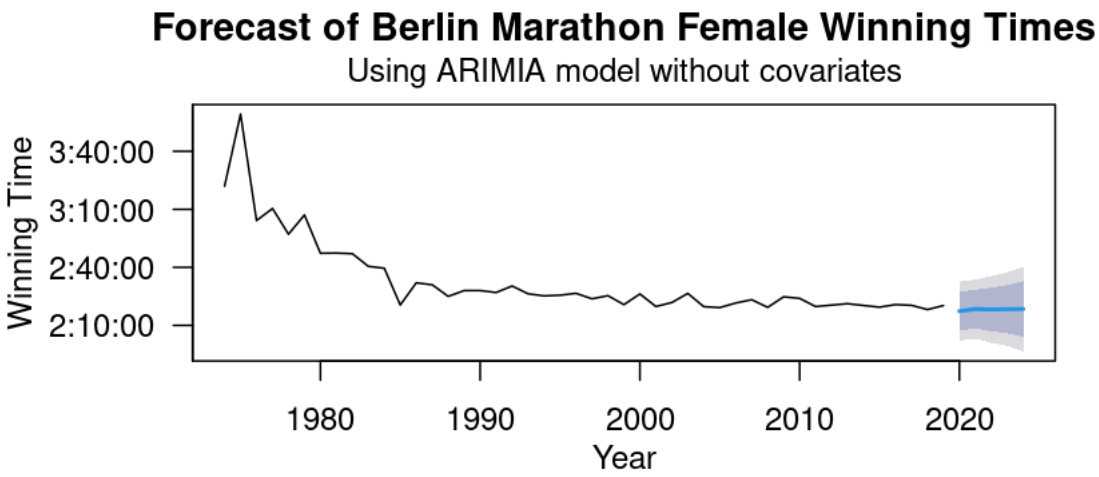

Figure 12: ARIMA(2,1,2) model for female winning times without covariates

Table 3: Comparison of actual vs. forecasted female winning times for the ARIMA model without covariates. (MAE = 00:03:08)

Year	Actual	Forecast	Error
2020	NA	02:17:20	NA
2021	02:20:09	02:18:28	00:01:41
2022	02:15:37	02:18:12	-00:02:35
2023	02:11:53	02:18:24	-00:06:31
2024	02:16:42	02:18:27	-00:01:45

The fourth model forecasted female winning times using weather covariates (Figure 13). This was an ARIMA(0,1,0) model. With a mean absolute error of 3 minutes and 22 seconds, the fourth model's performance was very similar to the third model's performance.

Table 4: Comparison of actual vs. forecasted female winning times for the ARIMA model with covariates. (MAE = 00:03:22.25)

Year	Actual	Forecast	Error
2020	NA	02:18:58	NA
2021	02:20:09	02:17:39	00:02:30
2022	02:15:37	02:19:12	-00:03:35
2023	02:11:53	02:18:03	-00:06:10
2024	02:16:42	02:17:56	-00:01:14

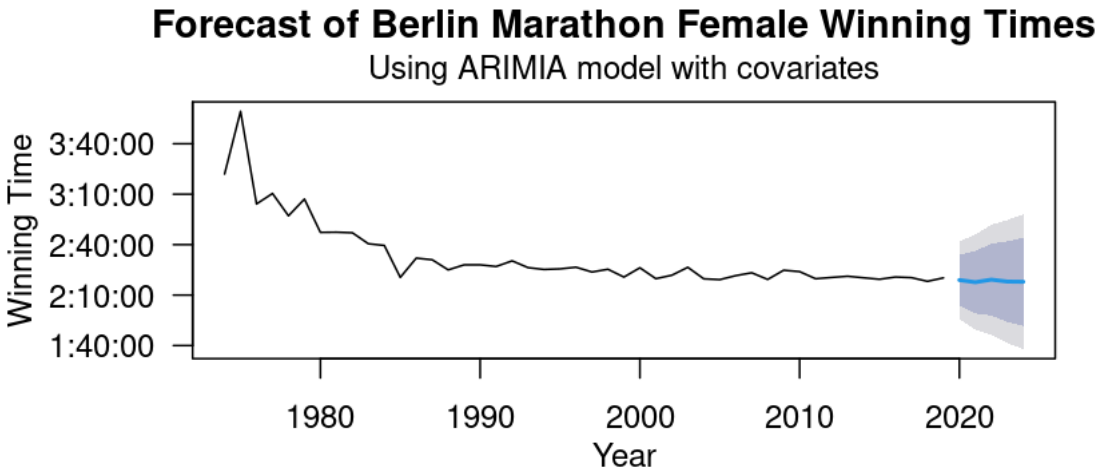

Figure 13: ARIMA(0,1,0) model for female winning times with covariates

Conclusion

This project sought to investigate finishing times and weather trends in the Berlin Marathon dataset and use time series analysis to predict future male and female winning times. After examining the feature correlations, I learned that there really wasn't much of a relationship between the weather variables and finishing times. However, I still wanted to incorporate weather features into the time series models to see how they would impact the model's performance. The forecasted values of the ARIMA models were fairly similar with and without covariates, although the models without covariates had slightly lower mean absolute errors. Overall, the ARIMA models offered smooth forecasts that weren't too far from the truth, but they failed to capture the year-to-year variation in winning times, particularly for female athletes. I'm thinking that a potential way to improve these models is by incorporating covariates related to competition (i.e., which elite athletes are in the race and what their personal best times are) along with variables related contract bonuses and prize money (which could provide an incentive for professionals to peak for the race). It would also be interesting to apply other machine learning methods, such as random forests, to predict future winning times at the Berlin Marathon.

References

Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2024). *forecast: Forecasting functions for time series and linear models*. R package version 8.23.0, https://pkg.robjhyndman.com/forecast/.

Knechtle, B., Valero, D., Villiger, E., Alvero-Cruz, J. R., Nikolaidis, P. T., Cuk, I., ... & Scheer, V. (2021). Trends in weather conditions and performance by age groups over the history of the Berlin Marathon. Frontiers in physiology, 12, 654544.

Scheer, V., Valero, D., Villiger, E., Alvero Cruz, J. R., Rosemann, T., & Knechtle, B. (2021). The optimal ambient conditions for world record and world class performances at the Berlin Marathon. Frontiers in physiology, 12, 654860.

Weiss, K., Valero, D., Villiger, E., Scheer, V., Thuany, M., Aidar, F. J., ... & Knechtle, B. (2024). Associations between environmental factors and running performance: An observational study of the Berlin Marathon. Plos one, 19(10), e0312097.