STAT 333 Final Project
Student: Duc Huy Nguyen (Thomas), Theo Schouten
Date: November 19, 2025
Subject: Deliverable 2, Descriptive Statistical Analysis

Overview:
This paper summarizes our preliminary findings of the dataset in order to serve two perposes:

- Predicting whether a given Super Mario run will be the new world record (WR)
- Predicting when top speed runners will be able to reach the theoretical tool-assisted speedrun time.

Descriptive Statistics and Key Findings
The datasets consists of two datasets:

- The first dataset records the World Record (WR) progression from the first WR ever recorded in June 25th, 2002 to the latest WR, which occurred October 22nd, 2025. It can be retrieved here: `https://www.speedrun.com/smb1/stats?h=Any-NTSC&x=w20p0zkn-onvvdymn.013zwgxq`
  - The first dataset contains the progression of the Super Mario Bros. Any% (NTSC) World Record (WR), beginning with the earliest recorded WR on June 25, 2002 and continuing through the most recent WR on October 22, 2025. It can be accessed here: `https://www.speedrun.com/smb1/stats?h=Any-NTSC&x=w20p0zkn-onvvdymn.013zwgxq`
  - The WR has been broken **46** times by **14** different players.
  - The first recorded WR is **325 seconds** (5 minutes 25 seconds), set on June 25, 2002 by Aaron Collins. The current WR is **294.448 seconds** (4 minutes 54.448 seconds), set on October 22, 2025 by Niftski.
  - The theoretical human-limit time, i.e., the fastest time a human could possibly achieve, is **294.032 seconds** (4 minutes 54.032 seconds).
  - The current WR is **0.186 seconds** slower than the TAS (Tool-Assisted Speedrun) time, which corresponds to approximately **11 frames** at 60 frames per second.
  - The median rate of WR improvement is **-0.00184 seconds per day**, while the most recent improvement rate is **-0.00243 seconds per day**.
  - The WR improvement rate reflects a standard pattern of diminishing returns: as players approach the theoretical TAS limit, achieving further WR improvements becomes increasingly difficult. In other words, the remaining room for improvement steadily decreases.
- The second dataset contains 1,353 attempts by the current WR holder, Niftski. Each attempt is split into eight major checkpoints, where 1 marks a successful completion and 0 marks a failure. The results at each checkpoint are then randomized. Refer to Table 1 in the Appendix for an overview of the data.

– The table below shows Niftski's success rate at each major checkpoint.

Checkpoint	1-1	1-2	4-1	4-2	8-1	8-2	8-3	8-4
Success	829	595	482	63	30	11	9	1
$\mathbb{P}$(Success)	61.27%	43.98%	35.62%	4.66%	2.22%	0.81%	0.67%	0.07%

– There are sharp decrease in the number of successful attempts at later checkpoints. For example, success falls from **482** at 4-1 to **63** at 4-2, an **87%** decrease. There is also a **63%** decrease between 8-1 and 8-2. A few factors can explain these drops:

* More pressure and stress in later stages, which can lead to mistakes.
* Higher difficulty in later stages, which increases the chance of errors.

Suggestion of Statistical Model

We suggest several models that fit the goal of the project and the data:

- An ARIMA model that uses WR improvement over time (seconds per day) to predict when the WR will reach the TAS time.
- A model based on Niftski's dataset to predict whether a run will be a WR.
- Other possible models.

Limitations:

- Research on games like *Super Mario Bros.* is limited, especially in academic settings. Because of this, data on the game and its time-attack attempts is limited.
- We do not have data from other competitive speedrunners and chose not to collect or publish any of their runs.
- We plan to address these limits by collecting data manually based on recorded results. These results are timed with reliable stopwatches to keep the measurements consistent.

Appendix:

- **R code:**

```
library(tidyverse)

# Data import and cleaning
wr <- read.csv("~/Downloads/history.csv")
wr$Time <- NULL
wr <- wr %>%
  mutate(Date = as.Date(Date, format = "%Y-%m-%d")) %>%
  mutate(Time = as.numeric(Date - min(Date)) + 1)
head(wr)
```

```
(wr)

# Draw WR progression plot
plot(wr$Time, wr$Seconds, type = "l",
     xlab = "Days since Jun 25 2002", ylab = "WR (sec)",
     main = "WR Progression 2002-2025", lwd = 2)
points(wr$Time, wr$Seconds, pch = 19, col = "red")
grid(nx = NULL, ny = NULL, col = "gray", lty = "dotted")

# Change in WR over change in time
delta_wr <- diff(wr$Seconds)
delta_time <- diff(wr$Time)
rate <- delta_wr / delta_time
rate
summary(rate)

plot(rate, type = "b",
     ylab="Rate of WR improvement (sec/year)",
     xlab="Record index")
abline(h = 0, col="red")



# 2nd data import and cleaning
niftski <- read.csv("~/Library/Mobile Documents/com~apple~CloudDocs/
                    Combined Niftski Runs.csv")
head(niftski)

# Some basic statistics
for (i in 1:8)
  print(mean(niftski[, i]))
```

- **Relevant visualizations:**

Run/Stage	X1.1	X1.2	X4.1	X4.2	X8.1	X8.2	X8.3	X8.4
1	1	1	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0
4	1	1	1	0	0	0	0	0
5	1	0	0	0	0	0	0	0
6	1	0	1	0	0	0	0	0

Table 1: Overview of Niftski's runs.

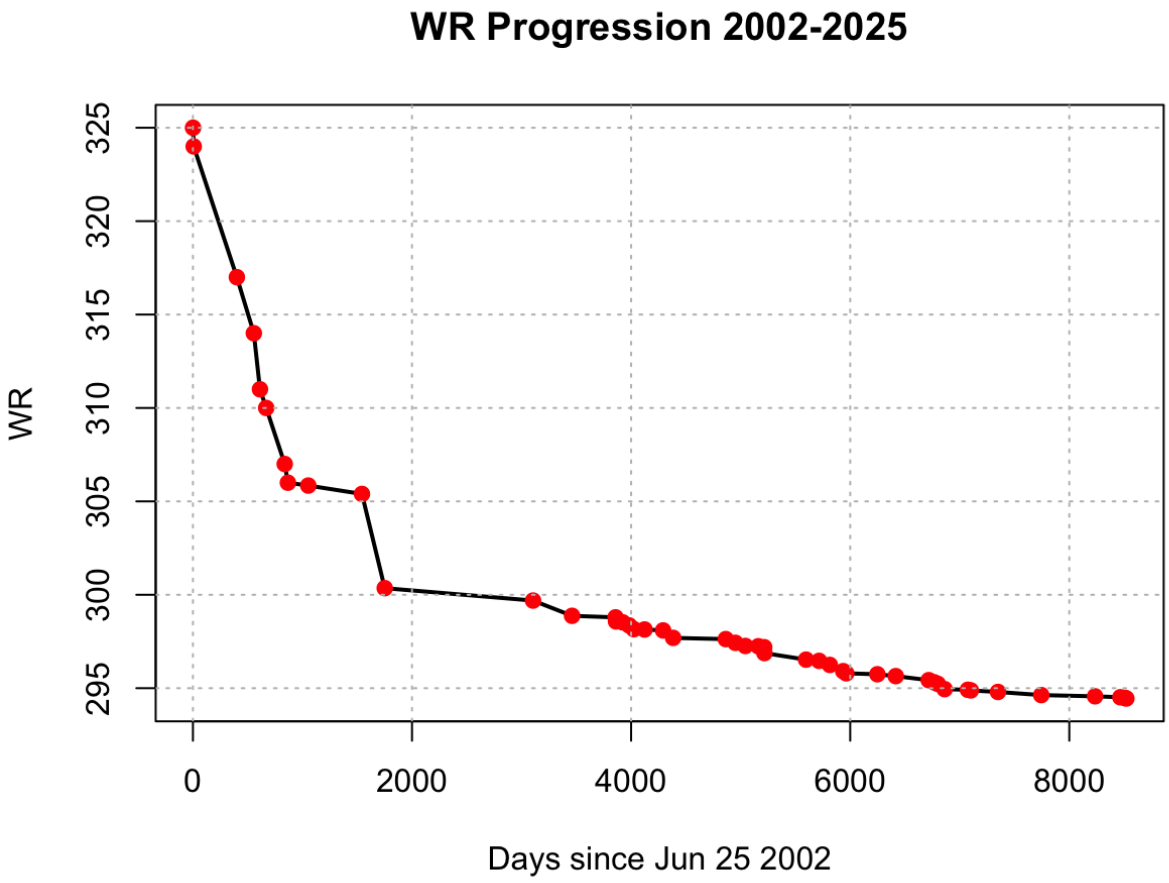

Figure 1: World Record progression from 2002 to 2025.

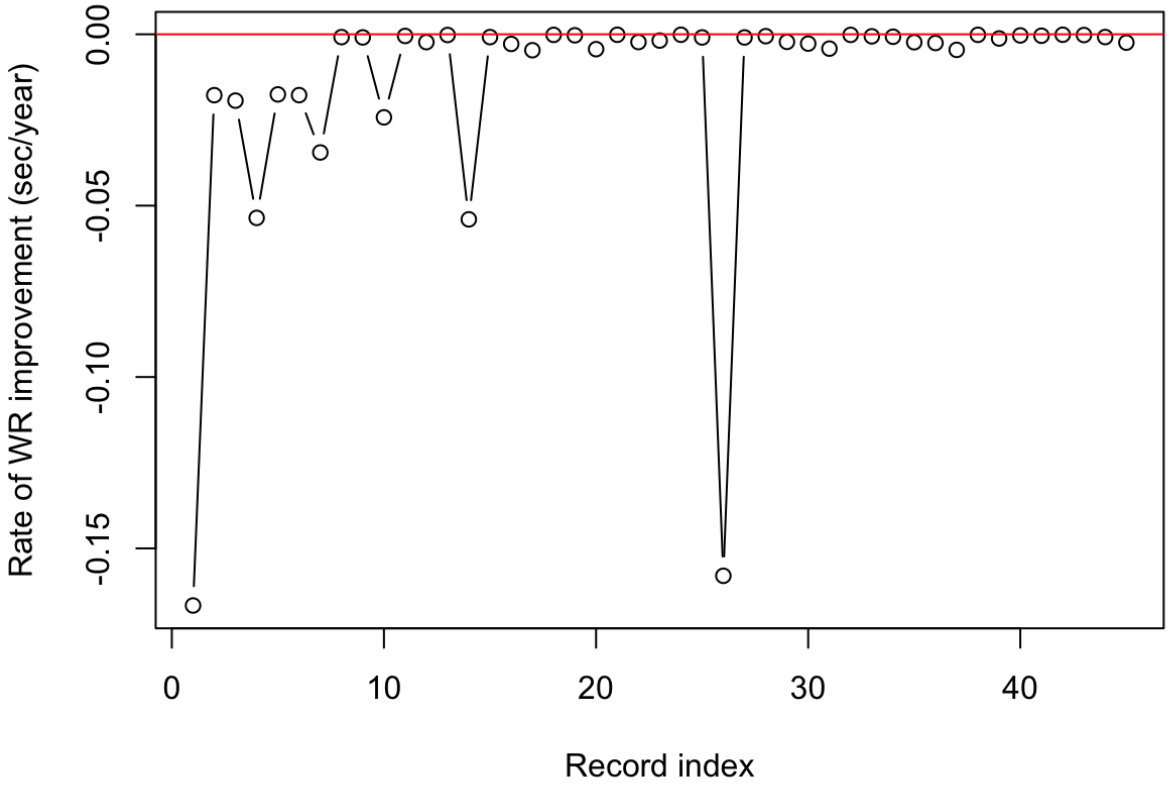

Figure 2: World Record average improvement over time.