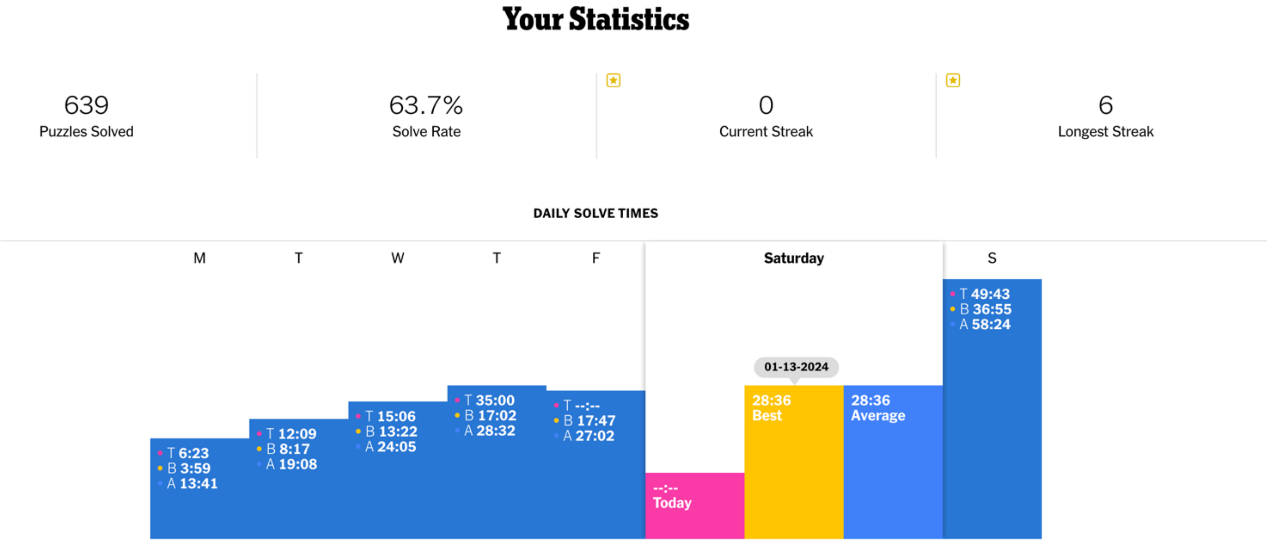
Crossword Data Statistics

Jordan Lerner
November 9, 2024

1. Project Description, Background, and Motivation

Easily recognized by their symmetric grids and numbered squares, crosswords have provided entertainment and fun for over a century. While crossword puzzles have existed in print media, such as newspapers, for most of their existence, the emergence of the internet and the digital world opened brand new doors of possibility for the future of crosswords.

The New York Times (NYT) is widely known and recognized for publishing high-quality and thought-provoking puzzles. They have three types of puzzles: Daily puzzles, which are the standard-size (15x15 grid Monday-Saturday and 21x21 grid on Sunday) puzzles that get released every day; Mini puzzles, which are a 5x5 grid and are released daily; Bonus puzzles, which are released on the first day of every month and typically are strongly tied to a theme. With such large amounts of data from their app, the NYT has an opportunity to provide users with statistics and visualizations about their solve times. However, nearly no data or insights are provided openly to users.



Your Statistics

639	63.7%	0	6
Puzzles Solved	Solve Rate	Current Streak	Longest Streak

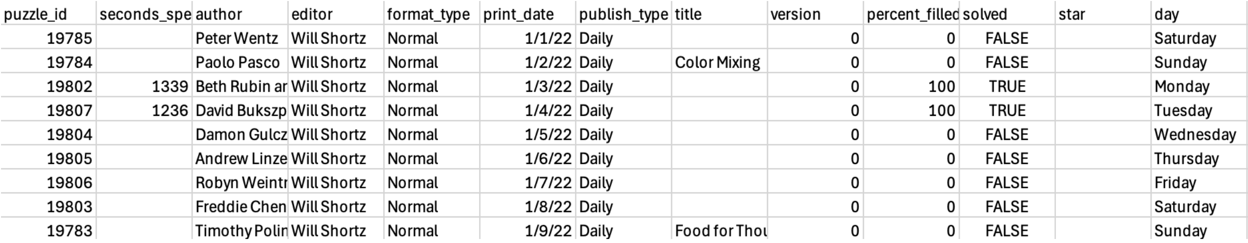
DAILY SOLVE TIMES

This is what the “Statistics” page looks like on the NYT Games website. You can see your overall solve rate (for all puzzles, including those in the archive) and then for each day of the week (puzzles get more difficult as the week goes on, starting on Monday as the easiest) they tell you about your time from that week, your best time, and your average time. There are no statistics provided for the Mini, however, there is a leaderboard that allows you to add friends and see their times.

I believe that there is a huge missed opportunity to provide some insightful statistics and analytics to users, and that is what I aim to provide with this project.

2. Data Description

The data that I am using for this project is my own crossword data that I obtained through the NYT website. I was able to access metadata for each crossword as well as records regarding my specific data. Since this data changes everyday, I wrote code that grabs all of the current data on the website starting at a specified date as well as code that grabs the previous days' data and adds it to the existing file.



puzzle_id	seconds_spe	author	editor	format_type	print_date	publish_type	title	version	percent_filled	solved	star	day
19785		Peter Wentz	Will Shortz	Normal	1/1/22	Daily		0	0	FALSE		Saturday
19784		Paolo Pasco	Will Shortz	Normal	1/2/22	Daily	Color Mixing	0	0	FALSE		Sunday
19802	1339	Beth Rubin ar	Will Shortz	Normal	1/3/22	Daily		0	100	TRUE		Monday
19807	1236	David Bukszp	Will Shortz	Normal	1/4/22	Daily		0	100	TRUE		Tuesday
19804		Damon Gulcz	Will Shortz	Normal	1/5/22	Daily		0	0	FALSE		Wednesday
19805		Andrew Linze	Will Shortz	Normal	1/6/22	Daily		0	0	FALSE		Thursday
19806		Robyn Weintr	Will Shortz	Normal	1/7/22	Daily		0	0	FALSE		Friday
19803		Freddie Chen	Will Shortz	Normal	1/8/22	Daily		0	0	FALSE		Saturday
19783		Timothy Polin	Will Shortz	Normal	1/9/22	Daily	Food for Thou	0	0	FALSE		Sunday

This is a screenshot of the Daily crossword data. It is stored in a CSV file. The columns are as follows:

puzzle_id – the unique ID for each NYT crossword
seconds_spent_solving – the amount of time (in seconds) that the puzzle has been solved for, whether it is complete or not
author – the author of the puzzle
editor – the editor of the puzzle
format_type – how the puzzle is displayed online
print_date – the date (YYYY-MM-DD) that the puzzle was released in print
publish_type – whether the puzzle is a daily, mini, or bonus puzzle
title – title of the puzzle, if there is one; typically only Sunday puzzles and Bonus puzzles have titles
version – if any changes were made to the puzzle after publishing
perfect_filled – the percentage of the grid that the user has filled in
soved – whether or not the puzzle is solved
star – NaN means that no star was given, blue means that the crossword was solved after 11:59pm PST of the print date and/or a hint was used when solving, gold means that the puzzle was solved on the print date and that no hints were used
day – day of the week that the puzzle was published

These columns are consistant for the CSV files for the mini and bonus puzzles.

As mentioned above, the data is refreshed every day. The data folder contains separate files for the daily, mini, and bonus crosswords. The naming convention for the file is (puzzle type)_(start date)_(end date).csv. For example, if a file contains data for the mini starting on August 21st, 2014 (the day that the first mini puzzle was published) and ending today (November 9th, 2024) the name of the file would be mini_20140821_20241109.csv.

3. Progress and Next Steps

Up to this point, I have written code that gets all past data for a date that is input as well as code that automatically runs every day to fetch the previous day's data and add it to each respective the CSV.

I need to create the visualizations that will be on the app and figure out how I want to configure the GUI. In regards to data, I would love to find a way to integrate global statistics into this somehow, but I have no leads on that right now.