Capstone Project: Analyzing the New York Times Crossword

Abby Spears

December 17 2023

Abstract

The New York Times crossword is one of the most prestigious and popular puzzle games in the United States. A new puzzle is published daily, with puzzles supposedly increasing in difficulty throughout the week. In this study, we hope to address the question of what makes a crossword puzzle difficult and to create a model to predict the difficulty of model. Using archived puzzles from the past 30 years, supplemented by an external source providing solve times and difficulties, we perform a series of visualizations as well as attempt to use a few logistic regression models in order to predict how difficult a puzzle will be to solve given the list of clues and answers.

Executive Summary

The primary goal of this project to understand the factors influencing difficulty of crossword puzzles based on clues and answers. This study makes use of lists of clues and answers from 30 years of archived puzzles as well as external solve time data in order to assess some of the factors that contribute to crossword difficulty.

In the data exploration step, we explore some ways that clues and answers can be quantified. This includes analysis of which clues and answers occur the most frequently. In particular, it was found that answers occurring most often are short words that make use of common letters, in order to act as glue between words in the crossword. For clues that occur frequently, we see two types - either straightforward fill-in-the-blank clues for more obscure terms, or short, vague clues that lend themselves to a variety of different answers. We compare our metrics across puzzles of different difficulties, and notice generally linear relationships between each metric and difficulty.

Finally, a few classification models were trained in order to distinguish between puzzles of different difficulties given some of the key puzzle metrics. When trying to distinguish between easy Monday puzzles and challenging Saturday puzzles, our classifier performed exceptionally well, with over 98% accuracy. This suggests that the editors and constructors at the New York Times take special steps when setting a Monday puzzle versus a Sunday puzzle. Similar models were constructed to distinguish between "hard" difficulty and easier difficulty on a fixed day of the week. These models did not perform quite as well, which presents some further questions for a future analysis.

Introduction

The first New York Times crossword was published in 1942, but the modern standard of the NYT crossword began when editor Will Shortz took over in 1993. Shortz aimed to change the puzzle to appeal to a more modern audience by focusing the puzzle away from obscure trivia and more towards wordplay and references to popular culture. For the purposes of this analysis, we begin with the first NYT crossword edited by Shortz on November 21, 1993, and end exactly 30 years later with the November 21, 2023 crossword.

The daily publication follows a similar structure each week. From Monday to Saturday, the crossword is on a 15x15 grid. The Monday puzzle is designed to be the easiest, with easy references and straightforward clues. Puzzles gradually get harder throughout the week, with Saturday as the hardest, often using more wordplay or obscure terms. The Sunday puzzle is usually the largest in terms of size and number of words, being on a 21x21 grid, but the difficulty level is aimed to be similar to a mid-week puzzle.

Crosswords are submitted by guest constructors, which are then chosen and edited by the New York Times staff. There are a few strict rules that crosswords must follow in order to be considered by the NYT - all cells must be part of both across and down answers, the entire grid must be interconnected, and every answer must be at least three letters long.

With that in mind, the vast archive of crosswords available online were scraped and saved into a large dataframe with columns for each clue and its corresponding answer, as well as the date the puzzle was published.

Section 2 contains data exploration and visualization, which reveals some of the most common words and clues used in the history of the New York Times. In Section 3, we build several different predictive models and find that harder puzzles tend to make use of metrics like more uncommon words, vaguer clues and more uncommon letters. We discuss the results of the model, including how to distinguish between puzzles published on different days of the week. Finally, we discuss conclusions, recommendations and ideas for future work in Section 4.

Data exploration and visualization

The main challenge of analyzing crossword puzzles is trying to quantify what makes a crossword puzzle challenging. Because crossword puzzles are primarily a language based game, there are certain nuances to clues that may depend on wordplay or alternative meanings, rather than simple definitions or synonyms. The challenge arises in how we can represent all of the "information" from the clue in a numerical way for the classification models.

The raw data was obtained from two main sources - XWordInfo, a puzzle archive, and XWStats, which allows fans to track their puzzle solve times. By web scraping these two pages, it was possible to get a list of all of the clue/word pairs, as well as a difficulty label ("Very Easy", "Easy", "Average", "Hard", "Very Hard") and average solve times for 30 years worth of puzzles.

Putting this together gave a very large dataframe (868200 rows!) full of every clue and answer pair ever found in the puzzle, as well. This was a great starting point, but did not lend itself well to analysis without the use of a sophisticated language model. Instead, it was necessary to come up with a series of metrics in order to determine how clues and answers might be different.

This is by no means a perfect list - and it misses out on some of the structure of the crossword puzzle. A real solver would have the crosses between across and down clues in order to help them fill in hard answers, which unfortunately was not considered in this study.

Metric	Description
Length	How long is the answer (in letters)
Percent vowels	What percentage of the word is vowels?
Scrabble score	The "scrabble score" of letters in the crossword answer word are added up and averaged by the length of the answer.
Answer frequency	How often has this specific answer appeared in the crossword?
Clue frequency	How often has this specific clue appeared in the crossword?
Real word	Is this answer a real word - i.e., can we find it the dictionary? These were decided using the `words` package, which provides a list of 175393 English words. If the answer is found in this list, we return True
Clue word count	How many words are in the clue?
Clue/Answer Pair	How many times has this specific clue paired with this specific answer appeared in the crossword?
Clue ratio	$\frac{\text{Clue/answer pair}}{\text{Clue frequency}}$ Given the clue, how many times was this exact clue used for the same answer?

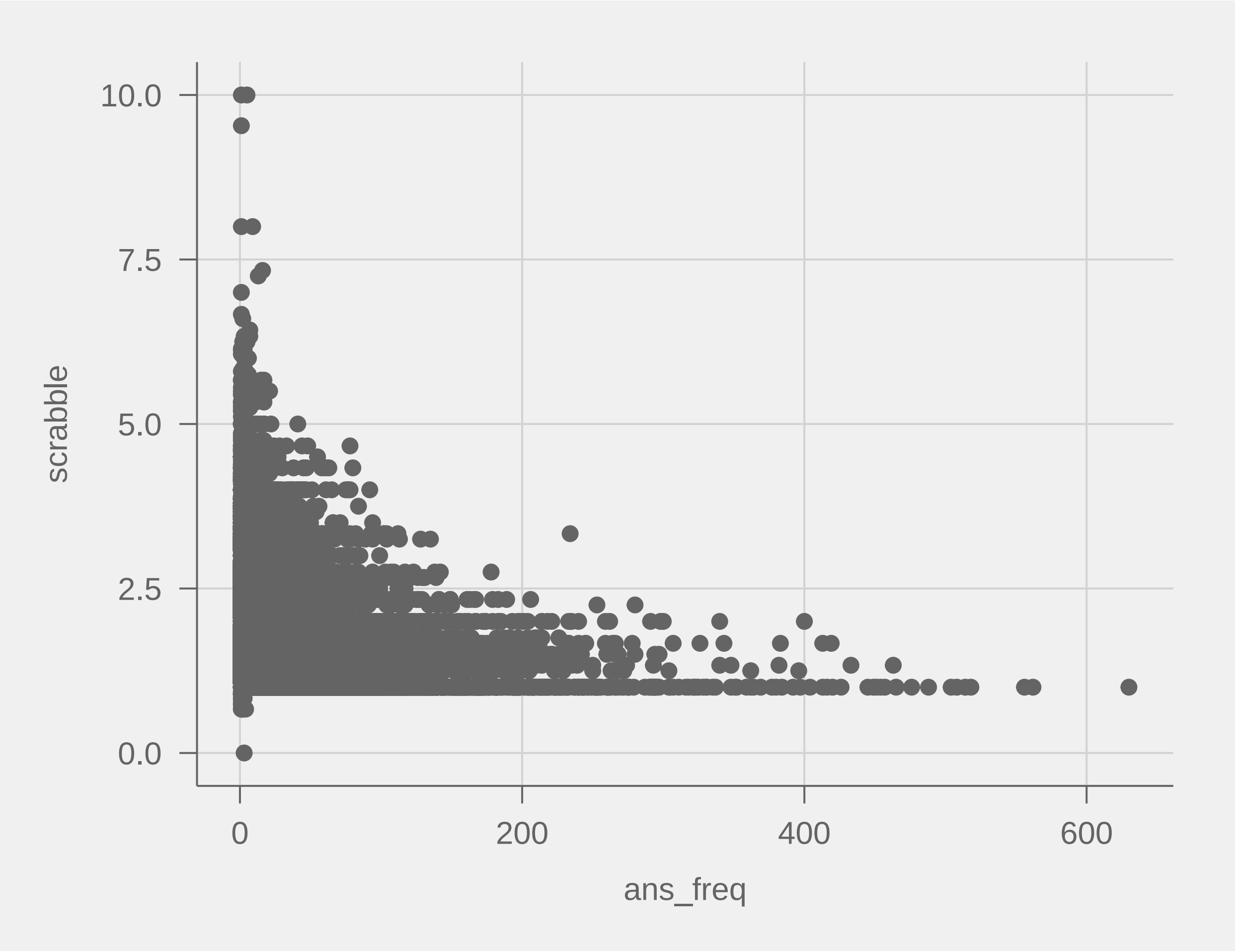
Metric	Description
Answer ratio	$\frac{\text{Clue/answer pair}}{\text{Answer frequency}}$ Given the answer, how many times was this exact clue used for the same answer?

My hope was these metrics when put together would help distinguish words into different groups that could be defined as "easier" or "harder". For instance, regarding answer length, we anticipated longer answers to be more likely to be unique to the puzzle and its specific theme, where as shorter answers might be more likely to be "filler" answers to bring the puzzle together. There are a few other nuances, for instance, in the "real word" category.

There were some interesting things to note looking at the clue and answer frequencies:

Answers	Frequency
ERA	630
AREA	562
ERE	556
ATE	518
ORE	514
ELI	508
ALE	504
ETA	488
ONE	476
ERR	465

It's easy to see that the words that show up most frequently in the crossword are short words with lots of vowels and other common letters - let's look briefly at the answer frequency compared with the average scrabble score. A lower score corresponds to more common letters, so obviously we see our lowest possible scores for our most commonly occurring answers.



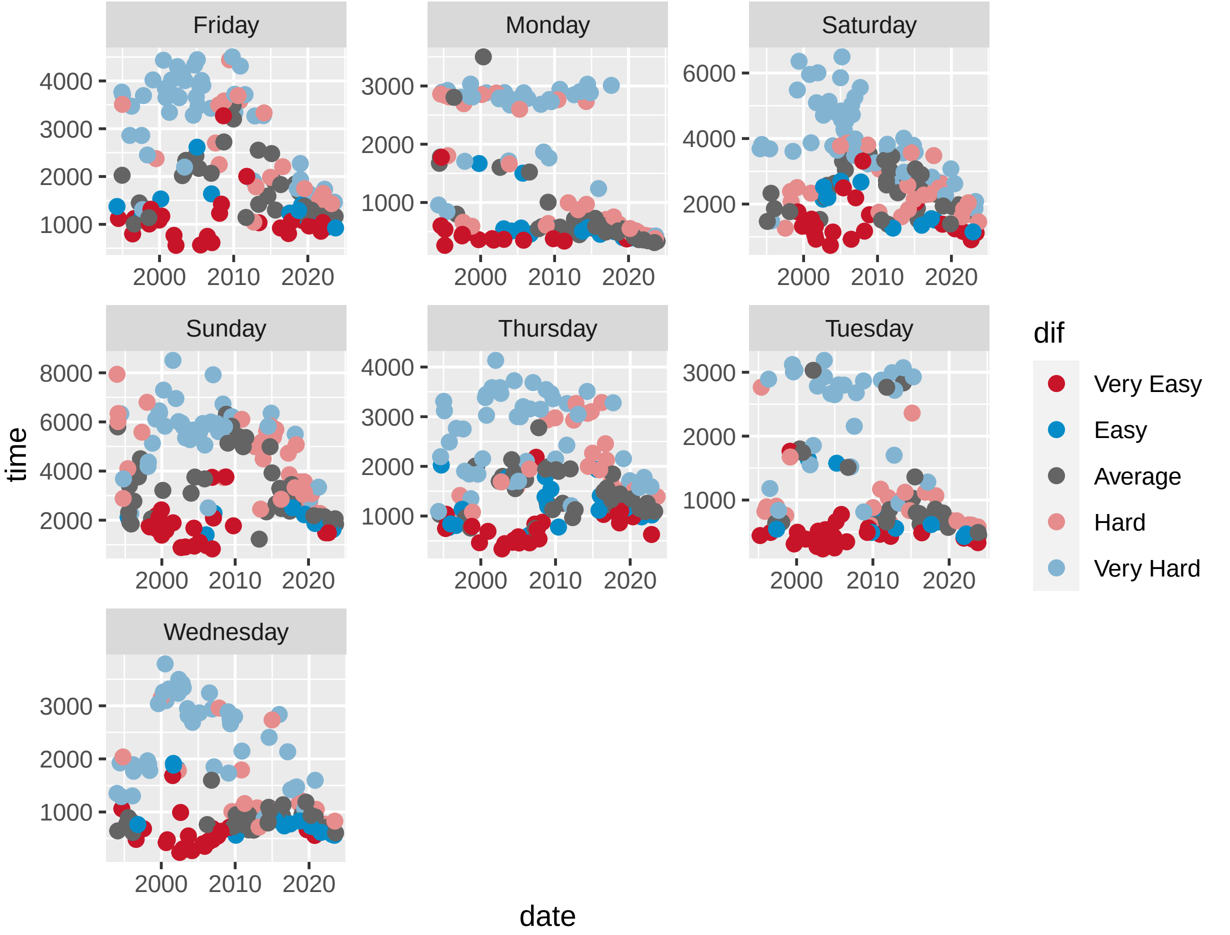
It’s even more interesting to look at the most common *clues* in the crossword.

clues	answers	clue_freq	ca_pair
Jai _____	ALAI	122	122
Mauna _____	LOA	116	66
Mauna _____	KEA	116	50
Up	RISEN	113	20
Up	ATBAT	113	18
Up	ARISEN	113	16
Up	ALOFT	113	14
Up	ASTIR	113	11
Up	AHEAD	113	9
Up	AWAKE	113	5
Up	RAISE	113	3
Up	ELEVATE	113	3
Up	NORTH	113	2
Up	INCREASE	113	2
Up	BOOST	113	2
Up	HIKE	113	2
Up	ADDTO	113	1
Up	HAPPY	113	1
Up	SKYWARD	113	1
Up	PERKY	113	1

clues	answers	clue_freq	ca_pair
Up	AWOKEN	113	1
Up	EACH	113	1

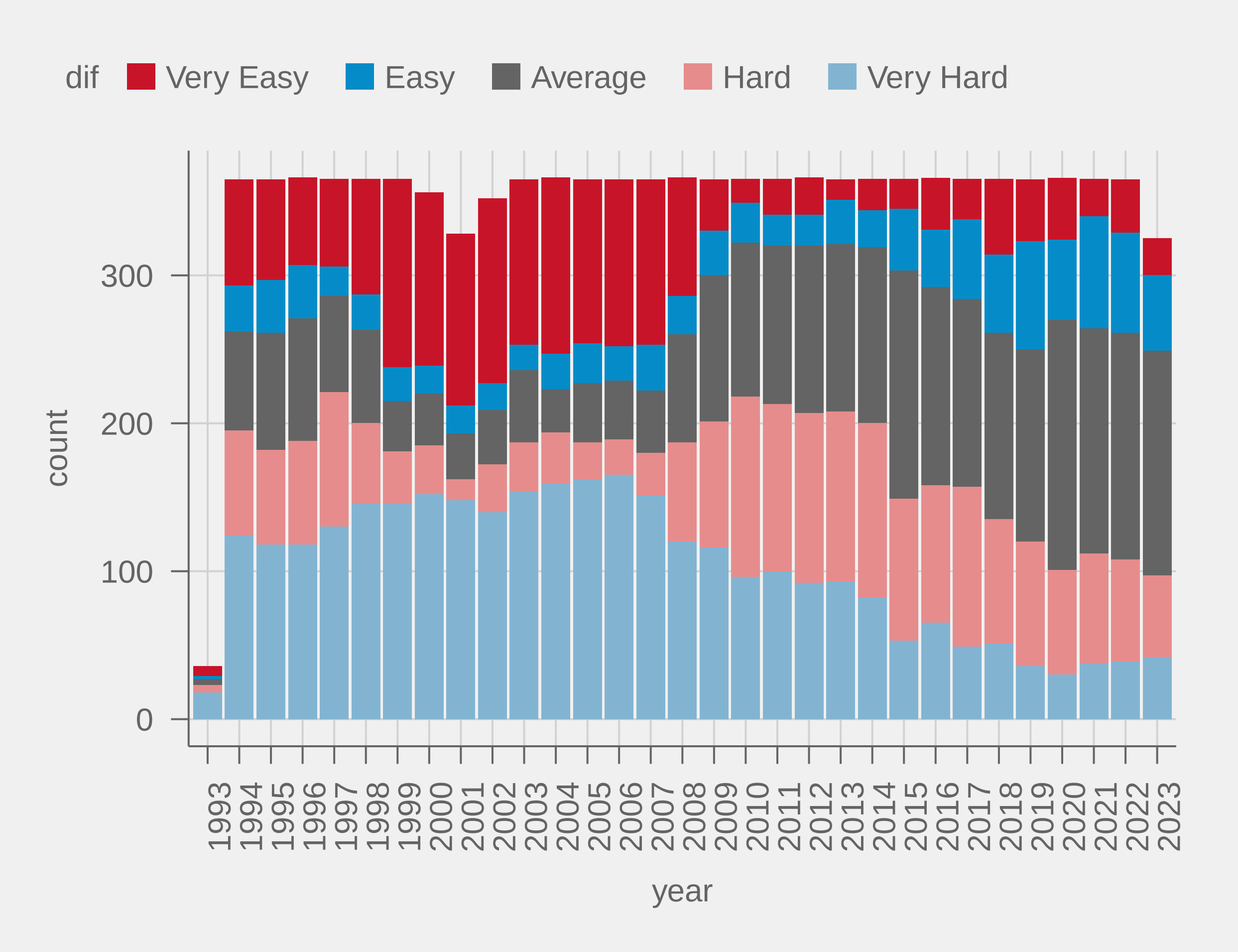
The most common clue in the puzzle is “Jai ______”, exclusively used to clue “ALAI.” This is an uncommon word that uses common letters, so it makes sense for it to be clued as a simple fill-in-the-blank. For an experienced crossword solver, this would be an automatic fill in and a very easy clue. Compare this with the third most common clue, “Up”. This is a vague, short clue and lends itself to 19 different answers - a combination of verbs, nouns, and adjectives. This would surely be considered a hard clue, so simply seeing what comes up often is not a surefire way to determine if a clue is easy or hard.

As stated previously, we combined the archived lists of clues and answers with some puzzle-wide statistics sourced from XWStats. XWStats is an opt-in tool for crossword fans to track their times to solve each puzzle. Puzzles are assigned difficulties based on how each solver’s solve time compares to their average. This is not a completely unbiased source - we must look at the difficulty and solve time metrics through the lens of an experienced daily crossword solver, who may be more familiar with the types of clues and words that commonly get asked in a crossword puzzle. It also means that unfortunately, we do not have a count for how many people solved each puzzle. Additionally, for older puzzles published before the XWStats tool was introduced, there is some degree of bias - one would expect only dedicated, serious solvers to go back and solve old puzzles and upload their solve times. But ignoring those factors just based on the availability of the data, let’s take a look at how solve times relate to difficulty on each day of the week.



Breaking it down by day of the week shows us mostly as expected, with harder puzzles taking longer to solve and easier puzzles taking less time, but there are a few outliers.

We also look at the breakdown over the years:



A very simple visual here shows that there appears to be a decrease in the number of puzzles described as "very hard" or "hard", but also a decrease in the number of puzzles described as "very easy" or "easy" over time, and a general increase in the number of "average" puzzle. This may be a result of more people solving and using the XWStats tool over time.

Modeling, Visualization, and Interpretation

In our exploration of the difficulties shown, we see there are really two different ways of measuring a difficult puzzle. We see that later in the week puzzles take longer to solve and are therefore constructed with the intent of being harder. We also see that within the same day of the week, sometimes puzzles are thought of as being "easier" or "harder" and may take a longer time to solve.

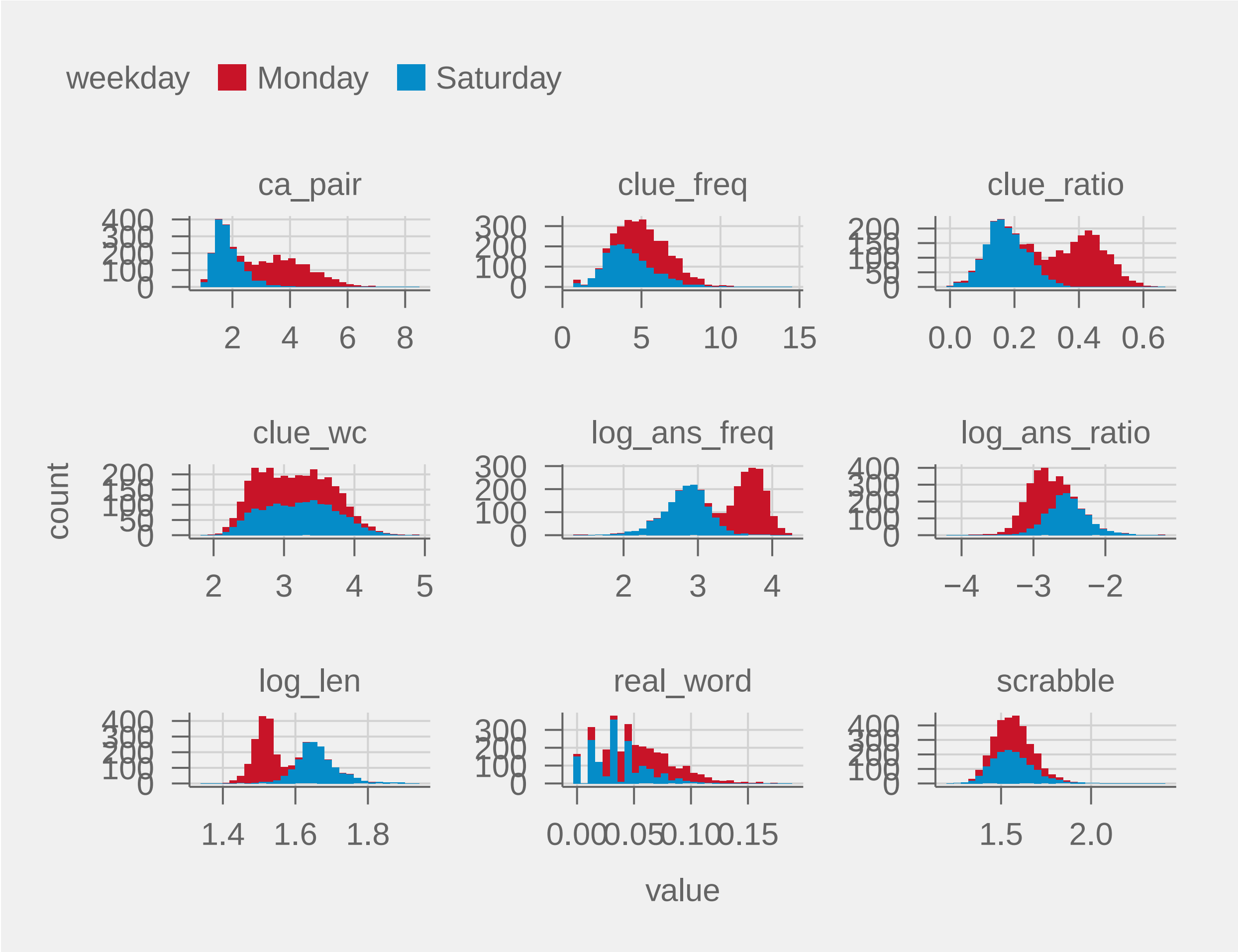
Thus my approach is to create two different types of classification model. One will be given the metrics descibed above, are we able to guess which day of the week this puzzle comes from? Second, given a fixed day of the week, are we able to guess whether it will be a hard or easy puzzle to solve?

We begin with a simple logistic regression model. Because the two measures of difficulty are on an ordered scale, either Very Easy-Very Hard or Monday-Saturday, it would be nice to preliminarily just try and binarily differentiate between the two groups.

Our first model aims to differentiate between a Monday puzzle and a Saturday puzzle.

In order to get our data ready for analysis, a few of the variables were log transformed in order to make the distributions closer to normal. Then, since each puzzle is made up of typically 60-75 clues, the average of each of the metrics for each day's puzzle was calculated. Each row in this model would have the average of each of the metrics for one such day, and then a final column "hardness", marked 1 for Saturday and 0 for Monday.

The histograms below compare our metrics between Monday and Saturday puzzles:



So now we fit our logistic regression model. The table below gives our coefficients for the model.

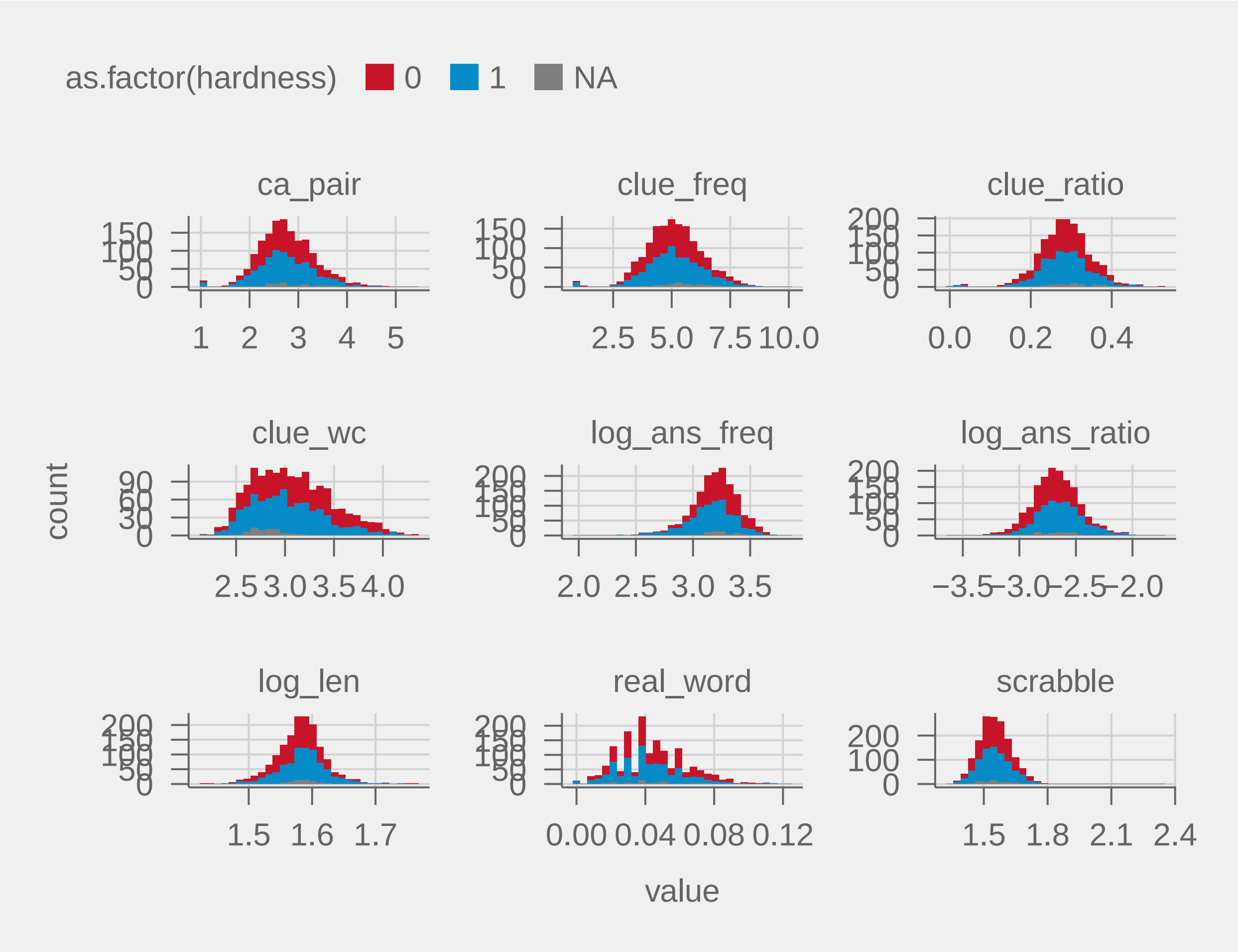
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	36.9794851	15.5710751	2.3748832	0.0175545
log_len	2.0398917	6.3339471	0.3220570	0.7474096
log_ans_freq	-16.2853421	5.8651203	-2.7766425	0.0054924
clue_freq	2.1782827	0.2732957	7.9704251	0.0000000
real_word	-15.0035236	6.7961742	-2.2076426	0.0272692
clue_wc	-0.2860841	0.3259696	-0.8776403	0.3801389
scrabble	-4.3857813	2.0588198	-2.1302405	0.0331518
ca_pair	-3.5272422	0.7110604	-4.9605377	0.0000007
clue_ratio	-2.3358333	6.8063283	-0.3431855	0.7314589
log_ans_ratio	-7.9853361	5.7689539	-1.3841913	0.1662999

We can see here that the most significant coefficients are the clue frequency, clue/answer pair frequency, and answer frequency. The negative coefficients for answer frequency and clue/answer pair frequency suggest that answers and clue/answer pairs that come up a lot, i.e. “gimme” clues, are more common on Mondays and less common on Saturdays. This makes a lot of sense! What this tells us is that the constructors and editors are using more obscure terms and harder things to guess on Saturdays than they are on Mondays. Likewise, the positive coefficient on clue frequency means that more common *clues* occur more often on Saturdays. This is interesting! My thoughts are that this relates to instances like we saw with the clue “Up” as discussed in the introduction, in which the same vague, short clue is used to hint for a variety of different terms.

If we assess the fit of our model, using a threshold of 0.5 to make our predictions, we get 98.978% accuracy, which is really great! This shows that our model does a really good job of differentiating between Mondays and Fridays. What this implies is that the editors and constructors of the crossword follow different guidelines when constructing a puzzle for Monday as opposed to Saturday, and these reasons make the puzzle more difficult for a solver.

Now let’s look at a slightly different approach. We have our difficulty measures and times given from the XWStats page. Can we try and predict those?

It makes sense to first limit ourselves to one particular day of the week. Let’s start with Sunday. We will use a similar logistic regression model, but this time we will try to predict whether the puzzle is hard to solve or not.
Since we are just using a binary classifier for now, let’s say “Hard” and “Very Hard” get a “hardness” of 1, and “Very Easy”, “Easy”, and “Average” get a hardness of 0.



Looking at the plot above, the hard and easy puzzles seem to overlap quite a bit. We would anticipate the model to perform less well than our day of the week classifier.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	21.3549927	5.2177785	4.0927365	0.0000426
log_len	-5.1032431	2.1871161	-2.3333207	0.0196313
log_ans_freq	-1.6908824	1.9630399	-0.8613591	0.3890403
clue_freq	0.0403670	0.0845004	0.4777141	0.6328537
real_word	-4.0724412	2.9036517	-1.4025240	0.1607588
clue_wc	-0.7622845	0.1520893	-5.0120862	0.0000005
scrabble	-0.7845016	0.7859822	-0.9981163	0.3182230
ca_pair	-0.2522926	0.2217726	-1.1376186	0.2552798
clue_ratio	-1.8073778	2.3159927	-0.7803901	0.4351613
log_ans_ratio	1.1603880	1.9504628	0.5949296	0.5518905

Here, the only significant coefficients are the clue word count and the answer length. With negative coefficients, it seems as if harder puzzles have shorter clues, possibly more vaguely written, and easier puzzles might be a bit more descriptive. Likewise, it seems as if harder puzzles may have longer, grid-spanning answers.

However, this model only performs with accuracy of about 60.359%, so it is not the highest performing model.

The same model was conducted for each day of the week, each with similar degrees of accuracy in prediction, around 60%.

	Accuracy
Sunday	0.6035857
Monday	0.6366539
Tuesday	0.6121406
Wednesday	0.6016624
Thursday	0.5907928
Friday	0.6035806
Saturday	0.5716113

This suggests to me that while the differences in difficulty between days of the week are a result of deliberate choices made by the editors and constructors, the differences in difficulty between puzzles on the same day

Conclusions and recommendations

The analysis above only begins to get into some of the nuances that make a crossword puzzle difficult. While we tried to explore in depth some of the frequencies of clues and answers as well as frequencies of letters in this study, there is much more that makes up a successful crossword puzzle. For one, we did not address at all the actual "crossing" structure of the puzzles with across and down clues, nor did we acknowledge the themed answers commonly found throughout the week.

Overall, there is a notable difference in the way crossword puzzles are set throughout the week. Easier puzzles, like on Mondays, tend to use common answers and familiar cluing to make it straightforward to solve. Harder puzzles, like on Saturdays, tend to use more obscure answers and vague cluing to make it more difficult. Harder puzzles also tend to have longer words and shorter clues.

In a future study it would be interesting to look at additional metrics, such as the frequency of words in English compared to the frequency of words in the crossword. Future modeling might consist of clustering analysis to look at answers and clues and put them into similar categories to see how they might be harder. It also may involve using some sort of network structure to show how hard clues might cross with easier clues. It would also be interesting to look at some level of vector similarity between clues to get a bit more nuance than when clues are used multiple times. Overall, it was interesting to think about how one might approach the problem at first, and this has opened up many new channels for future analyses.