Proposal

Kunning Shen

Section 1 - Introduction

In the digital age, the strength and security of passwords have become paramount in protecting personal and corporate information from unauthorized access. This study embarks on a comprehensive examination of password characteristics, their security implications, and the development of a password security assessment model.

I found this data in Kaggle, which is sourced from Information is Beautiful and is also twinned with a visualisation from the book, Knowledge is Beautiful by David McCandless (out Sep 2014). The dataset includes variables such as password popularity (rank), the actual text of the password, its category, the estimated time to crack the password online, and a measure of its strength.

It is worth noting that this data was collected before 2014. At that time, there were not as many requirements for setting passwords as there are now. Therefore, not only does this data give a true picture of people's password preferences, but I hope that the modeling of passwords and their strengths will help us figure out why there are so many requirements for setting passwords now, what constitutes a complex password, and the extent to which complex passwords can keep our information secure.

My investigation is structured around three key research questions:

1. Identifying common characteristics of popular passwords, providing insights into user behavior and preferences when selecting passwords.

2. Exploring the relationship between password popularity and strength, to discern whether a tendency exists towards the use of less secure passwords and to understand the potential risks associated with these choices.

3. Developing a model that considers various features of passwords to assess their security, aiming to pinpoint significant factors that influence password strength and thereby inform better password creation practices.

Section 2 - Data

The codebook generated by dataReporter package is placed at the end of this proposal.

Section 3 - Data analysis plan

Describe the key variables to answer your question

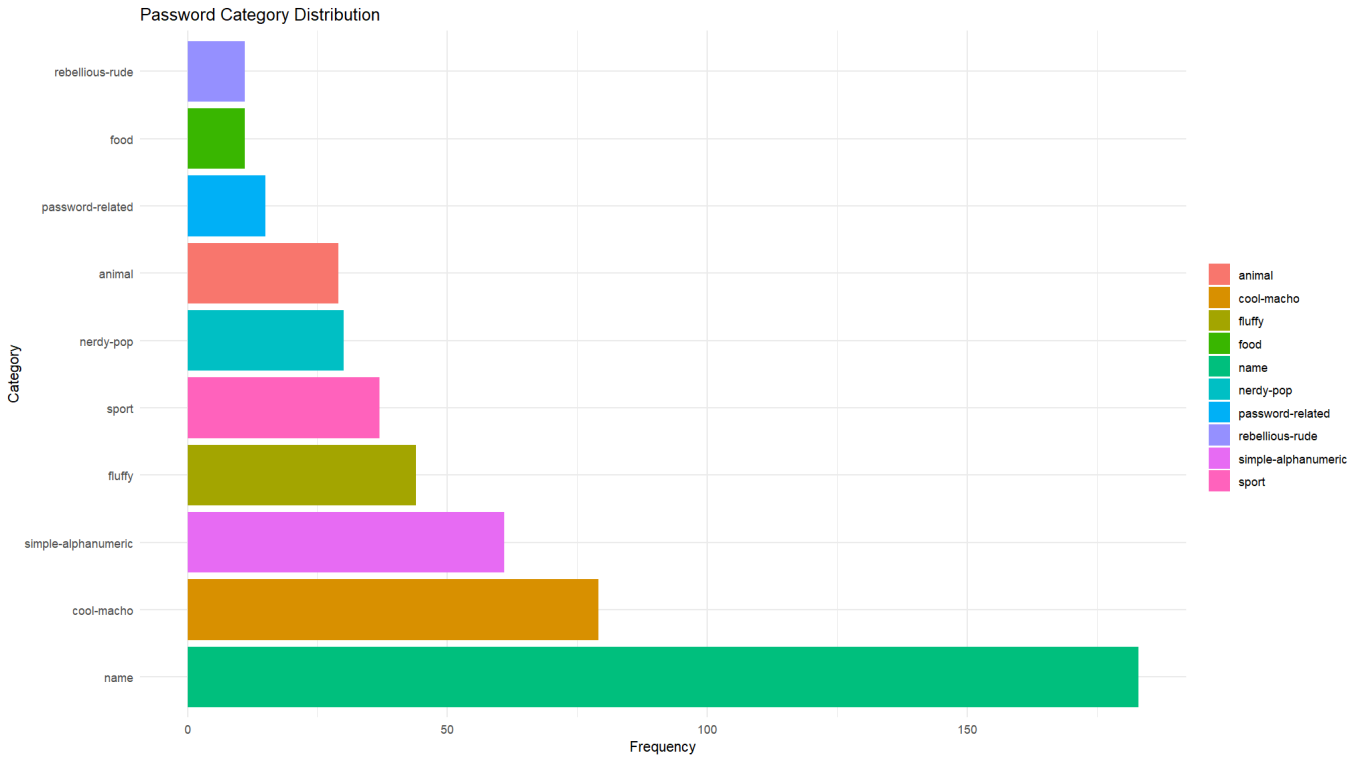
1. **Password**, **category** and **rank**. These three variables will reveal the preferences for passwords.

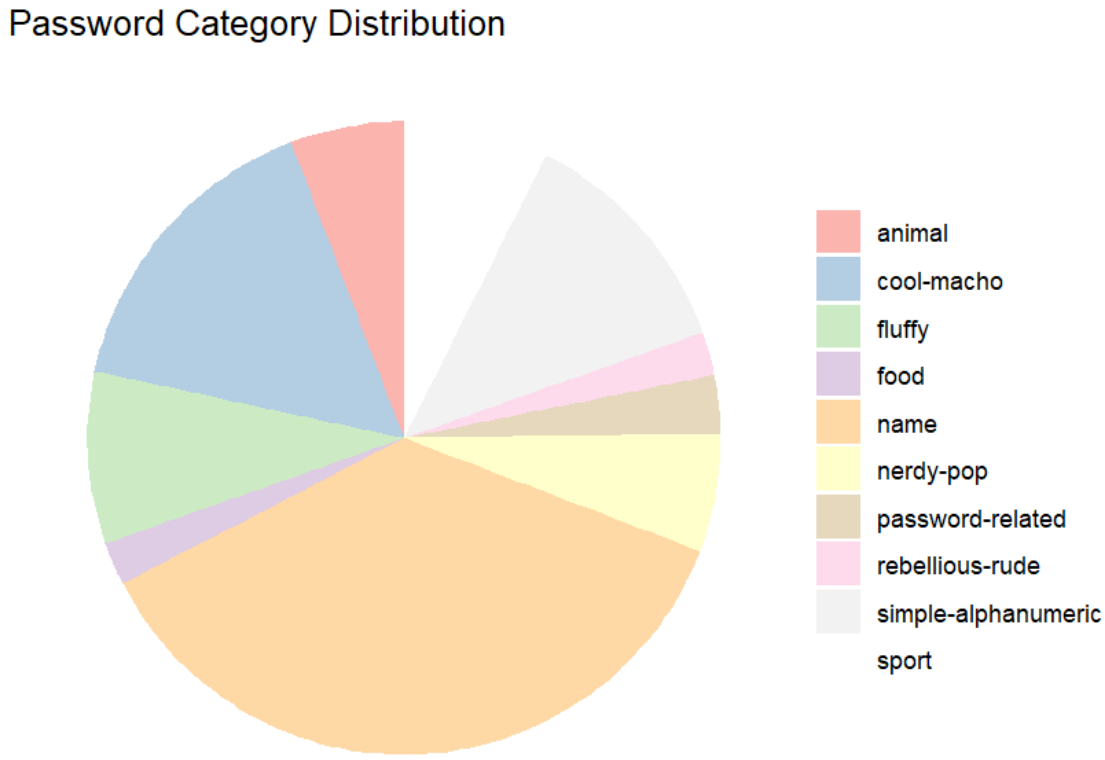
2. **Password** and **Strength**. This pair of variables will show the relationship between uses' password and strength. Here, the *password* is not only the text of users' passwords, but also includes many textual properties we can extract them. The password strength prediction model will also mainly rely on these two variables.

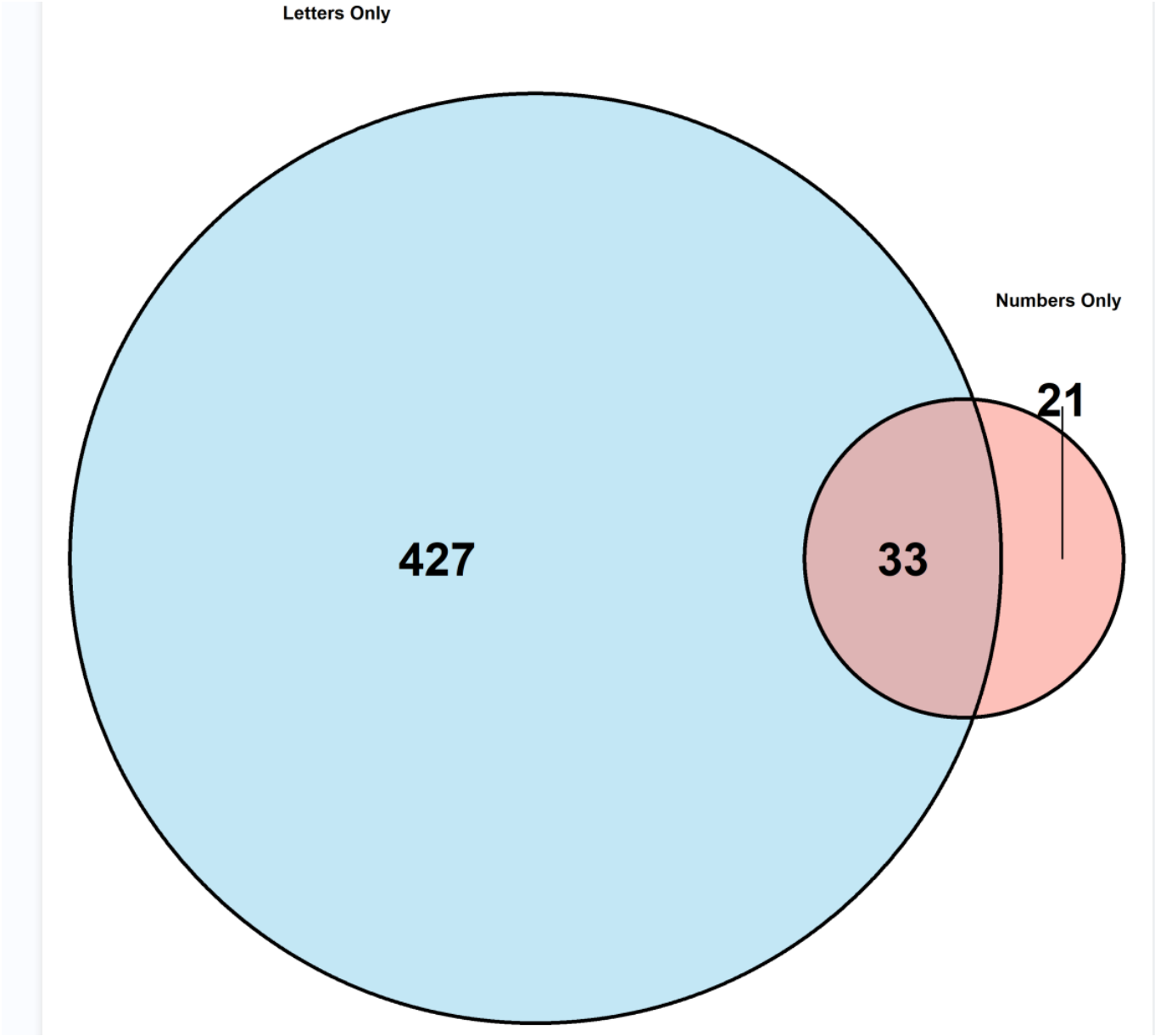
3. **Strength**, **value** and **offline_crack_sec**. These three variables will demonstrate what password strength actually means. Relating an untouchable strength to the actual time of cracking will provide a more intuitive picture of what it means to have a strong password.

Preliminary exploratory data analysis

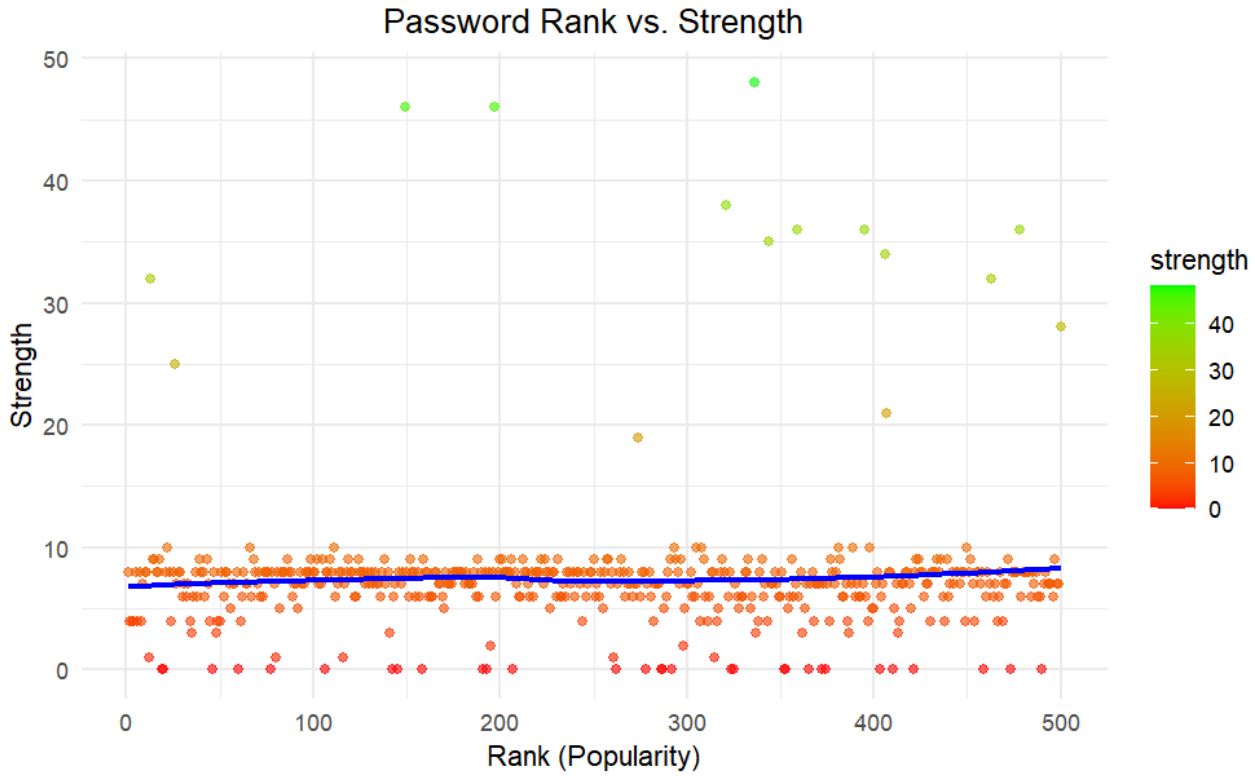
1. User Preferences for Passwords


Password Category Distribution
rebellious-rude
food
password-related
animal
nerdy-pop
sport
fluffy
simple-alphanumeric
cool-macho
name
Category
0
50
100
150
Frequency
animal
cool-macho
fluffy
food
name
nerdy-pop
password-related
rebellious-rude
simple-alphanumeric
sport

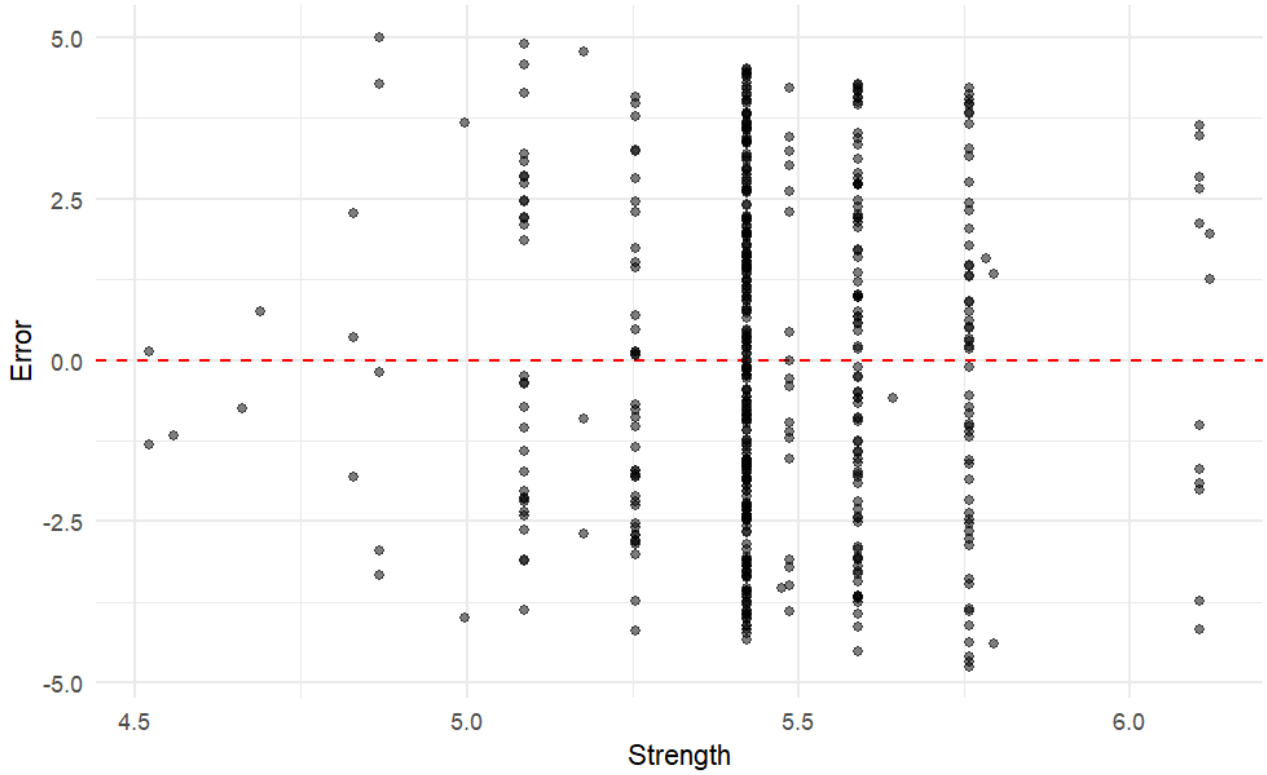

Password Category Distribution
animal
cool-macho
fluffy
food
name
nerdy-pop
password-related
rebellious-rude
simple-alphanumeric
sport


Letters Only
Numbers Only
427
33
21

2. The relationship between password popularity and strength



3. Predictions of password strength using category, password length, number of letters, number of digits using a linear regression model and plotting the errors.



The above preliminary exploratory work is only used to demonstrate that the research questions posed above are feasible on this dataset. Subsequent formal work will lead to more profound work as well as better visualization.

Methodology

1. **Statistical Analysis and Visualization:** Utilize descriptive statistics and visualizations to identify common characteristics of popular passwords. This can include frequency distributions, word clouds for textual analysis, and bar charts comparing categories.

2. **Regression Analysis:** To explore the relationship between password popularity and strength, I will apply regression analysis. I will try both linear and nonlinear regression models.

Codebook for password_data

Autogenerated data summary from dataReporter

2024-03-07 03:10:16.755357

Data report overview

The dataset examined has the following dimensions:

Feature	Result
Number of observations	500
Number of variables	8

Codebook summary table

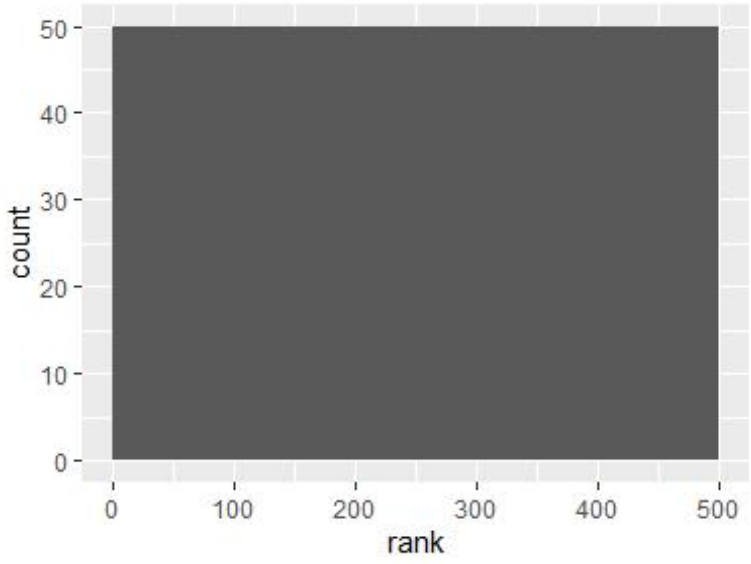
Label	Variable	Class	# unique values	Missing	Description
	rank	integer	500	0.00 %	Popularity in their database of released passwords
	password	character	500	0.00 %	Actual text of the password
	category	character	10	0.00 %	What category does the password fall in to?
	value	numeric	15	0.00 %	Time to crack by online guessing
	time_unit	character	7	0.00 %	Time unit to match with value
	offline_crack_sec	numeric	16	0.00 %	Time to crack offline in seconds
	strength	integer	22	0.00 %	Quality of password where 10 is highest, 1 is lowest, please note that these are relative to these generally bad passwords
	font_size	integer	19	0.00 %	Font size is related to popularity and is used to

Label	Variable	Class	# unique values	Missing	Description
					aid visualization

Variable list

rank

Feature	Result
Variable type	integer
Number of missing obs.	0 (0 %)
Number of unique values	500
Median	250.5
1st and 3rd quartiles	125.75; 375.25
Min. and max.	1; 500



---

password

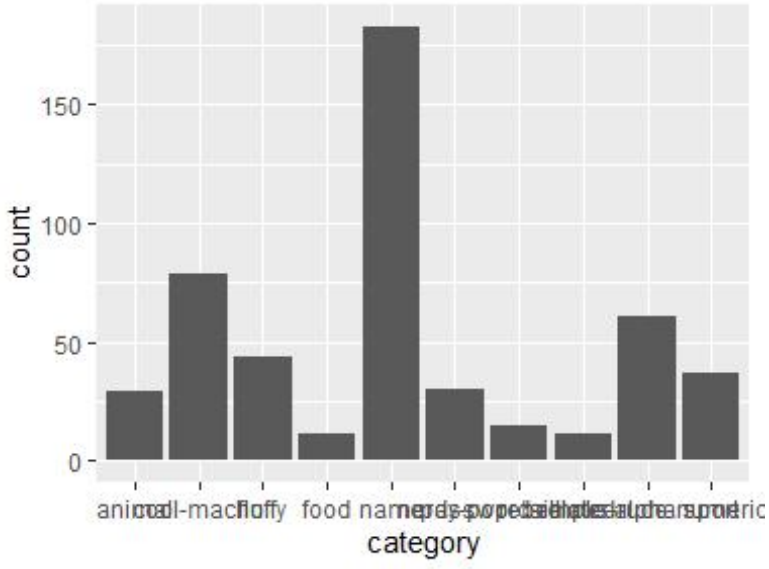
- The variable is a key (distinct values for each observation).

---

category

Feature	Result
Variable type	character
Number of missing obs.	0 (0 %)

Feature	Result
Number of unique values	10
Mode	“name”

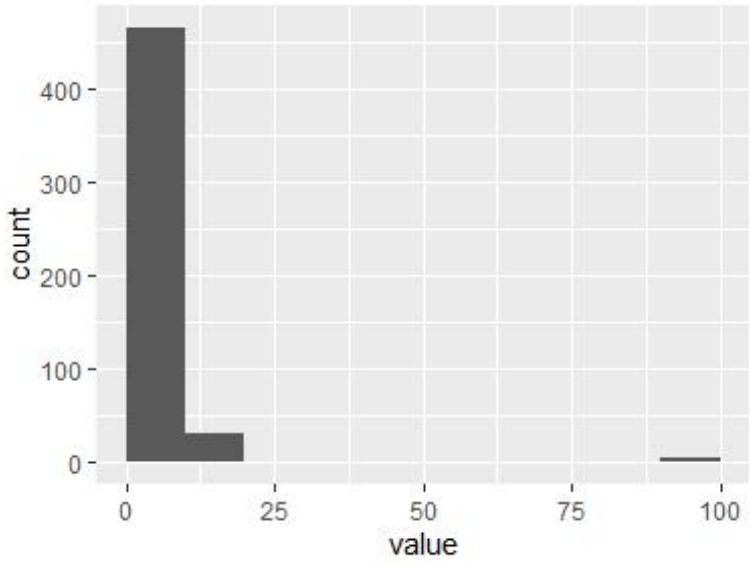


- Observed factor levels: "animal", "cool-macho", "fluffy", "food", "name", "nerdy-pop", "password-related", "rebellious-rude", "simple-alphanumeric", "sport".

---

value

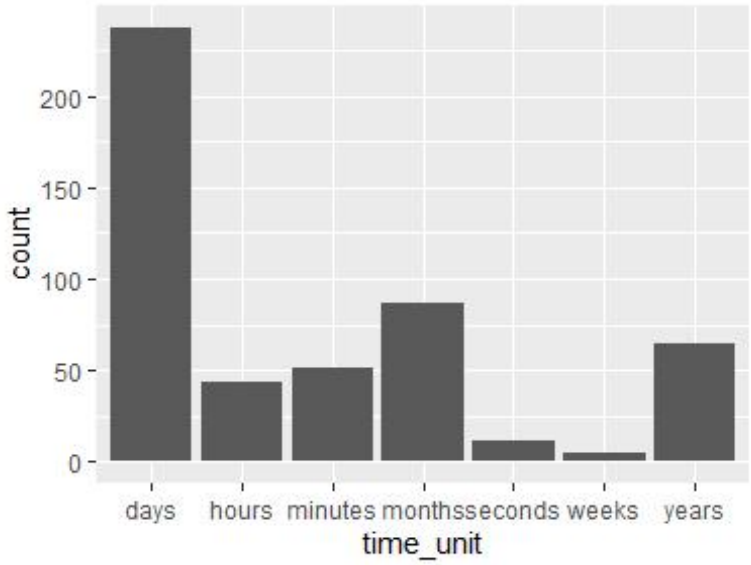
Feature	Result
Variable type	numeric
Number of missing obs.	0 (0 %)
Number of unique values	15
Median	3.72
1st and 3rd quartiles	3.43; 3.72
Min. and max.	1.29; 92.27



---

time_unit

Feature	Result
Variable type	character
Number of missing obs.	0 (0 %)
Number of unique values	7
Mode	“days”

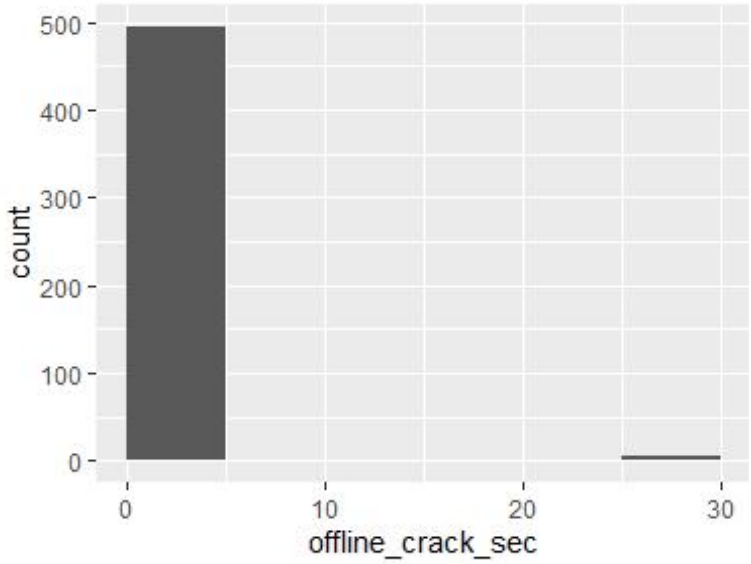


- Observed factor levels: "days", "hours", "minutes", "months", "seconds", "weeks", "years".

---

offline_crack_sec

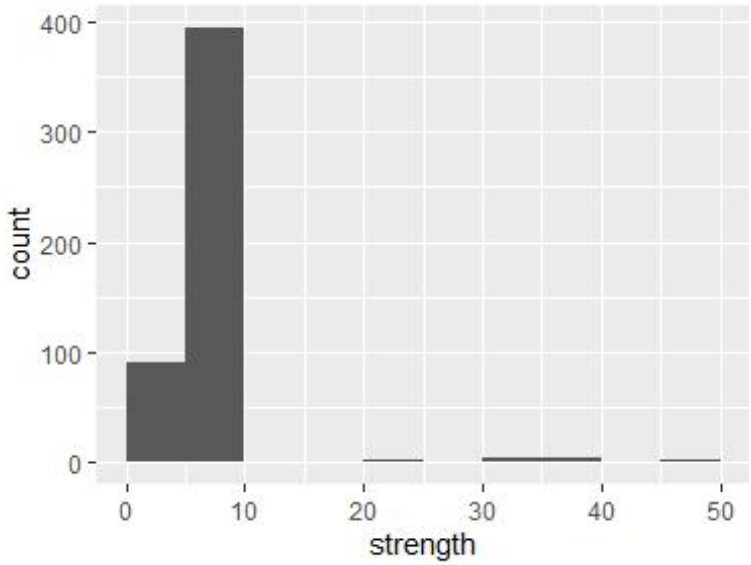
Feature	Result
Variable type	numeric
Number of missing obs.	0 (0 %)
Number of unique values	16
Median	0
1st and 3rd quartiles	0; 0.08
Min. and max.	0; 29.27



---

strength

Feature	Result
Variable type	integer
Number of missing obs.	0 (0 %)
Number of unique values	22
Median	7
1st and 3rd quartiles	6; 8
Min. and max.	0; 48



font_size

Feature	Result
Variable type	integer
Number of missing obs.	0 (0 %)
Number of unique values	19
Median	11
1st and 3rd quartiles	10; 11
Min. and max.	0; 28