Exploratory Data Analysis on Username-Password Dataset

Vanita Jain, Rishab Bansal, Mahima Swami

[1]Bharati Vidyapeeth's College of Engineering, INDIA
Emails: vanita.jain@bharatividyapeeth.edu; rishabbansal.it1@bvp.edu.in; mahima.it1@bvp.edu.in

*Correspondence: vanita.jain@bharatividyapeeth.edu

Abstract

Passwords act as a first line of defense against any malicious or unauthorized access to one's personal information. With the increasing digitization, it has now become even more important to choose strong passwords. In this paper, the authors analyze a 100 million Email-Password Database to perform Exploratory Data Analysis. The analysis provides valuable insights on statistics about the most common passwords being used, character set of passwords, most common domains, average length, password strength, frequencies of letters, numbers, symbols (special characters), most common letter, most common number, most common symbol, the ratio of letters, numbers, symbols in passwords which highlights the general trend that users follow while creating passwords. Using the results of this paper, users can make intelligent decisions while creating passwords for themselves, i.e., not opting for the most common features that will help them create robust and less vulnerable passwords.

Keywords: Data Analysis; Username-Password Dataset; Data Security

1. Introduction

Everything has a password [1][2] connected to it in this digital world, from doing online transactions to updating profile pictures on social media platforms; everything requires passwords as proof of authentication. Having a secure and robust password has become a significant necessity nowadays. A single user can have multiple passwords for multiple accounts. With the increasing number of passwords, it becomes challenging to remember complex passwords because of which people tend to select common and weak passwords. Often people use the same passwords for multiple accounts. This makes their multiple accounts vulnerable in the case of a data breach [3].

Some websites have now started encouraging [4] users to choose a strong password consisting of different character sets like a combination of lowercase letters, uppercase letters, symbols (special characters), numbers, and longer in length (typically, greater than or equal to 8). To increase security, websites have also started offering 2 Factor Authentication [5]. 2 Factor Authentication (2FA), as the name suggests, requires more than one form of verification. This means only a password is not enough to log in. An OTP or an Auth Code is also required to log in. It is always recommended to enable 2FA when possible.

There are many ways available by which hackers can break into an account, either by guessing or cracking passwords. Attacks like Dictionary Attack [6] and Brute Force Attack[7][8] can help break into an account with a weak password. Dictionary Attack involves automating a script that can try all the words of a dictionary or a word list provided. Brute Force Attack involves running a script that tries all the possible combinations of the supplied character set. Many passwords cracking wordlists like rockyou[9], CrackStation[10], Weakpass[11], and SkullSecurity [12] are also available on the web for hackers to launch an attack on accounts having weak passwords.

DOI: https://doi.org/10.54216/FPA.040101
Received January 07, 2021 Accepted May 10, 2021

Some passwords can also contain Unicode Characters. Unicode Characters [13] are special characters that are out of the ASCII[14] range of 128 and are commonly found on a general keyboard layout. The results of this analysis can also help readers/users make intelligent decisions about their passwords and choose a strong password for their accounts.

2. Methodology

2.1 Collecting Data

A dataset of emails-passwords used for the research is available at [15] was downloaded. The size of this dataset is 44.2 GB. It contains more than 100 million emails and passwords collected from different breaches and compiled for educational and research purposes.

2.2 Cleaning Data

The dataset had a lot of anomalies that did not align with the scope of the analysis, like some of the email-password combinations were blank, contained Unicode Chars, and duplicate entries were also present. All these anomalies were selected and deleted from the dataset to meet the scope of the analysis. The cleaning was performed with the code written by the authors in python [16] and is available at [17].

2.3 Sorting Data

The dataset is very large, and performing the analysis in one go is difficult. The dataset was broken into smaller sections to ease the process of analysis. The sorting was performed with the code written in python by the authors and is available at [17].

2.4 Performing Exploratory Data Analysis (EDA)

For performing EDA, some basic features were selected upon which the analysis was performed. The code for all the following subsections is available at[18], which is also written by the authors.

2.4.1. Password Strength

We defined three different categories of passwords: weak, moderate, and strong. To define the parameters, we used the following approach.

The score was incremented if one of the following conditions were met:

2.4.1.1 Length greater than or equal to 8

2.4.1.2 Lowercase letter used

2.4.1.3 Uppercase letter used

2.4.1.4 Number used

2.4.1.5 Symbol (special characters) used

Weak: If the score lies between 0 and 2

Moderate: If the score lies between 2 and 4

DOI: https://doi.org/10.54216/FPA.040101
Received January 07, 2021 Accepted May 10, 2021

Strong: If the score is greater than 4

2.4.2. Character set of passwords

All the passwords were run through a counter program which kept a count of the character set of all the passwords. The following are the seven-character set categories in which the passwords were categorized.

2.4.2.1 Small(lowercase) with Numbers

2.4.2.2 Small(lowercase) without Numbers

2.4.2.3 Big(uppercase) with Numbers

2.4.2.4 Big(uppercase) without Numbers

2.4.2.5 Small(lowercase) + Big(uppercase) with Numbers

2.4.2.6 Small(lowercase) + Big(uppercase) without Numbers

2.4.2.7 Small(lowercase) + Big(uppercase) + Numbers without Symbols (special characters)

2.4.3. Lowercase Letter Frequency

A counter program iterated through all the passwords and calculated the frequency of occurrences of all the lowercase letters. The results were presented in a bar graph along with their frequencies.

2.4.4. Uppercase Letter Frequency

A similar counter program iterated through all the passwords and calculated the frequency of occurrences of all the lowercase letters. The results were presented in a bar graph along with their frequencies.

2.4.5. Numbers Frequency

All the passwords were passed through an iterator which kept count of the frequency of all the numbers occurring in the passwords. The results were presented in a bar graph along with their frequencies.

2.4.6. Symbols (Special Characters) Frequency

To calculate the frequency of symbols like @, ! #, etc., all the passwords were checked for the occurrence of symbols. A counter function kept the count of the occurrences of symbols with their frequencies. The results were presented in a graph.

2.4.7. Most Common Password

A Python dictionary was created to keep a count of all the unique passwords along with their frequencies to find the most common password in the dataset. The top 10 most common passwords, along with their frequencies, were presented in a bar graph.

2.4.8. Most Common Domain

The emails present in the dataset had their domain name also present in the email id. All the unique domains along their frequencies were then plotted in a bar graph.

2.4.9. Most Common Unicode Characters

The dataset also contained a lot of Unicode Characters in the password. All the occurrences of Unicode Characters were plotted in a bar graph along with their frequencies.

DOI: https://doi.org/10.54216/FPA.040101
Received January 07, 2021 Accepted May 10, 2021

2.4.10. The ratio of the number of Alphabetic Letters to the length of the password

It signifies how much percentage of the password is composed of Alphabetic Letters. The ratio was calculated on an individual password level and then averaged for the entire dataset.

2.4.11. The ratio of the number of Numeric Digits to the length of the password

This ratio talks about how much percent of a password contains Numeric Digits. The ratio was locally calculated on a single password level and then adjusted for the complete dataset.

2.4.12. The ratio of the number of Symbols to the length of the password

This ratio talks about how much percent of a password contains symbols in it. This ratio was also calculated on an individual password level and then adjusted for the complete dataset.

2.4.13. Additional Features of the Data set

The following five additional features of the dataset were also observed:

2.4.13.1 Total Observations

2.4.13.2 Average Length of Passwords

2.4.13.3 Maximum Length of Passwords present

2.4.13.4 Minimum Length of Passwords present

2.4.13.5 Number of Unicode Characters present in the dataset

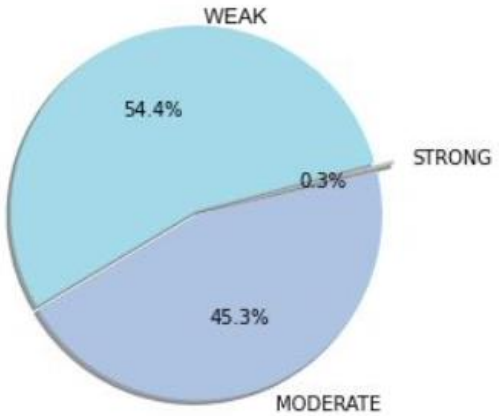
3. Results

3.1 Password Strength

Fig 1 shows the results of the Password Strength Analysis. Fig 1(a) indicates the frequency of weak, moderate, and strong passwords present in the dataset. Fig 1(b) shows the information in the form of a Pie Chart for easy visualization. It has been observed that only 0.3% of the passwords are strong and more than 50% of the passwords are weak in strength.

	strength	frequency
0	weak	58117532
1	moderate	48383138
2	strong	365179

a



b

Figure 1: Password strength statistics

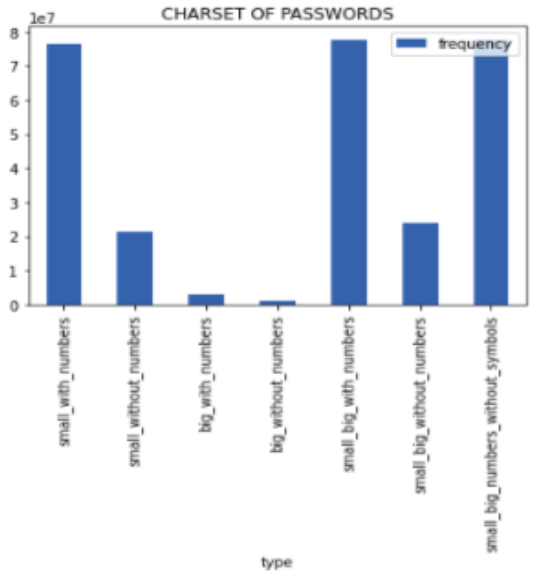
DOI: https://doi.org/10.54216/FPA.040101
Received January 07, 2021 Accepted May 10, 2021

3.2 Character Set of Passwords

Fig 2 shows the character set of passwords in the dataset. Fig 2(a) shows the frequency of seven categories of character set. A bar graph is also plotted in Fig 2(b) which indicates that the categories 'small(lowercase) + big(uppercase) + numbers' , 'small(lowercase) + numbers' and 'small(lowercase) + big(uppercase) + numbers + without_symbols' are most common character sets with almost 72% each.

	type	frequency
0	small_with_numbers	76626893
1	small_without_numbers	21448461
2	big_with_numbers	2821143
3	big_without_numbers	992278
4	small_big_with_numbers	77732205
5	small_big_without_numbers	24016251
6	small_big_numbers_without_symbols	77732205

a

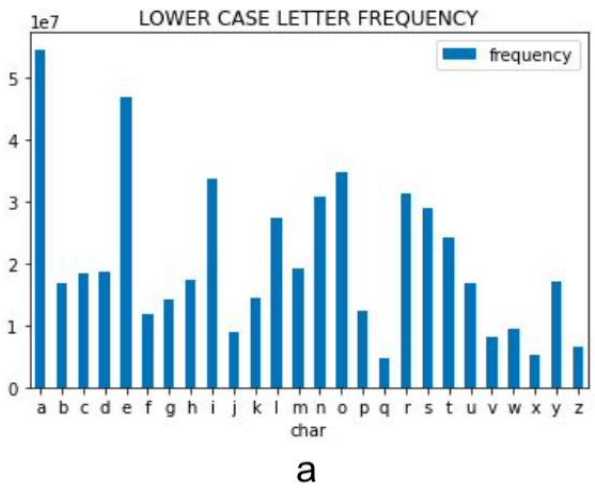


b

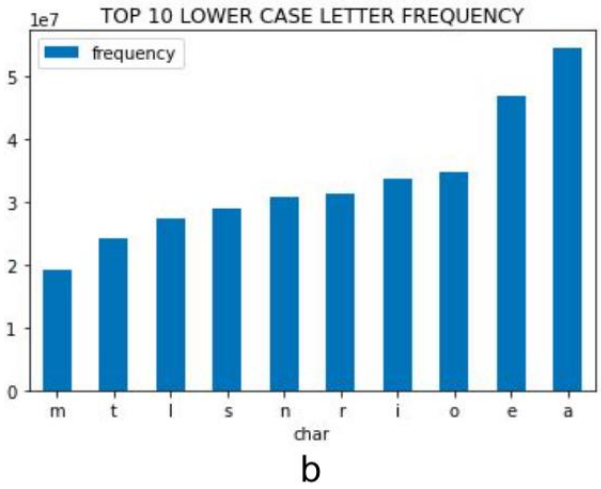
Figure 2: Character set statistics

3.3 Lowercase Letter Frequency

Fig 3(a) shows the frequency of all the 26 lowercase letters in the form of a bar graph. The top 10 most common lowercase letters found in the dataset are represented in Fig 3(b). It has been observed that the most common lowercase letter 'a' occurred 54517313 times and the least common lowercase letter 'q' occurred 4654476 times in the dataset.



a



b

Figure 3: Lowercase letter frequencies

DOI: https://doi.org/10.54216/FPA.040101
Received January 07, 2021 Accepted May 10, 2021

3.4 Uppercase Letter Frequency

Fig 4(a) shows the frequency of all the 26 uppercase letters in the form of a bar graph. The top 10 most common uppercase letters found in the dataset are represented in Fig 4(b). It has been observed that the most common uppercase letter 'A' occurred 2491065 times, and the least common lowercase letter 'X' occurred 501590 times in the dataset.

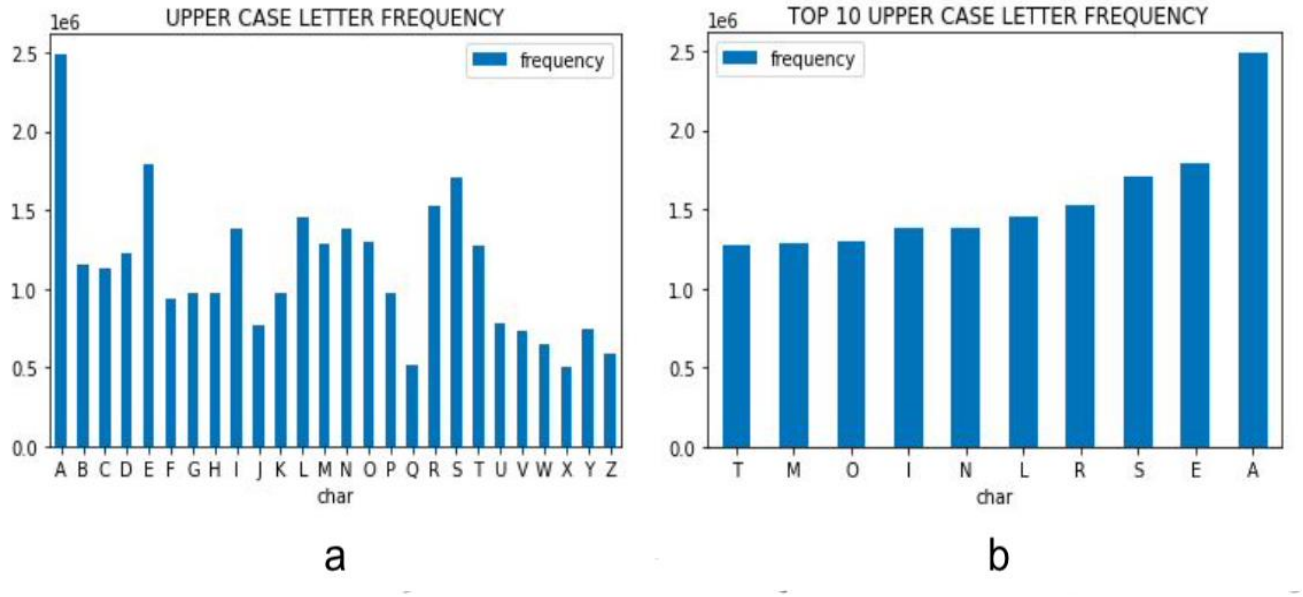


Figure 4: Uppercase letter frequencies

3.5 Numbers Frequency

Fig 5 shows the frequency of numeric digits occurring in the dataset. The most common numeric digit was found to be '1', occurring 61991073 times, and the least common numeric digit was found to be '7', occurring 22170243 in the dataset. Frequencies of numbers '6', '4', '8', and '5' were also found to be approximately equal.

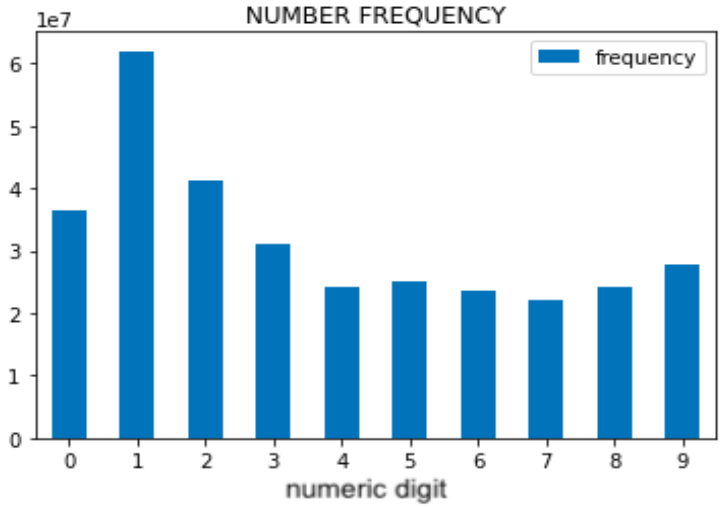


Figure 5: Numbers frequency

DOI: https://doi.org/10.54216/FPA.040101
Received January 07, 2021 Accepted May 10, 2021

3.6 Symbol (Special Characters) Frequency

Fig 6 shows statistics about the occurrence of symbols in the passwords of the dataset. Fig 6(a) shows twenty-five unique symbols present in the dataset with their frequency of occurrence. Fig 6(b) shows the top ten most common symbols along with their frequencies presented in the form of a bar graph. Symbol '.' (full stop) is the most common occurring symbol with a frequency of 1485430. Symbol '>' (greater than) is the least commonly occurring symbol with a frequency of 6500 only.

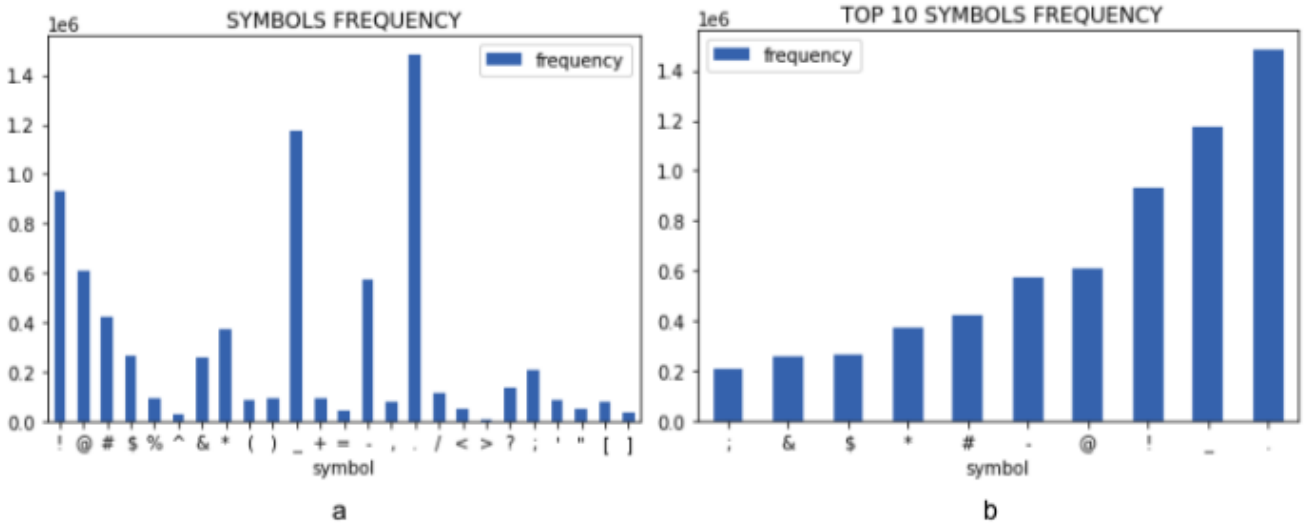


Figure 6: Symbols (special characters) frequency

3.7 Most Common Password

Fig 7 shows the top 10 most common passwords found in the dataset.Password '123456' is the most common password in the dataset with a frequency of 865098 i.e., 0.8% of the entire dataset. Passwords like '1234567', '000000', '123123', 'password1', '12345678', 'password', 'abc123' had almost same frequency of occurrence.

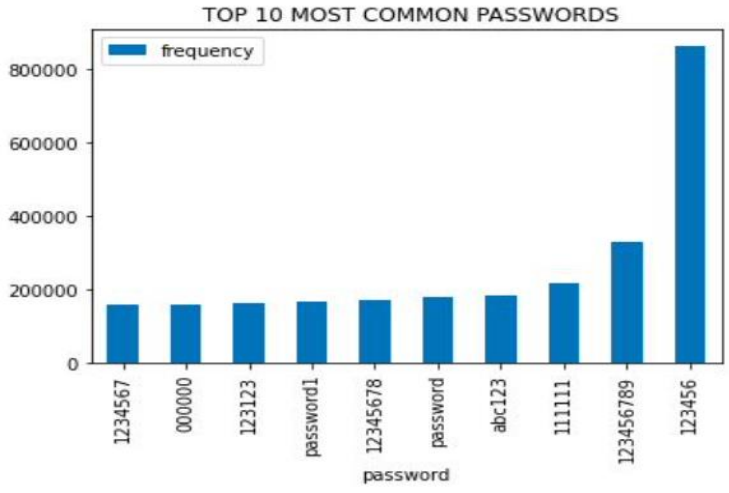


Figure 7: Most common password

DOI: https://doi.org/10.54216/FPA.040101
Received January 07, 2021 Accepted May 10, 2021

3.8 Most Common Domain

Fig 8 shows the top 10 most common domains found in the dataset. Domain 'yahoo.com' was found to be the most common domain in the dataset with a frequency of 18409390, i.e., 17% of the entire dataset. It was followed by 'hotmail.com', 'mail.ru', 'gmail.com' and 'aol.com' with the frequencies of 10895244, 8033848, 722,4411, 6650848 respectively.

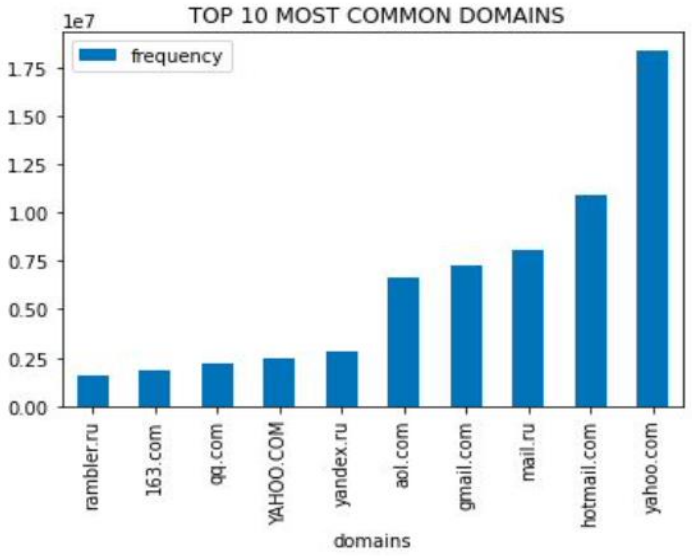


Figure 8: Most common domain

3.9 Most Common Unicode Characters

Some passwords in the dataset also contain Unicode Characters (characters whose ASCII value is greater than 128). Fig 9 shows the top 10 most common Unicode Characters along with their frequencies plotted in a bar graph. The most common Unicode character found in the dataset was 'a' (ASCII value 1072) with a frequency of 1942387, i.e., 19% of all the Unicode characters present in the dataset.

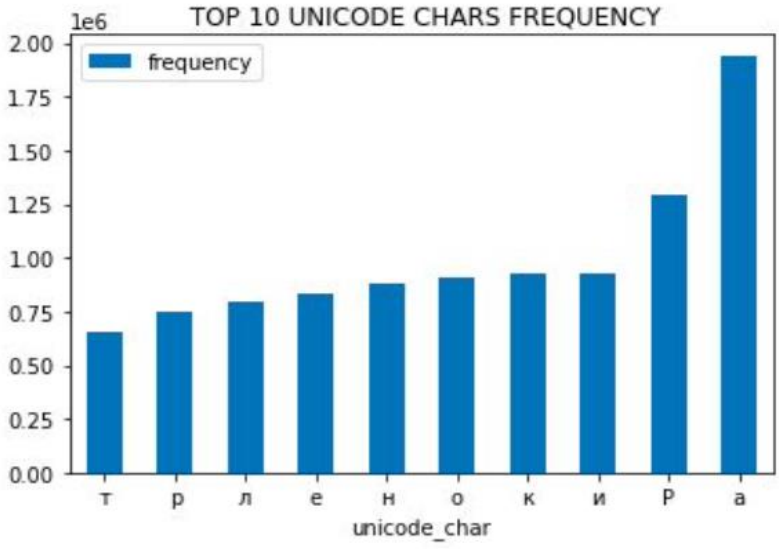


Figure 9: Unicode character frequency

DOI: https://doi.org/10.54216/FPA.040101
Received January 07, 2021 Accepted May 10, 2021

3.10 Ratios

- The ratio of the number of Alphabetic Letters to the length of the Password: 0.6304
- The ratio of the number of Numeric Digits to the length of the Password: 0.3622
- The ratio of the number of Symbols to the length of the Password: 0.0072

3.11 Additional Features

- Total Observations : 106,865,849
- Average Length of Passwords: 8.294
- Maximum Length of Passwords present: 255
- Minimum Length of Passwords present: 1
- Number of Unicode Characters present in the dataset: 10,138,047

4. Conclusion

EDA was performed on a 100 million email password dataset. It has been observed that only 0.3% of the passwords are strong and more than 50% of the passwords are weak in strength. The categories 'small(lowercase) + big(uppercase) + numbers' , 'small(lowercase) + numbers' and 'small(lowercase) + big(uppercase) + numbers + without symbols are most common character sets with almost 72% each. The most common lowercase letter 'a' occurred 54517313 times, and the least common lowercase letter 'q' occurred 4654476 times in the dataset. The most common uppercase letter 'A' occurred 2491065 times, and the least common lowercase letter 'X' occurred 501590 times in the dataset. The most common numeric digit was found to be '1', occurring 61991073 times, and the least common numeric digit was found to be '7', occurring 22170243 in the dataset. Frequencies of numbers '6', '4', '8', and '5' were also found to be approximately equal. Symbol '.' (full stop) is the most common occurring symbol with a frequency of 1485430. Symbol '>' is the least commonly occurring symbol with a frequency of 6500 only. Password '123456' is the most common password in the dataset with a frequency of 865098, i.e., 0.8% of the entire dataset. Passwords like '1234567', '000000', '123123', 'password1', '12345678', 'password', 'abc123' had almost same frequency of occurrence.Domain 'yahoo.com' was found to be the most common domain in the dataset with a frequency of 18409390, i.e., 17% of the entire dataset. It was followed by 'hotmail.com', 'mail.ru', 'gmail.com' and 'aol.com' with the frequencies of 10895244, 8033848, 722,4411, 6650848 respectively.The most common Unicode character found in the dataset was 'a' (ASCII value 1072) with a frequency of 1942387, i.e., 19% of all the Unicode characters present in the dataset. This analysis can help users make smart decisions while creating passwords for themselves, i.e., not opting for the most common features and creating strong and less vulnerable passwords.

References

[1] Chanda, Katha. (2016). Password Security: An Analysis of Password Strengths and Vulnerabilities. International Journal of Computer Network and Information Security. 8. 23-30. 10.5815/ijcnis.2016.07.04.

[2] Li, Yue & Wang, Haining& Sun, Kun. (2017). Personal Information in Passwords and Its Security Implications. IEEE Transactions on Information Forensics and Security. PP. 1-1. 10.1109/TIFS.2017.2705627.

[3] Cheng, Long & Liu, Fang & Yao, Danfeng. (2017). Enterprise data breach: causes, challenges, prevention, and future directions: Enterprise data breach. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 7. e1211. 10.1002/widm.1211.

[4] Yıldırım, M., Mackie, I. Encouraging users to improve password security and memorability. Int. J. Inf. Secure. 18, 741–759 (2019). https://doi.org/10.1007/s10207-019-00429-y

[5] De Cristofaro, Emiliano & Du, Honglu&Freudiger, Julien &Norcie, Greg. (2013). Two-Factor or not Two-Factor? A Comparative Usability Study of Two-Factor Authentication. USEC. 10.14722/usec.2014.23025.

[6] Pinkas, Benny & Sander, Tomas. (2003). Securing Passwords Against Dictionary Attacks. Proceedings of the ACM Conference on Computer and Communications Security. 10.1145/586110.586133.

[7] Bošnjak, Leon &Sres, J. &Brumen, B.. (2018). Brute-force and dictionary attack on hashed real-world passwords. 1161-1166. 10.23919/MIPRO.2018.8400211.

DOI: https://doi.org/10.54216/FPA.040101
Received January 07, 2021 Accepted May 10, 2021

[8] 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings (2018)
[9] https://www.kaggle.com/wjburns/common-password-list-rockyoutxt
[10] https://crackstation.net
[11] https://weakpass.com/download
[12] https://wiki.skullsecurity.org/Passwords
[13] Tull, L.. (2002). Library systems and Unicode: A review of the current state of development. 21. 181-185.
[14] Hahn, Brian & Valentine, Daniel. (2013). ASCII Character Codes. 10.1016/B978-0-12-394398-9.00026-5.
[15] https://github.com/hmaverickadams/breach-parse
[16] https://www.python.org
[17] https://github.com/rishab-rb/EDA_Passwords/blob/main/FINAL%20CODE.ipynb
[18] https://github.com/rishab-rb/EDA_Passwords/blob/main/EDA.ipynb

DOI: https://doi.org/10.54216/FPA.040101
Received January 07, 2021 Accepted May 10, 2021