7COM1079-0901-2024 - Team Research and Development Project

Final Report Title: Top 200 password 2021 by Country

Group ID: A 48

Dataset number: DS171

Prepared by:

Name	Student Id
Usama Bin Khalid	23091719
Muhammad Haris Bashir	23013854
Muhammad Ahmed Akram	23036804
Muhammad Arbaz Asif	23082709
Shoaib Latif	23086696

University of Hertfordshire
Hatfield, 2024

Table of Contents

1. Introduction

1.1. Problem statement and research motivation

- When it comes to cybersecurity, password security has always been a critical issue and a very common cause of data breach is using weak and predictable passwords. Many such cases have been found where the users use simple and common passwords, for example "123456" or “abcd” and because of this reason, the system becomes vulnerable to attacks (Burnett, 2020). This research is basically focused on examining the top 200 passwords from 2020 in Canada and Australia so as to find out regional differences in password strength. After knowing these differences, strategies can be made to improve password policies and users can be educated regarding security practices. Analyzing these patterns will be helpful to understand global password usage trends and ultimately, measures will be taken to strengthen cybersecurity.
- **Reference:** Burnett, M. (2020). Perfect Passwords: Selection, Protection, Authentication. Syngress.

1.2. The dataset contains the top 200 most common passwords from 2020 across various countries, including Australia and Canada. It includes data fields such as password rank, user count, estimated time to crack, and global rank. The dataset provides crucial insights into password strength and user behavior in each country, making it an ideal resource for analysing differences in password categories. The data enables statistical analysis to test hypotheses related to password strength and regional trends.

1.3. The research question investigates whether there is any difference in the password strength proportions categories between Canada and Australia. This question will be answered after categorizing passwords into different strength groups and conducting statistical tests so as to find out prominent differences and to provide details of regional password security behaviour’s.

1.4. Null hypothesis and alternative hypothesis (H0/H1)

- **Null Hypothesis (H0)**: There is no prominent difference in the proportions of password strength categories between Canada and Australia.
- **Alternative Hypothesis (H1)**: There is a significant difference in the proportions of password strength categories between Canada and Australia.

2. Background research

2.1. Research papers

- Passwords have always been an important aspect of cybersecurity, and different studies have been conducted regarding password behaviors to find out vulnerabilities and improve security measures. Das et al. (2014) conducted a comprehensive study on password reuse as well as predictability, uncovering patterns in frequently used passwords in different regions and focusing on the importance of analyzing password datasets so as to overcome the risks. Moreover, Bonneau (2012) worked on large-scale password datasets to uncover trends in user behaviour and password strength, highlighting the need for regional-specific studies. Another studies by Wang et al. (2021) researched password datasets by categorizing passwords into different strength levels so as to find out their vulnerability to attacks. These studies are extremely useful for understanding regional as well as global password behaviors.

The dataset used in this research aligns closely with the previous studies, and this dataset contains information regarding the top 200 passwords frequently used in Canada and Australia in 2020. The dataset mainly focuses on user count,global rankings and time-to-crack, and hence it enables a detailed analysis of password strength categories. Findings have been compared with previous research and this study is effective for addressing gaps in understanding regional variations in password security. Moreover, this study brings attention towards making targeted strategies to improve user practices regarding password setting.

References:

- Das, A., Bonneau, J., Caesar, M., et al. (2014). "The Tangled Web of Password Reuse." Network Security.
- Bonneau, J. (2012). "The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords." IEEE Symposium on Security and Privacy.
- Wang, D., Li, H., & Ma, Y. (2021). "Password Strength and its Implications on Security: A Dataset-Driven Study." Cybersecurity Journal.

2.2. Why RQ is of interest (research gap and future directions according to the literature)

The research question (RQ) investigates a critical gap in understanding how regional and cultural differences influence password choices. Previous studies have mainly focused on global password trends but the nuanced variations across countries have totally been neglected. By investigating Canada and Australia, this research puts stress on the need to tailor awareness campaigns and security policies specifically to regional behaviors. According to Bonneau (2012), region-specific studies can find out unique password patterns and hence enable much better interventions. This study plays prominent role in the field as it focuses on exploring the reasons for regional password strength variations and ultimately providing details for future work regarding strength of the cybersecurity globally.

3. Visualisation

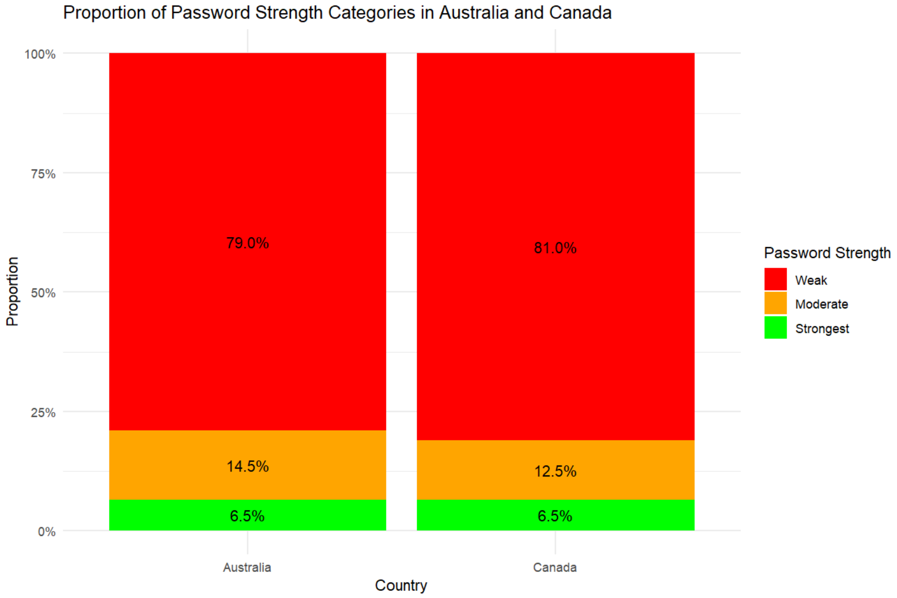
3.1. Appropriate Plot for the Research Question

Plot Description:

The stacked bar chart above represents the proportion of password strength categories (weak, moderate, and strong) for users in Australia and Canada. Each bar corresponds to a country and is divided into the three password strength categories. The x-axis represents the country, while the y-axis shows the proportion of each password strength category as a percentage. The colors (red, orange, green) denote weak, moderate, and strong passwords, respectively.

Why This Plot?

This plot is appropriate as it clearly illustrates the differences in the distribution of password strength categories between the two countries, aiding visual comparison. Stacked bar charts are effective for showing proportions and comparing categorical data across groups.



3.2. Additional Information Relating to Understanding the Data

This visualization highlights that the majority of passwords in both countries fall into the "weak" category. Canada has a slightly higher proportion of weak passwords (81%) compared to Australia (79%), while Australia has a slightly higher proportion of moderate passwords (14.5%) compared to Canada (12.5%). Strong passwords are equally low in both countries (6.5%).

3.3. Useful Information for Data Understanding

The stacked bar chart indicates a lack of strong password practices in both countries, as weak passwords dominate the distribution. While there are small differences in the proportions of moderate and weak passwords, these differences are minor and align with the chi-square test results, which confirmed no statistically significant difference between the countries.

4. Analysis

4.1. Statistical test used to test the hypotheses and output

The hypotheses were examined using the Pearson Chi-squared test. This test is useful because it compares the proportions of categorical data (password strength categories) in two nations (Canada and Australia), which are relevant to the research question. The Chi-squared test is often used to determine the independence of categorical variables. The test statistic and p-value were computed using a contingency table that included the password strength distributions for both nations.

4.2. The null hypothesis is rejected /not rejected based on the p-value

There is insufficient evidence in the Chi-squared test results to reject the null hypothesis (H0). It seems unlikely that the variations in the proportions of password strength between Australia and Canada are statistically significant, as the p-value of 0.841 is far higher than the conventional significance level of 0.05. Rather than reflecting significant differences, any observed variances in the proportions of weak, moderate, and strong passwords are more likely to be the result of chance. According to this analysis, the evidence denies the notion that a country's location significantly affects the strength of a password.

5. Evaluation – group's experience at 7COM1079

5.1. What went well

There was regular and effective communication between the group through Whatsapp and in university meetings and hence, all the members were collaborating at every step. A Trello board named "team-research-and-development" was created so as to streamline task management and to track the progress. Usama Bin Khalid assigned tasks to each of the team members efficiently, fostering clarity as well as accountability. This structured approach was effective to keep the team members organized, complete tasks on time, and work as a team. All these factors played an important role in the overall success of this project.

Trello:
https://trello.com/invite/b/67374bf0d9a2df1195dbc042/ATTIe48c518c1492e5e0ea9f43568f6edf63088C801B/team-research-and-development

5.2. Points for improvement

As members lived far apart and weren't able to meet outside of university so it was challenging for them to communicate and coordinate. In order to overcome this gap and to maintain collaboration, a Whatsapp group was created where members were regularly in contact with each other and discussed the tasks at every step. Ultimately, the team was able to maintain a friendly and cooperative relation and they succeeded to complete this project. The University of Hertfordshire's initiative of forming diverse groups proved effective in creating valuable learning opportunities and in improving teamwork even under challenging circumstances.

5.3. Group's time management

Time management was a plus point of the group, and they completed all assignments and quizzes on time. They made a Trello board to effectively track tasks and deadlines, connecting everyone on schedule. Trello really helped a lot to organize work and maintain accountability and it also contributed to the group's timely submissions.

5.4. Project's overall judgement

The group is feeling very confident about their performance and highly satisfied with the outcome of their project. The course served as an excellent opportunity for the team members to learn and apply new technologies for example GitHub, Trello, and R Studio, which were new for them. Overall, to work on this project was a great learning experience at the University of Hertfordshire.

5.5. Note any changes to group since submission of Assignment 1. Add new or amended GitHub Ids for new members

It was challenging to work with GitHub for the first time, but the team quickly adapted and collaborated effectively. Muhammad Haris Bashir and Usama Bin Khalid contributed most commits, while other members also played important roles. There were no changes in the group after Assignment 1, and the same group members worked together for the completion of this course. This teamwork helped all the team members to learn GitHub workflows and improve their technical skills at the individual level.

5.6. Comment on the GitHub log output

The GitHub log output, detailed in Appendix B, identifies three significant commits that shaped the project. Each commit reflects critical updates that enhanced the data analysis and visualization process.

Commit Message: Added filtered_dataset.csv file (Commit ID: adaba2e01b77491e3f614c1a00126aa084c7d43d)

Broader Impact: This commit was crucial in preparing the dataset for analysis. By filtering the dataset to focus on two countries (Australia and Canada), it ensured the analysis was relevant and manageable. The filtered data served as the basis for subsequent visualizations.
Commit Message: Code updated for stacked bar chart (Commit ID: cca045f3f0651610303d5b314d7adb6a703941b2)

Broader Impact: In this commit, the x-axis initially represented "weak," "moderate," and "strong" password categories, while the y-axis showed proportions. Countries were differentiated by colors (e.g., red and blue). However, this design was later improved for better clarity. This step laid the groundwork for the next significant modification to the visualization.
Commit Message: Changed side-by-side chart to stacked bar chart (Commit ID: 926f6f1c37dcd5c0a6b4d6d4cd63e98cb2f75f40)

Broader Impact: This commit introduced the final and improved visualization, where the x-axis now represents the countries (Australia and Canada) and the y-axis represents the proportion of password strength categories. Colors were updated to red, orange, and green to represent "weak," "moderate," and "strongest" password categories, respectively. This change significantly enhanced the interpretability and clarity of the visualization.
These iterative commits demonstrate the progression of the project from dataset preparation to a refined and informative stacked bar chart that accurately conveys the insights.

6. Conclusions

6.1. Results explained

The p-value of 0.841 for the chi-squared test indicated that there was no significant difference between Australia and Canada in the proportions of password strength categories. The stacked bar chart, which shows close distributions of strong, moderate, and weak passwords in the two countries, supports this conclusion. These results suggest that location has little influence on password strength, highlighting the need to investigate other factors that affect password security practices.

6.2. Interpretation of the results

The results indicate that password creation habits are not significantly affected by geographical or cultural differences, as password strength practices are similar in Australia and Canada. This finding suggests that nationwide cybersecurity awareness campaigns might work better than nation-specific strategies. For the general public and the context, it draws attention to the ongoing worldwide problem of weak password usage and highlights the importance of policies that prioritize strengthening passwords and enhancing education.

6.3. Reasons and/or implications for future work, limitations of your study

Future studies could utilize a larger dataset with more nations or user groups to examine the effects of factors like rank and user count on password strength. Deeper insights may also be obtained by looking into demographic or psychological variables like age or education. The use of just two countries for analysis is a drawback.

7. Reference list ***(not included in the work count)***

GitHub. (n.d.). GitHub: Collaborative coding platform. Available at: https://github.com

Trello. (n.d.). Trello: Project management tool. Available at: https://trello.com

R Studio. (n.d.). R Studio: Data analysis and visualization. Available at: https://posit.co/products/open-source/rstudio/

Prasert, T. (2021). Top 200 passwords by country (2021). Available at: https://www.kaggle.com/datasets/prasertk/top-200-passwords-by-country-2021

Herts365. (n.d.). Top 200 passwords dataset. Available at: https://herts365-my.sharepoint.com/:x:/g/personal/sb20agc_herts_ac_uk/EZSzTIq3JVpIi_fwDVq6a00BJ-TWa2JYEQr_WHZfO9CCTQ?rtime=KnCOkzgv3Ug

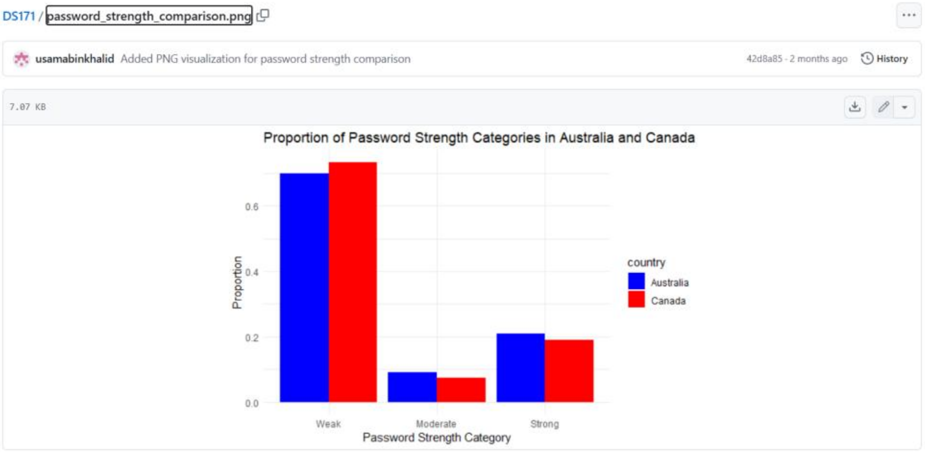
8. Appendices

A. R code used for analysis and visualisation

The following R code was utilized for analyzing the dataset and visualizing the proportions of password strength categories between Australia and Canada. The analysis included creating a contingency table, generating a stacked bar plot, and performing a chi-square test to determine statistical significance.

```
# Load required libraries
library(ggplot2)
library(dplyr)
library(scales)

# Load dataset
data <- read.csv("D:/Msc Cyber security/Semester A/Team Research & Project
development/Git/DS171/top_200_password_2020_by_country updated.csv")

# Filter dataset for Australia and Canada
filtered_data <- data[data$country %in% c("Australia", "Canada"), ]

# Categorize password strength
filtered_data$password_strength <-
cut(filtered_data$Time_to_crack_in_seconds,
                               breaks = c(-Inf, 60, 3600, 86400, Inf),
                               labels = c("Weak", "Moderate", "Strong", "Very
Strong"))

# Combine 'Strong' and 'Very Strong' into 'Strongest'
filtered_data$password_strength <- ifelse(filtered_data$password_strength
%in% c("Strong", "Very Strong"),
                                 "Strongest",
                                 as.character(filtered_data$password_strength))

# Set desired order for categories
filtered_data$password_strength <- factor(filtered_data$password_strength,
                                 levels = c("Weak", "Moderate", "Strongest"))

# Remove rows with missing values
filtered_data <- filtered_data %>%
  filter(!is.na(Time_to_crack_in_seconds) & !is.na(country))

# Save filtered dataset
write.csv(filtered_data, "filtered_dataset.csv", row.names = FALSE)

# Calculate proportions
proportions <- filtered_data %>%
  group_by(country, password_strength) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(country) %>%
  mutate(proportion = count / sum(count))

# Stacked bar plot
password_plot <- ggplot(proportions, aes(x = country, y = proportion, fill =
password_strength)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = scales::percent(proportion)),
            position = position_stack(vjust = 0.5)) +
  labs(title = "Proportion of Password Strength Categories in Australia and
Canada",
       x = "Country",
       y = "Proportion",
       fill = "Password Strength") +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_manual(values = c("Weak" = "red", "Moderate" = "orange",
"Strongest" = "green")) +
  theme_minimal()

# Display plot
print(password_plot)

# Create a contingency table
contingency_table <- table(filtered_data$password_strength,
filtered_data$country)
print(contingency_table)

# Perform Chi-square test
chi_test <- chisq.test(contingency_table)
print(chi_test)
```

B. GitHub log output.

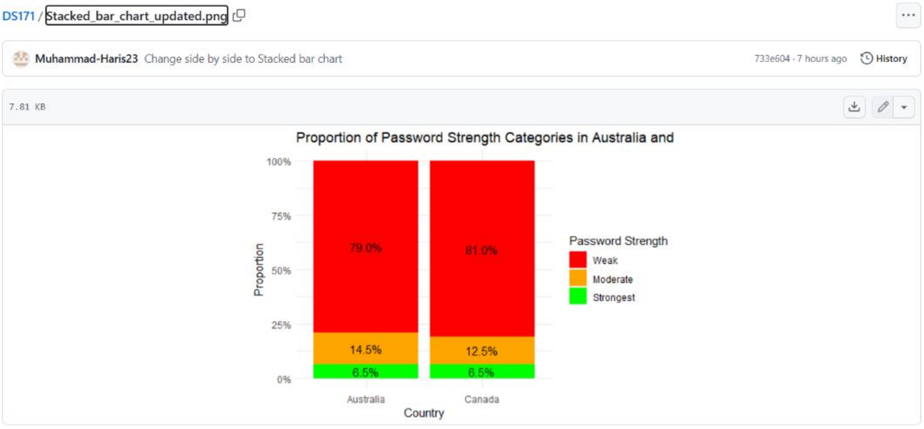


DS171 / password_strength_comparison.png

usamabinkhalid Added PNG visualization for password strength comparison 42d8a85 · 2 months ago History

7.07 KB

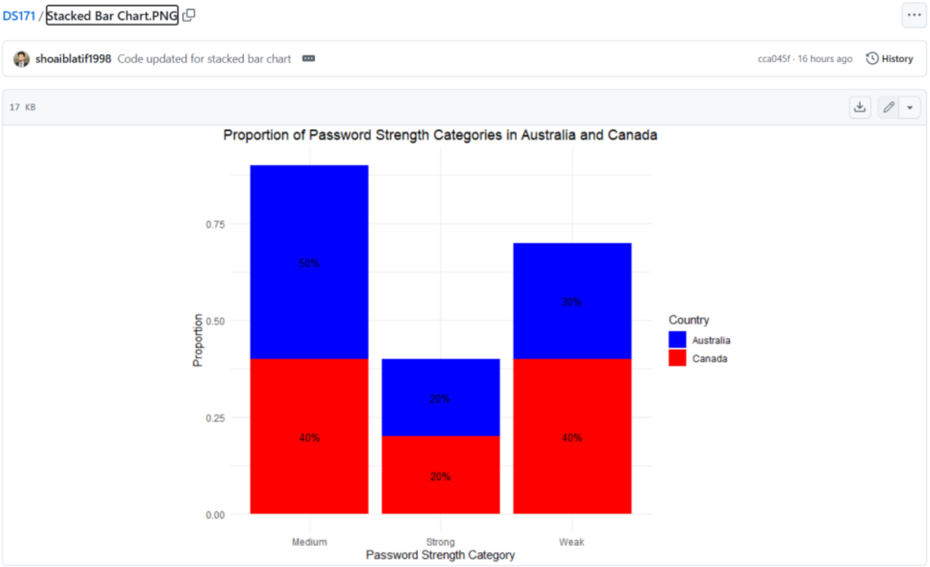
Proportion of Password Strength Categories in Australia and Canada
0.6
0.4
0.2
0.0
Proportion
Weak
Moderate
Strong
Password Strength Category
country
Australia
Canada



DS171 / Stacked_bar_chart_updated.png

Muhammad-Haris23 Change side by side to Stacked bar chart 733e604 · 7 hours ago History

7.81 KB



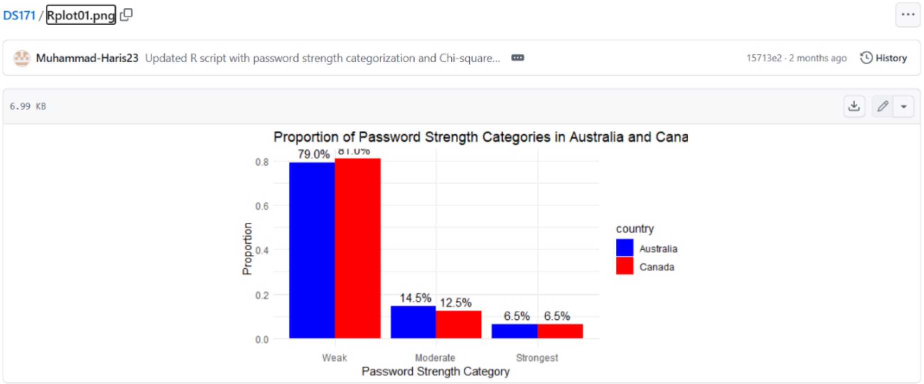
DS171 / Stacked Bar Chart.PNG

shoaiblatif1998 Code updated for stacked bar chart

cca045f · 16 hours ago History

17 KB

Proportion of Password Strength Categories in Australia and Canada
Proportion
0.00
0.25
0.50
0.75
50%
40%
20%
20%
30%
40%
Medium
Strong
Weak
Password Strength Category
Country
🟦 Australia
🟥 Canada



DS171 / Rplot01.png

Muhammad-Haris23 Updated R script with password strength categorization and Chi-square...

15713e2 · 2 months ago History

6.99 KB