Absenteeism Analytics

R Squad team



Supervised by: **Dr. Mazen Shawosh**

Prepared By: Abdulrahman Alghamdi, Muhammad Aljaali, Mohammed Alquraini

**dATA ANALYTICS**

# Executive Summary

We have investigated multiples scopes to approach the absenteeism challenge in an innovative way, while answering the main questions of which factors affect absinthium the most, Age being the most influential. Our produced final model achieved an accuracy (r2) of 73.91%. And have recommended a hiring policy that can reduce overall company absenteeism by 15% in 5 years, to be effective on 2 locations only. A companion live and interactive Power BI dashboard has been deployed to monitor changing absenteeism information to assist management and HR in their pursue to downsize absenteeism. Additionally, we suggested future works that can provide deeper insights and improve company data repositors.

# Table of Contents

[Executive Summary 1](#_Toc122954436)

[Table of Contents 2](#_Toc122954437)

[List of Tables 3](#_Toc122954438)

[List of Figures 3](#_Toc122954439)

[Introduction 4](#_Toc122954440)

[Problem Definition 4](#_Toc122954441)

[Problem Decomposition 5](#_Toc122954442)

[Data Prepossessing 10](#_Toc122954443)

[Analysis 11](#_Toc122954444)

[Descriptive Statistics 11](#_Toc122954445)

[Predictive Statistics 16](#_Toc122954446)

[Recommendations 19](#_Toc122954447)

[Future works 19](#_Toc122954448)

[Works Cited 21](#_Toc122954449)

[Appendix 22](#_Toc122954450)

[GitHub repository 22](#_Toc122954451)

[Power BI 23](#_Toc122954452)

# List of Tables

[Table 1 Data variables 4](#_Toc122953401)

[Table 2 Absenteeism by division breakdown 6](#_Toc122953402)

[Table 3 Absenteeism by department breakdown 7](#_Toc122953403)

[Table 4 Absenteeism by job title breakdown 9](#_Toc122953404)

# List of Figures

[Figure 1 Organizational chart 5](#_Toc122953405)

[Figure 2 The process and flow of the analytics 10](#_Toc122953406)

[Figure 3 Length Service and AbsHrsPerTenure regardless of business unit 11](#_Toc122953407)

[Figure 4 Length Service and AbsHrsPerTenure by Stores business unit 12](#_Toc122953408)

[Figure 5 Length Service and AbsHrsPerTenure by head office business unit 12](#_Toc122953409)

[Figure 6 The relationship between age & absHrsPerTenure 13](#_Toc122953410)

[Figure 7 The relationship between age & absHrsPerTenure 14](#_Toc122953411)

[Figure 8 Age and normal distribution of gender 15](#_Toc122953412)

[Figure 9 Top jobs split by gender 16](#_Toc122953413)

[Figure 10 ANOVA table of the model 18](#_Toc122953414)

# Introduction

This project is chartered to solve the absenteeism issue at Canada Stores ltd using data analytics. The absenteeism issue costs the company enormous expenses, and the use of data analytics can identify the most contributing factors that lead to absenteeism and can tap into recommendations to tackle this HR challenge. By conservative measures, the total absenteeism hours evaluated on the basis of minimum wage in Canada equates to $1,533,717 annually (US Dollars). This is deduced from the minimum wage of $11.42 (US Dollars) (Janisch, 2022).

Using the provided data, we'll define the underlying problem using root-cause analysis. Then, investigate the available data elements and features at our disposal to inspire insights. And finally, evaluate and enhance data quality.

## Problem Definition

We have a set of data features obtained from HR records about employees containing 8336 entries. We’re striving to achieve high quality data in order to better inform and feed our models, which ultimately improves accuracy and produce high quality result to better inform decision. Table 1 shows the data variables obtained from the raw data set and the added data variables we intend to include.

|  |  |
| --- | --- |
| **Feature** | **Type** |
| EmployeeNumber |  |
| Surname |  |
| GivenName |  |
| Gender |  |
| JobTitle |  |
| DepartmentName |  |
| StoreLocation |  |
| EmployeeType | 🆕 |
| EmployeeSector | 🆕 |
| StoreType | 🆕 |
| Division |  |
| Age |  |
| LengthService |  |
| AbsentHours |  |
| AbsentHoursPerTenure | 🆕 |
| BusinessUnit |  |

Table 1 Data variables

ProbableGenderByGivenName is a fetched from an extername names database to cross-check data quality. AbsentHoursPerTenure was added to better inform the annual absence for an employee regardless of their service years. EmployeeType This new variable will identify which type of employee we are studying, are they a working employee as a cashier or a back-end employee as a finance staff, this will help us identify which employee we can't afford to go absent. EmployeeSector This variable will help us identify absenteeism based on the sector the employee works in, it answers the question of "Are finance employees more likely to go absent in comparison to HR employees?". StoreType are classified into 3 types of stores. If a customer service manager exists and more than 40 cashiers.

Absenteeism in bigger stores may go unnoticed if another employee is present and will handle more load. Therefore, a higher responsibility factor may exist in smaller stores that is blind to the aggregate analytics. The number of store the best optimizes the absenteeism and optimal productivity before additional labor units begin to diminish marginal productivity (NDSU - North Dakota State University, 2022) and exsert pressure on operational expenses and company profit suitability can be elicited from the analysis. Cashiers per customer service manager ratio Meat cutter per Meat manager ratio Store manager per store employee ratio. And does having a customer service manager reduce absenteeism?

## Problem Decomposition

Multiple angles were examined to identify the most prominent congestions of absenteeism. We structured the organizational chart to evaluate the company from a helicopter view as in Figure 1.

Figure 1 Organizational chart



The below table investigates the breakdown of absentees by division.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Division | # Employees | Avg tenure | Total Absenteeism | Avg. Abs. / tenure | Absenteeism Culture |
| Stores | 8163 | 5 | 502722 | 16 | 29053 |
| FinanceAndAccounting | 73 | 12 | 2918 | 13 | 80 |
| HumanResources | 76 | 19 | 4137 | 5 | 20 |
| Executive | 11 | 11 | 532 | 8 | 8 |
| InfoTech | 10 | 12 | 401 | 4 | 3 |
| Legal | 3 | 11 | 154 | 6 | 1 |

Table 2 Absenteeism by division breakdown

The added metric Culture of Absenteeism is a synthetic measure for how strong is the absenteesm culture is in the division. It's the byproduct of the (number of employees x average absence hrs per tenure for employees in that division). if the avg. abs. per tenure increases, we assume that it's normalized in that division to be absent or late cumulatively, and when the number of employees is higher, it's more difficult to break and change this culture. More investments, policy, training, and time are required to influence that culture.

We will further break this down by DepartmentName and JobTitle in the 2 following tables, respectively.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DepartmentName | # employees | Avg tenure | Total Absintism | Avg. Abs. / tenure | Employees / Avg. Abs. tenure |
| Customer Service | 1737 | 5 | 109492 | 16 | 6103 |
| Bakery | 1449 | 5 | 88910 | 17 | 5384 |
| Dairy | 1515 | 5 | 95460 | 16 | 5343 |
| Meats | 1514 | 5 | 90222 | 15 | 5034 |
| Produce | 1163 | 5 | 69180 | 17 | 4218 |
| Processed Foods | 746 | 5 | 47468 | 18 | 2877 |
| Store Management | 39 | 5 | 1990 | 12 | 95 |
| Investment | 9 | 8 | 531 | 28 | 32 |
| Accounting | 18 | 9 | 739 | 15 | 30 |
| Accounts Receiveable | 16 | 12 | 477 | 9 | 12 |
| Audit | 15 | 17 | 660 | 11 | 10 |
| Accounts Payable | 15 | 11 | 511 | 6 | 9 |
| Executive | 11 | 11 | 532 | 8 | 8 |
| Training | 15 | 17 | 928 | 7 | 7 |
| Recruitment | 14 | 18 | 882 | 6 | 4 |
| Information Technology | 10 | 12 | 401 | 4 | 3 |
| Labor Relations | 12 | 21 | 608 | 5 | 3 |
| Employee Records | 12 | 18 | 722 | 4 | 3 |
| HR Technology | 14 | 19 | 674 | 4 | 3 |
| Legal | 3 | 11 | 154 | 6 | 1 |
| Compensation | 9 | 21 | 323 | 2 | 1 |

Table 3 Absenteeism by department breakdown

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **JobTitle** | **# employees** | **M** | **F** | **F%** | **Avg tenure** | **Total Absintism** | **Avg. Abs. / tenure** |
| Cashier | 1703 | 857 | 846 | 50% | 5 | 107431 | 16 |
| Dairy Person | 1514 | 785 | 729 | 48% | 5 | 95268 | 16 |
| Meat Cutter | 1480 | 711 | 769 | 52% | 5 | 88097 | 15 |
| Baker | 1404 | 711 | 693 | 49% | 5 | 86495 | 17 |
| Produce Clerk | 1129 | 586 | 543 | 48% | 5 | 67394 | 17 |
| Shelf Stocker | 712 | 368 | 344 | 48% | 5 | 45617 | 18 |
| Bakery Manager | 45 | 25 | 20 | 44% | 5 | 2415 | 13 |
| Meats Manager | 34 | 16 | 18 | 53% | 4 | 2126 | 17 |
| Customer Service Manager | 34 | 17 | 17 | 50% | 4 | 2061 | 14 |
| Store Manager | 39 | 20 | 19 | 49% | 5 | 1990 | 12 |
| Processed Foods Manager | 34 | 18 | 16 | 47% | 4 | 1851 | 13 |
| Produce Manager | 34 | 17 | 17 | 50% | 4 | 1787 | 14 |
| Trainer | 14 | 7 | 7 | 50% | 16 | 898 | 8 |
| Recruiter | 13 | 4 | 9 | 69% | 19 | 795 | 4 |
| Accounting Clerk | 17 | 8 | 9 | 53% | 10 | 688 | 15 |
| Benefits Admin | 11 | 4 | 7 | 64% | 17 | 680 | 4 |
| HRIS Analyst | 13 | 5 | 8 | 62% | 18 | 674 | 4 |
| Auditor | 14 | 6 | 8 | 57% | 16 | 629 | 12 |
| Labor Relations Analyst | 11 | 6 | 5 | 45% | 19 | 608 | 6 |
| Investment Analyst | 8 | 5 | 3 | 38% | 9 | 531 | 32 |
| Accounts Payable Clerk | 14 | 6 | 8 | 57% | 11 | 473 | 6 |
| Accounts Receiveable Clerk | 15 | 7 | 8 | 53% | 12 | 443 | 9 |
| Systems Analyst | 10 | 9 | 1 | 10% | 12 | 401 | 4 |
| Compensation Analyst | 8 | 5 | 3 | 38% | 21 | 298 | 2 |
| Dairy Manager | 1 | 0 | 1 | 100% | 5 | 192 | 37 |
| Corporate Lawyer | 3 | 2 | 1 | 33% | 11 | 154 | 6 |
| Exec Assistant, Human Resources | 2 | 1 | 1 | 50% | 9 | 119 | 12 |
| Director, Recruitment | 1 | 0 | 1 | 100% | 5 | 87 | 19 |
| CHief Information Officer | 1 | 0 | 1 | 100% | 12 | 87 | 7 |
| Exec Assistant, Legal Counsel | 1 | 0 | 1 | 100% | 7 | 78 | 12 |
| VP Stores | 1 | 0 | 1 | 100% | 10 | 75 | 7 |
| VP Human Resources | 1 | 1 | 0 | 0% | 3 | 71 | 28 |
| VP Finance | 1 | 1 | 0 | 0% | 7 | 66 | 10 |
| Director, Accounting | 1 | 0 | 1 | 100% | 2 | 51 | 23 |
| Director, Employee Records | 1 | 0 | 1 | 100% | 22 | 42 | 2 |
| Director, Accounts Payable | 1 | 0 | 1 | 100% | 4 | 38 | 10 |
| CEO | 1 | 1 | 0 | 0% | 14 | 36 | 3 |
| Director, Accounts Receivable | 1 | 0 | 1 | 100% | 14 | 34 | 3 |
| Director, Audit | 1 | 1 | 0 | 0% | 39 | 31 | 1 |
| Director, Training | 1 | 1 | 0 | 0% | 27 | 29 | 1 |
| Director, Compensation | 1 | 1 | 0 | 0% | 17 | 25 | 2 |
| Legal Counsel | 1 | 0 | 1 | 100% | 24 | 0 | 0 |
| Exec Assistant, VP Stores | 1 | 1 | 0 | 0% | 25 | 0 | 0 |
| Exec Assistant, Finance | 1 | 1 | 0 | 0% | 4 | 0 | 0 |
| Director, Labor Relations | 1 | 1 | 0 | 0% | 39 | 0 | 0 |
| Director, Investments | 1 | 1 | 0 | 0% | 1 | 0 | 0 |
| Director, HR Technology | 1 | 0 | 1 | 100% | 34 | 0 | 0 |

Table 4 Absenteeism by job title breakdown

We find out that the most difficult culture in the company is Customer Service. And the job title of Cashier is the most difficult under the customer service department. We also suggest that this job title is mission-critical and a bottle neck. It may be influencing other job titles in stores, i.e. if the cashier is late or absent, there's no need for it builds less importance with meat cutter if. It's also important to note that if the CRM/PoS system is connected to the cashier terminal, we can explain the high co-relation between cashier absenteeism and other store functions (job titles).

## Data Prepossessing

Our approach to processing this project is illustrated in Figure 2. The use of the prominent R programming language is the basis of operation. The data cleaning effort will then begin, then enriching the data and information before moving to the model generation segment.

Figure 2 The process and flow of the analytics



### Data Exploration

Locations may consist of multiple stores, e.g. Vancouver has 415 cashiers which may be considered is too high for a single store, or workers maybe a mix of part-time and full-time employees.

### Data Cleaning

We discovered age ranges below 18 years on payroll, as well as employees older than 65 years, and all records of this criteria will be trimmed.

In addition, duplicate values in the records will also be removed, cities with 2 employees will be removed. We have constituted a “dump” data set with proper justification of each record termination to provide a clear guide on data cleaning exercise commenced in this analytics effort, and to improve and inform future cleaning and data error preventions.

# Analysis

After we cleaned the data set, we commenced to the next segment in our project, which is the analysis of the data. the goal of the project is to identify factors that are most effective to absenteeism in the company, highlight them and offer our business recommendations on how to reduce them. Firstly, to reach that conclusion, we will do a descriptive analysis to understand what is currently happening. Secondly, we will use our findings from the descriptive analysis and conduct our predictive analysis to objectively project what will most likely to happen if we were to go through our recommendations.

## Descriptive Statistics

In this segment of the project, we will attempt to explore the data, and find correlations in our variables given. From this analysis, we will identify phenomenon and patterns that will justify the absenteeism causes within the company.

### Absenteeism and Length service

We wanted to test link between the connection between how long an employee has served the company and the absenteeism rate they have. The logic behind this is the fact that if someone served more in the company, are they more likely or less likely to go absent?

Figure 3 Length Service and AbsHrsPerTenure regardless of business unit

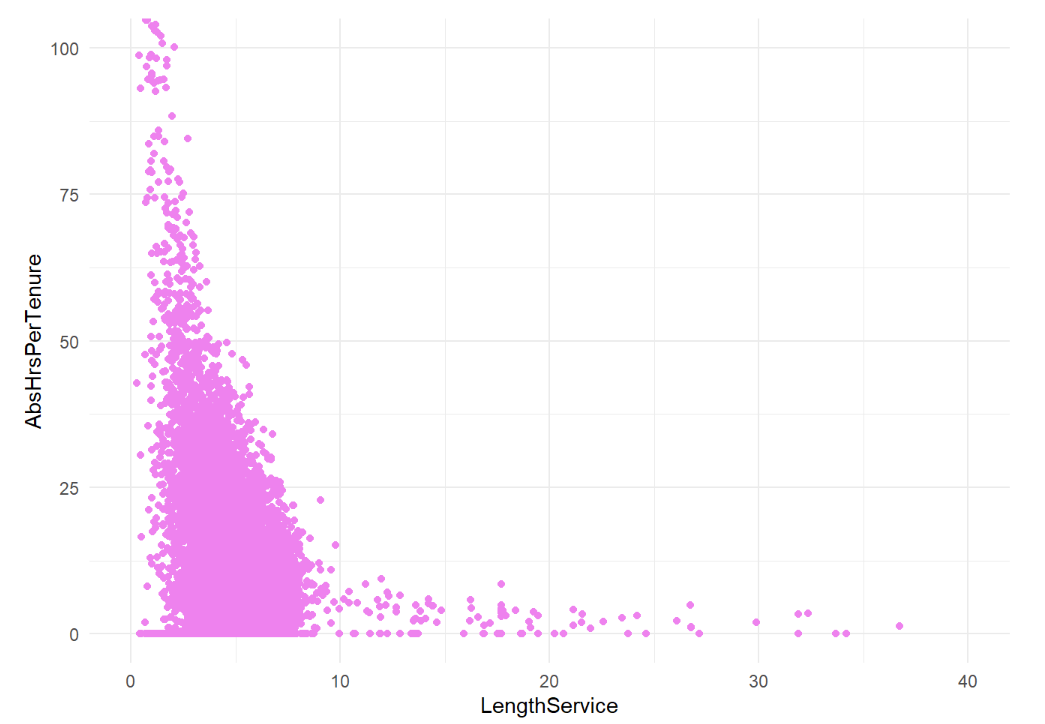


Figure 4 Length Service and AbsHrsPerTenure by Stores business unit

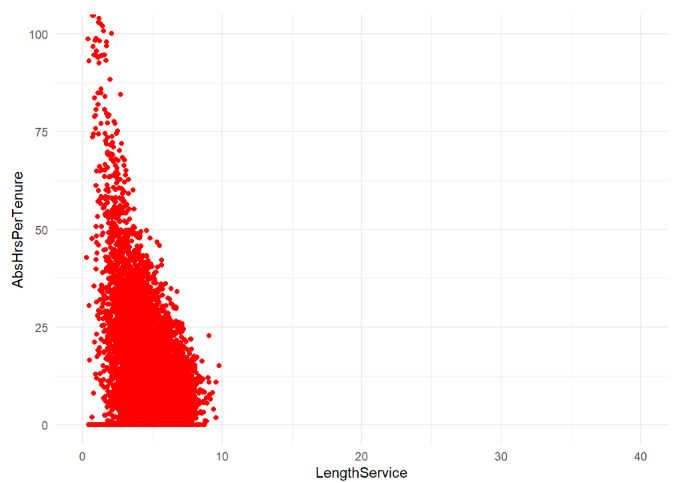


Figure 5 Length Service and AbsHrsPerTenure by head office business unit

Chart, scatter chart

Description automatically generated

### Findings

we can observe from the above graphs a great that there is a link between length service and absenteeism. The **purple graph** shows the overall relationship between absent hours per tenure for all employees. We observed a pattern and decided to do a deep dive as per business unit, and our findings were informative. The **red graph** shows the same graph but uniquely to stores, and the **blue graph** shows the relationship for head office employees. We can summarize from the graphs that for stores employees, they clutter absenteeism early service years, around 0 ~ 10 years of service. Whereas head office employees distribute their absences over their years service. This information is crucial because it gave us a new point of view to observe patterns, now we are considering the absenteeism culture over the whole business unit. Additionally, we will start to deep dive the importance of absences in Stores unit. We will do analysis on who we deem most important to not be absent. For example, if a high number of Cashiers go absent, customers will bottleneck on checkout, jeopardizing the in-store customer experience of the store.

### Absenteeism and Age

The one the most logical correlations is the age of the employee and how it affects the absenteeism rate they incur. This analysis answers the questions of “Is the employee more likely to be absent the older they are? does age affect absenteeism?”

Chart, scatter chart

Description automatically generated

Figure 6 The relationship between age & absHrsPerTenure

Figure 7 The relationship between age & absHrsPerTenure

Chart, scatter chart

Description automatically generated

### Findings

the findings of these graphs will shape our decision for the report. We can observe a very interesting phenomena when plotting age and absence hours per tenure, the graph shows how the absence hours per tenure increases referring to the age. Starting approximately the age 25, absenteeism increases around 0.9 hours. With this observation we can deduce that as for now, this is the strongest correlation we were able to find thus far. We will keep this graph in mind whilst making our recommendation, as well as explore it in the point of view of its effects per job title.

### Gender equality

Our plot for ages against the two genders show normal distribution on the full spectrum of ages. As shown in Figure 8 with a slight favor of males between ages 44-55.

Figure 8 Age and normal distribution of gender



It was also observed that the top jobs by tenure consist of near similar male-to-female percentage. With a slight favoring of females for the top 8 jobs as shown in Figure 9.

Figure 9 Top jobs split by gender



## Predictive Statistics

To begin with, it is advised to use several criteria for building an optimal model. Therefore, the start of our prediction will be through building our model using AbsentHours to be the dependent variable.

A correlation matrix is applied to determine the correlation between each of the independent variables and our response variable.

After computing the correlation between the response variable and other variables, a correlation will be applied between some variables, as we have an assumption of having multicollinearity between the variables. A test was done between BusinessUnit with Division and JobTitle with DepartmentName. The result came out be that a present a highly significant correlation between these variables, which indicates that the presence of the variables in our model will result in multicollinearity issues.

After computing the correlation matrix, models will be running for each variable to determine which variable has the highest impact on the response variable.

From the obtained results of the individuals’ models, it has been defined that the variable that has the highest impact on the response variable is Age.

After observing some knowledge from the individual models, a full model will be applied to see the effect of other variables. In the full model, it has been found that there are 8 significant variables with a coefficient of determination of 74.07666%, but a high number of variables may cause issues regarding implementing the model in the business world, such as a high cost of implementation.

To delve deeper into the analytics, a model will be created based on the variables that we believe it has an impact on the response variable. The result of this model was similar to the full model, which result in a lack of new information gained.

However, as the manual method of building the optimal model fails, some advanced methods for building the optimal model will be applied, such as Backward Elimination, and Forward Stepwise.

To enhance our modeling practice, we looked into build models using AbsHrsPerTenure as a response variable, but from the beginning of the correlation matrix up to the end of the prediction processes, the highest coefficient of determination observed was 23.73154%, which is not even close to the one that has been obtained using AbsentHours as a response variable.

After applying the backward elimination method, it was observed that five variables are approved to be in our model which are Gender, JobTitle, Age, LengthService, and AbsHrsPerTenure with a coefficient of determination of 73.91996%.

In the Forward Stepwise method, it has been defined ideally four optimal variables to be entered in the final model which are Gender, JobTitle, Age, and AbsHrsPerTenure with a coefficient of determination of 73.91046%.

Based on the obtained results, our optimal model using AbsentHours as a response variable will be the model that was obtained using the Forward stepwise approach, as it has a lower number of variables than the Backward Elimination approach. Additionally, although the full and the Backward Elimination models are having higher R-squared than Forward Stepwise, the difference between the R-squared is insignificant. furthermore, in the full model, we have 8 variables while in this model we have just four variables, which will minimize the effort when applying the regression.

To enhance our modeling, it has been tried to build models using AbsHrsPerTenure as a response variable, but from the beginning of the correlation matrix up to the end of the prediction processes, the highest coefficient of determination was observed is 23.73154%, which is not even close to the one that has been obtained using AbsentHours as a response variable.

To conclude, our optimal model will be the one that was obtained using the Forward stepwise method with AbsentHours as a response variable with the independent variables Gender, JobTitle, Age, and AbsHrsPerTenure and a coefficient of determination of 73.91046%. The ANOVA table of the model is shown below in Figure 10.

Figure 10 ANOVA table of the model

Text

Description automatically generated

# Recommendations

The Cashier has the highest Absenteeism Culture level. In the city of Vancouver, the age 24.51 y/o and beyond is where Absenteeism hours escalate with a strong correlation for both males and females at a slope of 0.9395x. Since our x-intercept is 24.51, or in other words, Cashiers below the age of 24.51 have no absenteeism hours (0 Hrs / year). The avg. tenure of the Cashier is 4.58 years. A policy to hire Cashiers of a maximum of 20 y/o will ensure most cashiers will clear out of the company before they turn 25 and reduce overall company absenteeism significantly. However, we have to validate if these young cashiers are full-time or part-time in order to control the cost associated with high-frequency and high-volume recruiting, plus investigate the compensation increase as a result, along with the total supply of such type of workers. If deployed in Vancouver, it can reduce the overall abs hrs by 17,307 hrs per year. And if deployed in Richmond and Vancouver only, and to Cashier job title only. The overall company absenteeism can go down by 3% each year. By year 5, this effect will be exhausted after having reduced total company absenteeism by 15% in 5 years.

## Future works

* In our next phase of the engagement, we can look into the time of day where most abs. hrs cluster. Couples with customer rush hours, it can inform when do absenteeism hrs make less impact on the organization. This impact can be quantified or estimated, if the impact in negligible, then may be a more relaxed policy can improve employee satisfaction in the organization. Employee satisfaction is proven to improve shareholders value.
* We can investigate how can we reduce absence hours as employees grow service years possibly by enhancing the promotion, since we have a bottleneck in the director position with an avg. tenure of 18 years. VPs have less than have that tenure of an 8 year avg. and managers 5 years tenure avg.
* We can quantify the exact effect of absenteeism by employee on the revenue to find the best optimization of effort that will drive sales.
* We can investigate the financial penalties to employees as they increase absenteeism or show up late to reduce the waste and effect on revenues and company operational costs. if high absence and no effect on work or rev. then consider terminating them.
* Answering questions about when do lateness and absentees happen? are they related to seasons? or mostly happen in low rush hours?
* Improving the data collection mechanism to enhance data quality and integrity, such as systems or employee attendance management in stores.
* We need to analyze the payroll system and look into the process of deboarding/retiring 65 y/o and above.

# Works Cited

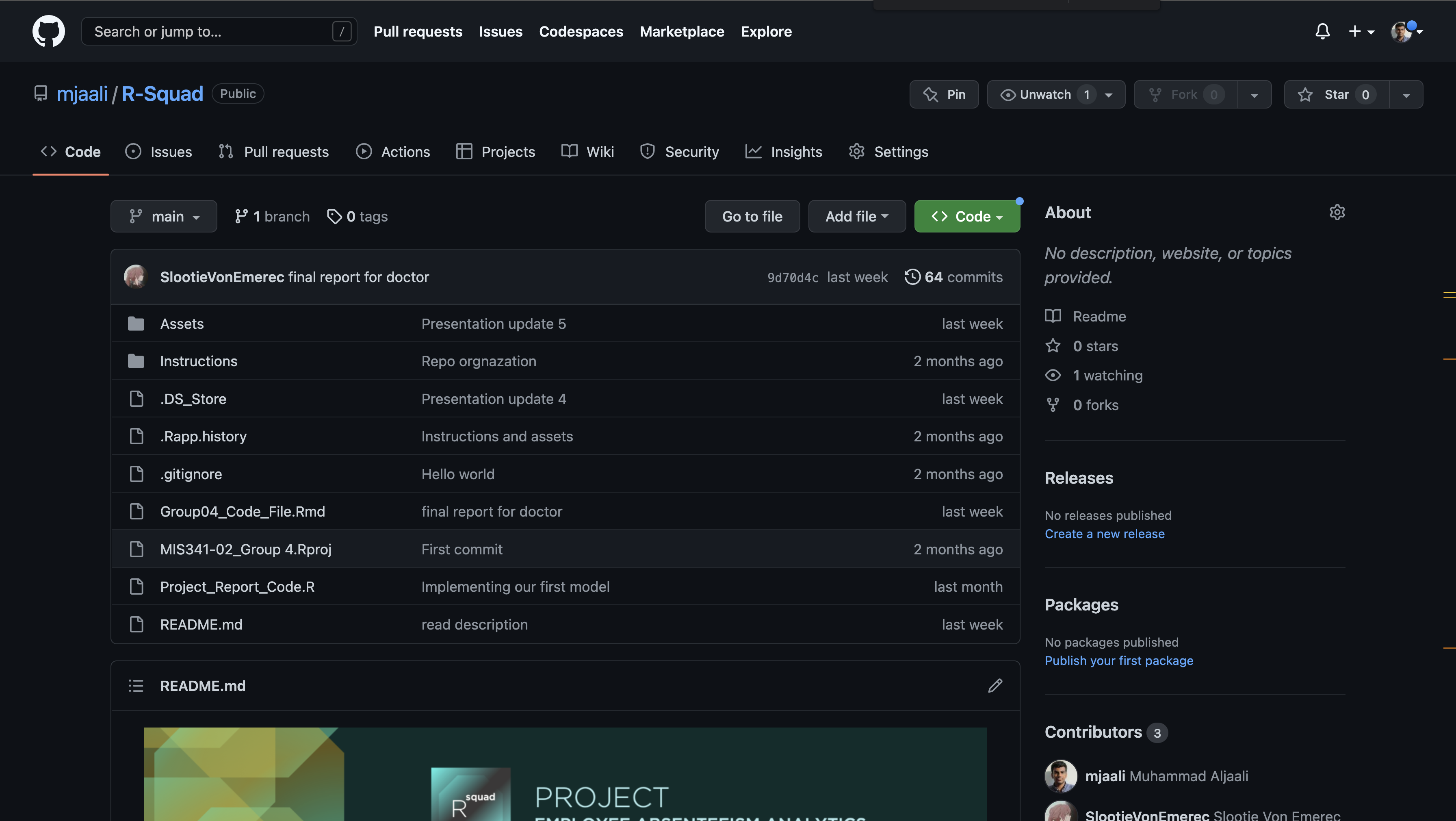
Janisch, K. (2022, 12 2). *Canada Minimum Wage* . Retrieved from EMPLOYMENT LAW: https://www.govdocs.com/canada-minimum-wage/#:~:text=Minimum%20wage%20in%20Canada%20mostly,transportation%20and%20federal%20Crown%20corporations.

NDSU - North Dakota State University. (2022, 12 2). *Diminishing Marginal Productivity* . Retrieved from AGRICULTURE LAW AND MANAGEMENT Accessibility : https://www.ag.ndsu.edu/aglawandmanagement/agmgmt/coursematerials/dimmargprod#:~:text=Diminishing%20marginal%20productivity%20is%20the%20concept%20that%20using%20increasing%20amount,eventually%20lead%20to%20decreasing%20productivity.

# Appendix

## GitHub repository

The repo contains all fills include R source codes with comments and R mark down files, and other related assets such as presentation files, instructions and others.



## Power BI

The developed Dashboard

A screenshot of a computer

Description automatically generated with medium confidence