IDENTIFICATION AND PREDICTION OF ABSENTEEISM IN EMPLOYEES

1. **Problem Definition:**

   What are the factors that can affect the absenteeism of an employee? How can the absenteeism of an employee be predicted?

2. **Data Collection:**

   Fictitious Employee Absence Dataset
   https://www.kaggle.com/datasets/HRAnalyticRepository/absenteeism-dataset

   This dataset was uploaded on Kaggle to serve as an attempt to dive into People/HR data with Analytical and Statistical tools. It is, however, a fake data as disclaimed by its owner.

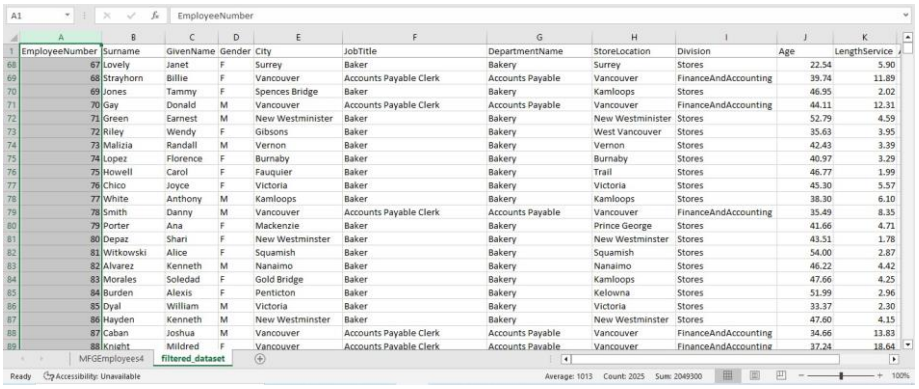
   The dataset consists of 13 columns (9 string, 3 decimal, and 1 integer), and 8336 rows of data, in a comma separated values (csv) file.

   The columns of the data are: Employee number, Surname, GivenName, Gender, City, JobTitle, DepartmentName, StoreLocation, Division, Age, LengthService, AbsentHours, BusinessUnit.

3. **Data Cleaning:**

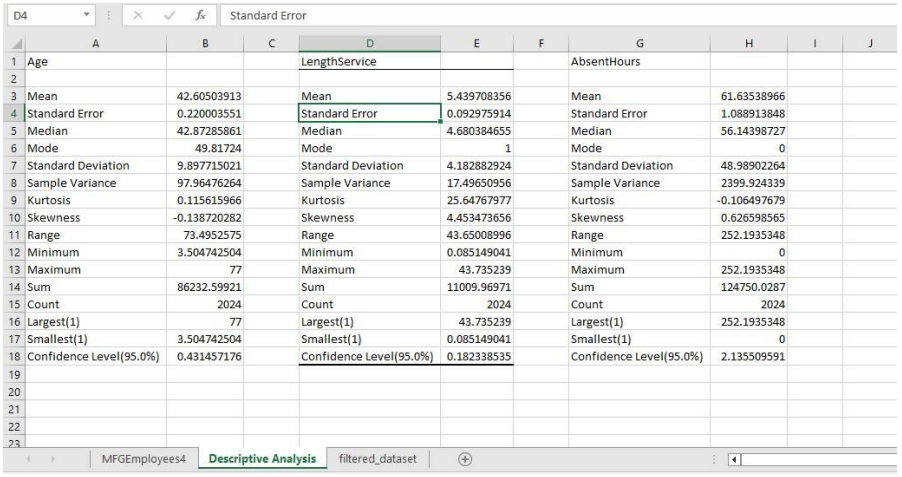
   To reduce the stress of working with a large dataset, I filtered the rows that have employee numbers 1 through 2024, which is a small dataset compared to the large 8336-row data. To enhance my visualisation of the data, I formatted the cells to autofit column width and height. I used TRIM on text columns to remove extra white spaces. I used Special Go-To to search for blank spaces in the data. I checked duplicate rows through the most unique column, which is the Employee Number column. The data does not contain any N/A value. I rounded the three number columns to two decimal places. I converted all columns to their correct data types - number columns to number data type and text columns to text data type.

   When I was not getting a quick-enough response from Excel, I filtered out another 100 rows to work with at some point in the project.



EmployeeNumber	Surname	GivenName	Gender	City	JobTitle	DepartmentName	StoreLocation	Division	Age	LengthService
67	Lovely	Janet	F	Surrey	Baker	Bakery	Surrey	Stores	22.54	5.90
68	Strayhorn	Billie	F	Vancouver	Accounts Payable Clerk	Accounts Payable	Vancouver	FinanceAndAccounting	39.74	11.89
69	Jones	Tammy	F	Spences Bridge	Baker	Bakery	Kamloops	Stores	46.95	2.02
70	Gay	Donald	M	Vancouver	Accounts Payable Clerk	Accounts Payable	Vancouver	FinanceAndAccounting	44.11	12.31
71	Green	Earnest	M	New Westminister	Baker	Bakery	New Westminister	Stores	52.79	4.59
72	Riley	Wendy	F	Gibsons	Baker	Bakery	West Vancouver	Stores	35.63	3.95
73	Malizia	Randall	M	Vernon	Baker	Bakery	Vernon	Stores	42.43	3.39
74	Lopez	Florence	F	Burnaby	Baker	Bakery	Burnaby	Stores	40.97	3.29
75	Howell	Carol	F	Fauquier	Baker	Bakery	Trail	Stores	46.77	1.99
76	Chico	Joyce	F	Victoria	Baker	Bakery	Victoria	Stores	45.30	5.57
77	White	Anthony	M	Kamloops	Baker	Bakery	Kamloops	Stores	38.30	6.10
78	Smith	Danny	M	Vancouver	Accounts Payable Clerk	Accounts Payable	Vancouver	FinanceAndAccounting	35.49	8.35
79	Porter	Ana	F	Mackenzie	Baker	Bakery	Prince George	Stores	41.66	4.71
80	Depaz	Shari	F	New Westminster	Baker	Bakery	New Westminster	Stores	43.51	1.78
81	Witkowski	Alice	F	Squamish	Baker	Bakery	Squamish	Stores	54.00	2.87
82	Alvarez	Kenneth	M	Nanaimo	Baker	Bakery	Nanaimo	Stores	46.22	4.42
83	Morales	Soledad	F	Gold Bridge	Baker	Bakery	Kamloops	Stores	47.66	4.25
84	Burden	Alexis	F	Penticton	Baker	Bakery	Kelowna	Stores	51.99	2.96
85	Dyal	William	M	Victoria	Baker	Bakery	Victoria	Stores	33.37	2.30
86	Hayden	Kenneth	M	New Westminster	Baker	Bakery	New Westminster	Stores	47.60	4.15
87	Caban	Joshua	M	Vancouver	Accounts Payable Clerk	Accounts Payable	Vancouver	FinanceAndAccounting	34.66	13.83
88	Knight	Mildred	F	Vancouver	Accounts Payable Clerk	Accounts Payable	Vancouver	FinanceAndAccounting	37.24	18.64

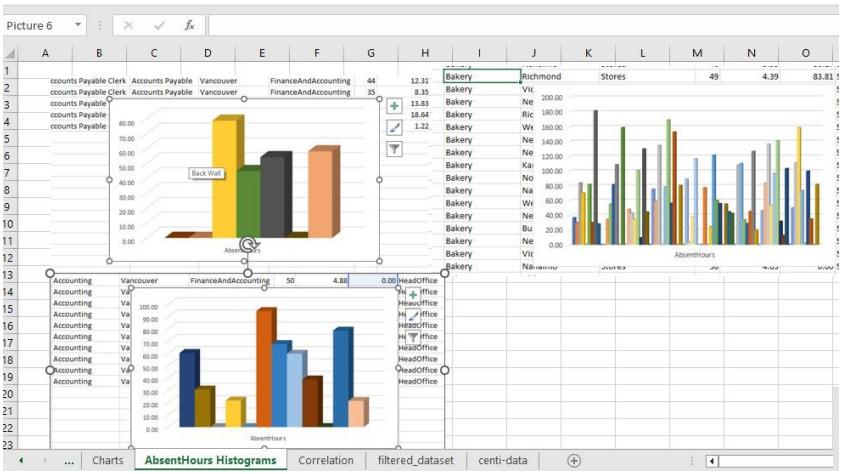
4. **Data Exploration:**



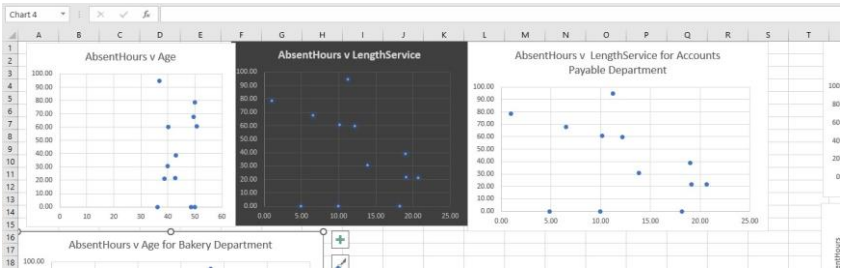
	Age		LengthService		AbsentHours	
Mean	42.60503913		Mean	5.439708356	Mean	61.63538966
Standard Error	0.220003551		Standard Error	0.092975914	Standard Error	1.088913848
Median	42.87285861		Median	4.680384655	Median	56.14398727
Mode	49.81724		Mode	1	Mode	0
Standard Deviation	9.897715021		Standard Deviation	4.182882924	Standard Deviation	48.98902264
Sample Variance	97.96476264		Sample Variance	17.49650956	Sample Variance	2399.924339
Kurtosis	0.115615966		Kurtosis	25.64767977	Kurtosis	-0.106497679
Skewness	-0.138720282		Skewness	4.453473656	Skewness	0.626598565
Range	73.4952575		Range	43.65008996	Range	252.1935348
Minimum	3.504742504		Minimum	0.085149041	Minimum	0
Maximum	77		Maximum	43.735239	Maximum	252.1935348
Sum	86232.59921		Sum	11009.96971	Sum	124750.0287
Count	2024		Count	2024	Count	2024
Largest(1)	77		Largest(1)	43.735239	Largest(1)	252.1935348
Smallest(1)	3.504742504		Smallest(1)	0.085149041	Smallest(1)	0
Confidence Level(95.0%)	0.431457176		Confidence Level(95.0%)	0.182338535	Confidence Level(95.0%)	2.135509591

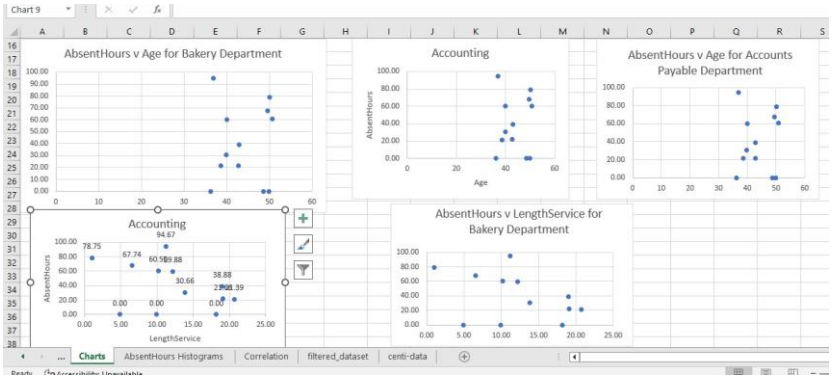
The majority of employees maintained a clean absence sheet since the mode of absenthours column is zero. This implies that a company can have more loyal employees than assumed.

For the three number columns, their means and medians are not far from one another. This means that there are no potential outliers in the dataset.



The 3-D histograms show the distribution of the number of absent hours across the three departments in the data - Accounting, Accounts Payable, and Bakery. These histograms show that each departmewnt has loyal employees who have maintained presence streak throughout the period of this data collection. They also show that these loyal employees are very small (25% or less) in each department.





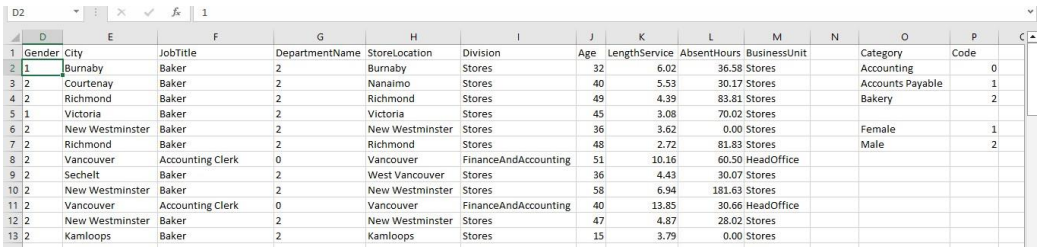
The scatter plot above shows the correlation between the length of service and age with number of absent hours across the three departments. The summary of these scatter plots are presented in the table below:

S/N	Column	Department	Degree	Correlation Type
1.	Age	Bakery	Low	Positive
2.	LengthService	Accounts Payable	Low	Negative
3.	Age	Accounting	Low	Positive
4.	Age	Accounts Payable	Low	Positive
5.	LengthService	Accounting	Low	Negative

6.	LengthService	Bakery	Low	Negative

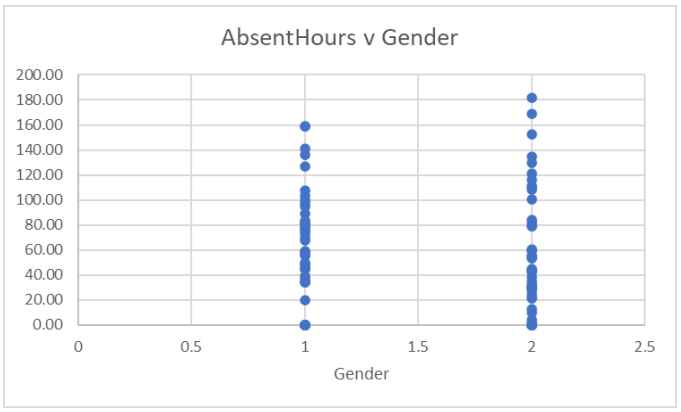
From this table, it can be deduced that the parameter that is most relevant to the number of absent hours is the age of the employee.

5. **Data Transformation:**

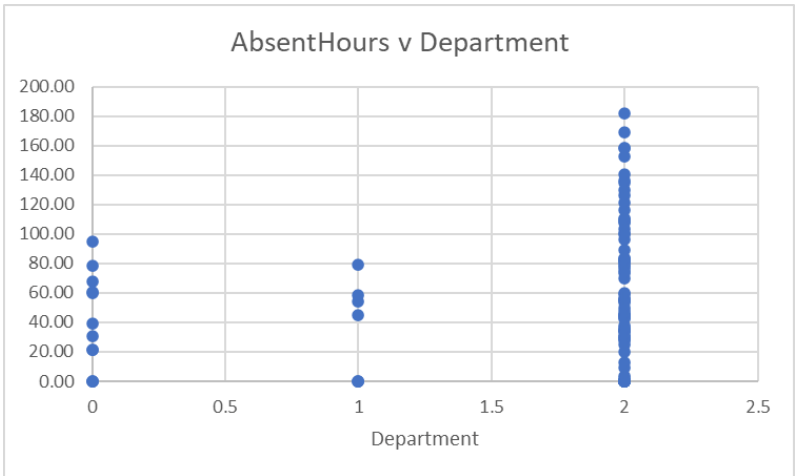


	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Gender	City	JobTitle	DepartmentName	StoreLocation	Division	Age	LengthService	AbsentHours	BusinessUnit		Category	Code
2	1	Burnaby	Baker	2	Burnaby	Stores	32	6.02	36.58	Stores		Accounting	0
3	2	Courtenay	Baker	2	Nanaimo	Stores	40	5.53	30.17	Stores		Accounts Payable	1
4	2	Richmond	Baker	2	Richmond	Stores	49	4.39	83.81	Stores		Bakery	2
5	1	Victoria	Baker	2	Victoria	Stores	45	3.08	70.02	Stores			
6	2	New Westminster	Baker	2	New Westminster	Stores	36	3.62	0.00	Stores		Female	1
7	2	Richmond	Baker	2	Richmond	Stores	48	2.72	81.83	Stores		Male	2
8	2	Vancouver	Accounting Clerk	0	Vancouver	FinanceAndAccounting	51	10.16	60.50	HeadOffice			
9	2	Sechelt	Baker	2	West Vancouver	Stores	36	4.43	30.07	Stores			
10	2	New Westminster	Baker	2	New Westminster	Stores	58	6.94	181.63	Stores			
11	2	Vancouver	Accounting Clerk	0	Vancouver	FinanceAndAccounting	40	13.85	30.66	HeadOffice			
12	2	New Westminster	Baker	2	New Westminster	Stores	47	4.87	28.02	Stores			
13	2	Kamloops	Baker	2	Kamloops	Stores	15	3.79	0.00	Stores			

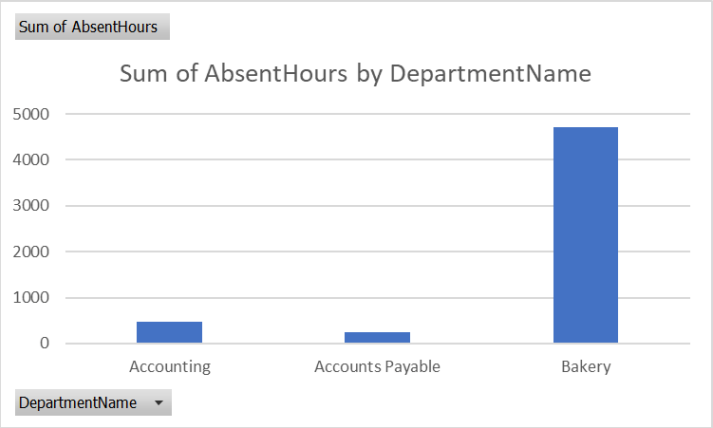
I transformed the Gender and DepartmentName columns into numerical columns so that I can view their correlations with the AbsentHours column.



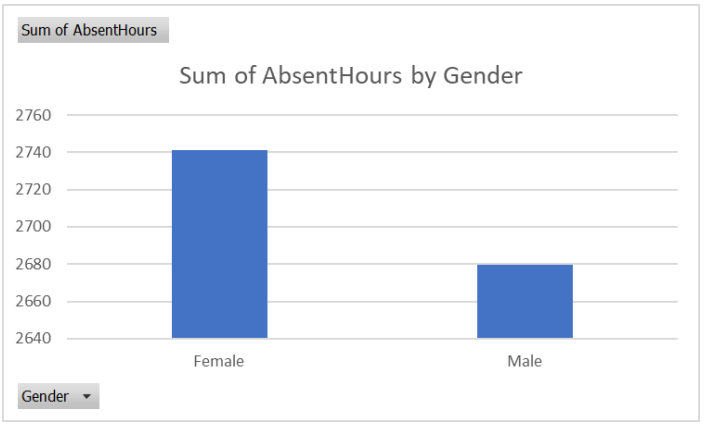
This scatter plot shows that there is no correlation between gender and AbsentHours columns. This implies that the amount of absent hours of an employee is not dependent on their gender.



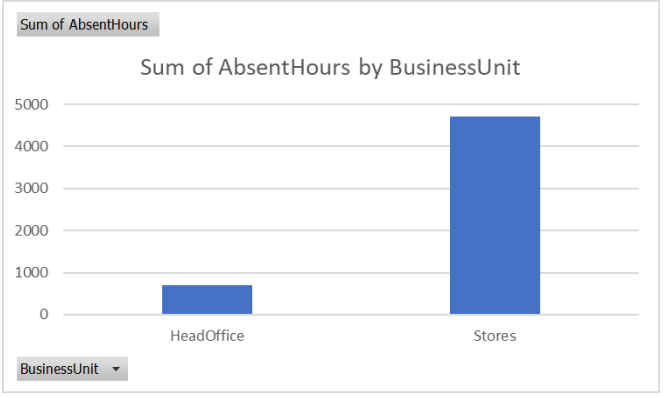
This scatter plot also shows that there is no significant correlation between the department of an employee and their amount of absent hours.



This chart, however, shows that the total amount of absent hours at the Bakery is more than that of its counterparts. This means that being in the Bakery department has a high impact on the amount of the absent hours of an employee. This also implies that there could be some internal problems in the Bakery department and it requires further investigation. Is it due to poor employee welfare management, or harsh work environment, or leadership and management issue, or other problems?



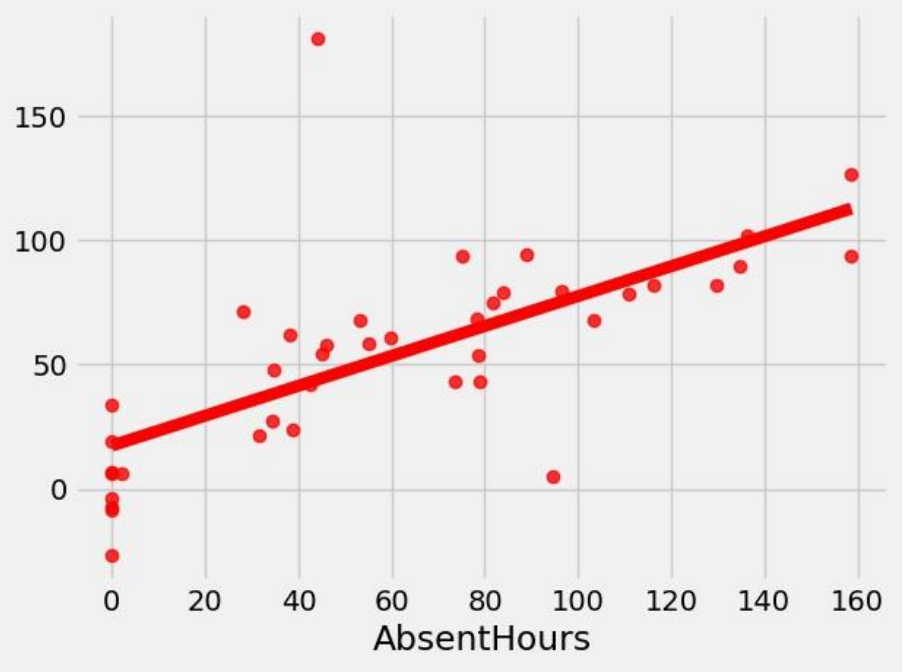
This chart also shows that being a female has an impact on the amount of absent hours. Females tend to be more absent at work than males. This requires further investigation to know why females are more absent at work than males. Is it due to a harsh work environment, harassment, poor management, family issues, or others?



From the chart above, it can be deduced that employees who are working at the Stores department tend to be more absent at work than their colleagues that are working at the HeadOffice. Could this be so because of the distance between work and home or some other reasons? It requires further investigation.

6. **Data Analysis:**

I trained the little centi-data with the machine learning's multiple linear regression model to predict AbsentHours based on the other numeric features including the categorical features that I converted to numericals. The numeric features that I used are: Gender, DepartmentName, Age, LengthService, and BusinessUnit.



This graph is the regression plot that shows the linear relationship between the actual values of AbsentHours and the values that were predicted by the regression model. It shows that there is a 'just good' relationship between the two values. They are not too widely separated and they are not clustered together.

```
[19] Accuracy=r2_score(y_test,y_pred)*100
     print(" Accuracy of the model is %.2f" %Accuracy)

     Accuracy of the model is 41.71
```

From the image above, the accuracy of this model is 41.71, which is quite a low value.

The script for the model development can be found at:

Absenteeism_ML_model.ipynb

https://colab.research.google.com/drive/1M__pODDMvQeRztqpOCQwObBO_9ot9Ly a?usp=sharing

7. **Interpretation of Results:**

The success of any company is ultimately related to the rate of work of her employees. Moreover, the rate of work of an employee can have different values depending on different conditions either within themselves, or their environments at work. But an employee will have a zero rate of work when he or she is totally absent at their on-site or remote workplace. It is therefore essential for a company to evaluate the factors that are contributing to the absence of employees at work and the probability of having an absent employee. This can actively assist the company to figure out the solution to the causing factors and inform them ahead of any possible absenteeism so that they can prepare for it, maybe by finding a supplementary work force and necessary replacement, or other means..

In lieu of those two problems, this project attempts to determine the factors that could cause the absenteeism of an employee at their workplace and develop a model that can predict a possible amount of absent hours of an employee, which can help in determining whether an employee will be absent or not.

This project has found that the age of an employee is a potential factor that causes absenteeism in employees. When the age of the employees were plotted against the number of absent hours, the plot resulted in a positive correlation for each of the three departments in the project's dataset. But this project did not find out whether absenteeism was more common in young or old workers.This can be determined in further research.

Another factor that was discovered in this project is gender. When the amount of absent hours of all employees was totalled and categorised into male and female, it was found that females have higher cumulative amount of absent hours. This means that females tend to be more absent at their workplace than their counterpart males. It also calls for the attention of the company's management to look into the causes of such absenteeism at

work. Are they being mistreated at work affecting their emotional well-being? Are they being exposed to harsh work environments that affect their mental or physical health? Among other factors to be considered by the management.

Finally, the third factor that was discovered in this research is the department of the employee. In the analysis process, the total amount of absent hours for all employees in each of the three departments was calculated. It was found out that the Bakery department has the highest number of absent hours. Although there is another column for the business unit that has Stores and other business units, it was found through filtering that 'bakery' is the only

department under the Stores business unit. This discovery could bring along some concerns about the leadership and management of this department and or business unit. Another thing to consider is the work conditions of this department., among other things to consider.

Regarding the model that can help the company in predicting a possible absenteeism from an employee, this project has trained a model that depends majorly on the three factors mentioned above to predict the possible amount of absent hours of an employee. The model is a multiple regression model. It was trained on five features to predict one continuous value, which is the amount of absent hours of an employee. The total number of data points is 100 and that is quite a small number. The accuracy of the model was found to be 41.71% and that is not a high value. But it is a good starting point for further analysis.

Therefore, in further analysis, the amount of data points to use might be increased, maybe 100 rows is too small to give a good result. Also, the model itself can be replaced with other models that can do the same task to see if the accuracy will improve.

In conclusion, this project discovered three factors that could have a strong impact on the rate of absence of an employee. These factors are the age of the employee, the gender of the employee, and the department or business unit of the employee. The data was, also, trained on multiple linear regression model using the discovered factors. Further investigations can be made to discover the root causes of the factors and further analysis can be done to get a better model that can predict the amount of absent hours of an employee with a higher accuracy.