# Situation understanding

## Situation objectives

More than 24 million employees quit their jobs in the United States between April and September 2021. This is an all-time record. According to the latest statistics from the U.S. Department of Labor's Bureau of Labor Statistics, mass turnover gained momentum again in November 2021, with a record 4.53 million job separations (NICE, 2021).

According to the MIT Sloan Business Review (Sull, 2022), it affects blue- and white-collar workers equally, although there are differences by industry. The hardest hit industries, apparel/retail, fast food, and specialty retail had the highest percentage of blue-collar workers of all industries surveyed.

The management consulting industry, on the other hand, had the second highest turnover rate. The enterprise software industry also had the highest turnover rate, with the highest percentage of engineers and technical employees.

Interestingly, even within the same industry, there is a wide division between companies that have high turnover and those that do not. For example, Boeing's turnover rate is 6.2% compared to 21.2% at SpaceX; at HSBC it is only 5.1%, while at Goldman Sachs it is 15.2% (Milano, 2021).

According to the United Nation (n.d.), this issue falls under number 8 of the 17 SDGs. The goal of this issue is to ensure that employees are satisfied with the company environment and not dissatisfied with their jobs. There must be a cause for this current phenomenon. It is necessary to solve the cause and create an environment where employees are happy and free from anxiety.

## Situation assessment

### Resource

### Requirement

In this project, we will analyze the employee attrition rate of a company. The company employs 4,000 people at any given time, and each year about 15% of the employees leave the company and the company needs to find available talent in the job market. The company needs to analyze the data to determine which factors are causing employees to leave and how to retain them. The focus here is on knowing which variables are important and how to address them. And can they be implemented immediately?

### Assumptions

This turnover rate is a detriment to the company and must be remedied immediately. A high turnover rate can lead to a shortage of talent, project delays, and a lack of ability to meet deadlines, which can damage the company's reputation with clients and partners. In addition, the HR department needs employees to hire new people and maintain a reasonable size department. This is because the HR department has to spend time on training and other education before new employees can become accustomed to the company.

### Constraints

Since this dataset is about an anonymous company, we do not know what kind of services they provide in what market. If this were known, comparisons with other firms might have been possible. For example, a comparison with a successful company in the same market with a low turnover rate would have revealed clear differences. However, even if the company with this dataset were identified, it would be nearly impossible to obtain the employee dataset for the other company.

### Risks & Contingencies

|  |  |
| --- | --- |
| Risk | Contingency |
| The results of the data analysis will not improve the current environment, but rather worsen it | The data analysis would have to be redone from scratch, using different modeling and algorithms to derive new results. |
| Delay of the project | Work on the project in overtime. |
| Software resources’ bugs | Change resources and continue the project using the same process. |

## Data mining Goals

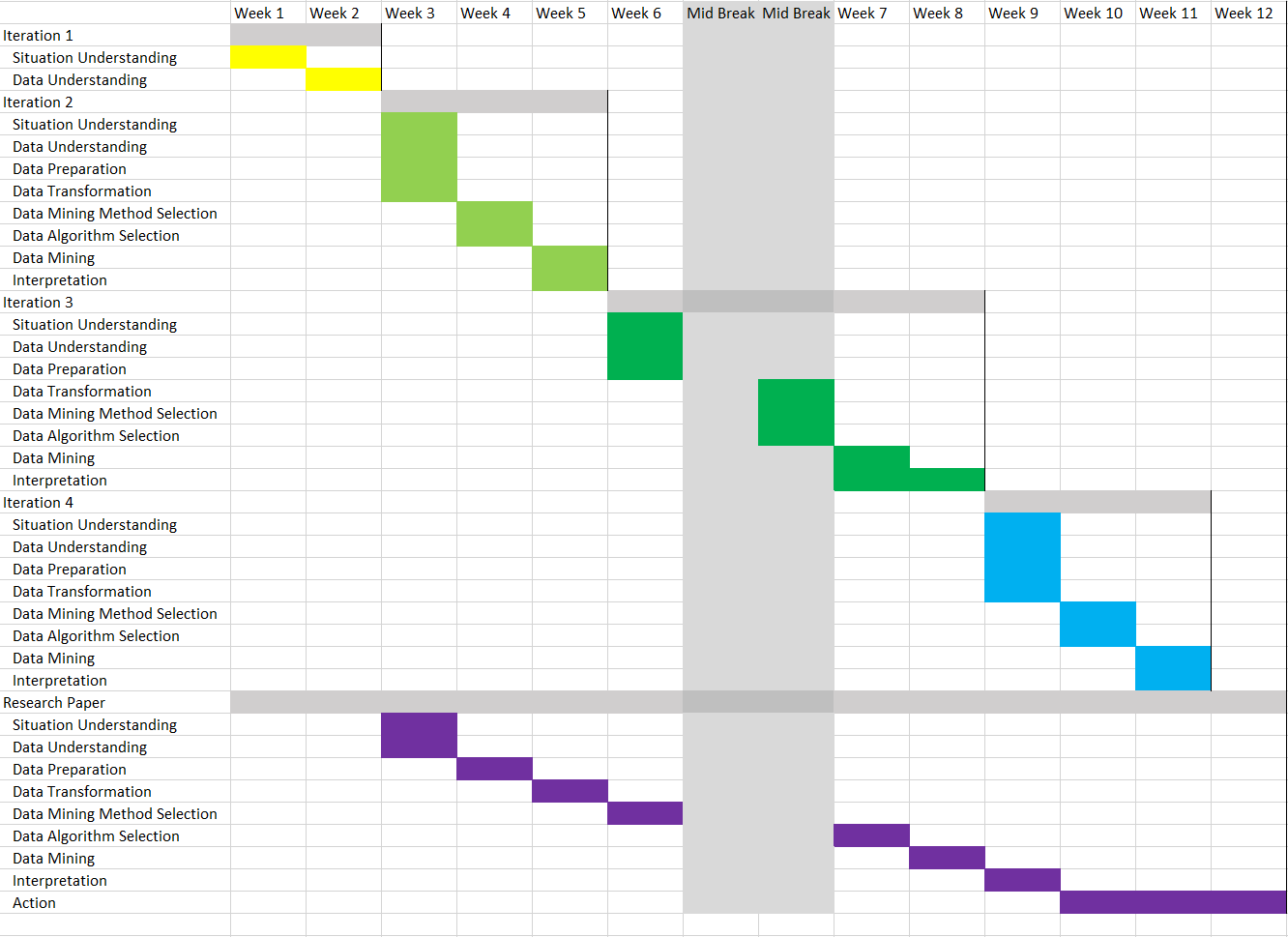
The goal of this data mining is to find the variables that cause turnover and to suggest remedies for them. There are 28 variables in the data, so we will use data mining to find out which of these variables are most related to the turnover rate. Then, using the results of the study, what kind of improvement measures will be developed? Is the idea feasible immediately or will it be a long-term improvement? Therefore, the three data mining objectives are as following.

1. Which fields have the most impact on Attrition.
2. What are the relationships between the most influential fields and Attrition?
3. Will the model created in this project be useful for future projections? And how will it be useful?

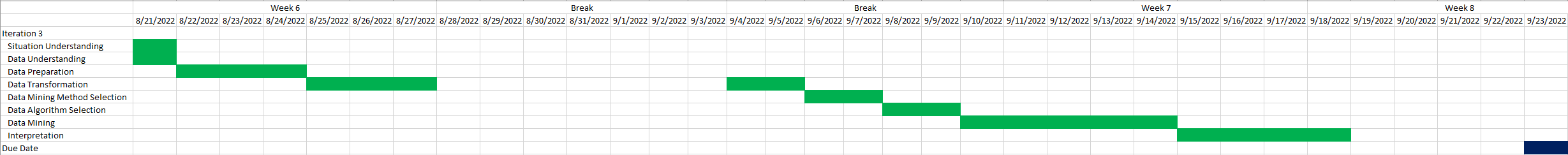
## Project Plan

This section indicates the weekly semester schedule and also the daily schedule for iteration two.

### Semester Gantt chart



### Iteration 3 Gantt chart



# Data understanding

## Initial data collection

This data set was collected from Kaggle. The first step was to check the websites of the companies I was interested in to see if there was any public data available there. However, none of the companies I checked had publicly available data. This is because, as a matter of course, it would be an invasion of privacy if they disclosed the details of their employees. Thus, I checked data websites such as NZStats and WorldinData to see if there was any data available. Kaggle had the most realistic and detailed data, so we decided to use it.

## Data description

The dataset I have collected contains 29 columns and 4410 rows. Each column is described as following.

* EmployeeID
* Age: Age of the employees
* Attrition: Whether the employee left in the previous year or not
* BusinessTravel: How frequently the employee travelled for business purpose in the last year
* Department: Department in company
* DistanceFromHome: Distance from home in kms
* Education: Education Level 1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'
* EducationField: Field of education
* EmployeeCount: Employee count
* Grnder: Gender of employee
* JobLevel: Job level at company on a scale of 1 to 5
* JobRole: Name of job role in company
* MaritalStatus: Whether the employee married, divorces, or single.
* MonthlyIncome: Monthly income in rupees per month
* NumCompaniesWorked: Total number of companies the employee has worked for
* Over18: Whether the employee is above 18 years of age or not
* PercentSalaryHike: Percent salary hike for last year
* StandardHours: Standard hours of work for the employee
* StockOptionLevel: Stock option level of the employee
* TotalWorkingYears: Total number of years the employee has worked so far
* TrainingTimeLastYears: Number of times training was conducted for this employee last year
* YearsAtCompany: Total number of years spent at the company by the employee
* YearsSinceLastPromotion: Number of years since last promotion
* YearsWithCurrManager: Number of years under current manager
* EnvironmentSatisfaction: Work Environment Satisfaction Level 1 'Low' 2 'Medium' 3 'High' 4 'Very High'
* JobSatisfaction: Job Satisfaction Level 1 'Low' 2 'Medium' 3 'High' 4 'Very High'
* WorkLifeBalance: Work life balance level 1 'Bad' 2 'Good' 3 'Better' 4 'Best'
* JobInvolvement: Job Involvement Level 1 'Low' 2 'Medium' 3 'High' 4 'Very High'
* PerformanceRating: Performance rating for last year 1 'Low' 2 'Good' 3 'Excellent' 4

The following figure shows the rough idea of how the data looks like before I manipulate it.

## Data exploration

## Data quality

### Errors

### Missing values

### Data quality patterns

# Data preparation

## Data selection

## Data cleaning

## Data construction

## Data integration

## Data formatting

# Data transformation

## Data reduction

## Data projection

# Data-Mining Methods Selection

## DM Methos Objectives

### DM Goal

### DM Method

## DM Methods Selection

# Data-Mining Algorithms Selection

## Exploratory Analysis

## Algorithms Selection

## Models Selection/Building

# Data Mining

## Logical Test Design

## Data Mining

## Model’s Output

# Interpretation

## Pattern’s Discussion

## Data Visualization & Results

## Results Interpretation

## Result Assessment & Evaluation

## Iterations

# Disclaimer

# Reference