**Group Members**

Zhongyuan Yan

Sara Zaheri

Camille Howard

**Study Overview**

This study contains a data analysis of a dataset (CaseStudy2-Data.xlsx). The goal is to conduct exploratory data analysis (EDA) to determine factors that lead to attrition. In an effort to further understand two datasets our team identify the top three factors that contribute to turnover. The business is also interested in learning about any job role specific trends that may exist in the data set. Our team provide any other interesting trends and observations from your analysis. Our result is backed by robust experimentation and (where applicable) appropriate visualization.

**Data Source/ Basic Statistics (EDA)**

CaseStudy2-data.csv the dataset is provided by the company that we are working with.

Our research team is using Generalized Linear Model to determine factors that lead to attrition.

**The Data**

CaseStudy2-data.csv

|  |  |
| --- | --- |
| Age | Age of Participants |
| Attrition | The unpredictable and uncontrollable, but normal, reduction of work force. |
| Business Travel | Travel Frequency |
| Daily Rate | Daily work rate |
| Department | Department of Participants |
| Distance from Home | Work Distances form home |
| Education | Education Level of Participant |
| Education Filed | Field of Education of the participant |
| Employee Count | Count of Employee |
| Employee Number | Employees’ number |
| Environment Satisfaction | Overall satisfaction of work environment |
| Gender | Gender of Participant |
| Hourly Rate | Hourly pay |
| Job Involvement | Involvement at work |
| Job Level | Level of work |
| Job Role | Working role of participant |
| Job Satisfaction | Overall job satisfaction |
| Marital Status | Marital status of Participant |
| Monthly Rate | Monthly pay |
| Number of Companies worked | Number of Companies worked by participant |
| Over18 | Is the Participant over the age of 18 |
| Overtime | Work overtime |
| Percent Salary Hike | Percent of salary increase |
| Performance Rating | Rating related to performance |
| Relationship Satisfaction | Overall satisfaction of the relationship at work |
| Standard Hours | Standard hours of work |
| Stock Option Level | Level of stock |
| Total Working Years | Total of working year of Participant |
| Training Times Last Year | Training time of Participant |
| Work Life Balance | Balance between work and life |
| Years At Company | Year work at the current company |
| Years In Current Role | Year work at the current role |
| Years Since Last Promotion | Year since the last promotion of the participant |
| Years With Current Manager | Year worked with the current manager |

**Workflow**

1. Formulate Repository though GitHub
2. Clean the dataset
   1. Read the csv into R
   2. Clean up the column names
   3. Convert the column data value into proper types
3. Preliminary Analysis
   1. Remove all observations where the participant is under age 18.
   2. Create a simple histogram
   3. Give the frequencies for Gender, Education, and Occupation.
   4. Give the counts of management positions.
4. Deeper Analysis and Visualization
   1. Groom the results
   2. Create bar charts in ggplot
   3. Create a scatterplot between Age and Income.
   4. Create scatterplot for Life Satisfaction

**Folder & File Information**

CaseStudy2-data.csv contains all the data files that is used in the research study.

CaseStudy2\_merged.Rmd is the Rmarkdown version of the entire report.

CaseStudy2\_merged.html is html page of the entire report of the research study.

Datacleaning.csv is the dataset after cleaning

Researcher Information & Responsibilities

|  |  |
| --- | --- |
| Researchers | Responsibilities |
| Camille Howard | Preliminary Analysis and PowerPoint creation |
| Sara Zaheri | Clean Raw Data, Deeper Analysis and Visualization |
| John Yan | Formulate your Repo |

**The Questions We're Interested in Answering**

a Remember to include things like a clean, informative title, axis labels that are in plain English, and readable axis values that do not overlap.

b Create bar charts in ggplot or similar. The bars should be in descending order, Use any color palette of your choice other than the default.

c Is there a relationship between Age and Income?

d What about Life Satisfaction