

**Final Report**

Robin Godinho

INST627: Data Analytics for Information Professionals

Professor Naeemul Hassan

December 15th, 2023

|                            |           |
|----------------------------|-----------|
|                            | 2         |
| <b>Table of Content</b>    |           |
| <b>Introduction</b>        | <b>1</b>  |
| <b>Research questions</b>  | <b>2</b>  |
| <b>Methodology</b>         | <b>3</b>  |
| <b>Data description</b>    | <b>4</b>  |
| <b>Solutions Framework</b> | <b>5</b>  |
| <b>Experiment results</b>  | <b>7</b>  |
| <b>Discussion</b>          | <b>14</b> |
| <b>References</b>          | <b>15</b> |

## **Introduction**

The purpose of this data analytics project is to use existing data from a company's human resources department in order to derive important quantitative insights. Analyzing HR data is paramount for a company's strategic decision-making and overall organizational effectiveness. In today's dynamic business environment, human resources play a pivotal role in shaping the success of an enterprise. By delving into HR data, companies gain valuable insights into employee performance, engagement levels, and talent acquisition. This analytical approach enables organizations to identify trends, forecast future workforce needs, and optimize resource allocation. Moreover, the analysis of HR data empowers businesses to enhance employee satisfaction, reduce attrition rates, and foster a more inclusive, equitable and diverse workplace. In essence, the careful examination of HR data equips companies with the tools needed to align their human capital strategy with broader business goals, fostering a culture of continuous improvement and adaptability in the ever-evolving corporate landscape.

Most organizations lack clear performance indicators to track and analyze HR performance and other metrics. With the assistance of powerful analysis tools and libraries such as Python and pandas modules, it is possible to perform relevant data extraction for decision making and forecasting performance.

The structure of this project report will begin with the research question and defining the hypothesis statements. This section will go over the proposed null hypothesis and alternative hypothesis. The next section will go over the data description, which is aimed at analyzing the qualitative and quantitative aspects of the dataset. This section will describe the dataset as a whole and the different data types it houses. Thereafter, we will go over the methodology and

code written to derive the key data points. This section goes over the different statistical methods used in python to generate key insights and visuals.

The solutions framework section will go over the tools and methods used to generate the visualizations in Python as well as the interactive data dashboard in PowerBI for the company's stakeholders. Lastly, this report will go over the experimental findings and how they were derived. This section will include factors such as dataset splitting, parameter settings and evaluation metrics.

## **Research questions**

The human resources dataset includes several data points that could be used to derive an array of insightful information about the company. The main research question are:

1. Is there a significant difference in monthly income between male and female employees?  
How does this vary across different departments or job roles?
2. What factors contribute most to employee attrition? Are there significant differences in attrition rates based on gender, age, department, or job role?
3. How diverse is the workforce in terms of gender, age, and educational background? How does diversity impact team performance and employee satisfaction?

The following research questions are sufficient to drive the quest for detailed information regarding this company. They ensure answers to key performance indicator questions such as employee attrition rate, which is a metric used to measure employees lost over a period of time. Identifying if there are any significant income disparities based on gender could be a good predictor of the attrition rate and subsequently job satisfaction. Assessing a company's diversity metrics will help determine if it incorporates diversity, equity and inclusion.

## Methodology

The research question that drove this project was “Is there a significant difference in monthly income between male and female employees?” The null hypothesis is that there is no significant difference in monthly income between the male and female employees. The alternative hypothesis states that there is a statistically significant difference between the male and female monthly income. In order to answer this hypothesis question, I decided to use the one-sided t-test to compare the mean salaries of male and female employees. This test helps determine if there are statistically significant differences between the two groups' means. The reason for selecting the one-sided t-test is because we want to determine a difference between the income in one direction, i.e. if one gender earns significantly more than the other.

For preliminary analysis, I performed a statistical summary on the ‘monthly income’ column, and extracted data on the mean, standard deviation, median, max, min, IQR and variance. This provided more insights on the varying income ranges and what gender those individuals belong to. To further analyze the gender, I developed a simple bar graph to showcase the no number of employees that are male or female.

The subsequent research question investigates the attrition rate. The question was “What factors contribute most to employee attrition? Are there significant differences in attrition rates based on gender, age, department, or job role?” For this I performed an ANOVA (Analysis of Variance) to compare the mean job satisfaction levels across various departments. This test helps to determine if there are any statistically significant differences between the means of three or more independent groups.

The final research question was in regards to diversity in the company. The question asks “How diverse is the workforce in terms of gender, age, and educational background? How does

diversity impact team performance and employee satisfaction?” This was answered using the interactive data dashboard visualization, generated using Power BI tool. This was achieved by filtering the data dashboard based on age, education background and gender. Using the dashboard was effective for the human resources department to have a live visualization of the diversity states as they hire new employees.

### **Data description**

The following fictitious human resources dataset was captured from Kaggle.com, depicting the extensive information of the workforce. The dataset consists of 1470 rows and 47 columns. Upon reviewing the dataset, I began the preprocessing step of cleaning to ensure consistency through the various columns and rows. Using the panda module, it was possible to run scripts to identify 4 discrepancies in the data. Firstly with the ‘job satisfaction’ column, containing a few null rows. I also identified a duplicate row in the dataset, and cleaned the error using excel.

The human resources dataset provides a comprehensive overview of various aspects of employee data within an organization. It includes a wide range of variables that cover both personal and professional attributes of the workforce. Key demographics include information such as age, gender, and education level, offering insights into the diversity of the employee base. Professional attributes encompass job roles, departments, and years of experience, highlighting the distribution of employees across different functional areas and their tenure within the company.

Crucially, the dataset also delves into compensation details, including monthly income and salary hikes, which are pivotal for analyzing pay structures and equity. Employee

performance metrics are captured through variables like performance ratings, providing a basis for understanding productivity and effectiveness. Additionally, the dataset addresses work-life balance and relationship satisfaction, key indicators of employee well-being and engagement.

Another significant aspect of this dataset is the inclusion of attrition data, which marks whether employees are currently active or have left the organization. This information is essential for examining turnover rates and identifying potential factors contributing to employee retention or departure.

Overall, this human resources dataset is a rich resource for analyzing various dimensions of workforce management, employee satisfaction, and organizational dynamics, making it an invaluable tool for HR analytics and strategic decision-making.

## **Solutions Framework**

To analyze the human resources dataset comprehensively, you can follow a structured framework that includes data cleaning, data analysis in multiple phases, and data visualization, leading to the creation of a data dashboard. Here's a detailed solution framework:

### **Data Cleaning: Phase 1**

1. Column and Row Counts: Determine the size of the dataset by counting the number of rows and columns. This helps in understanding the scope of data you're working with.
2. Missing Data Checks: Identify any missing or null values in the dataset. This can be done using functions like `isnull()` in Python.
3. Duplicates: Check for and remove any duplicate records to ensure the uniqueness of data.
4. Incorrect / Inaccurate Data Points: Verify the data for accuracy and consistency. For example, check for impossible values in age or income fields.

5. Outliers: Identify and handle outliers in the data. Outliers can be detected using statistical methods like the IQR (Interquartile Range) method.

### **Data Analysis: Phase 2**

1. Hypothesis Statements ( $H_0/H_a$ ): Formulate null ( $H_0$ ) and alternative ( $H_a$ ) hypotheses for statistical tests. For example,  $H_0$ : "There is no significant difference in average monthly income between genders."

2. Performing One-Sided t-test: Conduct a one-sided t-test to test the hypothesis. This will help to determine if there are significant differences in parameters like income between different groups.

### **Data Analysis: Phase 2**

1. Filter Using 'iloc': Utilize 'iloc' in Python to filter and manipulate the dataset for detailed analysis.

2. Extracting Insights for Key Performance Indicators (KPIs): Focus on KPIs such as employee turnover rate, average age, and job satisfaction levels to derive meaningful insights.

### **Data Analysis: Phase 3**

Preliminary Data Visualizations Using Pandas: Create basic visualizations like histograms, box plots, and scatter plots to understand data distributions and relationships.

### **Data Dashboard Flowchart**

#### KPIs:

- Overall Employee Count: Total number of employees in the dataset.
- Attrition Count: Number of employees who have left the company.
- Attrition Rate (Percentage): Attrition count divided by the overall employee count, expressed as a percentage.
- Active Employees: Number of current employees.
- Average Age: Mean age of the employees.



### Pie Chart:

- Department-wise Attrition: Display the proportion of attrition in each department.
- Age-wise Attrition: Show attrition distribution across different age groups.
- Stacked Bar Graph: Create a graph for Gender filtered by age group to visualize the distribution of employees.

Table: Use a heat map to show job satisfaction by job role, highlighting areas of high and low satisfaction.

This comprehensive framework allows for a thorough exploration and understanding of the human resources dataset, providing valuable insights into workforce dynamics and enabling data-driven decision-making.

## Experiment results

Having a look at the results from the data analysis conducted throughout this project, there are some key insights that were derived. Starting off with some preliminary analysis figures, that will later enhance our understanding of the research questions.

```
Gender Analysis

Female = hr['Gender'].value_counts()['Female']
print(f'The number of female employees in the company is {Female}')

Male = hr['Gender'].value_counts()['Male']
print(f'The number of Male employees in the company is {Male}')

Total = Female + Male
Total

Female_percentage = Female/Total*100
print(f'The females in the company is {Female_percentage}%')

Male_percentage = Male/Total*100
print(f'The males in the company is {Male_percentage}%')

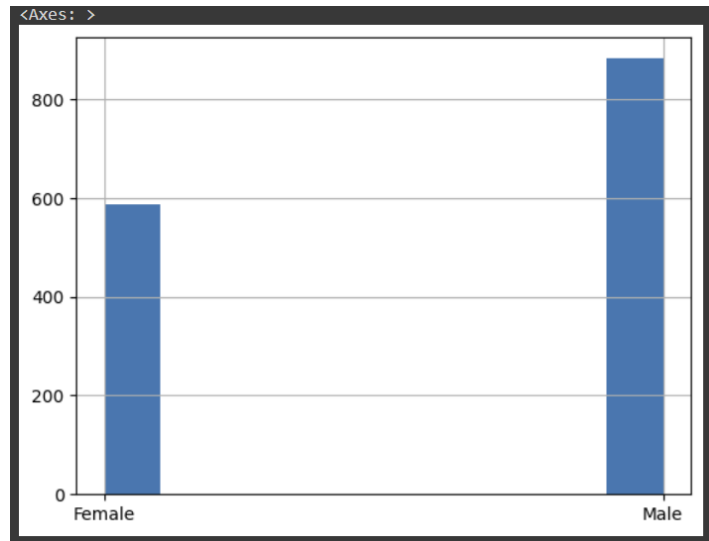
hr['Gender'].hist()

The number of female employees in the company is 588
The number of Male employees in the company is 882
The females in the company is 40.0%
The males in the company is 60.0%
<Axes: >
```

**Figure 1: Gender Analysis**

In fig.1 we analyzed the employee count based on the distinctive gender. It was derived that the female employees amount to 588 and male employees are 882, making the company male dominant. This discrepancy averages to 40% females and 60% male employees. The bar graphs

below, is a clear indication of this gender imbalance at the company.



**Figure 2:** Bar graph of employees based on gender

Thereafter, an analysis based on the 'education level' was conducted. According to the data extracted from fig.3 below, it was evident that most employees have a Bachelor's degree at 572 employees. Thereafter, a significant number of employees have obtained a Master's degree with about 398. Lastly, we see that the minority education level is a Doctoral degree, with only 48 employees. This analysis can enhance our understanding of the different income levels.

```
# Different levels of Education
Associates = hr['Education'].value_counts()["Associates Degree"]
print(f'The number of employees with a Associates degree is {Associates}')

Bachelors = hr['Education'].value_counts()["Bachelor's Degree"]
print(f'The number of employees with a Bachelors degree is {Bachelors}')

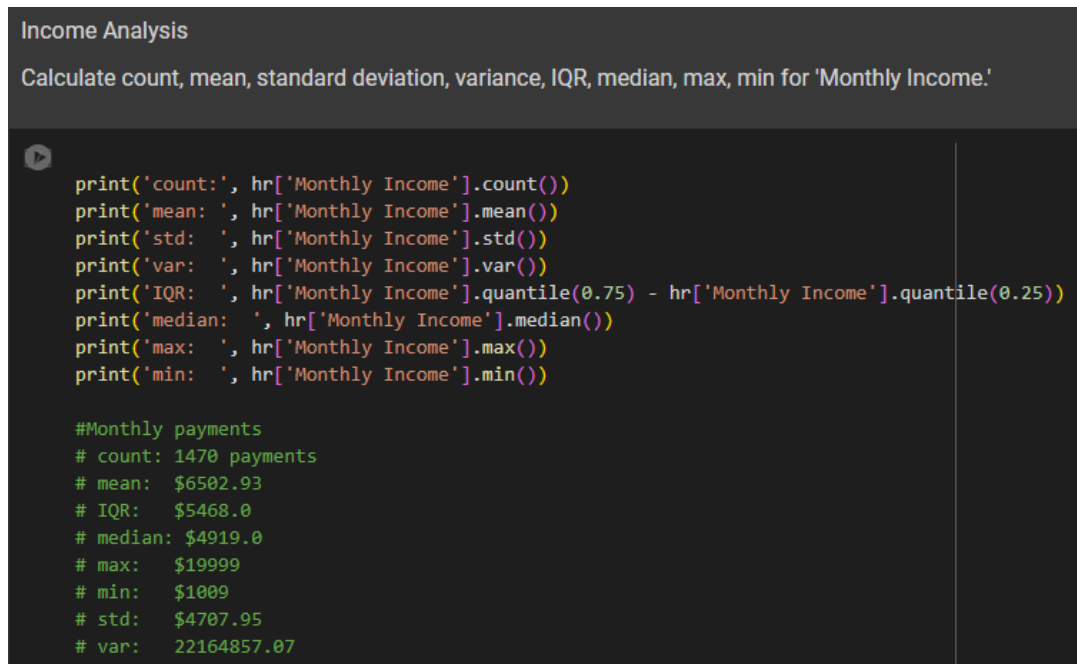
Masters = hr['Education'].value_counts()["Master's Degree"]
print(f'The number of employees with a Masters degree is {Masters}')

Doctoral = hr['Education'].value_counts()["Doctoral Degree"]
print(f'The number of employees with a Doctoral degree is {Doctoral}')
```

```
The number of employees with a Associates degree is 282
The number of employees with a Bachelors degree is 572
The number of employees with a Masters degree is 398
The number of employees with a Doctoral degree is 48
```

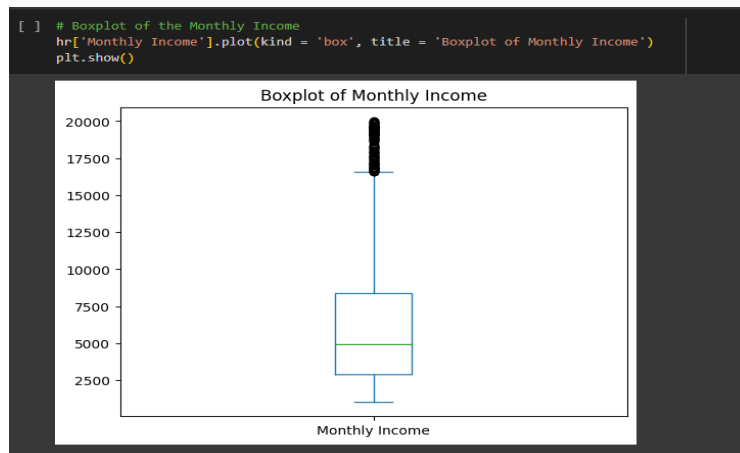
### Figure 3: Educational Level Analysis

Subsequently, we derived some statistical summary data on the 'monthly income' attribute of the dataset. This summary consisted of mean, median, max, min, IQR, and standard deviation of the monthly income at this company. Below is a fig.4 with the results from the summary analysis on the monthly income.



**Figure 4: Income Summary Analysis**

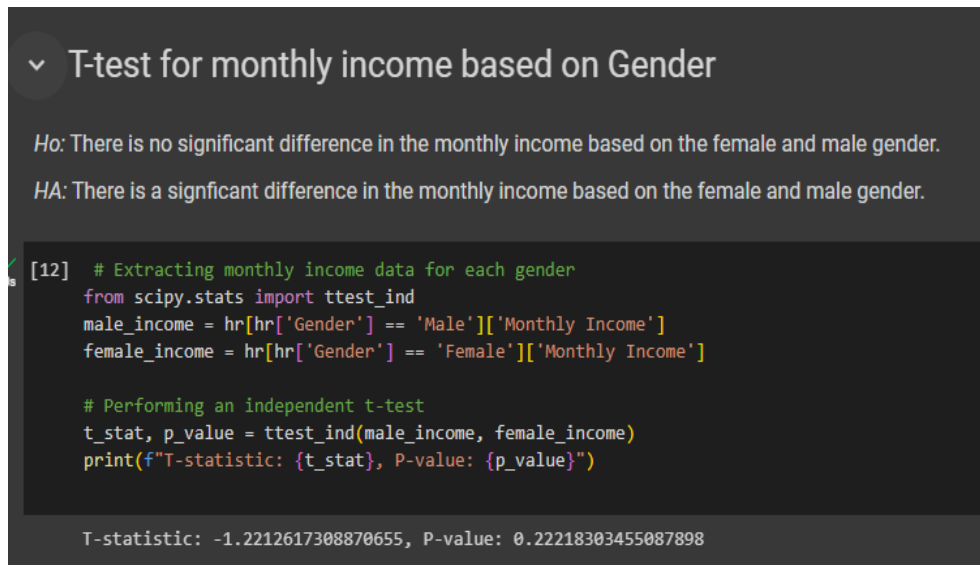
Based on the analysis, it was possible to determine that the mean (average) salary at the company is about \$6502.93 per month, the minimum salary is \$1009 and the maximum salary is about \$20,000 per month. These ranges vary significantly and may be difficult to interpret numerically. I decided that using a boxplot would help further one's interpretation of these income ranges. The figure below is a boxplot on the monthly income range.



**Figure 5:** Boxplot of Monthly Income

By performing the boxplot visualization, it was easier to see the variability with regards to the income. It is clear to see that a number of employees earn significantly more, in the range of (\$17,500 - \$20,000). This could be a reflection of the 48 employees who have Doctoral degrees, working in the R&D department.

To answer the research question that drove this analytics project “Is there a significant difference in monthly income between male and female employees?” I performed a one-sided t-test to determine if there is a significant difference between the male and female monthly income ranges. The null hypothesis is that there is no significant difference in monthly income between the male and female employees. The alternative hypothesis states that there is a statistically significant difference between the male and female monthly income. Fig. 6 below showcases the t-test analysis.



The screenshot shows a Jupyter Notebook cell with the following content:

```
✓ [12] # Extracting monthly income data for each gender
      from scipy.stats import ttest_ind
      male_income = hr[hr['Gender'] == 'Male']['Monthly Income']
      female_income = hr[hr['Gender'] == 'Female']['Monthly Income']

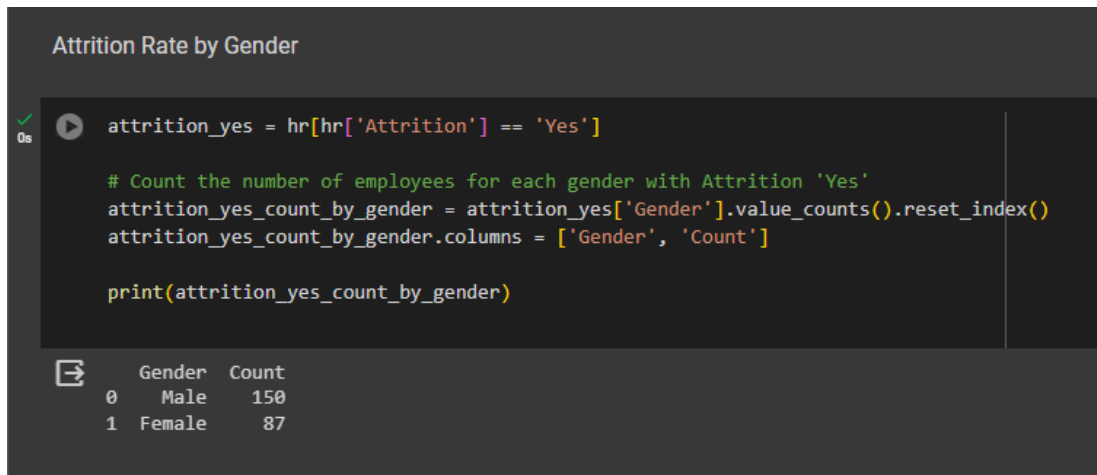
      # Performing an independent t-test
      t_stat, p_value = ttest_ind(male_income, female_income)
      print(f"T-statistic: {t_stat}, P-value: {p_value}")
```

T-statistic: -1.2212617308870655, P-value: 0.22218303455087898

**Figure 6:** One-sided t-test on Monthly Income filtered by gender

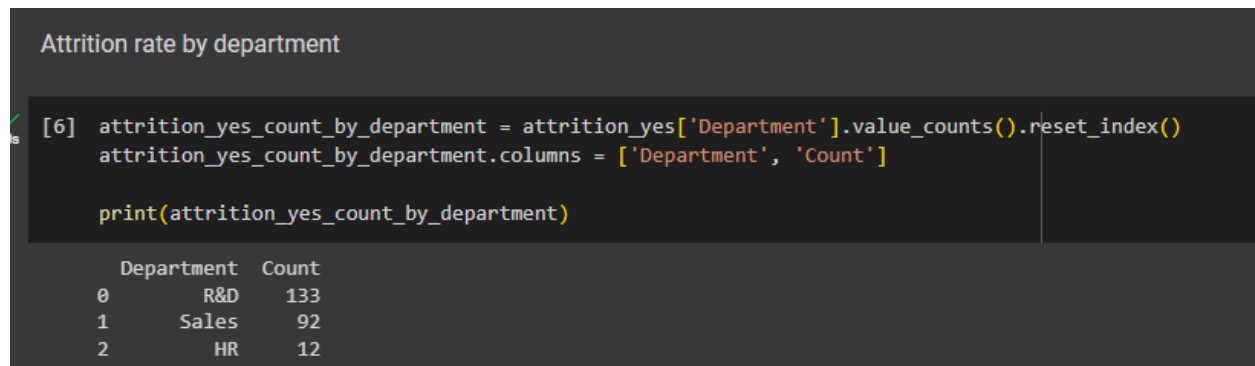
Since the p-value of 0.22218303455 is much higher than the standard confidence level of 0.05, we fail to reject the null hypothesis that there is no significant difference between the male and female gender based on their monthly income. There is insufficient evidence in support of the alternative hypothesis that the difference in proportion is statistically significant.

The subsequent analysis was conducted regarding the attrition rate, meaning the number of employees leaving the company during a given period. The figures below showcase the attrition rate filtered by gender, department and job satisfaction.



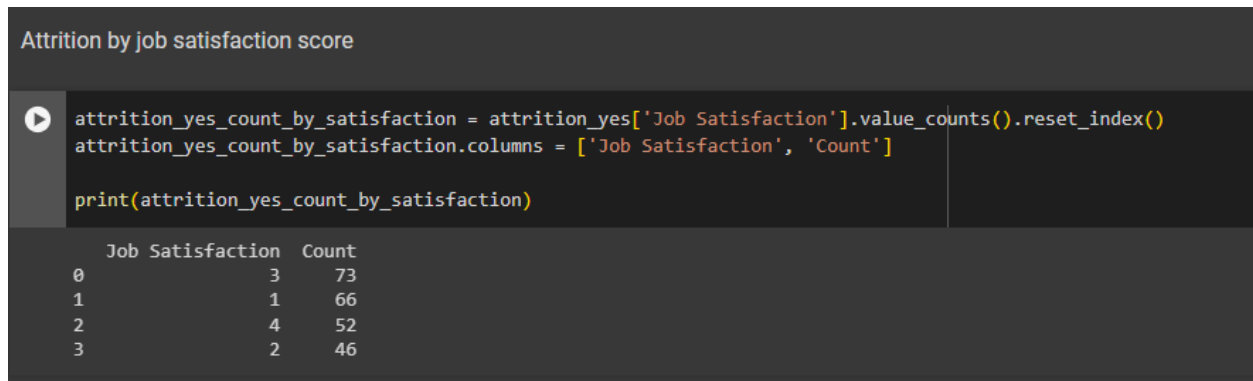
**Figure 7.1:** Attrition by Gender

In this instance we determine that males are leaving at a higher rate of attrition than females. This could be due to the fact that there are more males in the company than females.



**Figure 7.2:** Attrition by Department

The department with the largest attrition rate is the R&D department. Given that this is one of the most important departments, this signifies that it has to re-evaluate its efforts to retain its employees. This also signifies that the HR department will need to hire several new employees to fill up the vacant positions.



**Figure 7.3:** Attrition by job satisfaction

This is a significant indicator of the attrition rate. Given that the employees that had a lower job satisfaction rating, tended to have a higher probability of leaving the company. This is an indication that the HR department needs to analyze what factors are contributing to the high unsatisfactory scores from its employees.

Lastly, to analyze the diversity of the workforce in terms of gender, age, and educational background, I developed an interactive dashboard. The dashboard was filtered based on gender, age and education background. Below the figure below showcases the different KPIs and visualizations that were included in the interactive dashboard.



**Figure 8:** HR Analytics dashboard

The HR analytics dashboard incorporates different visualizations such as pie charts for the department-wise attrition, age-wise attrition, and a stacked bar graph filtered by age group to visualize the distribution of employees. The dashboard also had a heatmap displaying the different job satisfaction ratings filtered by job roles.

## **Discussion**

The human resources department could leverage this dataset effectively to enhance various aspects of workforce management and strategic planning. By meticulously analyzing data pertaining to employee demographics, compensation, performance, and attrition, HR can uncover key insights that inform policy and practice improvements. For instance, the identification of pay gaps or discrepancies in performance ratings across different groups could lead to more equitable compensation structures and appraisal processes. Similarly, understanding patterns in attrition rates and job satisfaction can guide the development of targeted retention strategies and employee engagement initiatives. Furthermore, the data can be instrumental in workforce planning, helping to anticipate staffing needs, identify skill gaps, and tailor recruitment efforts. Overall, this dataset serves as a critical tool for the HR department, enabling data-driven decisions that can foster a more productive, satisfied, and equitable workplace environment.



## References

HR Dataset: <https://www.kaggle.com/datasets/koluit/human-resource-data-set-the-company>  
[https://docs.google.com/spreadsheets/d/1GI7UkgkXZW9kKejwrIRxKHRPB3oi05CK/edit?usp=drive\\_link&ouid=116622864834462641553&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1GI7UkgkXZW9kKejwrIRxKHRPB3oi05CK/edit?usp=drive_link&ouid=116622864834462641553&rtpof=true&sd=true)

Literature:

Chong, Jason. “*Build Your First Interactive Dashboard from Scratch Using Power BI.*” Medium, 10 May 2022, [towardsdatascience.com/building-your-first-interactive-dashboard-from-scratch-using-power-bi-af7a3e0203d4](https://towardsdatascience.com/building-your-first-interactive-dashboard-from-scratch-using-power-bi-af7a3e0203d4).

Diez, David, et al. *OpenIntro Statistics*. 4th Edition ed., CreateSpace, 1 May 2019.

*Filtering with “Loc” and “Iloc” Methods - Mastering Data Analysis with Python Pandas.*” Educative, [www.educative.io/courses/mastering-data-analysis-python-pandas/filtering-with-loc-and-iloc-methods](https://www.educative.io/courses/mastering-data-analysis-python-pandas/filtering-with-loc-and-iloc-methods). 10 Dec. 2022.

Frost, Jim. “When Can I Use One-Tailed Hypothesis Tests?” *Statistics by Jim*, 12 Nov. 2018, [statisticsbyjim.com/hypothesis-testing/use-one-tailed-tests/](https://statisticsbyjim.com/hypothesis-testing/use-one-tailed-tests/).

“Pandas.DataFrame.boxplot — Pandas 1.3.4 Documentation.” *Pandas.pydata.org*, [pandas.pydata.org/docs/reference/api/pandas.DataFrame.boxplot.html](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.boxplot.html).