Core C Team 9
Professor Francis Pereira
DSO 545: Statistical Computing and Data Visualization
December 9, 2024

What Factors Drive Job Satisfaction?

Introduction

The business question driving this analysis is: *What are the factors that drive job satisfaction?* Understanding job satisfaction is vital for businesses as it directly influences employee engagement, productivity, and retention, all of which are critical for organizational success. When employees are satisfied with their jobs, they are more likely to perform at their best, remain committed to the organization, and foster a positive work culture. This, in turn, reduces turnover costs, enhances team cohesion, and contributes to the long-term sustainability and profitability of the company.

The dataset analyzed in this report contains detailed employee information, including demographic, economic, and workplace-related indicators. These indicators were carefully selected as they theoretically correlate with job satisfaction. For example, factors such as monthly income are linked to financial security, while work-life balance contributes to overall well-being. Other variables like company size, remote work, and distance from home offer insights into how structural and personal contexts shape employee satisfaction. By visualizing and analyzing these indicators, the report aims to identify patterns and provide actionable insights for businesses seeking to create thriving work environments.

This technical report focuses on answering the research question through robust data visualizations and statistical analyses, aligning theory with practical implications. These insights can help organizations prioritize factors that matter most to employees, ultimately driving success and sustainability in a competitive business landscape.

Dataset Overview

Before we dive into deep analysis, we would like to use our initial instincts to create possible assumptions to assist us in discovering the most appropriate, logical, comprehensive, analytical dataset. Through brainstorming, our initial assumptions about the factors that would affect job satisfaction include but are not limited to Salary (Monthly Income), Distance to Work, Company Size, Work-life Balance, Remote Work, Job Role, Job Level, Company Culture, and so on. Among all these possible factors, we do not expect perfect correlations between job satisfaction, but we are expecting some possible strong relationships with it.

After we had settled our direction, we jumped into various popular database sources like Kaggle to search for related datasets. Through multiple rounds of selections, the final dataset we chose is the following *'Employee Attrition Classification Dataset'* on Kaggle, and the link is https://www.kaggle.com/datasets/stealthtechnologies/employee-attrition-dataset/data?select=train.csv. Below is the catalog of the columns that this dataset contains:

```
Catalogue of DataFrame Columns:
--------------------------------------------------
1   Age                    13  Marital Status
2   Gender                 14  Number of Dependents
3   Years at Company       15  Job Level
4   Job Role               16  Company Size
5   Monthly Income         17  Company Tenure
6   Work-Life Balance      18  Remote Work
7   Job Satisfaction       19  Leadership Opportunities
8   Performance Rating     20  Innovation Opportunities
9   Number of Promotions   21  Company Reputation
10  Overtime               22  Employee Recognition
11  Distance from Home     23  Attrition
12  Education Level
```

There are four major reasons why we believe this is the best dataset. First and foremost, we wish to utilize this dataset to help business organizations drive more successful and sustainable company cultures; therefore, we better step into the angle of the companies themselves to conduct the analysis, which exactly fits the original purpose and sources of information of this dataset. Secondly, we want a dataset that is diverse enough to represent the population. For example, within the Education Level variable, this dataset includes degrees from high school until PhD and at the same time maintains significant amounts of counts for each categorical level. Thirdly, the dataset has high quality and a good quantity of columns, which means that all the data are intuitive, closely related to the topic, and multi-dimensional. Last but not least, this dataset contains by far the largest observations and it contains 59,598 distinct employees' data. The more data points we have, the more convinced our analysis will be.

The key clarification for this dataset is that this is an artificial simulated dataset created by the author based on real-life data. The only reason for getting data through simulation is because HR-related analysis always faces legal issues and privacy concerns; therefore, the simulated dataset will avoid the trouble and will give us the pathway for conducting the analysis.

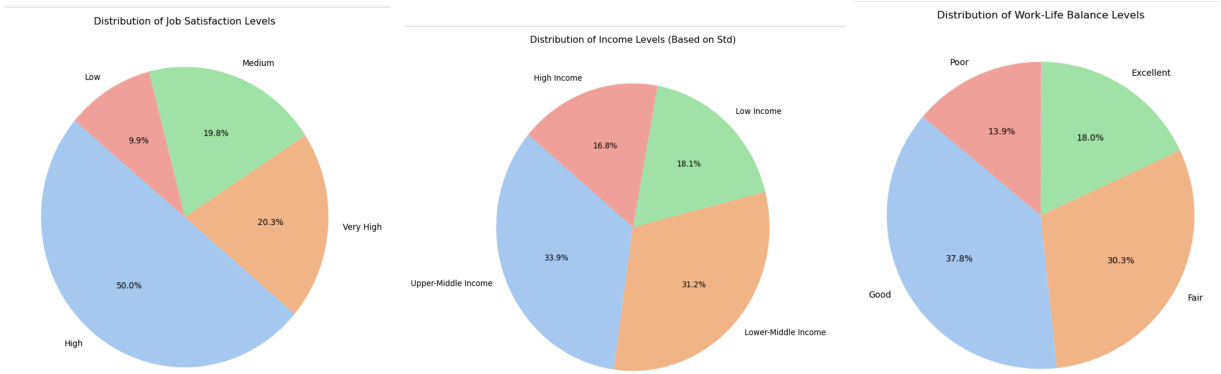
The main usage of this data will be trying to investigate the possible factors or combinations of factors that affect employees' job satisfaction levels. To make every factor valuable and useful, we have done data cleaning beforehand. The first and basic move is to detect if there are any null values and drop them off. For categorical variables, we use dummy variables to make them significant. For example, the Remote Work variable contains values 'Yes' and 'No'. We change it to 1 (Yes) and 0 (No) to make the data analyzable. For certain numerical variables, we wish to make them meaningful and distinguishable; therefore, we use bins to separate them into several sections for further analysis. For example, under the Age column, the dataset includes employees from 17 years old to 59 years old. Each single year of age is not meaningful but age groups are another story. We created four bins 17-27, 28-38, 39-48,

49-59 and labeled them as ‘Young Adult’, ‘Adult’, ‘Middle-Aged’, and ‘Senior’. We believed that utilizing age groups would provide us with more visible analysis results.

The final issue about this dataset is that there are still limitations and space for improvement. First and foremost, this is a simulated dataset and it is hard to check the original resources from the author. If in the future we have a chance to self-collect data and create a more refined dataset will be better. Moreover, the dataset only focuses on the US job market; however, intuitively demographic factors should be important and valuable to investigate.

Exploratory Data Analysis

The first step of EDA is to check the categorical variable's unique values and distribution. Using pie charts to generate the distribution of different groups and observe certain dominant groups for further analysis.



(Pie charts of three categorical variables)

From the above graphs, half of employees report high job satisfaction, highlighting overall positive sentiment while leaving room for improvement among others. For the work-life balance and income groups charts, we observe a similar distribution which suggests a potential relationship. Further analysis of their interaction effect on Job Satisfaction could uncover key insights.

Next, we use shape() and describe() to get each variable's mean, median, standard deviation, and IQR. This helps us better understand which factors we are going to investigate in the next step.

None

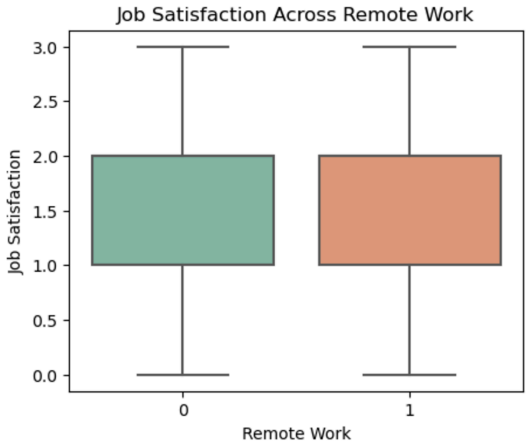
	Employee ID	Age	Years at Company	Monthly Income
count	59598.000000	59598.000000	59598.000000	59598.000000
mean	37227.118729	38.565875	15.753901	7302.397983
std	21519.150028	12.079673	11.245981	2151.457423
min	1.000000	18.000000	1.000000	1316.000000
25%	18580.250000	28.000000	7.000000	5658.000000
50%	37209.500000	39.000000	13.000000	7354.000000
75%	55876.750000	49.000000	23.000000	8880.000000
max	74498.000000	59.000000	51.000000	16149.000000

	Number of Promotions	Distance from Home	Number of Dependents
count	59598.000000	59598.000000	59598.000000
mean	0.832578	50.007651	1.648075
std	0.994991	28.466459	1.555689
min	0.000000	1.000000	0.000000
25%	0.000000	25.000000	0.000000
50%	1.000000	50.000000	1.000000
75%	2.000000	75.000000	3.000000
max	4.000000	99.000000	6.000000

	Company Tenure
count	59598.000000
mean	55.758415
std	25.411090
min	2.000000
25%	36.000000
50%	56.000000
75%	76.000000
max	128.000000

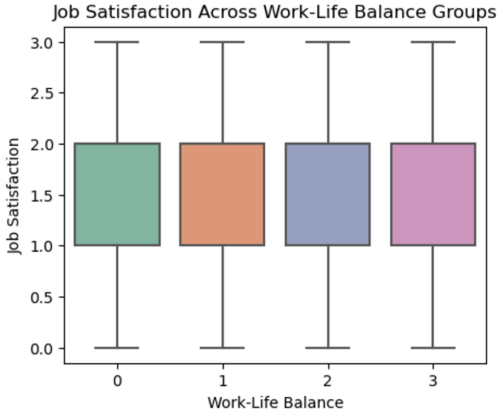
(descriptive summary of numerical variables)

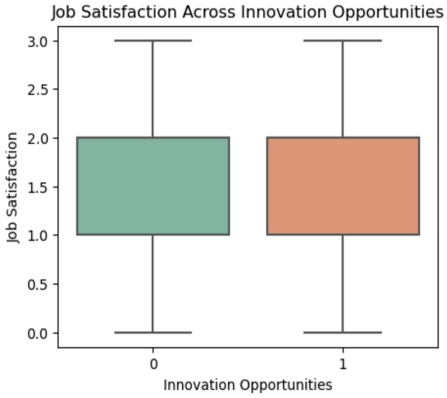
In order to study the company size, remote work, Work-Life Balances, Income levels, and innovation opportunities' effect on job satisfaction, we first group by each variable. For instance, variables like "Work-Life Balance" might include categories such as Poor, Fair, Good, and Excellent, while "Remote Work" could distinguish between Yes and No groups. We want to examine whether different groups have the same result on job satisfaction. By applying a boxplot, we can see the distribution of job satisfaction across different remote work status groups.

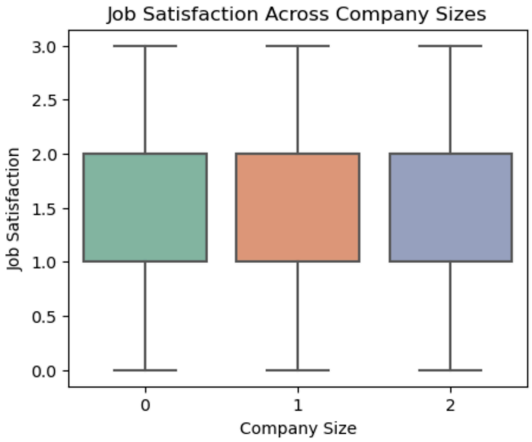


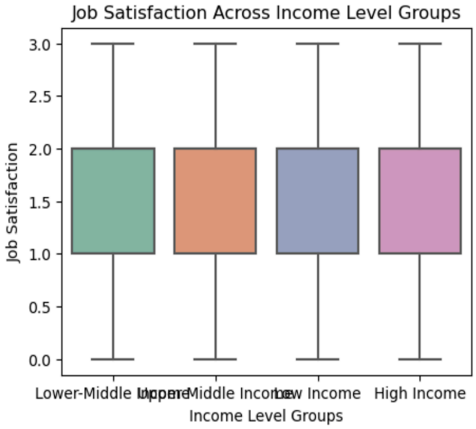
From the graph, we can see that job satisfaction across different remote work statuses is the same. This result differentiates from our initial assumption, as we might have anticipated a difference in job satisfaction between employees who work remotely and those who do not. However, this outcome can be explained by the characteristics of our dataset. It is possible that the dataset is highly assimilated, meaning that there may not be a significant variance between the two groups regarding other factors that influence job satisfaction. The same situation happens

in other variables like monthly income, work-life balance, and company size. Here are boxplots of these factors.

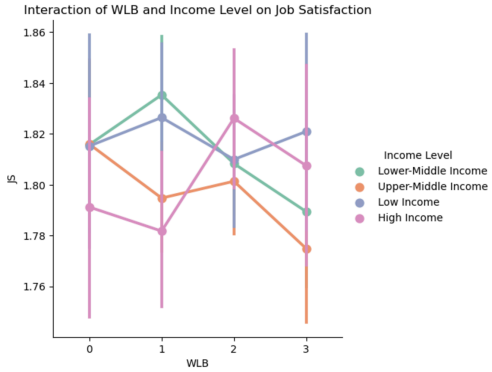






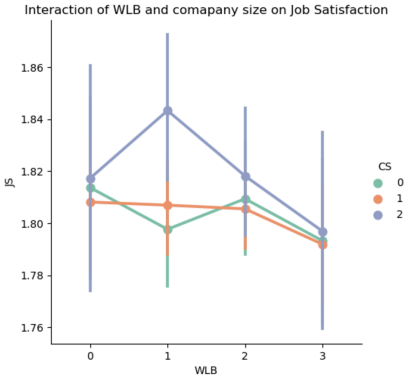


To further analyze the interaction term between two variables on job satisfaction, we utilize a point plot to explain the effect of interaction terms. With a point plot, we can easily identify the difference across different income level groups.

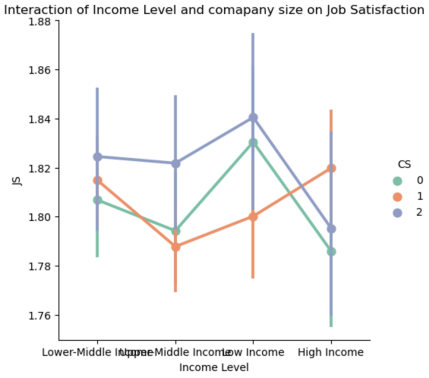


The interaction suggests that the relationship between work-life balance and JS is influenced by income level. Some income groups (e.g., high-income individuals) are more

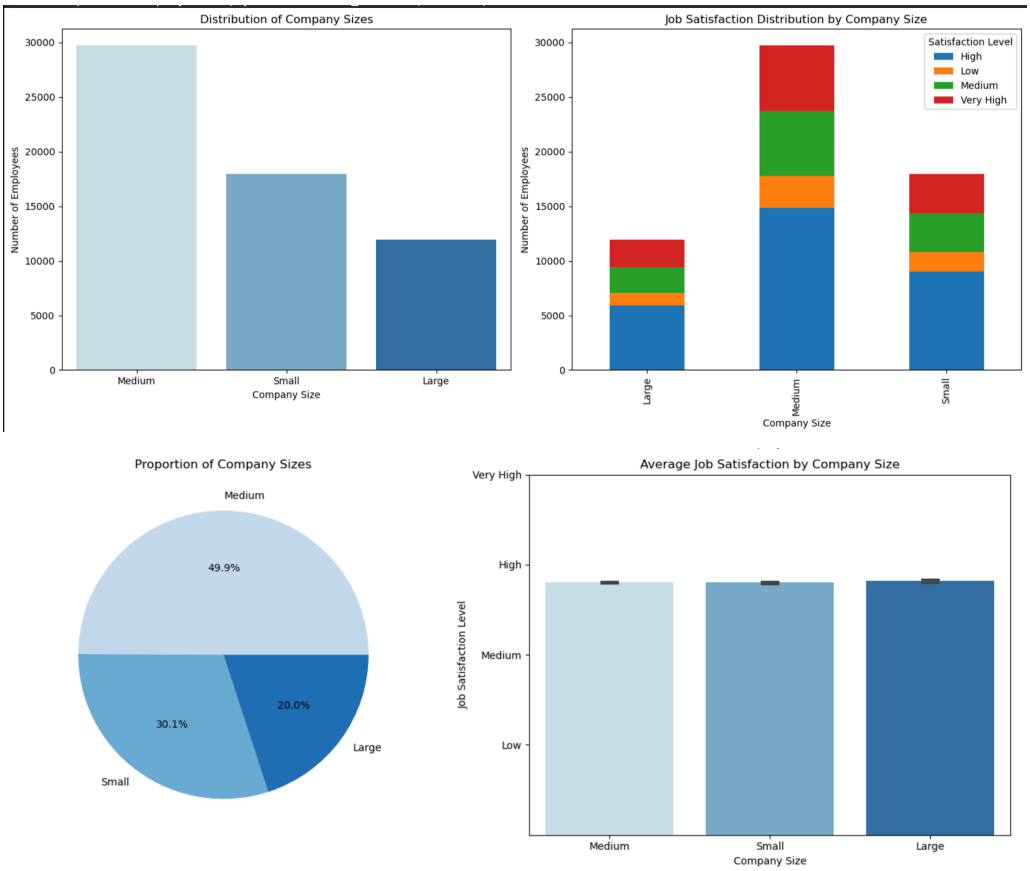
sensitive to changes in work-life balance, which could show improving workplace policies may improve job satisfaction.



Large companies (CS = 2) show the most sensitivity to changes in WLB, with job satisfaction peaking at WLB = 1 but declining sharply afterward. Small and medium companies (CS = 0, CS = 1) have more consistent trends in job satisfaction across WLB levels. This result surprisingly reflects that employees from large companies with higher work-life balance actually have lower satisfaction compared to those who have lower work-life balance.

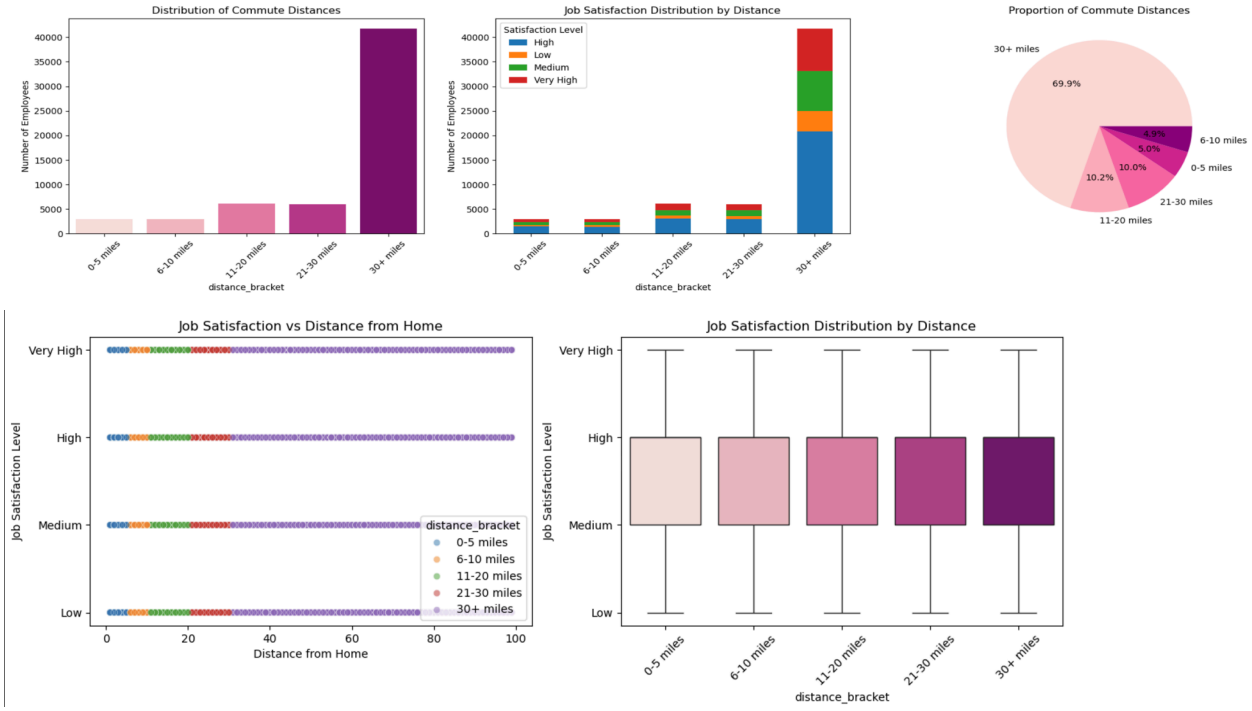


Lastly, we want to study the interaction between income level and company size. The graph shows both small companies (CS=0) and large companies (CS=2) have a significant drop in job satisfaction while medium companies (CS=1) have relatively increased satisfaction. Employees in large companies (CS = 2) report the highest variability in job satisfaction, especially at the High Income level.



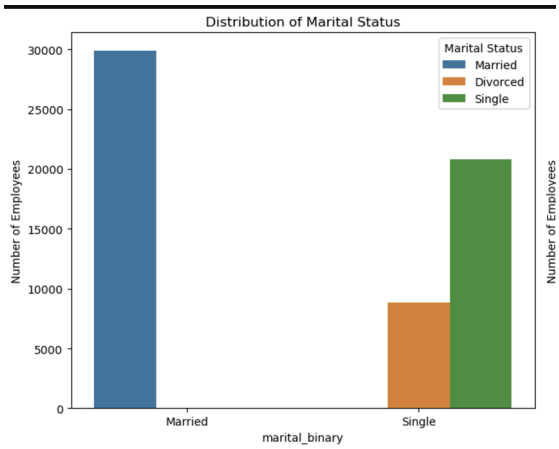
In this analysis of job satisfaction across different company sizes, we created a comprehensive visualization using four distinct plots. The count plot and pie chart reveal the distribution of employees across company sizes, showing that medium-sized companies employ the largest portion of the workforce. The stacked bar chart breaks down job satisfaction levels within each company size, allowing us to see the proportion of employees reporting different satisfaction levels (Low, Medium, High, Very High) in each category. Finally, the bar plot with error bars displays the average job satisfaction levels with 95% confidence intervals, enabling statistical comparison between company sizes.

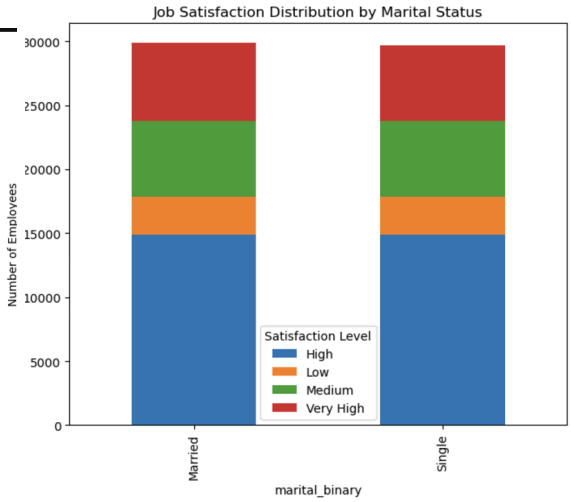
To ensure accurate analysis, we converted categorical satisfaction ratings to numeric values (1-4) and calculated both raw counts and percentage distributions. The statistical analysis shows minimal variation in satisfaction across company sizes, which means hovering around 2.8 (between Medium and High satisfaction) for all categories, suggesting that company size has little impact on overall employee satisfaction levels.

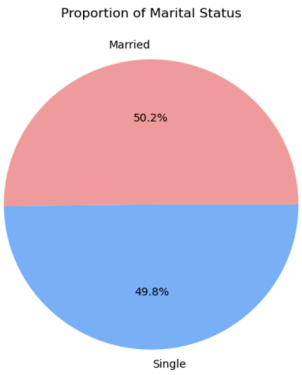


In this analysis of commute distance and job satisfaction, we developed a comprehensive visualization featuring five different plots to explore the relationship. First, we categorized commute distances into five brackets (0-5, 6-10, 11-20, 21-30, and 30+ miles) for clearer analysis. The count plot and pie chart reveal the distribution of employees across these distance brackets, with a notable finding that a significant majority of employees commute more than 30 miles. The stacked bar chart breaks down job satisfaction levels within each distance bracket, while the scatter plot shows the raw relationship between exact commute distances and satisfaction levels. The box plot provides statistical distributions of satisfaction across distance groups, showing consistent medians and similar spreads.

We converted categorical satisfaction ratings to numeric values (1-4) and calculated both raw counts and percentage distributions. Interestingly, despite the varying commute distances, the analysis reveals remarkably consistent satisfaction levels across all distance brackets, with minimal variation in means (around 2.8) and medians, suggesting that commute distance has surprisingly little impact on overall job satisfaction.







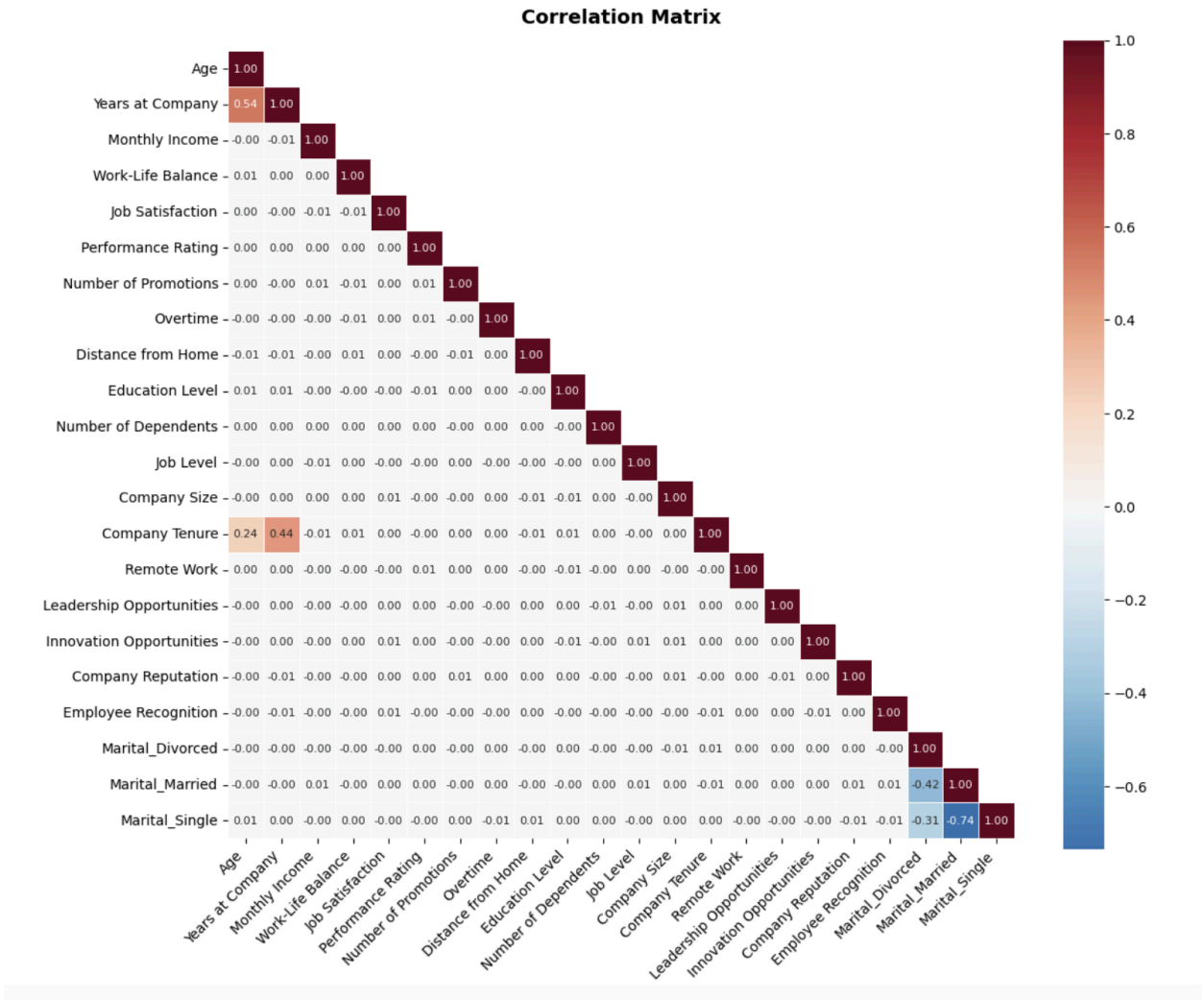
In this analysis of marital status and job satisfaction, we created three distinct visualizations to explore the relationship. The first count plot provides a detailed breakdown of the distribution of employees across marital statuses, distinguishing between married, single, and divorced employees. For deeper analysis, we simplified these categories into a binary classification (married vs. single/divorced combined). The stacked bar chart illustrates the distribution of job satisfaction levels within each marital status category, while the pie chart shows the overall proportion of married versus single employees in the workforce.

We converted categorical satisfaction ratings to numeric values (1-4) and calculated both raw counts and percentage distributions. The analysis reveals an almost equal split between married and single employees (approximately 50% each), with remarkably similar satisfaction distributions between the groups. Both categories show comparable means (around 2.8) and similar patterns across satisfaction levels, indicating that marital status has minimal influence on job satisfaction levels in our workforce.

Methodologies

To further analyze the variables we have and to pick out the best set of variables, we decided to run 3 tests: Correlation analysis, Linear regression, and Random Forest.

First, we wanted to examine the relationship between each variable and our target variable, Job Satisfaction. The purpose of conducting correlation analysis is to determine the strength and direction of the association between each independent variable and the target variable. By identifying these relationships, we can gain insights into which factors may have a significant influence on job satisfaction and prioritize them for further analysis or action.



Here are the top correlated features (absolute value of correlation):

```
Top features, Correlation:

Innovation Opportunities    0.009090
Monthly Income              0.007708
Company Size                0.006573
Work-Life Balance           0.005977
Employee Recognition        0.005952
Number of Dependents        0.004815
Marital_Married             0.004621
Company Reputation          0.004451
Marital_Single              0.004436
Distance from Home          0.004039
```

Next, we performed linear regression on each independent variable separately to analyze its direct relationship with the target variable, Job Satisfaction.

For this analysis:

Standard Scaling: Each independent variable was standard-scaled to have a mean of 0 and a standard deviation of 1. This ensures that the coefficients are directly comparable, as they represent the effect of a one-standard-deviation change in the independent variable on job satisfaction.

Set up: Each independent variable was treated as the sole predictor (independent variable) in a separate regression model. The target variable, job satisfaction, was used as the dependent variable in all models.

The ultimate objective of this regression model was to examine the coefficients of each variable and determine its direct influence on job satisfaction.

Here are the top coefficient features (absolute value of coefficient):

```
Top features: abs(coefficient)

Innovation Opportunities    0.0079
Monthly Income              0.0067
Company Size                0.0057
Work-Life Balance           0.0052
Employee Recognition        0.0052
Number of Dependents        0.0042
Marital_Married             0.0040
Marital_Single              0.0039
Company Reputation          0.0039
Distance from Home          0.0035
```

We selected Random Forest as one of our last testing models because it excels in capturing complex, non-linear relationships and provides insights into feature importance, which helps identify key drivers for decision-making. The result indicates that **Monthly Income**, **Company Tenure**, and **Distance from Home** are the most influential factors, with importance scores of 0.170747, 0.130969, and 0.128594, respectively. These variables significantly contribute to the model's ability to predict outcomes, suggesting they have a strong relationship with the target variable.

```
Monthly Income              0.170747
Company Tenure              0.130969
Distance from Home          0.128594
Age                         0.105749
Years at Company            0.102125
Number of Dependents        0.048446
Education Level             0.043842
Work-Life Balance           0.035820
Number of Promotions        0.034582
Performance Rating          0.032584
Employee Recognition        0.031373
Company Reputation          0.031163
Company Size                0.027245
Job Level                   0.026529
Overtime                    0.015819
Remote Work                 0.014611
Innovation Opportunities    0.012222
Leadership Opportunities    0.007579
dtype: float64
```

Given the set of correlations, coefficients, and importances and our initial hypothesis, we have narrowed down into these 5 features:

1. Monthly Income
2. Work-Life Balance
3. Distance From Home
4. Company Size
5. Marital Status

Statistical Testing

Two Sample Testing:

Note: We considered variance is similar:
- when the ratio of the larger variance to the smaller variance is less than 2:1
- When the ratio of the larger sample size to the smaller sample size is less than 1.5:1

Monthly Income

Purpose: To see if different income groups have different mean Job Satisfactions:

Null H_0: No difference in the mean job satisfaction scores between the "high monthly income" group and the "low monthly income" group.

Alternative H_a: There is a difference in the mean job satisfaction scores between the "high monthly income" group and the "low monthly income" group.

T-statistic: -2.081

P-value: 0.037

The difference in mean Job Satisfaction between the two Monthly Income groups is statistically significant.

Distance from Home

Purpose: To see if different distance from home groups have mean different Job Satisfactions
Null H_0: No difference in the mean job satisfaction scores between the "long distance from home" group and the "short distance from home" group.
Alternative H_a: There is a difference in the mean job satisfaction scores between the "long distance from home" group and the "short distance from home" group.
T-statistic: 0.629
P-value: 0.528
The difference in mean Job Satisfaction between the two Distance from Home groups is not statistically significant.

Work-Life Balance

Purpose: To see if different work-life balance groups have mean different Job Satisfactions
Null H_0: No difference in the mean job satisfaction scores between the "good work-life balance" group and the "bad work-life balance" group.
Alternative H_a: There is a difference in the mean job satisfaction scores between the "good work-life balance" group and the "bad work-life balance" group.
T-statistic: -0.9976
P-value: 0.3185
The difference in mean Job Satisfaction between the two work-life balance groups is not statistically significant.

Company Size

Purpose: To see if different company size groups have mean different Job Satisfactions
Null H_0: No difference in the mean job satisfaction scores between the "small company size" group and the "large company size" group.
Alternative H_a: There is a difference in the mean job satisfaction scores between the "small company size" group and the "large company size" group.
T-statistic: 0.7078
P-value: 0.4790
The difference in mean Job Satisfaction between the two company size groups is not statistically significant.

Marital Status

Purpose: To see if different Marital: Married groups have mean different Job Satisfactions
Null H_0: No difference in the mean job satisfaction scores between the "Married" group and the "Not Married" group.
Alternative H_a: There is a difference in the mean job satisfaction scores between the "Married" group and the "Not Married" group.

T-statistic: 1.1280
P-value: 0.2593
The difference in mean Job Satisfaction between the different marital groups is not statistically significant.

Chi-Square:

Since categorical data often lack normal distribution or meaningful measures like means and standard deviations. The chi-square test is a non-parametric test, making it ideal for analyzing relationships in such data.

Work-Life Balance

Chi-Square value: 11.4453
P-value: 0.2464
The p-value is greater than 0.05, indicating that the relationship between work-life balance and the dependent variable is not statistically significant.

Company Size

Chi-Square value: 10.2524
P-value: 0.1144
The p-value is greater than 0.05, suggesting that the association between company size and the dependent variable is also not statistically significant.

Marital: Married

Chi-Square value: 3.4282
P-value: 0.3302
Similarly, the p-value exceeds 0.05, meaning that marital status (whether married or not) is not significantly associated with the dependent variable

		Work-Life Balance	Company Size	Marital: Married
Two Sample Test	**Test Statistic**	**-0.9976**	**0.7078**	**1.1280**
	p-value	**0.3185**	**0.4790**	**0.2593**
Chi-Square	**Chi-Square**	**11.4453**	**10.2524**	**3.4282**
	p-value	**0.2464**	**0.1144**	**0.3302**
	Result	**Not Statistically Significant**		

Further Analysis

We only found `Monthly Income` to be significant for Two Sample Tests. So we decided to filter the data and perform the tests again.
We will break it into 4 parts.

1.`Company Size`

- We filtered the data and created three different groups:
  - Small Company, Medium Company, Large Company
- Results:
  - **None of the features were statistically significant**

2. `Years at Company`

- We filtered the data and created three different groups:
  - Low year, Mid year, High year
- Results:
  - `Monthly Income` for Mid Year group is **statistically significant** - Two Sample
  - `Work-Life Balance` for high year group is **statistically significant** - Chi Square
  - `Company Size` for higher year group is **statistically significant** - Chi Square

3. `Remote Work`

- We filtered the data and created two different groups:
  - Remote, Non-remote
- Results:
  - `Company Size` for Remote workers is **statistically significant** - Chi Square

4. `Age Group`

- We filtered the data and created four different groups:
  - Young_adult, Adult, Middle_age, Senior
- Results:
  - **None of the features were statistically significant**

Findings and Further Improvement

With the application of statistical tools, we derived from the unsegmented two-sample test that there is a significant difference between the average job satisfaction level between those of different income levels.

Upon dividing the dataset into groups, different features were tested for each group to find associations. After applying two-sample and chi-squared tests to the segmented data, four associations were found: By grouping employees into remote and non-remote groups, chi-squared testing showed that there is an association between company size and remote work, as company size is significant for the group that works remotely. When data is segmented by

tenure, Monthly Income for the mid-year group is significant under the two-sample test. Work-life balance and Company size has an association with Job Satisfaction for the higher-year group under the chi-squared test. However, one should note that the association between factors in the segmented datasets does not necessarily indicate causation. Further testing, such as Cramér's V test, can be done to undiscover the strength of the associations.

In a practical sense, these associations can be focus points in HR performance analysis. Deducting the exact interpretation of these associations requires domain expertise and guidance, which can help HR departments discover whether improving areas such as employee work-life balance and income level can positively contribute to job satisfaction for employees at their organization.

Additionally, the limitations of the dataset call for further studies. The dataset's simulated nature might have an impact on the feature's effect on job satisfaction and may reduce the generalizability of the findings to real-world contexts, where complexities like organizational culture or interpersonal dynamics play a more prominent role. Some results might reflect inherent biases in the data rather than true population-wide trends.