uOttawa

Finding the Keys to Job Satisfaction

SDS3386 - Group Project

Elli Traboulsi #300175764

Coralie Ostertag #300174530

Tanner Giddings #300172545

Tamara Micic #300163312

Prof. Tanya Schmah

December 17, 2022

Table of Contents

Introduction

This report presents the results of a study on job satisfaction among employees at IBM (International Business Machines), a global information technology company as well as a fictional Employee Satisfaction Index (ESI) dataset. Our study was conducted to find out which factors most influence job satisfaction to guide job hunters in their job search phase. The analysis was based on survey responses from a representative sample of employees across various departments and job roles. The findings of this report will be valuable for job seekers who are considering different career paths and want to make an informed decision about their potential job. By examining factors such as salary, overtime hours worked, career advancement and the sex of the employee, this report aims to provide a comprehensive overview of the current state of job satisfaction across different industries and job roles. By understanding these trends, job seekers can better tailor their job search to find positions that are most likely to provide them with satisfaction and fulfillment.

Previous Studies

Many studies have already been conducted and have shown a wide variety of results. We read many articles and found a few very interesting ones. In 2022, *Tracking Happiness* conducted a large-scale study on 12,455 employees spread across the world and found that the ability to work from home increases employees' work happiness rating. Another study, posted by the *Harvard Business Review*, used the survey results of 2,285 Americans from 26 different industries to conclude that 9 out of 10 workers are willing to cut back their salary if it means they get to do more meaningful work. A third study, performed by Barbara Sypniewska at the University of Finance and Management in Warsaw, Poland, highlighted that the most important factors influencing job satisfaction are the atmosphere at work, stability of employment, good relations with co-workers and good relations with superiors. The least important factors are the content of work and the possibility of development. Keeping all of this in mind, we started analyzing our datasets to come up with our own conclusions for our main question: What are the main factors that influence job satisfaction?

Methodology

To be successful in our analysis, we first cleaned the data and separated the qualitative and quantitative data. Our second step was to perform linear and multiple regression on the quantitative data to evaluate the relationships between different variables, calculate the R-squared values and find strong correlations between variables. We continued by creating correlation matrices and contingency tables. Finally, we tried to do principal component analysis (PCA) and TSNE to identify patterns or clusters or even outliers in our data.

Exploratory Data Analysis

IBM Dataset

This study was conducted with an IBM Watson dataset. It was made available to us on two different platforms: Kaggle and GitHub. There is unfortunately no information on the data collection method or why and when this information was gathered. This dataset has 30 variables and 19,478 rows. Variables include relative monthly income, age, sex, monthly income, job satisfaction, department and more variables that consider environmental factors, relationship factors and time factors.

In order to analyze the data collected from IBM, we first visualized all the different variables in order to get a better understanding of the relationships at play in this data. In order to simplify this, we ordered the graphs by the R-squared value of the variable. At first, we found some relationships between job satisfaction and variables that went against our preconceptions. We decided to look more deeply into these variables since they could give us unexpected insights into job satisfaction at IBM.

The most notable variable we found went against our preconceptions was monthly income, this was also the variable with the highest R-squared. We noticed a negative correlation between mean monthly income with job satisfaction, where we expected the opposite (see Figure 1).

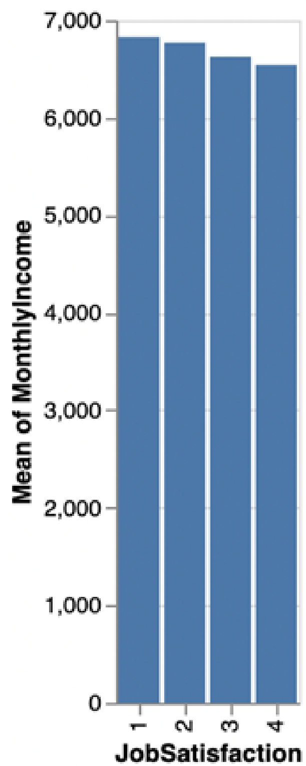

Figure 1 Graph showing the relationship between mean monthly income and job satisfaction.

We also plotted the relationship between environmental satisfaction and the mean job satisfaction and found a negative correlation (see Figure 2). Following this, we did a Kruskal-Wallis test and got a p-value of 9.889702611251235e-05, which is far below our level of significance of 0.05, which indicates that there are at least two means that are different. Following this, we did a Tukey's HSD to test which of the means are different. We found that the means when environment satisfaction is equal to 1 and 2 are the same and the means when environment satisfaction is equal to 3 and 4 are the same, but those two means are different (see Figure 3).

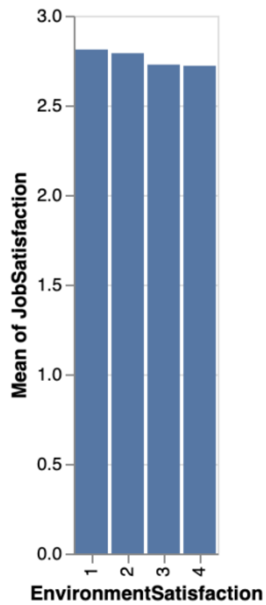

Figure 2 Graph showing the relationship between mean of environment satisfaction and job satisfaction.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
1	2	-0.0204	0.8543	-0.086	0.0452	False
1	3	-0.0843	0.0014	-0.1434	-0.0251	True
1	4	-0.0907	0.0005	-0.1499	-0.0314	True
2	3	-0.0639	0.026	-0.1224	-0.0053	True
2	4	-0.0702	0.0112	-0.1288	-0.0116	True
3	4	-0.0064	0.9888	-0.0577	0.0449	False

Figure 3 Table showing the results of Tukey's HSD test for the environment factor.

Following this, we created a correlation map in order to get an idea of which variables could be combined, removed or adjusted (see Figure 4). By combining, removing or adjusting variables, we would be able to avoid multicollinearity and improve the efficacy and accuracy of our model. On further analysis of the heatmap, we found with this that there were strong correlations between the variables JobLevel and TotalWorkingYears, TotalWorkingYears and MonthlyIncome and that a number of variables were a function of time (see Figure 4: YearsAtCompany, TotalWorkingYears, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager).

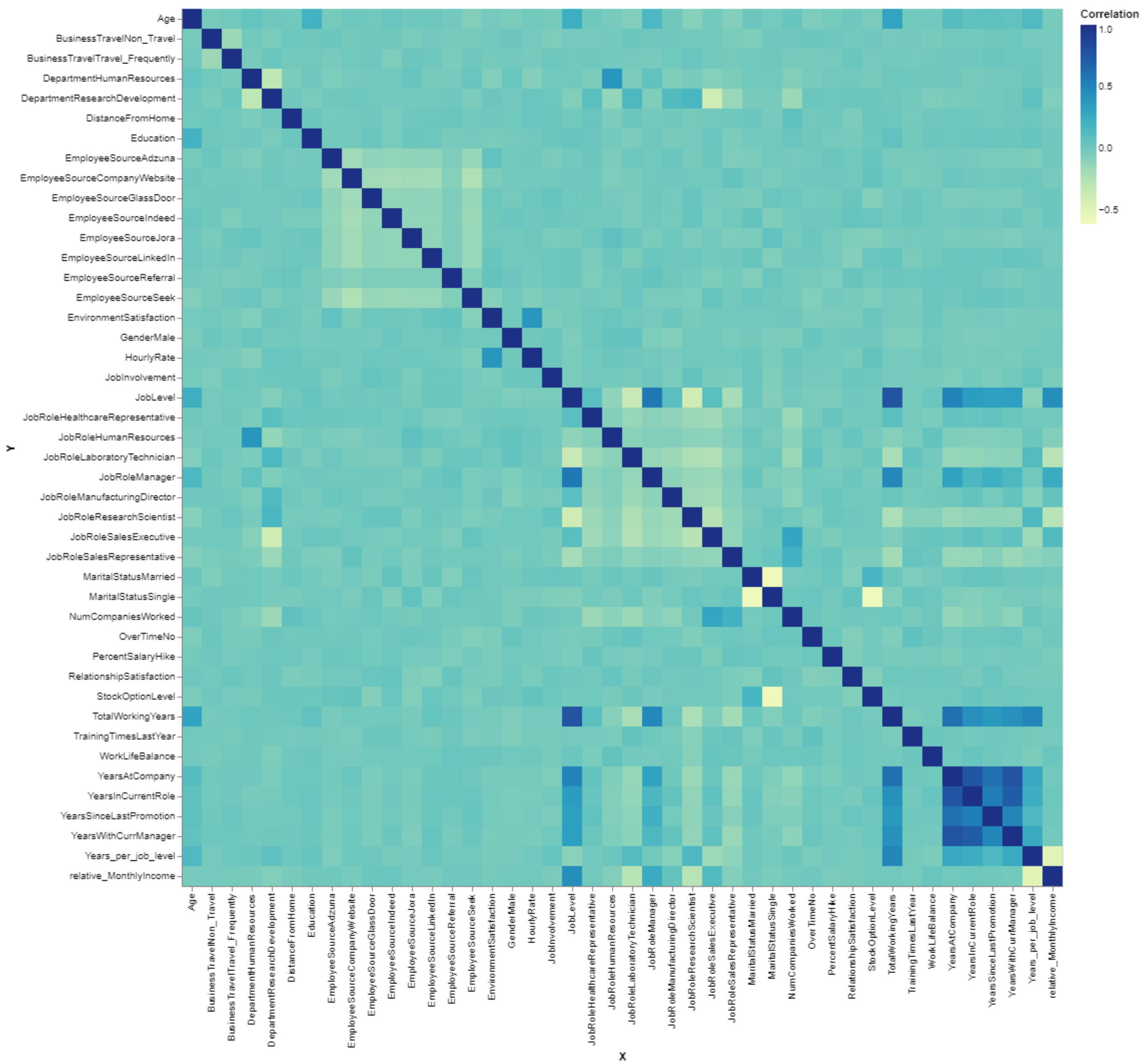

Figure 4 Correlation heatmap showing the correlations between different variables in the IBM dataset.

Following our analysis, we decided to combine TotalYearsWorked and JobLevel into Years_per_job_level, in order to model the speed of a worker's career advancement, MonthlyIncome and TotalWorkingYears, as relative_MonthlyIncome, in order to model pay compared to their coworkers of a similar age, and MonthlyIncome, TotalWorkingYears, Education and JobLevel, as relative_MonthlyIncome_education_and_level, in order to better model a worker's pay compared to their peers. Following this, we graphed the relationships between these new variables and JobSatisfaction (see Figure 5).

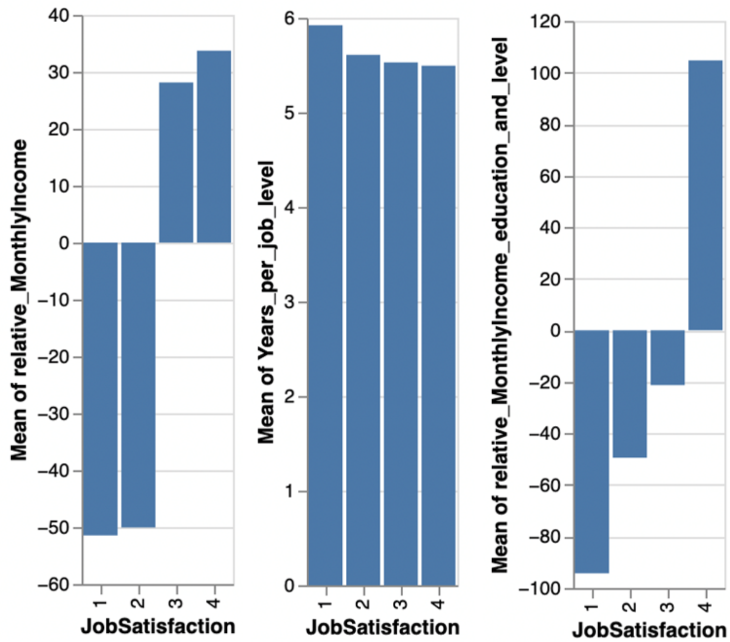

Figure 5 Graphs showing the relationships between JobSatisfaction and mean of relative_MonthlyIncome, Mean of Years_per_job_level and mean of relative_MonthlyIncome_education_and_level.

In the graph plotting JobSatisfaction and Mean of relative_MonthlyIncome, we can see that people who make more than people of their age are more satisfied at work. In the graph plotting JobSatisfaction and Mean of Years_per_job_level, we can see that people who progress in their careers fastest are on average the most satisfied at work. Lastly, in the graph plotting JobSatisfaction in regards to Mean of relative_MonthlyIncome_education_and_level, we can see that workers at IBM who make more than their peers in regards to age, job level and education, are more satisfied than their counterparts who make less in regards to their peers.

Following this, we created a multiple regression model with all the relevant variables (see Figure 6). This is in order to be able to capture all possible relationships with job satisfaction in our model. Following this, we do backward variable selection in order to remove variables which have a high p-value, until all p-values are below our designated significance threshold of 0.05 (see Figure 7). Also, in order to simplify the interpretation of the coefficient in relation to the variable's impact on job satisfaction, the variables were standardized, therefore all the variables are in the interval 0:1. This allows us to directly rank the impact of the variables on job satisfaction by the absolute value of their coefficients.

Variable	**Coefficient**	**Standard error**	**P-value**
DistanceFromHome	0.0066	0.009	0.475
Relative_MonthlyIncome	0.6105	0.027	0.000
Education	-0.0132	0.010	0.197
EnvironmentSatisfaction	-0.0139	0.008	0.081
HourlyRate	-0.0367	0.010	0.000
JobInvolvement	-0.0153	0.011	0.171
MonthlyRate	0.0140	0.009	0.125
NumCompaniesWorked	-0.0124	0.011	0.255
PercentSalaryHike	0.0694	0.010	0.000
RelationshipSatisfaction	-0.0192	0.007	0.009
StockOptionLevel	0.0786	0.012	0.000
TrainingTimesLastYear	0.0159	0.012	0.192
YearsAtCompany	0.1206	0.030	0.000
YearsInCurrentRole	-0.0115	0.021	0.588
YearsSinceLastPromotion	-0.0612	0.016	0.000
Years_per_job_level	0.1402	0.041	0.001
BusinessTravelTravel_Frequently	0.0507	0.007	0.000
BusinessTravelNon_Travel	0.0513	0.008	0.000
DepartmentResearchDevelopment	0.0107	0.007	0.108
DepartmentHumanResources	0.0017	0.015	0.911
GenderMale	0.0275	0.005	0.000
JobRoleResearchScientist	0.1862	0.012	0.000
JobRoleLaboratoryTechnician	0.1103	0.012	0.000
JobRoleSalesRepresentative	0.1601	0.016	0.000
JobRoleSalesExecutive	0.1847	0.012	0.000
JobRoleManager	0.0949	0.014	0.000
JobRoleHumanResources	0.1185	0.019	0.000
JobRoleHealthcareRepresentative	0.1521	0.013	0.000
MaritalStatusMarried	0.0405	0.007	0.000
MaritalStatusSingle	0.0933	0.009	0.000
OverTimeNo	-0.0272	0.006	0.000
EmployeeSourceSeek	0.0463	0.011	0.000
EmployeeSourceIndeed	0.01350	0.011	0.236
EmployeeSourceReferral	0.1155	0.022	0.000
EmployeeSourceCompanyWebsite	0.0335	0.010	0.001
EmployeeSourceAdzuna	0.0666	0.012	0.000
EmployeeSourceGlassDoor	0.0213	0.012	0.073
EmployeeSourceJora	0.0310	0.012	0.008
EmployeeSourceLinkedIn	0.0144	0.012	0.214

Figure 6 Table showing the coefficients, standard errors and p-values of variables in the full multiple regression model.

	coef	std err	t	P>|t|	[0.025	0.975]
PercentSalaryHike	0.0710	0.010	7.057	0.000	0.051	0.091
RelationshipSatisfaction	-0.0181	0.007	-2.485	0.013	-0.032	-0.004
StockOptionLevel	0.0797	0.012	6.535	0.000	0.056	0.104
YearsAtCompany	0.1453	0.024	6.155	0.000	0.099	0.192
YearsSinceLastPromotion	-0.0640	0.015	-4.148	0.000	-0.094	-0.034
relative_MonthlyIncome	0.6272	0.021	30.415	0.000	0.587	0.668
Years_per_job_level	0.1036	0.038	2.750	0.006	0.030	0.177
BusinessTravelTravel_Frequently	0.0510	0.007	7.189	0.000	0.037	0.065
BusinessTravelNon_Travel	0.0504	0.008	5.938	0.000	0.034	0.067
GenderMale	0.0277	0.005	5.155	0.000	0.017	0.038
JobRoleResearchScientist	0.1954	0.011	17.854	0.000	0.174	0.217
JobRoleManufacturingDirector	0.1213	0.012	10.087	0.000	0.098	0.145
JobRoleLaboratoryTechnician	0.1757	0.011	15.661	0.000	0.154	0.198
JobRoleSalesRepresentative	0.1858	0.015	12.331	0.000	0.156	0.215
JobRoleSalesExecutive	0.1484	0.011	13.940	0.000	0.128	0.169
JobRoleManager	0.0949	0.014	6.831	0.000	0.068	0.122
JobRoleHumanResources	0.1203	0.017	7.207	0.000	0.088	0.153
JobRoleHealthcareRepresentative	0.1601	0.012	12.814	0.000	0.136	0.185
MaritalStatusMarried	0.0389	0.007	5.676	0.000	0.025	0.052
MaritalStatusSingle	0.0960	0.009	10.301	0.000	0.078	0.114
OverTimeNo	-0.0276	0.006	-4.577	0.000	-0.039	-0.016

Figure 7 Table showing the coefficients, standard errors, t-values, p-values and confidence intervals for variables in the reduced multiple regression model.

Subsequently, we performed PCA on the entire dataset to try to get more insights on the data. Unfortunately, this analysis method did not give us any new insights as the points seemed to be randomly distributed in the scatterplots without any noticeable clustering or patterns with regard to job satisfaction (see Figure 8 for an example of the PCA scatterplots).

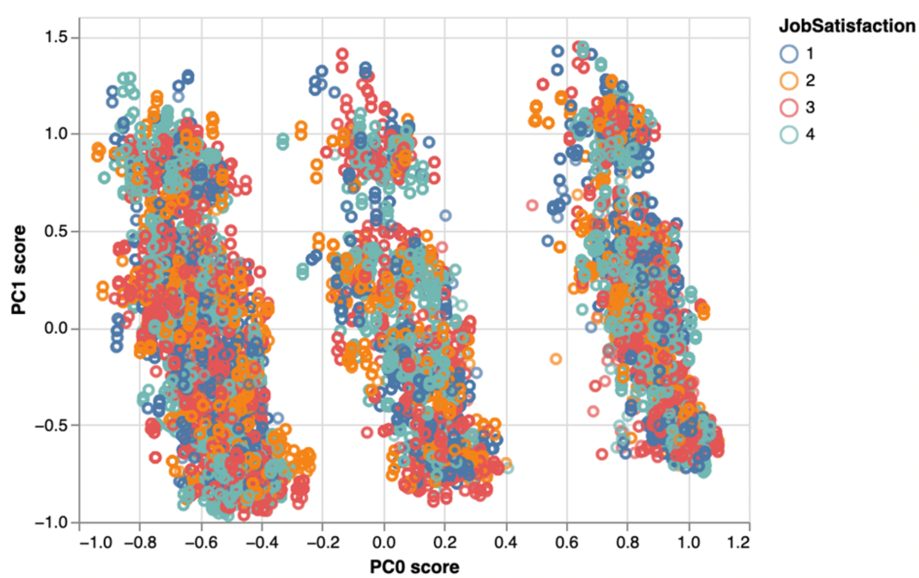

Figure 8 Example of the scatter plot plotting the relationship between the PC0 score and PC1 score for the IBM dataset.

Finally, to try to get more insights from dimensionality reduction, we tried using TSNE to identify possible clusters, patterns or relationships we missed with principal component analysis. However, this was also inconclusive as the points seemed once again randomly dispersed (see Figure 9).

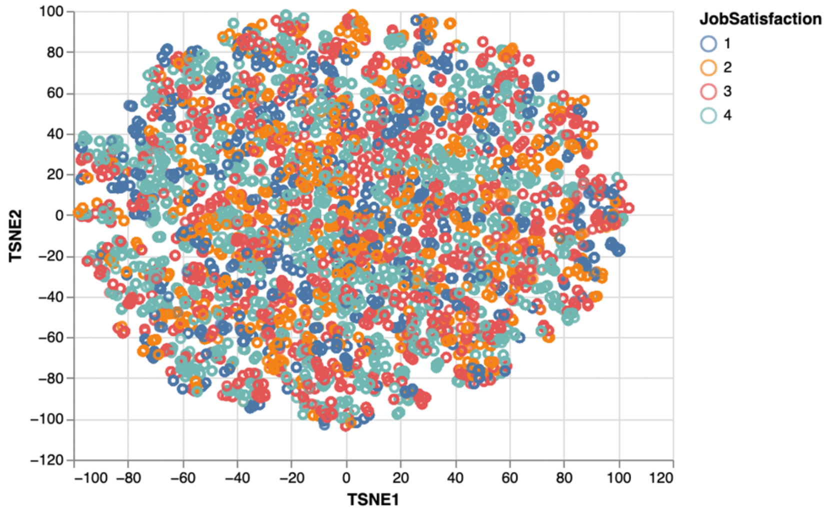

Figure 9 Scatter plot showing the dimensionality reduction using TSNE of the IBM dataset.

Employee Satisfaction Index Dataset

The Employee Satisfaction Index has 500 rows, each corresponding to one employee. It has 14 columns, with attributes describing the employee (ID, age, etc.) and data related to their employment, such as job level, which is measured from 1 to 5, 1 being the lowest position and 5 being the highest position. Another column is 'rating', which is the previous year's rating of the employee, also measured from 1 to 5. The column 'satisfied' contains values 0 or 1, with 0 corresponding to job dissatisfaction, and 1 corresponding to job satisfaction. Unfortunately, we realized in the process of writing this report, that the Employee Satisfaction Index Dataset is fictional. Nevertheless, we believe that the analysis we did on the dataset provides valuable insights into employee satisfaction.

To start, we made contingency tables for 'satisfied' and all the other variables. We will summarize 3 tables with interesting results (see Figure 10). Note that in this dataset, 52% of people report being satisfied, and 48% report being unsatisfied.

satisfied	0	1
Dept		
HR	53	53
Marketing	51	44
Purchasing	47	62
Sales	41	51
Technology	45	53

satisfied	0	1
job_level		
1	41	54
2	57	53
3	41	41
4	53	57
5	45	58

satisfied	0	1
awards		
0	20	32
1	20	21
2	29	19
3	32	25
4	22	29
5	25	27
6	19	27
7	18	32
8	30	26
9	22	25

Figure 10 Contingency tables for 'satisfied' vs 'Dept', 'job_level', and 'awards', from left to right.

Looking at the leftmost contingency table, we see that in Purchasing, Sales, and Technology departments, more people are satisfied than unsatisfied. But in Marketing, more people are unsatisfied, so it seems that working in Marketing might be a factor that contributes to dissatisfaction.

Looking at the middle contingency table, we see that people with job levels 1, 3, 4, or 5 are more satisfied or evenly split. For job level 2, more people are unsatisfied.

Looking at the rightmost contingency table, we see that among people with 2, 3, or 8 awards, more report being dissatisfied. Among people with 0, 1, 4, 5, 6, 7, or 9 awards, more report being unsatisfied.

Next, we created bar graphs to look at the relationship between satisfaction and each of the other variables.

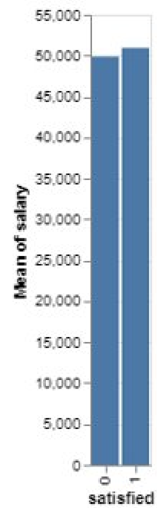

Figure 11 Graph showing the relationships between 'satisfied' and the mean of 'salary'.

In Figure 11 above, we see that employees who report being satisfied have a higher salary on average than employees who report being unsatisfied. However the difference seems very small, so to test the significance, we performed a t-test. The p-value is 3.55e-259, therefore the test rejects the null hypothesis, meaning that there is a statistical difference between the two means.

We did similar graphs for other variables. That is, instead of the mean of 'salary', we compared the mean of 'rating', the mean of 'certifications', and the mean of 'age' between satisfied and dissatisfied employees (see figure 12).

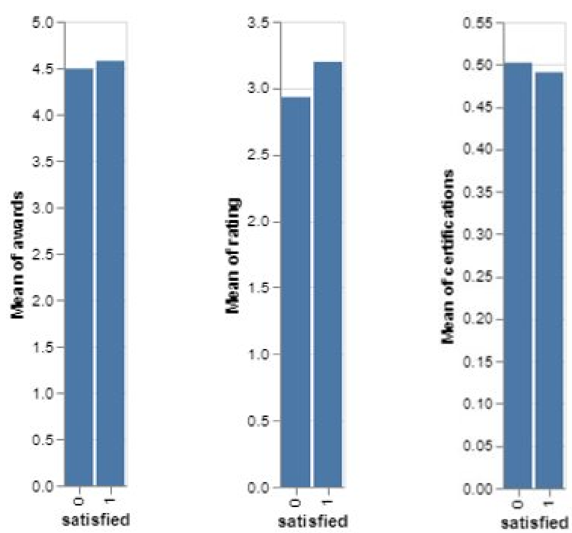

Figure 12 Graph showing the relationships between 'satisfied' and the mean of 'awards', 'rating', and 'certifications' from left to right.

The differences seem small once again. The t-tests show that the differences are significant in 'awards' and 'rating', but not for 'certificates'.

Since we noticed in the IBM data set that relative_MonthlyIncome was an important factor in predicting job satisfaction, we decided to do an analogous analysis for the Employee Satisfaction Index. We looked at 'salary' relative to 'age'. We found that employees who report being satisfied have a higher salary on average than employees who are the same age as them, and employees who report being unsatisfied have a lower salary on average than employees who are the same age as them (see figure 13).

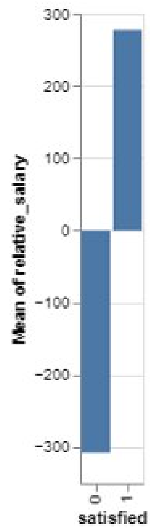

Figure 13 Graph showing the relationships between 'satisfied' and mean of relative_salary.

We also created a correlation map. See figure 14 below.

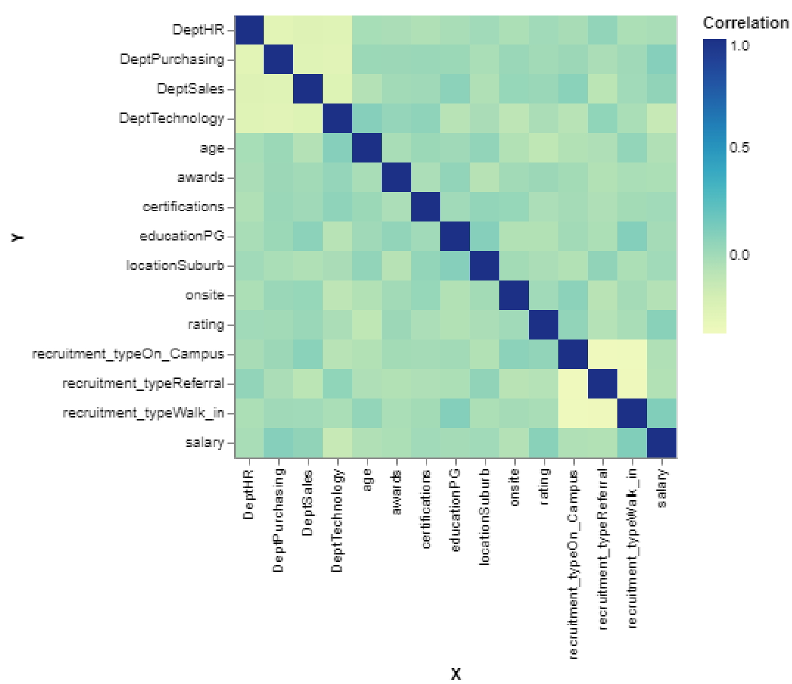

Figure 14 Correlation heatmap showing the correlations between different variables in the Employee Satisfaction Index dataset.

There are no strong correlations between any variables except the dummy variables. Note, however, that this is after the 'job_level' column was removed. We removed it because we realized that 'job_level' and 'salary' are perfectly correlated. The unique salaries are 24076, 29805, 42419, 65715, and 86750. This is a flaw in the dataset, which may be attributed to the fact that it is fictional.

We also tried doing a linear regression for this data set, which gave an R-squared value of -0.17, which is quite alarming because the model is worse at predicting than the mean. So instead, we did a logistic regression, which models data with a binary dependent variable. This is true for this dataset because 'satisfied' values are 0s and 1s. And indeed, using logistic regression, the R squared value is 0.552, which is much better.

We also looked at the residuals (see figure 15 below). We see that the model is wrong 60% of the time at predicting unsatisfied people and 30% wrong at predicting satisfied people. So still, the model is flawed.

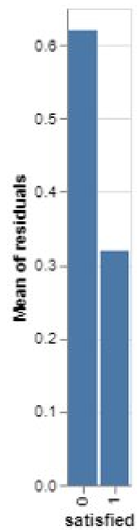

Figure 15 Residuals of the logistic regression.

Discussion and Conclusion

Discussion: Comparing the Class Survey Results to our Results

Closer to the start of our project, we decided it could be interesting to see what factors our classmates thought were the most important. During our project proposal presentation, we asked everyone to rank five factors on a scale of one to five; one being the most important factor and five being the least important factor. We chose our five factors based on some of the existing studies we looked at and based on some of the preparation work we had done. These five factors seemed to encapsulate the main points that seemed to come up in these studies: Salary, Hours Worked per Week, Environment, Company Valuation and Size, and Vacation Time.

After receiving the survey answers, we inputted and organized the data in a jupyter notebook so that we could visualize and analyze it. We then averaged out those scores (by using a formula for each column similar to the following one: =SUM(B2*1, B3*2, B4*3, B5*4, B6*5)/22). We finally got the following results (reminder: the lower the score, the more important the factor):

	Number of Students Who Ranked Each Factor				
Ranking	Salary	HoursPerWeek	Environment	VacationTime	CompanyValuationAndSize
1	9	2	10	0	1
2	8	4	6	3	1
3	2	11	5	2	2
4	1	3	1	12	5
5	2	2	0	5	13
Mean Ranking Score:	*2.045454545*	*2.954545455*	*1.86363636*	*3.863636364*	*4.272727273*

Figure 16 Table showing the mean ranking score of each factor.

Based on these results, we can see that our class finds Environment and Salary are the most important factors, while Company Valuation and Size and Vacation Time are the least important.

Now, let's compare these class survey results to our results. First, let's look at what our datasets say regarding the environment factor. When we described this factor to the class, we said it could include anything from working onsite versus working remotely, to working individually versus working in a group setting. Our Employee Satisfaction Index dataset shows us that people who work onsite are more satisfied with their job. This is very interesting as some other studies we looked at say that remote workers are happier and more productive. However, these studies came out during the COVID-19 pandemic, and our dataset comes from pre-pandemic times. This could indicate that the overall state of the world has an impact on whether people are happier working onsite or are more comfortable working from home. For a more global portrait of this factor, we saw that our IBM dataset suggests that people who are least satisfied with their work environment are the happiest (see Figure 2). So, though our class found that environment is a very important factor, our IBM dataset seems to say otherwise.

Next, let's take a look at what our datasets say regarding salary. We saw in both datasets that there is a link between salary and job satisfaction. However, the link between them is more than just about the amount of money that an employee makes. Both datasets indicate that people who make more money compared to their coworkers who are around their age are more satisfied with their job.

Lastly, let's take a look at the correlation between roles and job satisfaction. Though role was not a factor we had the class rate, we thought it could be interesting to look at this part of the data to see if people with roles similar to the possible future roles of our classmates are satisfied with their jobs! By looking at the IBM dataset's contingency table for satisfaction versus role (see figure 17 below), we can see that overall, people in STEM roles are pretty satisfied with their jobs compared to people with jobs in Human Resources or Sales.

JobSatisfaction	1	2	3	4
JobRole				
Healthcare Representative	348	260	602	575
Human Resources	122	218	172	176
Laboratory Technician	654	659	942	1092
Manager	291	286	381	485
Manufacturing Director	358	468	730	513
Research Director	231	230	383	311
Research Scientist	663	657	1214	1304
Sales Executive	889	702	1111	1517
Sales Representative	120	232	305	277

Figure 17 Contingency table for 'JobSatisfaction' vs 'JobRole'

After looking at all these factors, we decided to additionally look into whether or not being male or female has an effect on job satisfaction. Unfortunately, one of the two datasets we worked with had information on sex. However, based on the IBM dataset (which used a satisfaction scale of 1 to 4), we could see that the average satisfaction rating increases by about 0.03 for male employees. Which provides us with some good insight on the topic. To fully answer this, we'd have to dig deeper and look at more datasets to see if this is a trend that's seen across the board, or only at IBM. We could look into seeing if the satisfaction difference is because of a pay gap, because of role access or other possible reasons.

Ethical Concerns

In data analysis, ethical concerns may arise, whether intentional or not. We will take a closer look at possible ethical concerns in this report. In fact, we will look at privacy concerns and bias.

Let's begin by looking at our two datasets, the IBM dataset and the ESI dataset. Unfortunately, there is no information on the data collection of the IBM dataset so it is difficult to discuss possible privacy concerns without making assumptions. However, it is safe to say that there is bias in our report since we are only analyzing one dataset from a high-tech company and no other real establishments. Perhaps employees working at other companies have different factors of importance or are more satisfied than those at IBM. As for the fictional ESI dataset, there are no privacy concerns since there was no data collection. As for bias, since the data is fictional, there is bias in that alone.

Before our analysis, we collected data from our classmates. While we decided to keep our classmates' answers anonymous by collecting the data on two different sheets of paper, privacy was still a concern. Our classmates were sitting in their "unassigned assigned seats", all using different coloured pens, writing down their answers. This is a concern because, with less than thirty students participating, we may have been able to recognize one's handwriting or ink colour and even figure out who wrote which answer depending on where they were sitting and where that answer lies on our sheet of paper. However, after the initial collection, we transferred the results to a Microsoft excel sheet and discarded the paper copies which minimized those concerns. We can also express concerns about bias in this process. Due to our data collection method, survey respondents may have been influenced by the answers of prior respondents that appear on the same sheet of paper.

In short, our data collection method was flawed. This may lead to bias in our conclusion when evaluating the differences between what the class deemed as the most important factor that influences job satisfaction and the results we obtained from the analysis of our two other datasets. Perhaps changing the data collection method to a Google Forms survey, and making sure the responses are anonymous, would help eliminate these privacy concerns.

Conclusion

To summarize, we saw that the most important factor influencing job satisfaction is "how much money you are making relative to people your age". This was seen in both the IBM and Employee Satisfaction Index datasets. The IBM dataset also showed us that the position or role along with "how long you've been working at a company" are also interesting factors that influence job satisfaction. The Employee Satisfaction Index also showed us that in general, the people who are most satisfied with their jobs have a higher salary.

In the end, we were able to identify the main factors influencing job satisfaction. Along the way, we were also able to find answers to some of our sub-questions.

If we were to continue this research project, we could use our models on more datasets to keep investigating and to keep looking for certain trends. We could also take a look at how the importance of the identified factors has changed over time. It could be interesting to explore the changes that have happened over the past few years, during the COVID-19 pandemic. As we discussed earlier on in the report, onsite workers were more satisfied with their jobs, but this was based on data collected before the pandemic. By finding some more recent datasets to use, we could open the door to an infinite amount of possible sub-questions we could try to answer with further analysis.

References

Tracking Happiness team. "Remote Work is Linked to Happiness: Study of 12,455 Respondents." *Tracking Happiness,* June 20, 2022, www.trackinghappiness.com/remote-work-leads-to-happiness-study/#remote-work-increases-happiness

Achor, Shawn, et al. "9 Out of 10 People Are Willing to Earn Less Money to Do More-Meaningful Work." *Harvard Business Review. November 6, 2018,* hbr.org/2018/11/9-out-of-10-people-are-willing-to-earn-less-money-to-do-more-meaningful-work

Sypniewska, Barbara. "Evaluation of Factors Influencing Job Satisfaction". *Contemporary Economics, Vol. 8, No. 1, pp. 57-72*. March 31, 2014, ssrn.com/abstract=2435040.

Datasets found at:
https://www.kaggle.com/datasets/mohamedharris/employee-satisfaction-index-dataset
https://www.kaggle.com/datasets/itssuru/hr-employee-attrition

Contributions

Coralie took the lead on the preliminary research with support from Elli.
All members played a role in looking for datasets to use for this project (specifically on Kaggle).
Elli took the lead on the introduction with support from Coralie.
Tanner took the lead on modeling, using it on the IBM dataset.
Tamara took the lead in using the model to analyze the Employee Satisfaction Index.
Coralie managed the class survey, compiled the results (table used in our final report) and created the visualizations (used in our presentation). She then used them to compare those results with the results we got from analyzing the datasets.
Elli monitored for ethical concerns.
Coralie summarized the results for our conclusion.
Coralie and Elli took the lead on creating our presentations and took the lead on the final report organization.