**Overview:**

The goal of this project is to analyze correlations of factors in an employee’s career. We collected data from various sources and converted them into a standardized format, and then used various statistical and visualization techniques to explore the relationships between different variables, such as education level, gender, age, and salary. We also performed hypothesis testing to determine the significance of the observed patterns.

**Decompose the Ask:**

Using a data driven approach, we broke down this task into the following steps to create an analysis to figure out the happiness of employees. Based on the data that we collected, we aimed to answer the following questions:

• How does the education level affect employee salary?

• Which recruitment type has the highest retention rate of employees?

• What is the correlation between employee age and salary?

• What is the correlation between education level and salary?

**Identify Data Sources:**

* Kaggle
* Data.gov
* Github
* Data.world
* Tableau

**Retrieving the Data**

Our first step was to find data that we could use to fit the parameters of our project. Any data set that we found would have to have adequate data that was high quality and quantity for our analysis so we could draw a reasonable inference to support a valid conclusion. We researched data sets across a few different platforms and through our research we encountered many data sets were incomplete or irrelevant for our analysis, making for a challenge, but eventually we did manage to come across some data that looked good enough to fit out parameters.

**Assemble and Clean Data**

We needed to use two sets of data (Employee Attrition Data and Salary Data) that came from different sources to draw our conclusions and we needed to merge the data using the overlapping categories from both data sets, so we used age, annual salary, gender, years of experience. Once we merged the data, we were able to start the cleanup process. In the employee Attrition Data set, there were some fields that needed to be modified to fit our merger process. For example, the monthly salary needed to be converted into yearly salary to match the Salary data on the other dataset. Some of the qualitative and quantitative data were scaled1 so created columns with the converted data.

**Analysis of Trends**

A picture containing text, screenshot, diagram, font

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing screenshot, display, text, line

Description automatically generatedA picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, screenshot, diagram, line

Description automatically generatedA picture containing text, screenshot, diagram, font

Description automatically generated

**Acknowledge the Limitations**

Some of the problems that we encountered when we did the analysis from two different sets of data were as following:

- Data inconsistency: The data from the two different sources had different formats, units of measurement and definitions. For example, the monthly salary vs the yearly salary, the

- Data duplication: The data from different sources may have overlapping or redundant records that need to be identified and removed.

- Data quality: These datasets from different sources had varying levels of, completeness, and validity. For example, there were a lot of blanks in the Employee Attrition Data sets. The data sets we used also fictional, so the conclusions we derived from them may not match the reality. The conclusion is specific to these data sets only.

- Data complexity: These datasets had high dimensionality, meaning they have many attributes or variables. For example, Environment Satisfaction, Job Involvement, Job Satisfaction, Performance Rating Relationship Satisfaction and work life balance all have scaled data where the scales are broad and open for interpretation. 1 = Bad ; 2 = Good ; 3 = better ; 4 =Best ‘ 1 = Low ; 2 = Medium ; 3 = Excellent ; 4 = Outstanding

- Data variability: The data from different sources may have different distributions, means, standard deviations, or outliers. This can affect the statistical analysis or comparison of the data³.

To overcome these problems, we did the following:

- Data cleaning: We merge the data using the overlapping categories from both data sets, age, annual salary, gender, years of experience. We converted monthly salary needed to be converted into yearly salary to match the Salary data on the other dataset

- Data transformation: This involves converting the data from different sources into a common format, standard, definition, or unit of measurement. You can use tools such as Excel or SQL to help you with this process.

- Data analysis: This involves organizing, modeling, and interpreting the data to uncover useful insights and draw conclusions. You can use tools such as R or Python to help you with this process.

**Sources:**

<https://www.kaggle.com/datasets/mohithsairamreddy/salary-data>

<https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset>

**Index:**

1.

Education  
1 'Below College'  
2 'College'  
3 'Bachelor'  
4 'Master'  
5 'Doctor'  
EnvironmentSatisfaction  
1 'Low'  
2 'Medium'  
3 'High'  
4 'Very High'  
JobInvolvement  
1 'Low'  
2 'Medium'  
3 'High'  
4 'Very High'  
JobSatisfaction  
1 'Low'  
2 'Medium'  
3 'High'  
4 'Very High'  
PerformanceRating  
1 'Low'  
2 'Good'  
3 'Excellent'  
4 'Outstanding'  
RelationshipSatisfaction  
1 'Low'  
2 'Medium'  
3 'High'  
4 'Very High'  
WorkLifeBalance  
1 'Bad'  
2 'Good'  
3 'Better'  
4 'Best'