**Employee Attrition Rate Analysis**

Data Visualization (INFO-I 590)

Project Proposal

**Ameya Dattaram Parab**

Master of Science in Data Science

Indiana University Bloomington

**Abstract**

As per defined in the Gartner’s Glossary, Attrition is the departure of employees from the organization for any reason (voluntary or involuntary), including resignation, termination, death or retirement. Employee retention is essential for a company’s success, and the attrition rate is the metric that provides insight into how well the employees are retained in a company. Measuring and analyzing the attrition rate will help in identification of number of employees leaving the organization within a certain period and the reasons behind their exit.

Since late 2020 resignation rates have been increasing in every organization, which has led to terming this global historic event as the Great Resignation. According to a research by the Bureau of Labor Statistics, 4 million people quit their jobs last February. By observing this noticeable trend, there arises a question: why we are seeing a record number of employees who quit their jobs to pursue other opportunities.

To address and answer the above concern, this visualization project will consider various factors related to the attrition of employees for creating an informative dashboard. The inferences from the visualizations created will help in HR analytics and to understand how employee turnover can benefit an organization during unpredictable times such as the Great Resignation.

**Index**

1. **Introduction** 4
   1. Overview of Existing Visualizations 5
   2. Objectives 7
2. **Process** 8
   1. Data 8
      1. Job Openings and Labor Turnover Survey 8
      2. IBM HR Analytics Employee Attrition & Performance 9
   2. Preprocessing 12

* 1. Visualization Libraries 13
  2. Visualization Methods 14
     1. Basic Visualizations 14
        1. Violin Plot 14
        2. Parallel Co-ordinates Plot 17
        3. Chord Diagram 18
     2. Interactive Visualizations 20
        1. Histogram 20
        2. Binned Scatter Plot 22
        3. Ridgeline Plot 24
        4. Radar Chart 26
     3. Animated Visualizations 27
        1. Bubble Chart 27
  3. Aggregating Visualizations 29
     1. Colors 29
     2. Gestalt Principles of Design 29
        1. Similarity 29
        2. Proximity 30
     3. Alignment 31

1. **Results**  32
   1. Insights 32
   2. Observations 35
2. **Conclusion** 38
3. **References** 39
4. **Introduction**

It’s usual for any organization to have employees leaving them. However, when the frequency of employees quitting increases within a short period, it is an alarming situation that needs the immediate attention of higher management. The higher-level management may have other priority matters to look into as well. Hence, by combining the employee data and attrition analytics, we can get insights into potential issues that could help in finding solutions. A dashboard of visualizations will aid the management of the organization to better understand workforce distribution, diversity goals, employee satisfaction, and forecast budgets.

The motivation behind gathering attrition data and creating such visualizations is to help an organization analyze and maintain a low attrition rate. A higher attrition rate can cause several losses to an organization such as:

1. Financial losses
2. Delayed deliverables
3. Hiring new talent

Firstly, there’s a huge cost involved in the recruiting process. It could take weeks to months to find the right talent and train the new employees, along with arranging necessary resources for them. This eventually adds up a certain amount of delay in the delivery of ongoing work. When a skilled employee quits, work is interrupted until a replacement is found, as it could take time for the new talent to match the productivity level and expertise of prior employee. Moreover, the attrition rate represents work environment and company culture to outside world. A higher attrition rate makes it difficult to find new talent, that leads to a gruesome process of advertising the profile, sorting through potential candidates and the interview process.

* 1. **Overview of Existing Visualizations**

There are illustrations already implemented using the selected dataset. Few of them are described below:

Chart, diagram

Description automatically generated

Figure 1 (<https://public.tableau.com/app/profile/shuham.paliwal/viz/SampleOne_15933605582560/Dashboard1>)

The dashboard in Figure 1 has been created using Tableau and highlights the various aspects of the data. However, these visualizations have been built using the raw data and features which were present in the dataset. In other words, the plots use the data and merely represent what happened, which is known as descriptive analytics.

Chart

Description automatically generated

Figure 2 (<https://public.tableau.com/app/profile/mo.with.data/viz/IBMHRAnalytics_16451331222160/Dashboard>)

Another Tableau dashboard depicted in Figure 2 which is created using the same dataset emphasizes on the descriptive variables of the dataset. Although this dashboard also serves the same purpose, it is better in certain aspects than the former one. Each figure in this dashboard is based on at least two dimensions. Apart from that, the arrangement of plots and the choice of color scheme used are pleasing to the viewer. Whereas in the previous dashboard, the figures are unorganized, disproportionate, and do not follow a color scheme.

There are very few inferences that can be derived from descriptive visualizations to analyze and solve a particular problem. The goal of this project is not only to create visualization for overview of the data but also to dive in further and understand why the issue under analysis happened. By performing diagnostic analysis to establish correlation between various features and by illustrating interactive visualizations with multiple dimensions, we would be able to get better inferences from the data.

* 1. **Objectives**

The project aims to create a data visualization dashboard using python-based libraries such as Seaborn and interactive visualizations using Vega-Altair. The dashboard will feature descriptive and diagnostic visualizations that will help the higher management of an organization to get a brief overview of the situation. It will also focus on the aesthetics of the dashboard by using a uniform color scheme across all the figures and adhering to the Gestalt Principles of design.

1. **Process**
   1. **Data**

The data that will be used for this project is been taken from Kaggle**’s** [IBM HR Analytics Employee Attrition & Performance](https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset)as well as the [U.S. Bureau of Labor Statistics, Job Openings and Labor Turnover Survey](https://www.bls.gov/jlt/data.htm).

**2.1.1. Job Openings and Labor Turnover Survey**

The Job Openings and Labor Turnover Survey program provides national estimates of rates and levels for job openings, hires, and total separations. Total separations are further broken out into quits, layoffs and discharges, and other separations. All data series are available from December 2000 to August 2022 on a monthly basis. Annual estimates are available for all industries for hires, quits, layoffs and discharges, other separations, and total separations.. The data is divided into separate file and each file contains month and year wise frequency for all the US states including overall US data. The files present in this dataset are divided into three categories:

1. Series file
2. Mapping file
3. Data file

The series file contains a set of codes which, together, compose a series identification code that serves to uniquely identify a single time series. The mapping files are definition files that contain explanatory text descriptions that correspond to each of the various codes contained within each series identification code. The data file contains one line of data for each observation period pertaining to a specific time series. Each line contains a reference to the following:

* Series identification code
* Year in which data is observed
* Period for which data is observed
* Value
* Footnote code

The categorical columns such as Series ID and Period are encoded using a string, the values of which can be found in the respective mapping files of Industry, State and Period.

The Series ID (JTS000000000000000JOR) can be broken out into:

|  |  |
| --- | --- |
| Code | Value |
| Survey abbreviation | JT |
| Seasonal (code) | S |
| Industry code | 000000 |
| State code | 00 |
| Area code | 00000 |
| Size class code | 00 |
| Data element code | JO |
| Rate level code | R |

In this project, the visualization that is based on this dataset is a timeseries graph. For this purpose, only the Job Openings, Hires and Quits data will be considered along with the mapping files. The Job Openings and Labor Turnover data will be used to depict an overview of the industry and year-wise employee attrition, job openings, and layoffs observed in the US to give an idea about the importance of the attrition rate.

* + 1. **IBM HR Analytics Employee Attrition & Performance**

The dataset from Kaggle will be utilized to create a dashboard of visualizations consisting of basic informative, correlation, and causation graphs and interactive plots to help in the analysis of the attrition issue.

The dataset contains 35 columns out of which 13 numerical features and 17 categorical features can be deemed as necessary for creating visualizations and analytics. There are about 1470 observations in the csv. Below is a snapshot of the dataset obtained from Kaggle:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sr No. | Column Name | Type | Possible Values | Key |
| 1 | Age | Numerical | [18,60] |  |
| 2 | Attrition | Categorical | Yes/ No |  |
| 3 | Business Travel | Categorical | Travel\_Rarely |  |
| Travel\_Frequently |  |
| Other |  |
| 4 | DailyRate | Numerical | [102,1499] |  |
| 5 | Department | Categorical | Sales |  |
| Research & Development |  |
| Other |  |
| 6 | DistanceFromHome | Numerical | [1,29] |  |
| 7 | Education | Categorical | 1 | Below College |
| 2 | College |
| 3 | Bachelor |
| 4 | Master |
| 5 | Doctor |
| 8 | EducationField | Categorical | Life Sciences |  |
| Medical |  |
| Other |  |
| 9 | EmployeeCount | Numerical | 1 |  |
| 10 | EmployeeNumber | Numerical | [1,2068] |  |
| 11 | EnvironmentSatisfaction | Categorical | 1 | Low |
| 2 | Medium |
| 3 | High |
| 4 | Very High |
| 12 | Gender | Categorical | Male |  |
| Female |  |
| 13 | HourlyRate | Numerical | [30,100] |  |
| 14 | JobInvolvement | Categorical | 1 | Low |
| 2 | Medium |
| 3 | High |
| 4 | Very High |
| 15 | JobLevel | Categorical | 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |
| 16 | JobRole | Categorical | Sales Executive |  |
| Research Scientist |  |
| Other |  |
| 17 | JobSatisfaction | Categorical | 1 | Low |
| 2 | Medium |
| 3 | High |
| 4 | Very High |
| 18 | MaritalStatus | Categorical | Married |  |
| Single |  |
| Other |  |
| 19 | Monthly Income | Numerical | [1009, 20,000] |  |
| 20 | MonthlyRate | Numerical | [2094, 27,000] |  |
| 21 | NumCompaniesWorked | Numerical | [0,9] |  |
| 22 | Over18 | Categorical | Y |  |
| N |  |
| 23 | OverTime | Categorical | Yes |  |
| No |  |
| 24 | PercentSalaryHike | Numerical | [0,100] |  |
| 25 | Performance Rating | Categorical | 1 | Low |
| 2 | Good |
| 3 | Excellent |
| 4 | Outstanding |
| 26 | RelationshipSatisfaction | Categorical | 1 | Low |
| 2 | Medium |
| 3 | High |
| 4 | Very High |
| 27 | StandardHours | Numerical | [0,80] |  |
| 28 | StockOptionLevel | Categorical | 0 |  |
| 1 |  |
| 2 |  |
| 3 |  |
| 29 | TotalWorkingYears | Numerical | [0,40] |  |
| 30 | TrainingTimeLastYear | Numerical | [0,6] |  |
| 31 | WorklifeBalance | Categorical | 1 | Bad |
| 2 | Good |
| 3 | Better |
| 4 | Best |
| 32 | YearsAtCompany | Numerical | [0,40] |  |
| 33 | YearsInCurrentRole | Numerical | [0,18] |  |
| 34 | YearsSinceLastPromotion | Numerical | [0,15] |  |
| 35 | YearsWithCurrentManager | Numerical | [0,17] |  |

Each categorical feature is independent from another and contains finite categories. The numerical variables are discrete and consist of discrete values for every observation. Also, there are no missing values and very few outliers present in the data set. Moreover, each variable represents a column, and each observation forms a row in the dataset. Hence, we can say that the dataset is tidy along with appropriate number of categorical and numerical features for illustrating the data. Therefore, we can use this dataset to create insightful visualizations for analysis in this project.

* 1. **Preprocessing**

The JOLTS has a lot of data separated industry and state wise. To be able to represent this data on the time series plot, the data needs to be preprocessed and combined to generate meaningful insights. The purpose of the visualization from this data is to compare job openings, quits and hires ranging in certain time period.

* The files contain encoded data for state and month. To provide a better readability, two new columns are added to each data frame for state and month respectively.
* The value column is normalized by multiplying the observations by 1000
* The footnote codes column is dropped since it won’t be useful in the visualization
* Since the data is recorded from December 2000 till August 2022, for these two particular years there won’t be any information for the rest of the months. Hence, the data belonging to these two years is dropped and only the observations in the range 2001 – 2021 are considered.

The HR analytics data is tidy; however, it does contain columns which are not much useful for creating visualizations. These columns will be dropped:

* EmployeeCount: The value for this column is 1 since every observation represents a single employee.
* EmployeeNumber: This is the Employee ID which will not be of much help for any insights.
* Over18: It is a binary value column where all the records have same value.
* StandardHours: The hours for all the employees are same i.e. 80 hours.

Also, there exists similar set of columns such as HourlyRate, DailyRate, MonthlyRate and MonthlyIncome, all of which have the meaning and trend. Hence, for uniformity across all the visualization MonthlyIncome column is being used. Further, depending on the purpose of the plots, data is filtered based on attrition and used in the respective plots.

* 1. **Visualization Libraries**

The visualizations in this project have been created using the following Python based libraries:

* Matplotlib: Matplotlib is the most famous library for data visualization with python. It allows to create literally every type of chart with a great level of customization.
* Seaborn: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
* Vega-Altair: It is a declarative statistical visualization library for Python, based on Vega and Vega-Lite. Altair offers a powerful and concise visualization grammar that enables you to build a wide range of statistical visualizations.
* Plotly: It is a javascript library for data visualization. It is based on the famous d3.js library, and provides a python wrapper allowing to build stunning interactive charts directly from Python.
* MNE-Connectivity: It is an open-source Python package for connectivity and related measures of MEG, EEG, or iEEG data built on top of the MNE-Python API. It includes modules for data input/output, visualization, common connectivity analysis, and post-hoc statistics and processing.
  1. **Visualization Methods**

A wide range of visualization have been developed in this project to create a final dashboard of set of visualizations. This section consists of different types of plots created and the methods used for creating them.

* + 1. **Basic/ Static Visualizations** 
       1. **Violin Plot**

A violin plot allow to visualize the distribution of a numeric variable for one or several groups. They are similar to histograms and box plots in that they show an abstract representation of the probability distribution of the sample. Rather than showing counts of data points that fall into bins or order statistics, violin plots use kernel density estimation (KDE) to compute an empirical distribution of the sample.

Violin plots can be useful to compare summary statistics of different groups.

For creating a Violin Plot below steps were followed:

|  |  |
| --- | --- |
| Boxplots for different education levels were created ranging across the X-axis and the Years in the Company variable on the Y-axis. These boxplots clearly represent the summary of the data. |  |
| To further get an idea about the probability distribution of the data along the Y-axis, Violins are added for each category. However, this plot doesn’t give an idea about the number of data points in each category. |  |
| To solve this issue, points are plotted belonging to each category with a different color. Since there is a single axis for each category, it leads to overlapping points. |  |
| The scatter points are jittered along with adjusting transparency to represent the number of points in different categories. However, the category labels are numbers which could be difficult to interpret. A final figure can be generated by converting these values to respective labels and naming the axes. |  |

Diagram

Description automatically generated

The final plot thus created depicts distribution of the employees quitting the company after certain number of years belonging to different categories of education level. The variables from the data are represented using different visual encodings:

* Education level category is represented by the use of different colors from Tab10 colormap.
* Years in Company is plotted across the Y-axis.
* The frequency distribution of the employees can be interpreted through the Violin plot and jittered points.
* Summary statistics can be found using the underlying Boxplot.

Violin plot is an example in which by combining several univariate visualizations a meaningful plot with multiple dimensions can be created.

* + - 1. **Parallel Plot**

Parallel plot or parallel coordinates plot allows to compare the feature of several individual observations (series) on a set of numeric variables. Each vertical bar represents a variable and often has its own scale. (The units can even be different). Values are then plotted as series of lines connected across each axis.

The parallel coordinates have been created using columns belonging to psychological observations recorded of the employee such as job Involvement, job satisfaction, work-life balance, relationship satisfaction, and environment satisfaction. Plotly was used to create the below parallel plot.

A picture containing shape

Description automatically generated

The plot created consists of many lines crossing each other. Since, there are many lines and certain lines may overlap due to presence of discrete values. The ordering of the columns can be changed to minimize the number of cross between series. However, that would also not be effective to read the data. A parallel plot works best with columns having continuous numeric data range.

* + - 1. **Chord Diagram**

A chord diagram represents flows or connections between several entities (called nodes). Each entity is represented by a fragment on the outer part of the circular layout. Then, arcs are drawn between each entities. The size or color of the arc is proportional to the importance of the flow.

Chord diagrams can be created using Matplotlib, Seaborn, Plotly, Bokeh and MNE. Although MNE library has been written for neuroscience purposes, it comes with some handy functions when it comes to building chord diagrams.

To solve the issues of Parallel Plot for this data a chord diagram can be used. Since the data is categorical in nature a chord diagram would be the suitable visualization method in this scenario. For creating a Chord Diagram, the columns selected belonged to psychological observations recorded of the employee such as job Involvement, job satisfaction, work-life balance, relationship satisfaction, and environment satisfaction. The data was tidy enough that all these columns had values in the same range. Hence, no need for scaling the data of any column.

As each link in the Chord Diagram is governed by a weight assigned to it, a 2d matrix consisting of count of employees belonging to the two column values was created.

|  |  |
| --- | --- |
| Creating the initial connectivity plot with all the labels. The purpose of representing the frequency between each node is being done by the Sequential Colormap. However, the color scale for the nodes is not appropriate to segregate into variables that were defined. |  |
| This can be solved by changing colormap. Tab20 is the best suited for this data, since there are 20 nodes and each column has 4 nodes. Also , the variables need to be labeled. The node groups belonging to a column can further be segregated by adding a separation between each column. This can be done by assigning angles to each node. |  |

Chart, radar chart

Description automatically generated

Hence, by using Sequential and Qualitative colormaps different categories and the frequency between each node is represented in the above Chord Diagram. The visual encodings present in this diagram are as follows:

* Categorical columns are represented by using different colors on the edges of the circle.
* Each categorical value of a column has a shade of the color for that particular column.
* The frequency of employees between two columns can be visualized by the color from sequential colormap.
  + 1. **Interactive Visualizations:**
       1. **Histogram**

A Histogram represents the distribution of a numeric variable for one or several groups. The values are split in bins, each bin is represented as a bar.

Histogram can be used to view various features present in this dataset.

Two such features selected for visualizing in this plot are the numerical variable of Distance from Home and the categorical variable of Business Travel. It was created by following the below steps:

|  |  |
| --- | --- |
| A stacked histogram is created first. Although stacking of bars allows comparison of different categories, it does not give an appropriate perspective for comparing the frequencies in different bins. |  |
| Even though stacking is set to none, it is not clear to the viewer where the bars behind start from. Also, some bars might hide other bars. |  |
| To avoid this, transparency is added so that the bars behind would be visible. Adding transparency doesn't help much as it is still not clear where the bars start from. |  |

Altair allows creation of interactive plots by adding a drop down or radio button to select the variable to be displayed on graph. Using this feature, a drop down is added to this plot to select the Business Travel category.

Chart, bar chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Thus, the visibility of the Histogram improved by using Altair's selection and binding features allowing us to get better insights. The Histogram created has following visual encodings:

* Distance from home values are binned along the X-axis.
* Frequency of the employees who resigned is shown on Y-axis.
* The categories of Business travel are represented using colors.
  + - 1. **Binned Scatter Plot**

A binned scatter plot is a more scalable alternative to the standard scatter plot. The data points are grouped into bins, and an aggregate statistic is used to summarize each bin. Here we use a circular area encoding to depict the count of records, visualizing the density of data points. Apart from that color can also be used to represent groups.

In this project, a binned scatter plot is created to visualize the relationship between job satisfaction and years since last promotion and their effect on the employees leaving the organization. Below are the steps in which the binned scatter plot was created:

|  |  |
| --- | --- |
| Creation of a binned scatterplot using circle. In this case, the area of the circle represents the frequency of employees. |  |
| Since the circle is filled there might be a chance that a smaller circle of another categorical value may get hidden behind the bigger circle. To avoid this, only outline of circle can be used in the plot. |  |
|  |  |

The label values are encoded in numbers which can be converted to their original values. Further, the graph can be made interactive and a radio button for selection of categories can be added to visualize individual frequencies.

Chart

Description automatically generated

A picture containing chart

Description automatically generatedChart

Description automatically generated

Binned Scatterplot combines the usage of a scatter plot and histogram. As compared to a scatter plot or histogram, this combined plot can represent an additional dimension. The visual encodings considered for the above graph are as follows:

* Job satisfaction categories are plotted on X-axis.
* Years since last promotion is a numerical variable which is shown on Y-axis.
* The frequency of the employees is denoted by the area of the circle.
* The attrition category is encoded using two colors.
  + - 1. **Ridgeline Plot**

A Ridgeline plot (sometimes called Joy plot) shows the distribution of a numeric value for several groups. Distribution can be represented using histograms or density plots, all aligned to the same horizontal scale and presented with a slight overlap. It works well when there is a clear pattern in the result, like if there is an obvious ranking in groups. Otherwise group will tend to overlap each other, leading to a messy plot not providing any insight.

In this scenario, a ridgeline plot is drawn to compare the attrition rate across different departments. It was created by following below steps:

|  |  |
| --- | --- |
| The base for a ridgeline plot is the KDE plot. KDE plots of all the departments are visualized together in a single plot. Although there is transparency added, the distributions for different departments get overlapped and it is difficult to analyze the chart. |  |
| Ridgeline plot solves this issue by minimizing the overlap among categories by combining the charts of different departments using facets. |  |

The final plot was created by adjusting size as per the dashboard, formatting of labels and hiding the axis.

Chart, line chart

Description automatically generated

Therefore, instead of using a stacked density plot to compare several groups, a ridgeline plot can be drawn which gives better readability of the chart. Following dimensions are being shown in this plot:

* Work experience is distributed along X-axis.
* The frequency distribution of the employees leaving the organization is plotted on the Y-axis.
* Different departments are stacked above each other minimizing the overlap and are distinguished using colors.
  + - 1. **Radar Chart**

A radar or spider or web chart is a two-dimensional chart type designed to plot one or more series of values over multiple quantitative variables. Each variable has its own axis, all axes are joined in the center of the figure.

This chart has been implemented in the project to give an idea about the current situation of the job market in US and why is it essential to study about the attrition of employees. It is used to plot series of year wise frequencies for job openings, hires and quits in the 20 years range. Radar charts can be easily created using Plotly. The chart has been implemented using Plotly library.

Chart, radar chart

Description automatically generated

As compared to a horizontal graph the radial graph can be used to compare multiple year values with each other. In a horizontal graph the comparison between start and end years is difficult to visualize.

Following dimensions and encodings are represented in the chart:

* The years on the edges of the circle.
* Frequency of people on each radian.
* The categories of quits, hires and job openings by colors.
  + 1. **Animated Visualizations**
       1. **Bubble Chart**

A bubble plot is a scatterplot where a third dimension is added: the value of an additional numeric variable is represented through the size of the dots. Three numerical variables as input: one is represented by the X axis, one by the Y axis, and one by the dot size.

An animated chart displays a sequence of static charts, often in a gif or movie format. It can be really useful to describe the evolution of a dataset, or to denote the difference between two states (going from one to another and backward).

A bubble chart with animation has been created to display the correlation between monthly income and hike percentage by animating frames over the years spent in the company. This way the data distribution can be compared for various years.

Initially a static bubble plot was created.

Chart, bubble chart

Description automatically generated

There are too many points in the plot and the entire data at once does not make any sense for analysis. The plot can be made dynamic by adding another variable as frame for each iteration. The additional variable could be a time series to visualize the change of data distribution over a time frame.

Chart, scatter chart, bubble chart

Description automatically generated

Chart, bubble chart

Description automatically generated

This visualization contains the maximum number of dimensions out of all the visualizations. Following information has been conveyed through this plot:

* Monthly income is represented using a log scale on X-axis as the data distribution seems to be skewed towards the lower values.
* Percent salary hike is distributed along Y-axis.
* Attrition category is colored with red and green colors.
* The size of the circle is proportional to the work experience of the employee.
* The value of the companies worked before can be changed using the slider.
* Also, by hovering on each point the designation of employee is visible.
  1. **Aggregating Visualizations**

As defined in the objectives of this project a dashboard is created using HTML to display the aggregated visualizations for storytelling. A viewer can change play with the interactive plots, run animation and analyze the static plots with ease on a single web page.

* + 1. **Colors**

Usage of color maps to denoted categories, quantities, separations, and frequencies is essential for pre-attentive processing. All the visualization on this dashboard utilizes the qualitative color maps of tab10 and tab20c which contain similar shades of color to represent several groups in the data. This makes the dashboard aesthetic and pleasing to the viewer.

A picture containing chart

Description automatically generated

* + 1. **Gestalt Principles of Design**

Gestalt principles are an important set of ideas for representing designs and their implementation can greatly improve the aesthetics of a design as well as its functionality and user-friendliness. There are six individual principles commonly associated with gestalt theory: similarity, continuation, closure, proximity, figure/ground, and symmetry & order.

* + - 1. **Similarity**

It’s human nature to group like things together. In gestalt, similar elements are visually grouped, regardless of their proximity to each other. They can be grouped by color, shape, or size.

|  |  |
| --- | --- |
| In the following Chord Diagram, the values within a single category are grouped together in divided section of arc and also represented by varying shades of same color. |  |
| Set of points belonging to two different categories is represented by different colors in Bubble Chart. |  |
|  |  |
|  |  |

* + - 1. **Proximity**

Proximity refers to how close elements are to one another. The strongest proximity relationships are those between overlapping subjects, but just grouping objects into a single area can also have a strong proximity effect.

|  |  |
| --- | --- |
| The stacking of plots on each other in the Ridgeplot makes use of proximity principle. It becomes easier to read and interpret that the aim is to compare the distribution across departments even if the plots are separate. |  |

* + 1. **Alignment**

The figures are organized on a web page in a manner for story telling purpose. Bootstrap has been used for arranging the figures in a layout along with making the web page responsive to be viewed on any device.

Chart

Description automatically generated

The entire implementation of the process followed in this project is documented in a notebook on [Github](https://github.com/ameyaparab1996/employee-attrition-rate-analysis).

1. **Results**
   1. **Insights**

From the analysis performed on the data visualized on the dashboard by modifying the categories for plots, viewing animated frames and playing with interactive plots following observations can be extracted:

|  |  |
| --- | --- |
| The job openings and hiring has significantly increased in the past two years. This has led to an increase in attrition rate throughout the US. |  |
| Distance doesn’t affect the attrition rate; as with less distance from home there are higher number of employees leaving the company. |  |
| However, the business travel requirement could be a major deciding factor for an employee to stay in the organization. | Chart, bar chart, histogram  Description automatically generated |
| Recently promoted employees with low job satisfaction responses are likely to quit. One possible assumption could be that most of the employees wait till their promotion so that they get better opportunities outside their organization. |  |
| Research & Development is the department having highest number of quits for less experienced employees. |  |
| There is a strong association among level 3 value of Job Involvement, Job Satisfaction and Work-life Balance. This means even though the working conditions are better in a company the employees tend to leave. |  |
| The highest attrition rate is among the below college and bachelor’s degree holders who have spent few years in the company. A possible explanation for this could be that the employees want to pursue further education because of which they don’t want to continue working. |  |
| Employees who have worked in a single company before tend to quit for better opportunities. |  |
| Whereas, the attrition rate is low among the employees who do not have any prior work experience. This means there are learning opportunities for freshers in the company and they really want to gain some experience rather than quitting. |  |
| The employees who have worked for more than one company are willing to settle in the same organization. Low monthly income and low hikes are the driving factor for resignations. |  |

* 1. **Observations:**

By studying and implementing various types of plots in this project, the benefits and limitations of each is summarized below:

* By adding a selectable category, an additional dimension can be added in the plots.

A picture containing application

Description automatically generated

* The links between discrete values of data can be better represented by a Chord Diagram than a Parallel Plot.

A picture containing shape

Description automatically generated Shape, icon

Description automatically generated

* Two or more uni-variate visualizations when combined together by appropriate encoding can represent multiple variables.

Chart

Description automatically generated

* Creating interactive visualizations can help in better exploration of the data.

Chart

Description automatically generated

Chart, bubble chart

Description automatically generated

* Furthermore, when animated an additional dimension can be added.

Chart, bubble chart

Description automatically generated

* Since area is not perceived very well, the binned scatter plot can be improvised by changing the shape to denote frequency or by converting it into a 2d heat map.

Graphical user interface, application

Description automatically generated with medium confidence

* Although there are caveats associated with the Radar chart, it can be made effective to visualize the time series data by animating it over months represented on the edges of circles for each years. Since the months are cyclic in nature, a Radar chart would be the best representation of such time series animation.

1. **Conclusion:**

With consideration of various numerical and categorical features from the data, a visualization dashboard which provides significant insights has been created. The source code of Altair plots can be easily accessed and modified to enhance the visualizations.

The insights generated above could certainly help an organization take necessary actions for solving the problem of attrition.

1. **References**
2. Kaggle: IBM HR Analytics Employee Attrition & Performance Dataset

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

1. U.S. Bureau Of Labor Statistics

<https://www.bls.gov/news.release/jolts.toc.htm>

1. Skeeled: What is an Employee Attrition Rate and Why it Matters

<https://www.skeeled.com/blog/what-is-an-employee-attrition-rate-and-why-it-matters#:~:text=Attrition%20rates%20help%20you%20understand,for%20longer%20periods%20of%20time>.

1. Forbes: ‘The Great Resignation’ Migration And What This Means For Your Career

<https://www.forbes.com/sites/bryanrobinson/2021/06/11/the-great-resignation-migration-and-what-this-means-for-your-career/?sh=43528e5b69aa>

1. Tableau Dashboards:

IBM HR Analytics by Shubham Paliwal

<https://public.tableau.com/app/profile/shuham.paliwal/viz/SampleOne_15933605582560/Dashboard1>

IBM HR Analytics by mowithdata

<https://public.tableau.com/app/profile/mo.with.data/viz/IBMHRAnalytics_16451331222160/Dashboard>

1. From Data to Viz

<https://www.data-to-viz.com/>

1. Python Graph Gallery

<https://www.python-graph-gallery.com/>

1. Seaborn: Statistical Data Visualization Tool

<https://seaborn.pydata.org/>

1. Plotly Open Source Graphing Library for Python

<https://plotly.com/python/>

1. Vega – A Visualization Grammar

<https://vega.github.io/vega/>

1. Vega-Altair: Declarative Visualization in Python

<https://altair-viz.github.io/>

1. MNE-Connectivity

<https://mne.tools/mne-connectivity/stable/index.html>

1. Choosing Colormaps in Matplotlib

<https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e>

1. Creating Interactive Radar Charts with Python

<https://towardsdatascience.com/creating-interactive-radar-charts-with-python-2856d06535f6>

1. Guide to Animated Bubble Charts using Plotly

<https://www.kaggle.com/code/aashita/guide-to-animated-bubble-charts-using-plotly>

1. Animating Data Visualization: Bringing Analytics to Life

https://www.analyticsinsight.net/animating-data-visualization-bringing-analytics-to-life/

1. HR Analytics : Drawing HR insights from widely used IBM Watson’s dataset

<https://thebrew.in/hr-analytics-quick-snapshot-ibm-dataset/>

1. “IBM” HR Analytics Data (EDA)

<https://blog.jovian.ai/ibm-hr-analytics-data-eda-35728dec9d84>

1. IBM HR EDA

<https://www.kaggle.com/code/rahmaabdelkader/ibm-hr-eda>

1. Exploring the Gestalt Principles of Design

<https://www.toptal.com/designers/ui/gestalt-principles-of-design#:~:text=There%20are%20six%20individual%20principles,order%20(also%20called%20pr%C3%A4gnanz)>.