A Report on Visualizing HR Data

By: Evan Sant

Abstract

This report is based on a dataset found on kaggle. It was uploaded to the site to serve as a case study to try and figure out why the company had an unusually high attrition rate. The goal of this report is to determine a proper way to visualize the dataset. In it there are discussions about why each choice was made. The breakdown of the report is as follows. It begins by looking at the data itself, then goes over the evolutions of each visualization, then a deep dive into each of the final visualizations that were chosen, and it concludes with a review of the discoveries and observations that were made throughout this process.

Introduction

While the dataset was introduced to try and determine the cause of the attrition rate in the company, that was not the focus of this report. Instead it highlights the differences and similarities between how gender is represented in the workplace. It was decided that exploring this aspect of the data was more interesting. Does this company do a good job of treating both genders equally or is there a clear skew? These are some of the questions that this report will explore.

The Data

The HR dataset consists of 24 features and 4,411 observations. The majority of them are self explanatory such as Age, Gender, Over18, Department, MonthlyIncome, etc. Others seem to require further explanation. Some examples of this are BusinessTravel, EnvironmentSatisfaction, and PrecentSalaryHike. The following table is a detailed breakdown of the variable name, what it represents, and a breakdown of how the variable is represented in the data, if necessary.

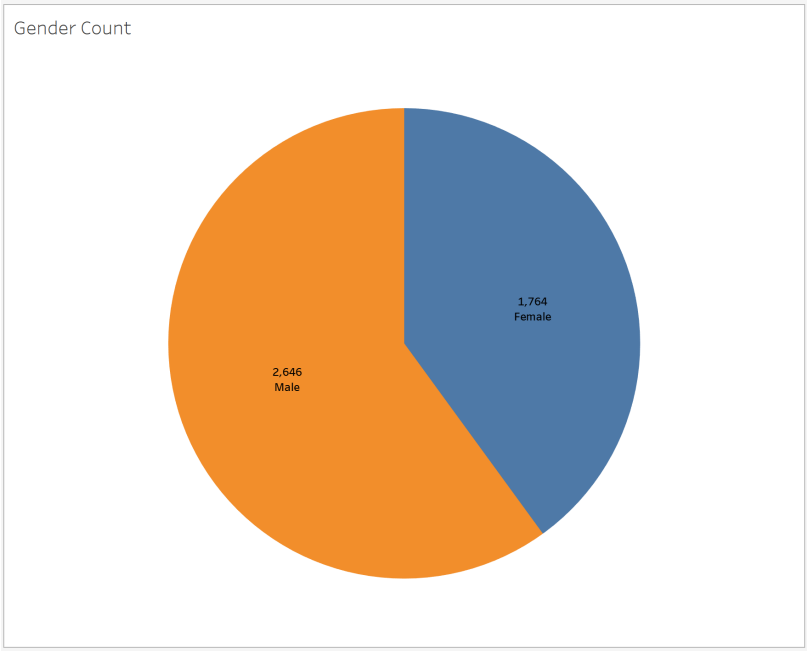
Variable	Meaning	Levels
Age	Age of the employee	
Attrition	Whether the employee left in the previous year or not	
BusinessTravel	How frequently the employees traveled for business purposes in the last year	
Department	Department in company	
DistanceFromHome	Distance from home in kms	
Education	Education Level	1 'Below College'
		2 'College'
		3 'Bachelor'
		4 'Master'

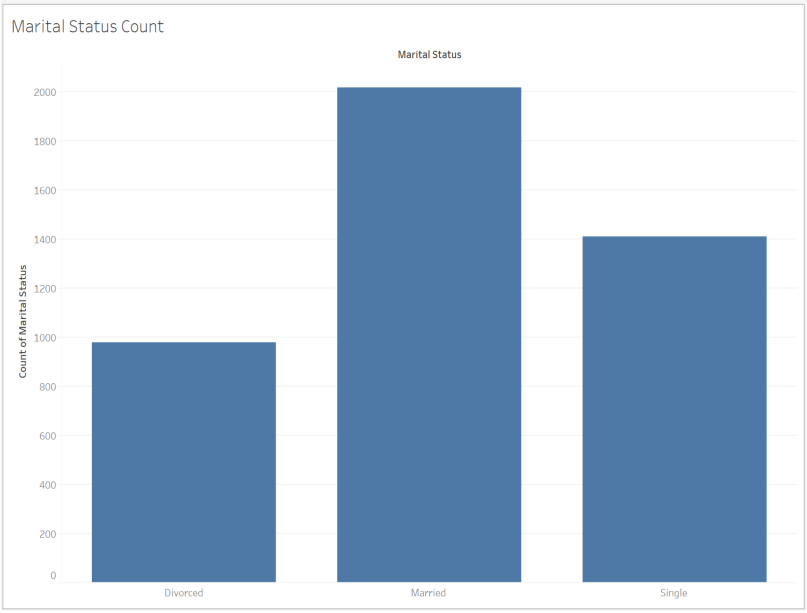
		5 'Doctor'
EducationField	Field of education	
EmployeeCount	Employee count	
EmployeeNumber	Employee number/id	
EnvironmentSatisfaction	Work Environment Satisfaction Level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
Gender	Gender of employee	
JobInvolvement	Job Involvement Level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
JobLevel	Job level at company on a scale of 1 to 5	
JobRole	Name of job role in company	
JobSatisfaction	Job Satisfaction Level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
MaritalStatus	Marital status of the employee	
MonthlyIncome	Monthly income in rupees per month	
NumCompaniesWorked	Total number of companies the employee has worked for	
Over18	Whether the employee is above 18 years of age or not	
PercentSalaryHike	Percent salary hike for last year	
PerformanceRating	Performance rating for last year	1 'Low'
		2 'Good'
		3 'Excellent'
		4 'Outstanding'
RelationshipSatisfaction	Relationship satisfaction level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
StandardHours	Standard hours of work for the employee	
StockOptionLevel	Stock option level of the employee	
TotalWorkingYears	Total number of years the employee has worked so far	
TrainingTimesLastYear	Number of times training was conducted for this employee last year	
WorkLifeBalance	Work life balance level	1 'Bad'

		2 'Good'
		3 'Better'
		4 'Best'
YearsAtCompany	Total number of years spent at the company by the employee	
YearsSinceLastPromotion	Number of years since last promotion	
YearsWithCurrManager	Number of years under current manager	

Exploratory Analysis

When first exploring the data this idea for this report was still unknown. There wasn't a question that stuck out to us right away and the assumption was that the question posed by the user on Kaggle was just what this report would end up becoming. However, these two visualizations from the milestone two submission are what paved the way for this project now. This pie chart was the first one created that showed some interesting information. Essentially it shows that the company is 60% male and 40% female. Then the bar chart of marital status was created and it also showed intriguing results. The vast majority of the company either was or had been married. As such, It was decided that it would explore these two variables further in the report. Marital status made its way into many of the visualizations created in milestone 3 as it was quite interesting to see the effects. However, It was decided that using both gender and marital status caused the visualizations to be too cluttered and unfocused. By the end, the marital status variable had been almost completely dropped and focused on the gender variable from the dataset.

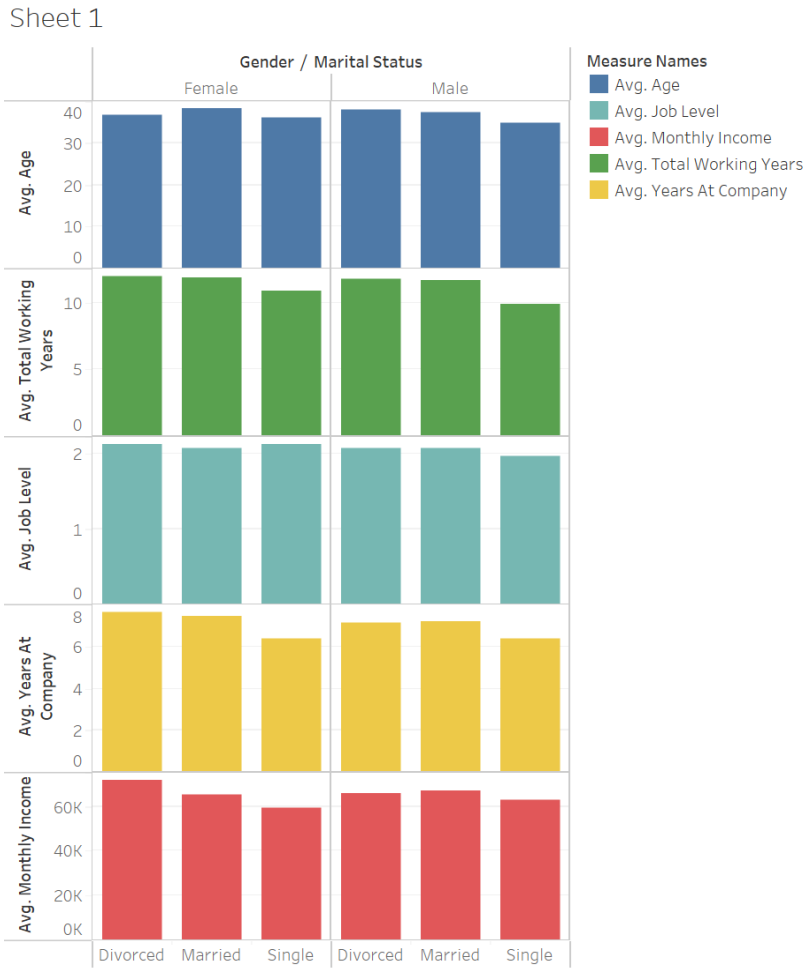




A Breakdown of Each Visualization

The Evolution of the Isotype Visualization

Pretty early on into the process the visualizations gravitated towards the gender variable. That is when the decision was made to focus on the gender variable as a whole. The first part of this visualization was this bar chart. There are a few things wrong with this visualization. First, the legend doesn't need to be present as the colors just represent each variable and as such just repeat the variable names. Additionally, it is a confusing visualization because each bar seems to be a similar height, but the y-axis scale is different for each one. This creates a false sense of scale in the viewer's head. Finally, there is also just too much information packed into this visualization. By this point the decision to make an infographic had already been made and this one visualization contains gender, marital status, average age, average total working years. average job level, average years at the company, and average monthly income. By itself it takes a bit to read, compare, and understand the message that it is sending, and alongside other visualizations this would be too much.

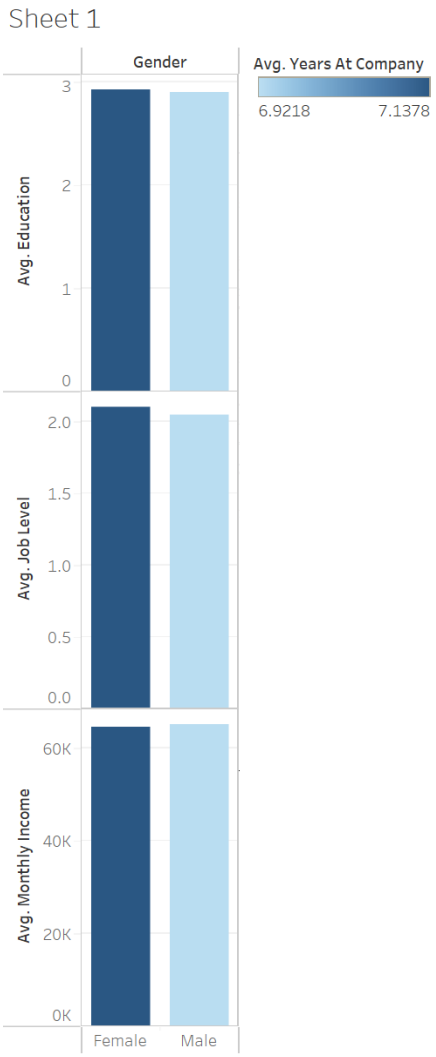


Avg. Age, Avg. Total Working Years, Avg. Job Level, Avg. Years At Company and Avg. Monthly Income for each Marital Status broken down by Gender. Color shows details about Avg. Age, Avg. Total Working Years, Avg. Job Level, Avg. Years At Company and Avg. Monthly Income.

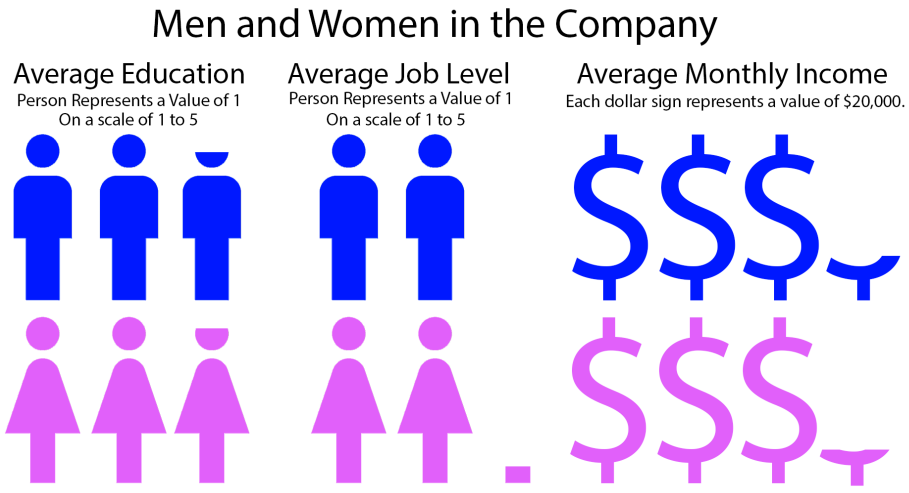
The next evolution of this visualization was this bar chart. It was simplified to show average education, average job level, and average monthly income. Because there are substantially more men at the company than women, the average is the best statistic to use to account for this discrepancy. Additionally, the marital status aspect was dropped in order to focus the visualization more onto the gender aspect. To account for the loss of information the decision was made to color the bars by average years working at the company. However, it is clear that this ends up being misleading. Based on the colors it seems like there is a large difference in the

average years at the company between men and women. However, there is only a difference of about 0.2. This causes confusion and is misrepresenting the data.

The next iteration is then the final iteration of the visualization. The first change is quite noticeable in that it has gone from a bar chart to an isotype visualization. This decision was made for a few reasons. The point of this visualization has always been to highlight the comparison between men and women in the company. To do this effectively, the decision was made to keep the amount of variables to a minimum. This is what influenced the change from iteration 1 to iteration 2. However, the bar chart doesn't present itself in an interesting fashion. While it is important to have visualizations that prioritize accuracy, readability, and are reflective of the data, it is still important to make sure the reader is interested in the visualization. Because the final product is an infographic, the decision was made to make this the visualization that would draw in the viewer. It still conveys the information that is there while also being eye catching. As mentioned before this visualization is the culmination of multiple iterations of bar charts. This section will be more about the visualization itself, rather than how the visualization was made. The first thing to note is that there is seemingly a bit of redundancy in this visualization. There are shapes for men and women and there are also colors for men and women. While the message could still be conveyed if the symbols were the same and there was only color, or if there were only shapes and the colors were the same, both serve their own purpose. First, without the color the dollar signs would be indistinguishable. While it may be true that most would assume that the dollar signs on the bottom correspond with the women as women are on the bottom row in this visualization, the addition of color gets rid of even the few cases where that confusion would occur. Additionally, the difference in shapes for men and women are included to further show the difference between the two. The variables in this visualization are also clearly mapped. Average education is for the first



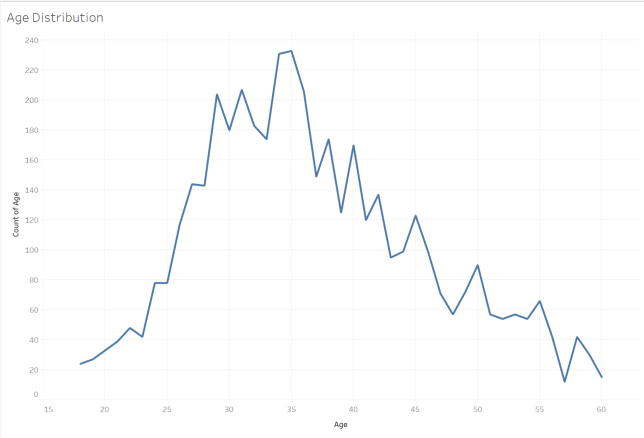
Average of Education, average of Job Level and average of Monthly Income for each Gender. Color shows average of Years At Company.

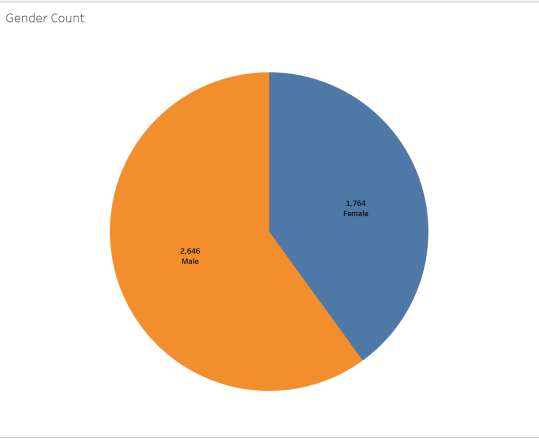


grouping of men and women, average job level is for the second grouping, and average monthly income is for the grouping of dollar signs. This fits into our story by being our introduction. To compare our visualization to an essay this would be the hook. It's simple there are bright colors, and isotype visualizations are shown to stick in people's short term memory better. It does a good job of representing and introducing what the more in depth and complicated visualizations will show.

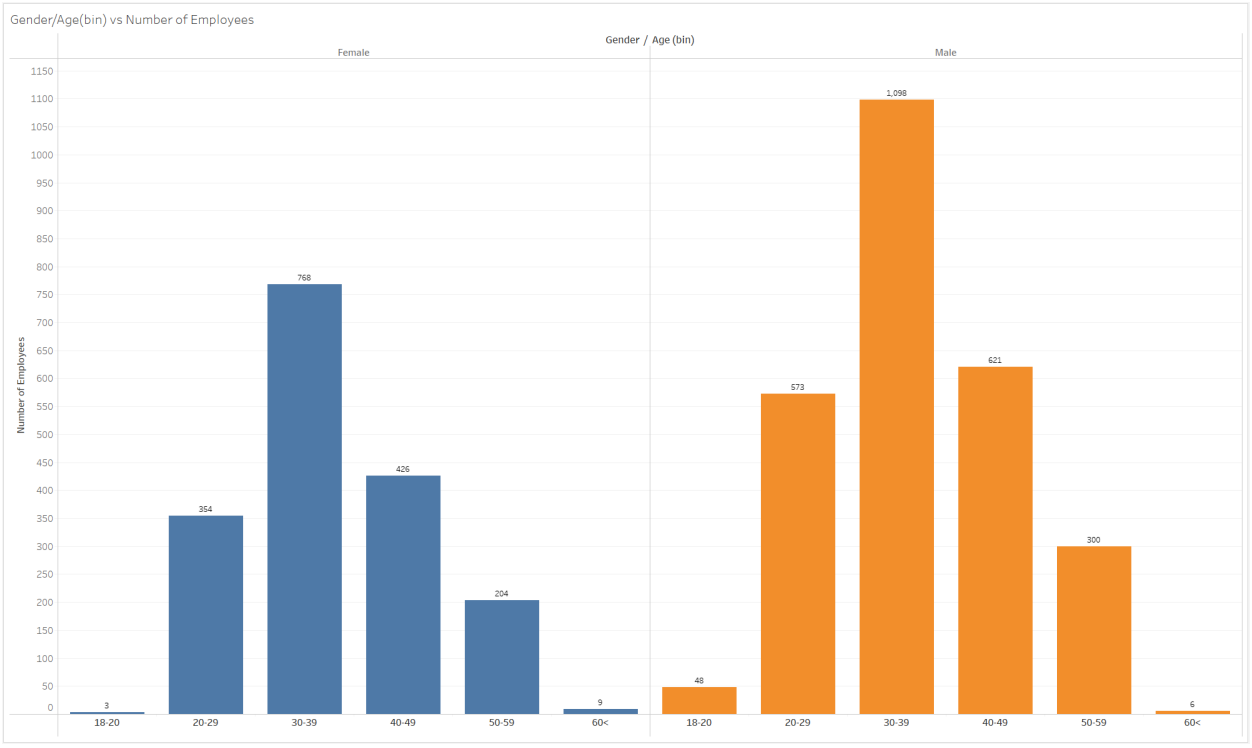
Count of Employees w.r.t Gender and Department/Marital Status

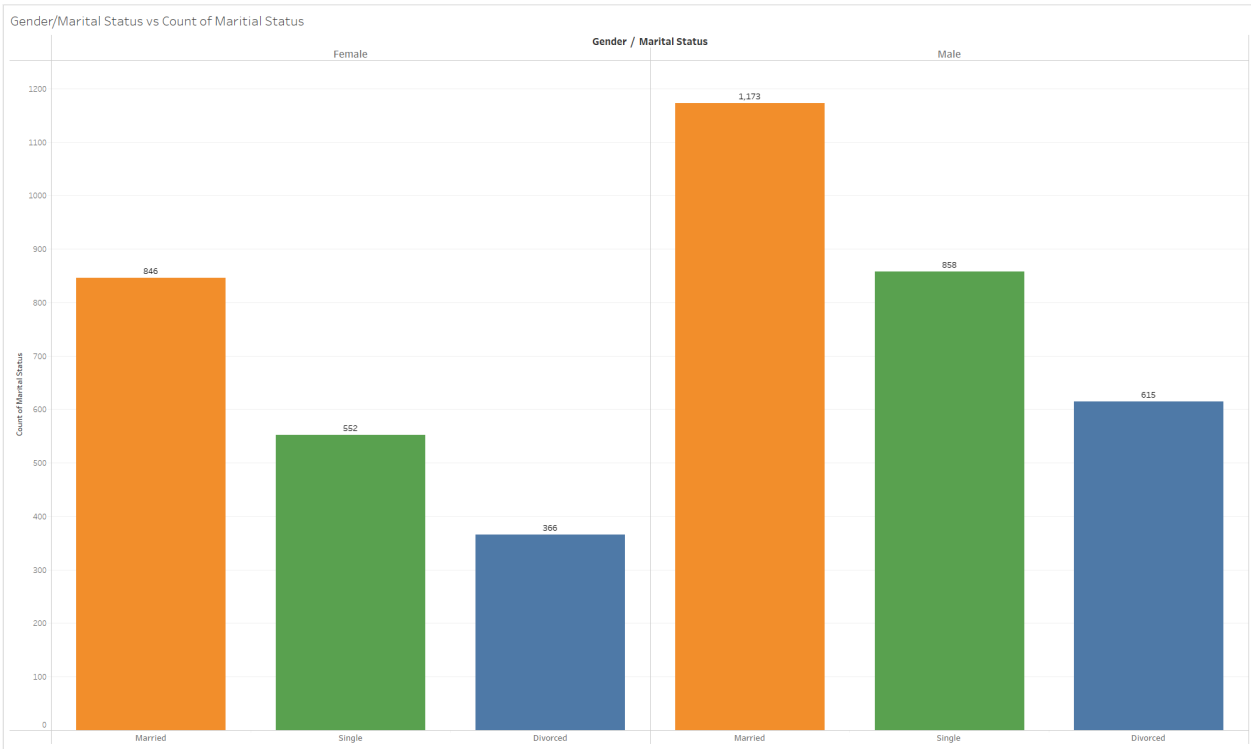
One focus was to examine different facets of the company's workforce in the analysis. In order to gain insight into potential age-related trends, the relationship between age and count of employees of that age was looked into. Then, more closely at the gender distribution of the workforce. Next was to illustrate the proportion of male to female employees using a pie chart, emphasizing the company's overall gender distribution.





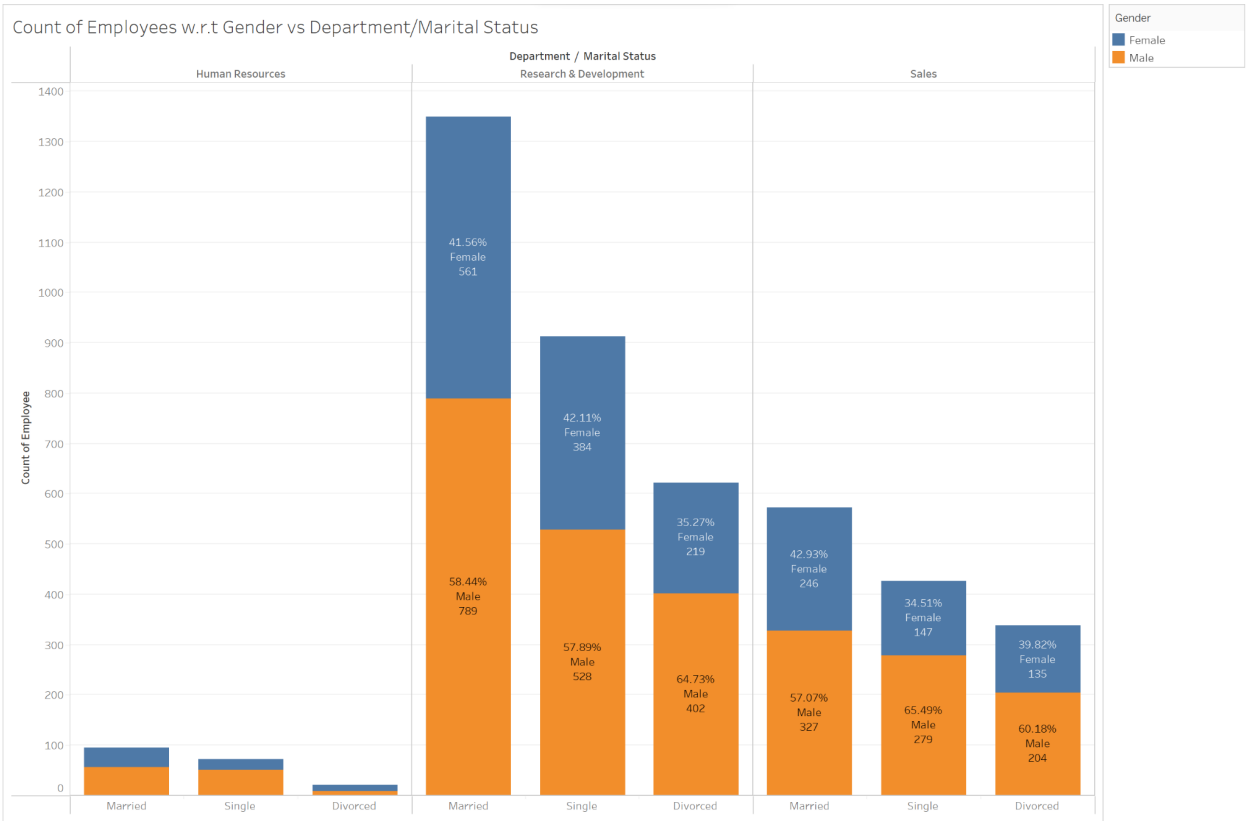
To better understand the marital status of the employees, a bar plot was developed. The number of male and female employees for each marital status category was represented by coloring the bars according to marital status and adding employee counts. As an alternative, this data could have been displayed using a stacked bar plot.



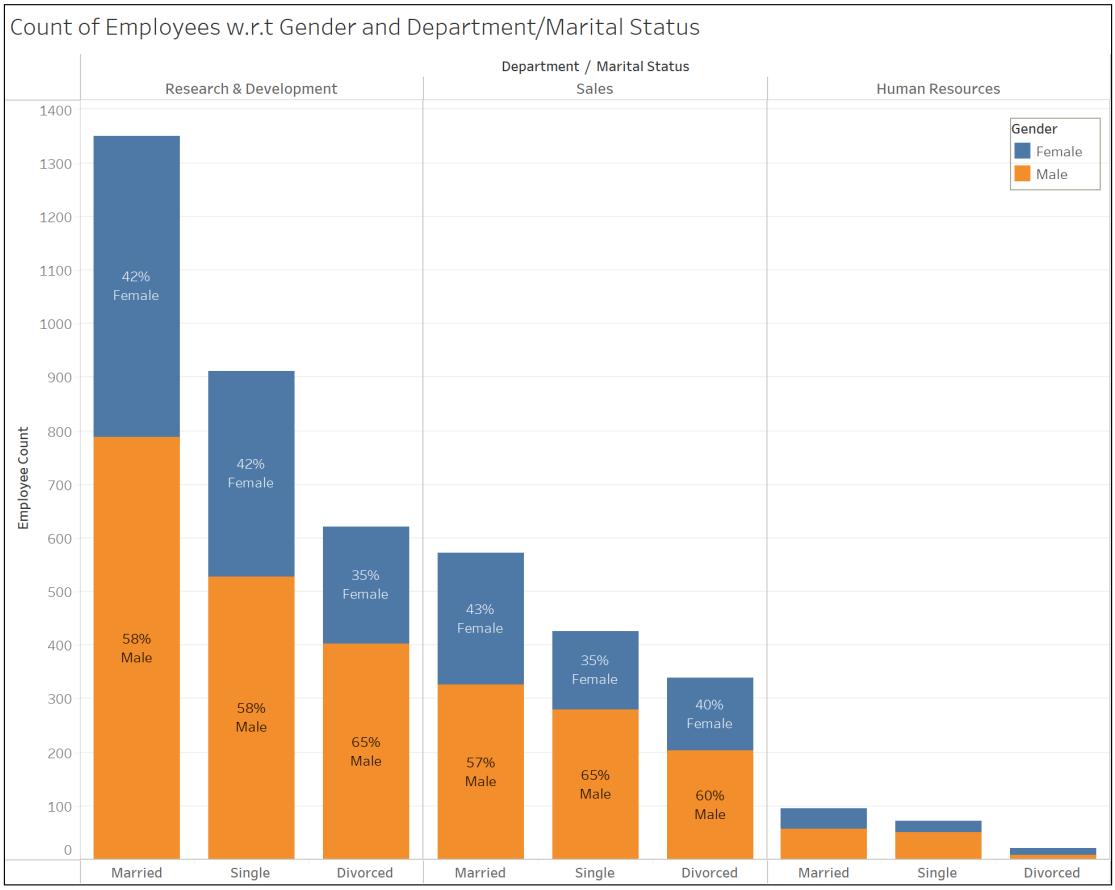


The distribution of employees by gender across various age groups was then investigated. The number of employees in each age group was represented using a bar plot with age bins and colored the bars according to gender. The significant differences in the age distribution of the male and female employees are easy to spot from this visualization.

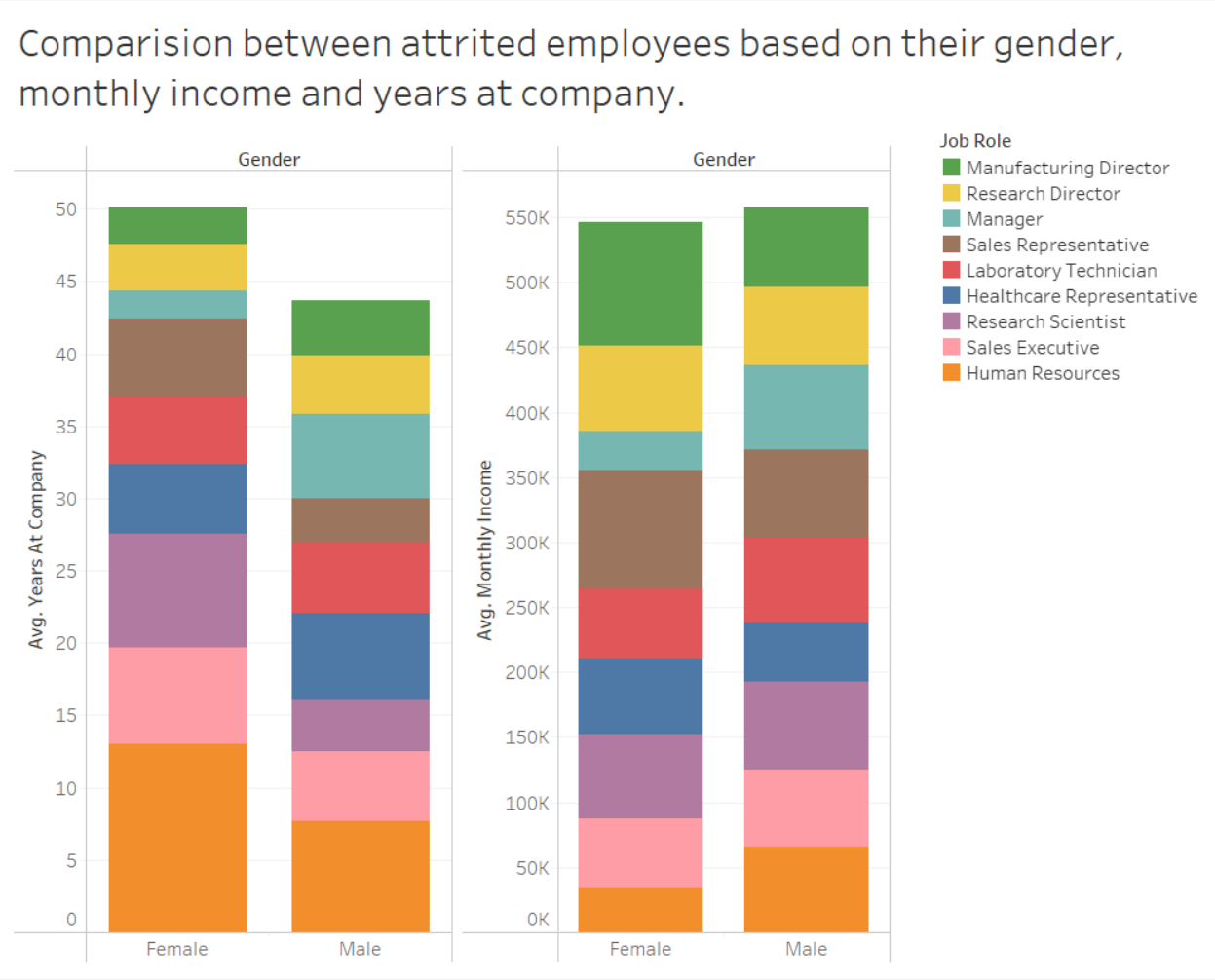
As part of the ongoing analysis, a new variable, "Department" was added to the visualization of the gender distribution. By stacking the genders, more was able to be discerned about the department gender ratio based on marital status. For better clarity, the bars were colored according to gender and added the percentage of each gender. A floating legend could have further improved this visualization.



After receiving feedback from Professor Brown, the visualization was improved by sorting in descending order, rounding up the percentage and adding a floating legend. The final iteration of this visualization is below.


Count of Employees w.r.t Gender and Department/Marital Status
Department / Marital Status
Research & Development
Sales
Human Resources
Gender
Female
Male
Employee Count
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
42% Female
58% Male
42% Female
58% Male
35% Female
65% Male
43% Female
57% Male
35% Female
65% Male
40% Female
60% Male
Married
Single
Divorced
Married
Single
Divorced
Married
Single
Divorced

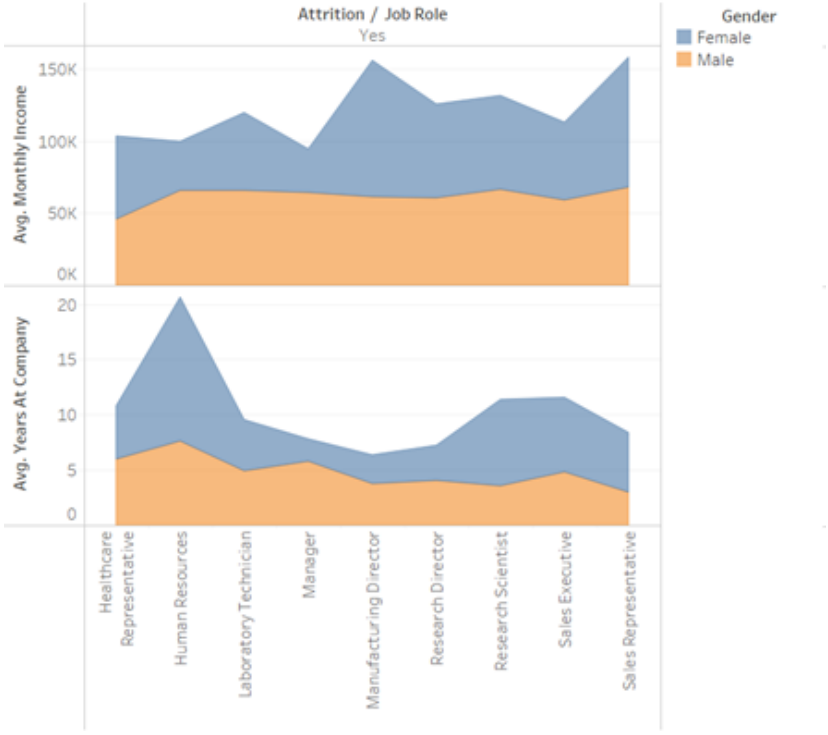
Stacked bar graphs:



These are two stacked bar graphs side by side. The first graph has Gender in the x-axis and the Average of years worked at company. The second bar graph has Gender in x-axis and Average of Monthly Income. In both the graphs, Attrition is set to yes.

The dataset has information mainly to analyze attrition trends of employees at XYZ Company. During the exploratory phase, upon running correlation, it has been observed that both monthly income and years at company are highly correlated to Attrition. After deciding on these variables, it was time to choose the kind of visualization to do. Area chart was chosen. But the problem was that, even though the area chart in theory should work, it was difficult to distinguish the values.

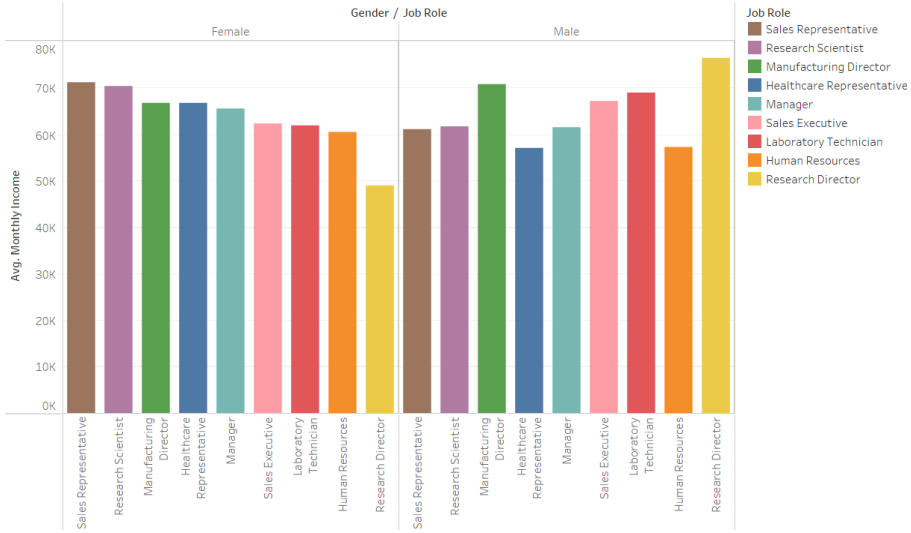
Area chart on attrition trends based on Gender, Role, Income and Years at company.



And then the idea of an area chart had to be dropped and the concept was transformed into a stacked bar chart as it makes it easier to distinguish and understand the data clearly.

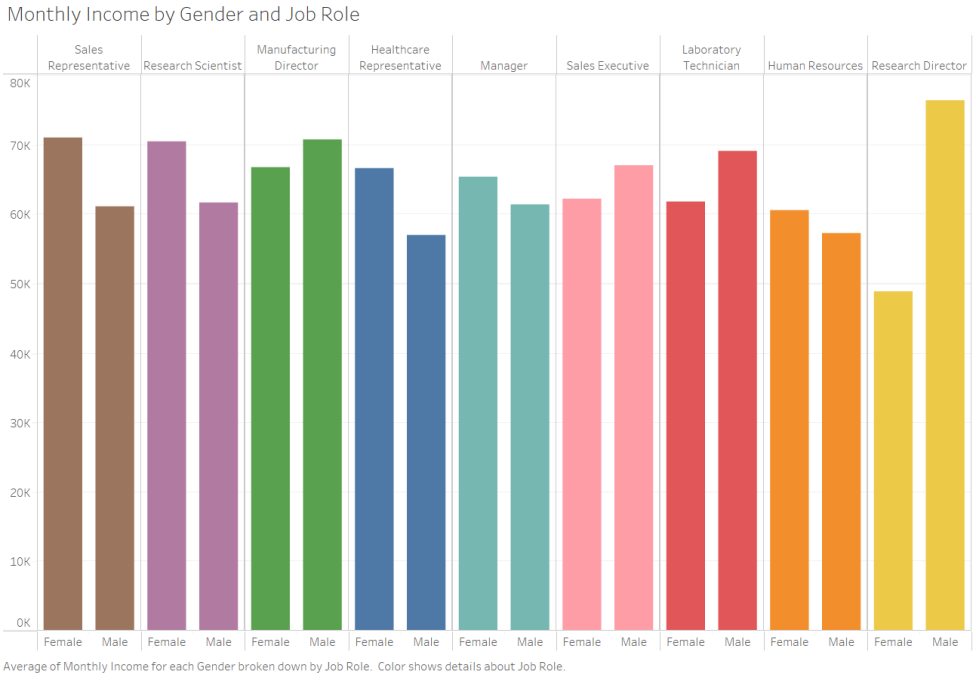
Salary Breakdown by Job Role and Gender

Part of the exploratory analysis was trying to determine if there was a discrepancy in income between male and female employees. When considering how to further break down the data, job role seemed to be a natural choice. Comparing raw salaries with no other identifying information is not very useful, but if a difference can be shown between salaries of men and women who occupy the same job role, that is a much more compelling story.

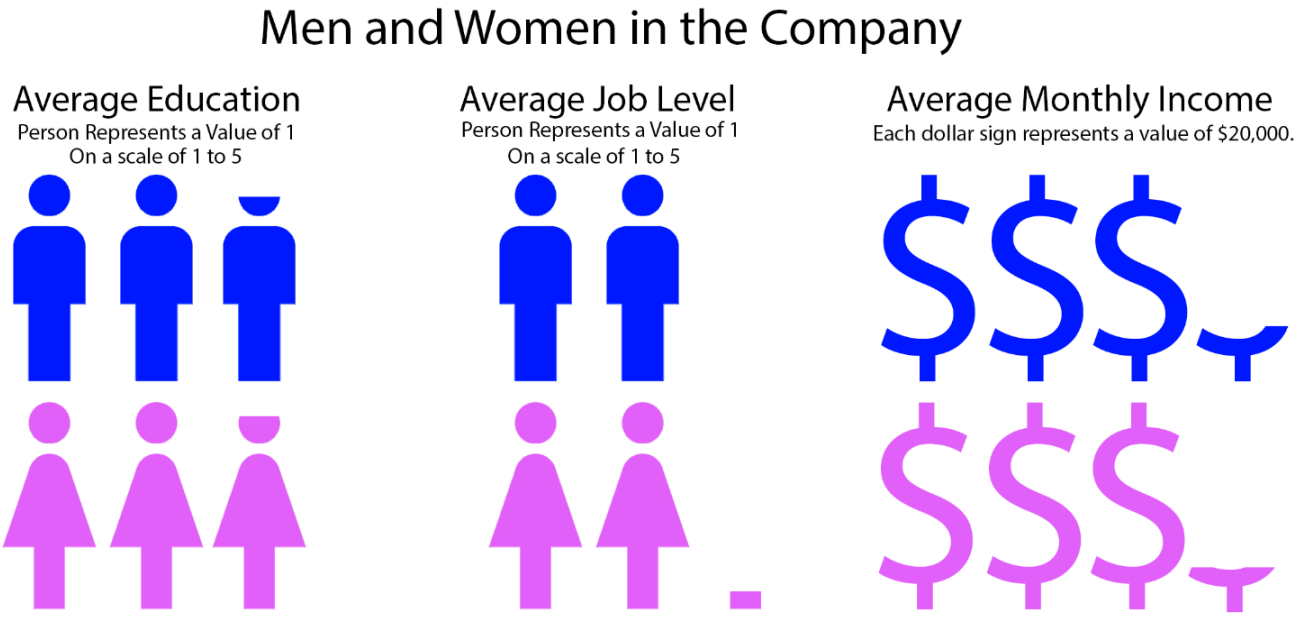


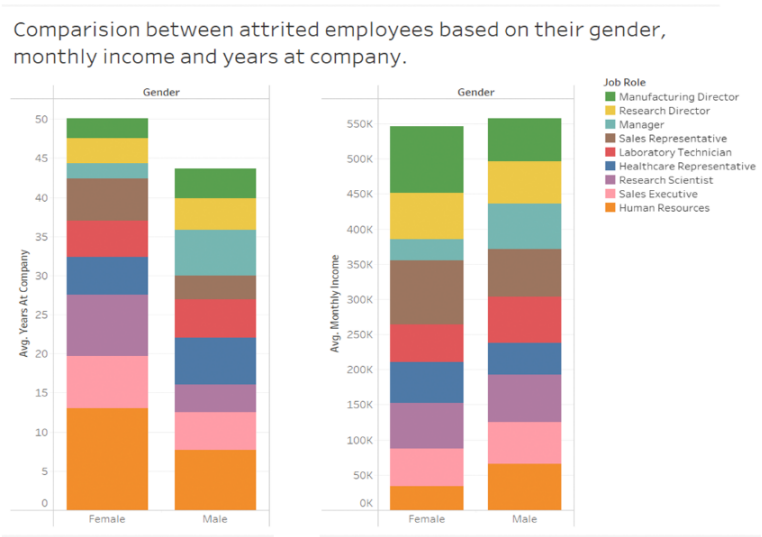
Here is a side-by-side bar chart to visualize this difference. This type of plot was chosen to clearly show the differences in salary between men and women for the same job role. It was color-coded by job role in order to easily compare salaries for the same job role. Once the data was plotted in this way, it quickly became apparent that men had higher salaries for all senior/executive level roles.

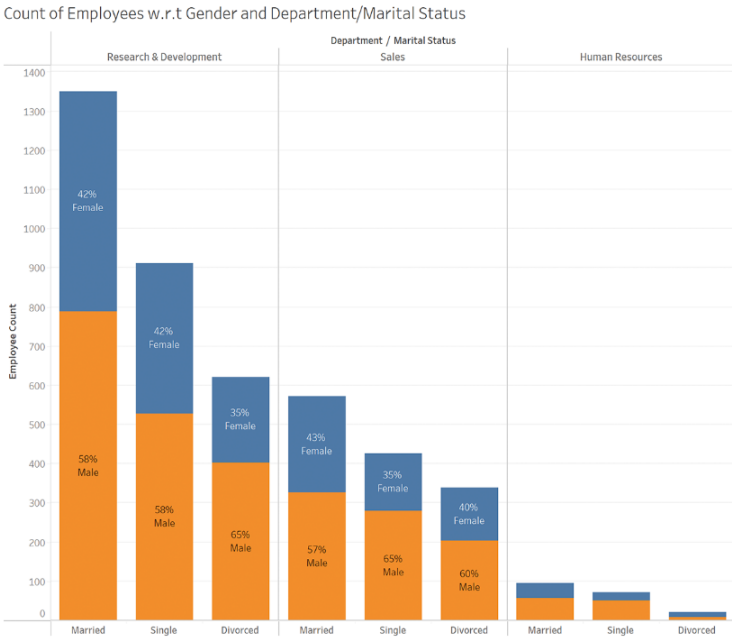
There were still weaknesses with this visualization, however. Grouping by gender first made it unnecessarily difficult to compare the different salaries. There was also redundant information on display with the legend and axis labels. Taking the Professor's feedback into account, this final visualization was produced.

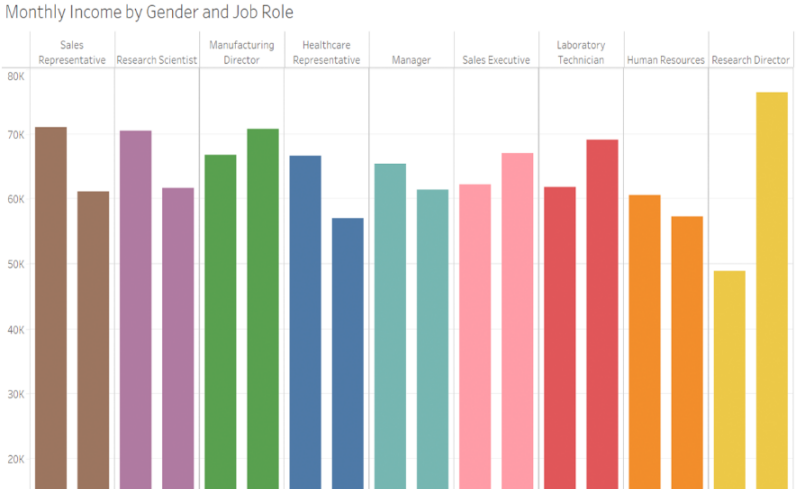


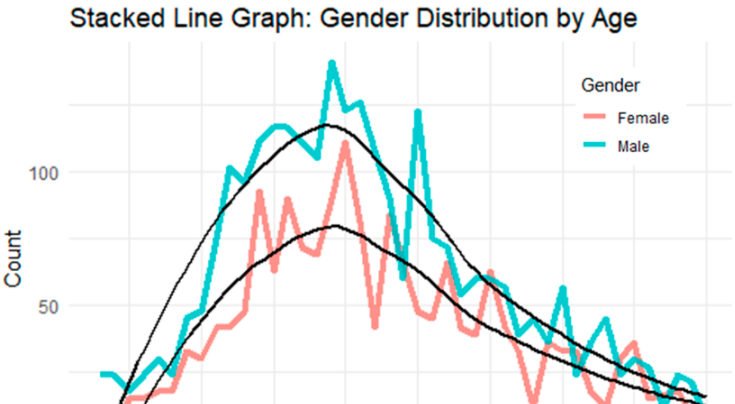
Final Visualization Analysis and Discussion











Analysis/Discussion

As mentioned earlier there was a lot of good feedback that was given on the first version of this visualization. It was initially meant to be a static view of an HR dashboard but after receiving some criticism the concept was changed to be an infographic. The topline containing the HR Dashboard view was the first thing to go. As a result, the focus of what is important changes. With an HR dashboard the user is someone who is working with the data so the important data is the most important aspect. Additionally, the inclusion of an isotype visualization does not really make sense in the context of an HR dashboard. However, in an infographic it does. This type of visualization is an attention grabber. It helps to hook an audience in with colors and shapes that one can immediately recognize and understand. It may seem simple, but that is the point. It's been shown that these types of visualization help with short term memory. As such, the choice was made to make this the top of the visualization. It is meant to get someone to look at the visualization and get hooked, then look down to see the more detailed information. Additionally, it serves as an introduction to what the rest of the visualization will be about.

After that, there is a stacked bar chart. This is placed here to get the viewer to see more of what the visualization will be. This stacked bar chart shows the breakdown of job roles between males and females compared to average years at the company and average monthly income. This helps to get the viewer immediately familiar with a lot of different things about the company. It shows them what roles are present at the company, what income is like for the roles between men and women, and how long men and women have been in those roles.

Next there is another stacked bar chart that shows employee counts broken down by men and women, department, and marital status. Once again this allows the viewer to get even more familiar with the makeup of this company. By now the viewer knows that men make more than women on average, women on average spend longer at the company, women have a higher education on average, the makeup of men and women in each department, the marital status of men and women in each department, and the makeup of job roles by men and women.

Moving on there is a side by side bar chart that shows the average monthly income by job role compared with men and women. This may seem a lot like the stacked bar chart from the second visualization but the important difference here is that the values are much more comparable. Because income, role, and gender are the sole focus, the viewer can compare the income between men and women to a finer degree than possible in the stacked bar chart. The decision to keep both took some time, but in the end they serve two different purposes. The stacked bar chart shows a general idea of the income of each role by gender while the side by side bar chart shows a precise figure. One to get an overall understanding, and one to be exact.

Finally, there is a stacked line graph that shows the gender distribution by age. This allows the audience to see the overall what age ranges make up this company. Doing this further helps the viewer understand the makeup of the company between men and women.