IBM Employee Performance and Attrition

Emily Darigo CS 544 Final Project

Dataset Details

The IMB Employee Performance and Attrition dataset is a fictional dataset obtained through kaggle. The data is developed by IBM data scientists to test their own models on real employee data. The data represents employees over a 10 year period. Parts of data, such as job role, department and standard hours are likely based off of real IBM employees. The dataset description does not specify how the IBM data scientists created each record.

Objective

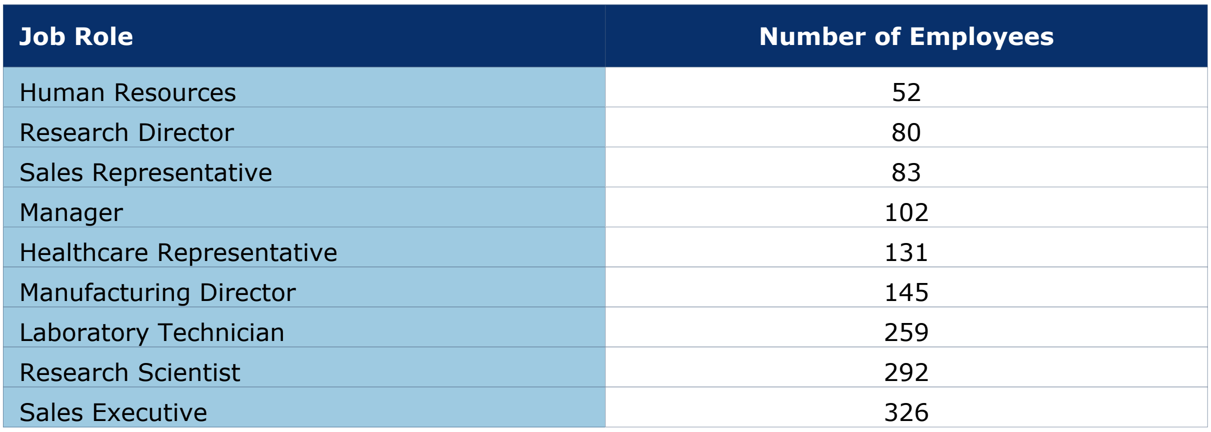
The purpose of this project is to analyze connections between various employee factors and if they have an influence on employee performance and attrition. The code used on the fictional data can be applied to live data in other organizations.

Variable Focus

There are a total of 35 variables in the dataset. In addition to performance and attrition, this project will focus on employee job roles, job satisfaction, number of years in their current role, and monthly income.

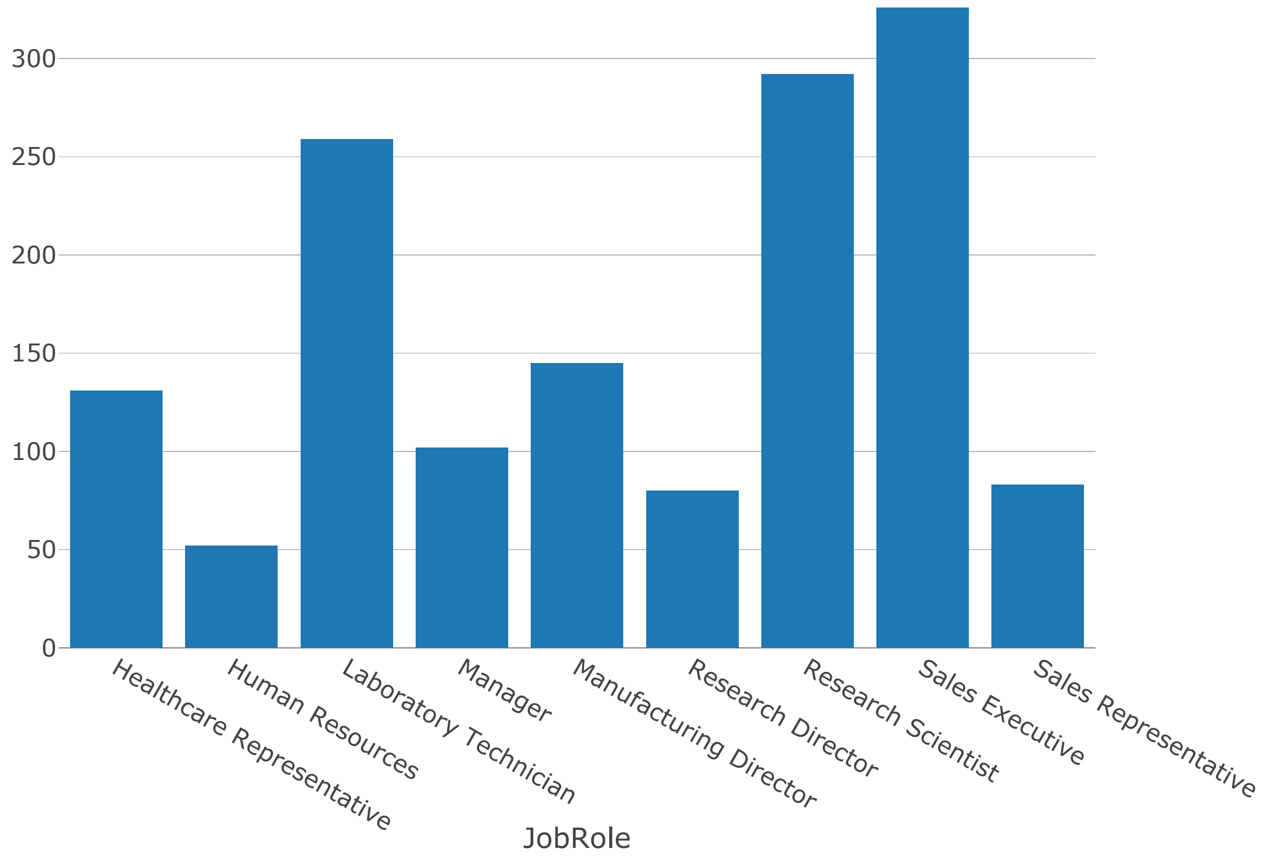
Analysis for one categorical variable

One influence on employee performance and attrition may come from the type of job. From the sorted table, most of the surveyed employees came from a sales executive, research scientist and laboratory technician position. Human resources represents the smallest group in the survey.



Job Role	Number of Employees
Human Resources	52
Research Director	80
Sales Representative	83
Manager	102
Healthcare Representative	131
Manufacturing Director	145
Laboratory Technician	259
Research Scientist	292
Sales Executive	326

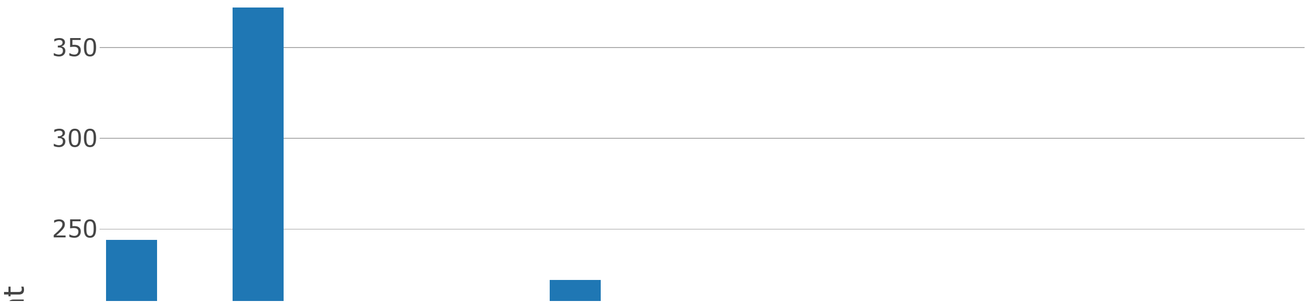
The bar chart is another representation of employees that are in each job at IBM. This chart shows how much more employees are a part of the Sales Executive, Research Scientist, and Laboratory Technician roles compared to the other roles.

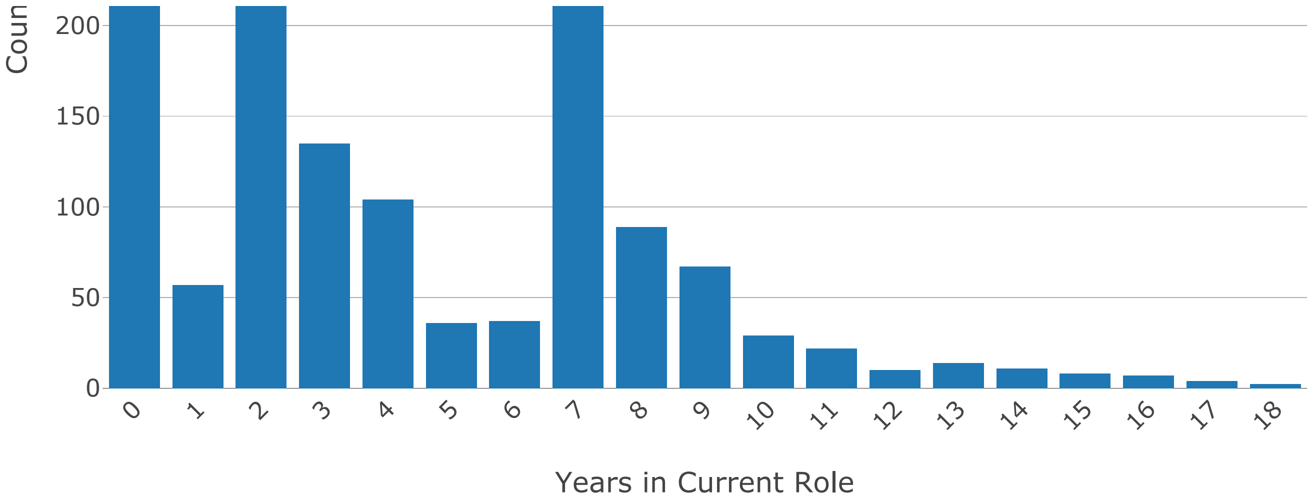


Analysis for one numerical variable

The length of employement at IBM can be another factor in attrition and performance rates. This dataset only has employees that have just started (have not yet reached 1 year) up to 18 years. In a real employee dataset, it is likely that there would be employees who have been with the company longer than 18 years, depending on the age of the company.

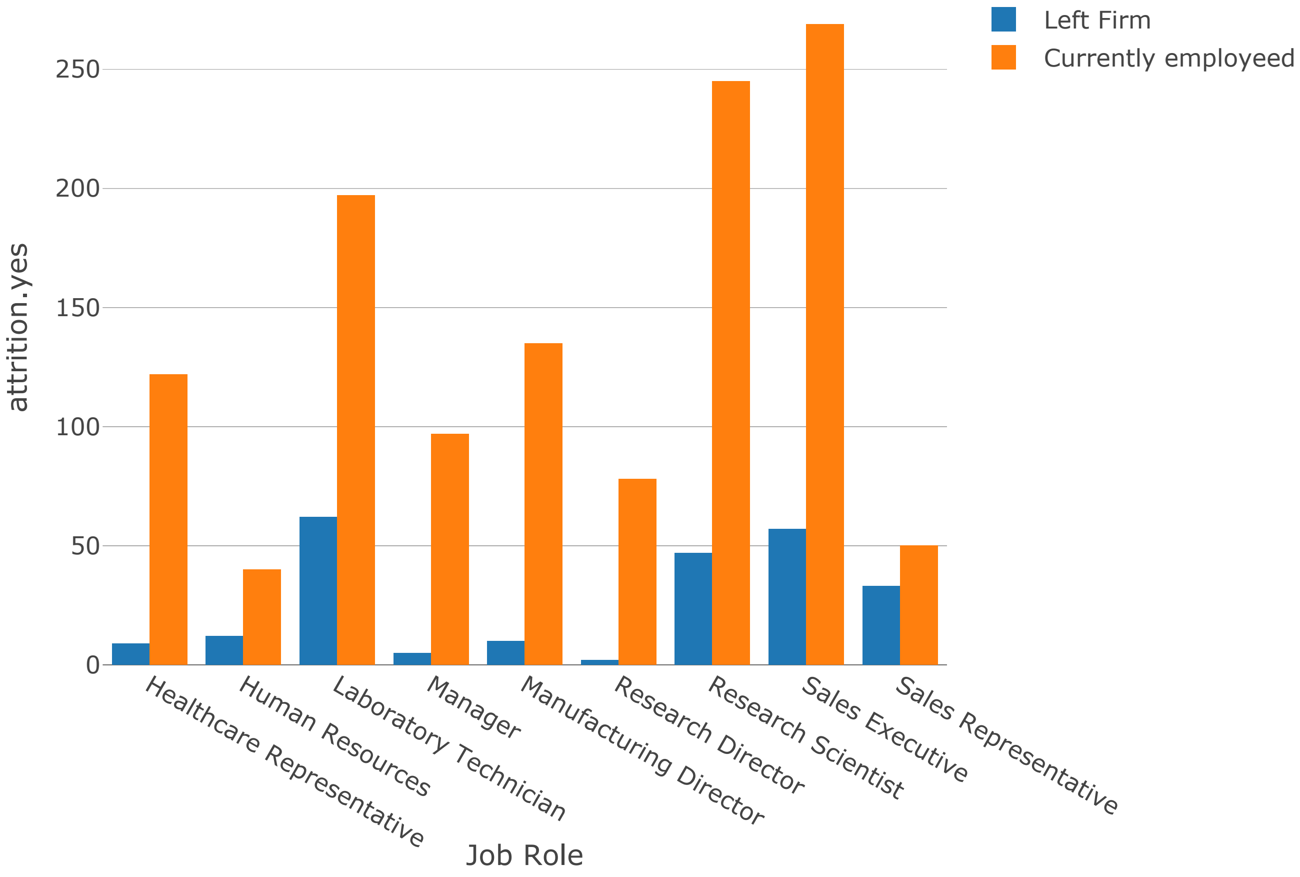
There are three noticable spikes at 0, 2, and 7 years. There are significantly less employees from the 10 year mark to the 18 year mark.





Job Roles and Attrition

To continue with the first categorical analysis, there may be some insight between the current role of an employee and if they left the firm or not.

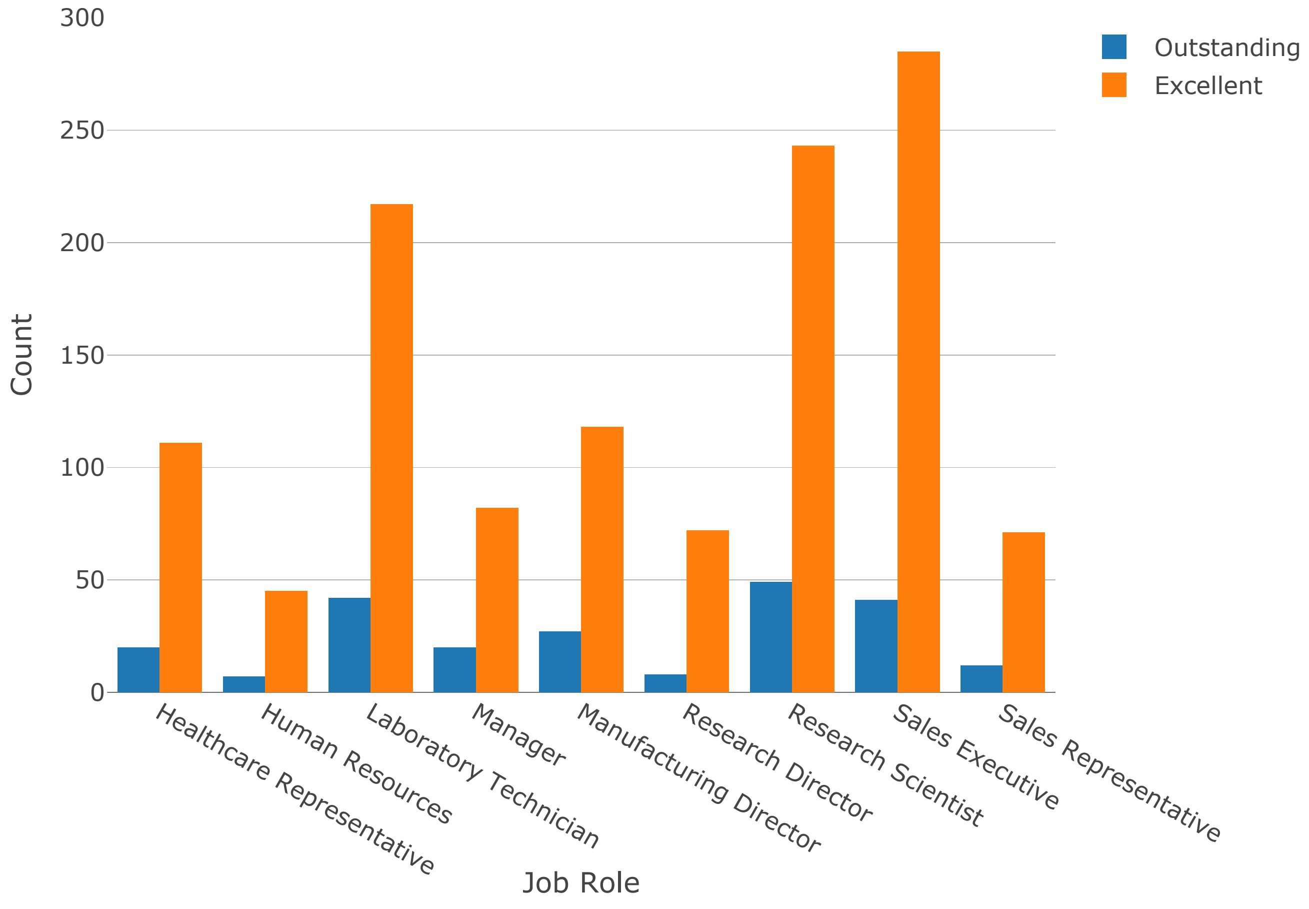


Overall, the attrition rate between job roles seems proportional, with the exeption of the role of Sales Representative. The table below shows the attrition rate for each job role. The roles with the most employees have similar rates. Research Director and Manager have the least amount of employees leaving. Sales Representatives have the highest rate, with almost 40% of their employees leaving.

```
##
## Healthcare Representative           Human Resources
##                0.06870229                0.23076923
##     Laboratory Technician                   Manager
##                0.23938224                0.04901961
##    Manufacturing Director         Research Director
##                0.06896552                0.02500000
##        Research Scientist           Sales Executive
##                0.16095890                0.17484663
##      Sales Representative
##                0.39759036
```

Job Roles and Performance

With job roles with performance rating, most employees received an “Excellent” rating (rating 3). Only about 10-20% of employees received an “Outstanding” rating (rating 4). Research Directors had the least amount of outstanding ratings (10%) and Managers had the most amount of outstanding ratings (19.6%).

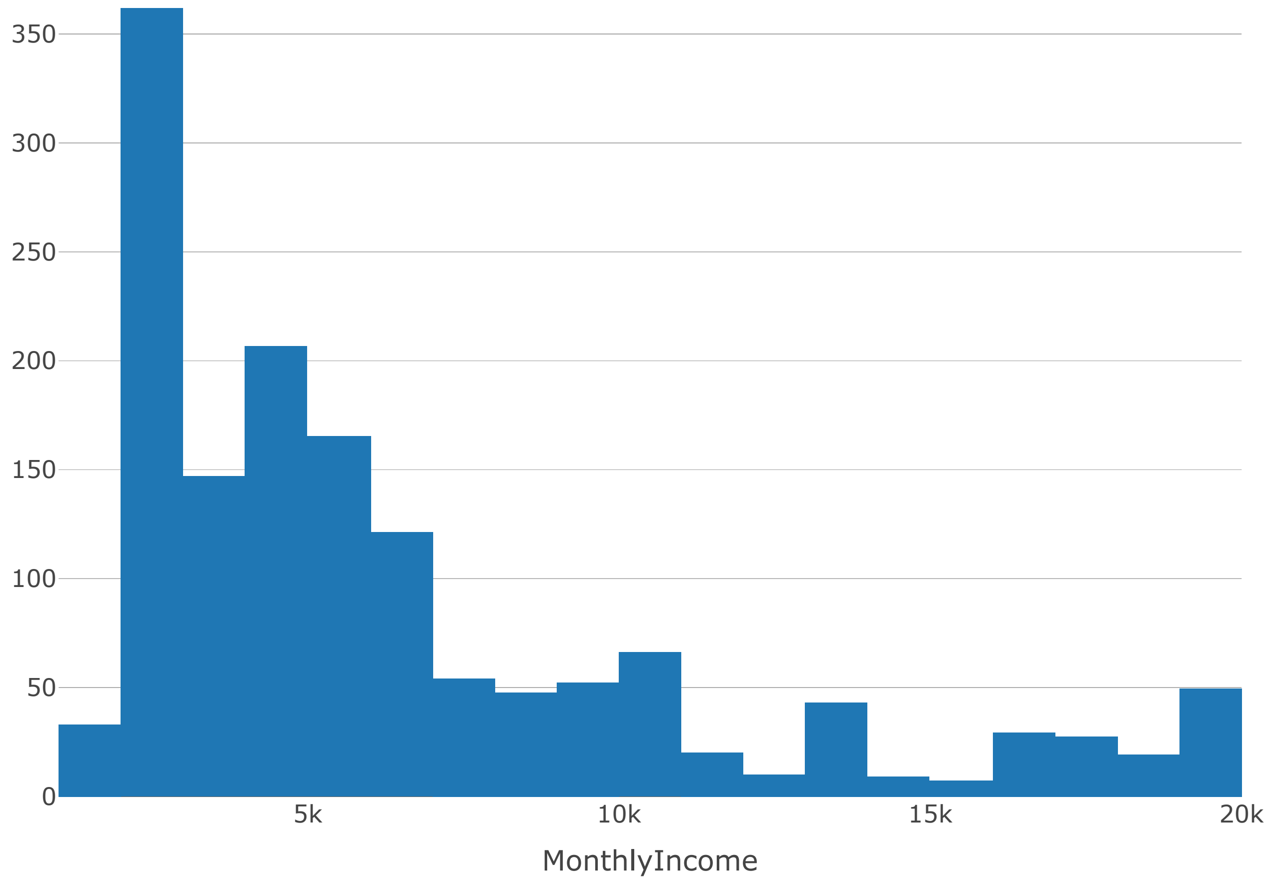


```
##
## Healthcare Representative            Human Resources
##                 0.1526718                  0.1346154
##     Laboratory Technician                    Manager
##                 0.1621622                  0.1960784
##    Manufacturing Director          Research Director
##                 0.1862069                  0.1000000
##        Research Scientist            Sales Executive
##                 0.1678082                  0.1257669
##      Sales Representative
##                 0.1445783
```

There may be a connection between an employees role, performance and attrition. Those roles where people received more outstanding ratings tended to have lower attrition rates.

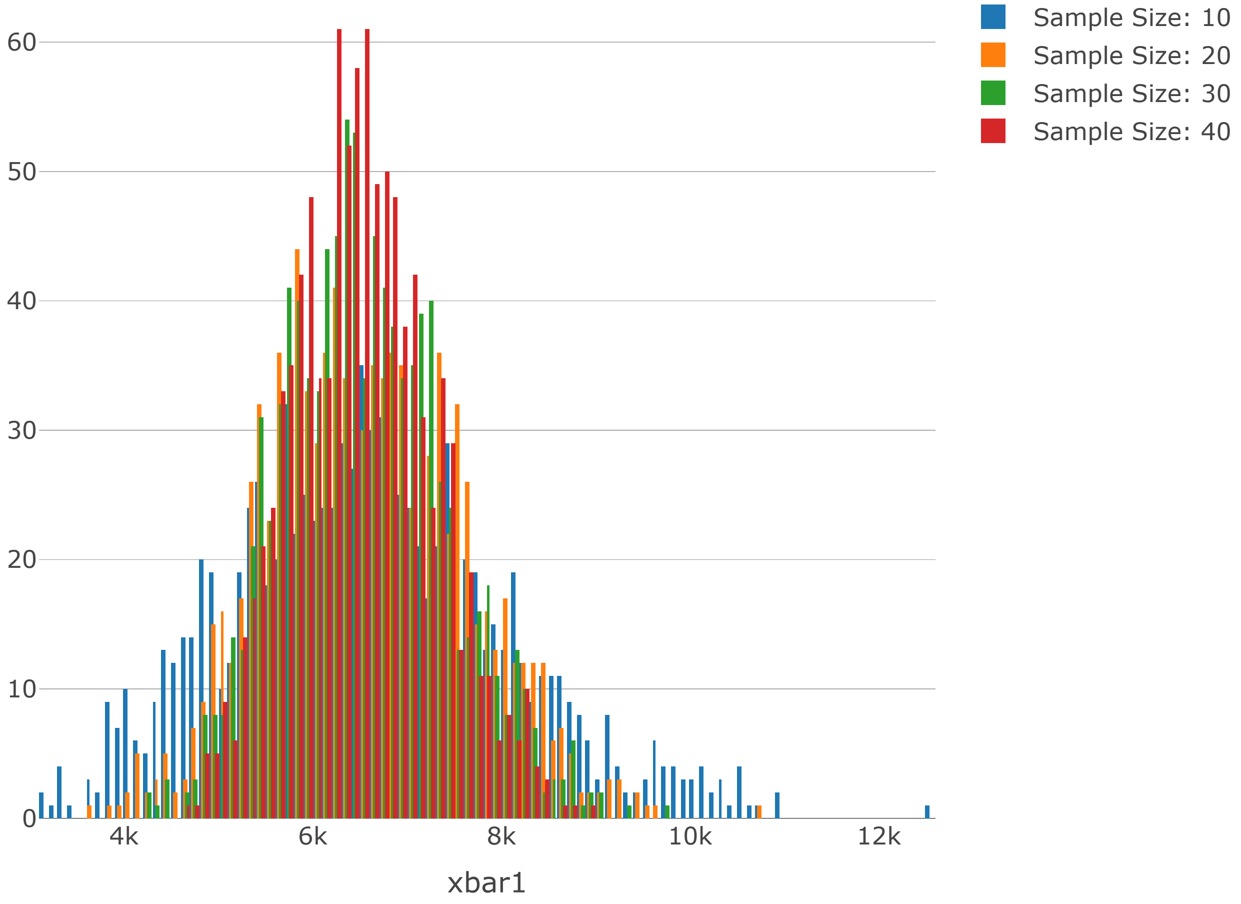
Distribution of numerical data

The current distribution of employees' monthly income is right skewed, where most people are making less than $10,000 per month. This is expected, as there are less employees in upper management, who make more than $10,000 per month.



Random Sampling, Central Limit Theorem

The central limit theorem can be applied to the monthly income variable. With a sample size of 10, more values are closer to the mean than the original distribution. The original standard deviation was 4707.957 and the size 10 sample is 1489.324. With each increase in sample size, the mean continues to stay about the same, but decreases in standard deviation, bring the values closer to the mean. The sample size of 40 has the most similar mean compared to the initial data.



```
##   Initial data  Mean =  6502.931  SD =  4707.957
##   Sample Size = 10  Mean =  6602.255  SD =  1489.324
##   Sample Size = 20  Mean =  6566.372  SD =  1055.739
##   Sample Size = 30  Mean =  6542.483  SD =  872.6016
##   Sample Size = 40  Mean =  6523.685  SD =  738.8162
```

Sampling Methods

The following sections will look at the different kinds of sampling for the employee dataset. To see how sampling affects the data, the mean and standard deviation will be calculated for each type of sampling method. This will be compared to the entire dataset.

Simple Random Sampling

For this method of sampling, the number of years in an employee's current role will be compared.

Frequencies of Years in Current Role from Sample

```
## 
##  0  1  2  3  4  5  7  8  9 10 14 15 
## 14  1 24  9  8  7 16 10  6  2  1  2 
```

Frequencies of Years in Current Role from Dataset

```
## 
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17 
## 244  57 372 135 104  36  37 222  89  67  29  22  10  14  11   8   7   4 
##  18 
##   2 
```

The mean and standard deviation are similar to each other. However, due to the sampling method, years 16, 17 and 18 were not included in the sample. This may affect other analysis if not all years are included.

```
##  Initial data  Mean =  4.229252  SD =  3.623137
##  Simple Random Sampling  Mean =  4.53  SD =  3.459594
```

Systematic Sampling

The Job Satisfaction variable will be compared between the sample and the whole dataset.

```
## [1] 15
```

```
## [1] 7
```

```
## [1] 0 0 0 1 0 0 1 0
```

Frequencies of Job Satisfaction from Sample

```
## 
##  1  2  3  4 
## 22 20 33 25 
```

Frequencies of Job Satisfaction from Dataset

```
## 
##   1   2   3   4 
## 289 280 442 459 
```

Proportions from Systematic Sampling

Similar to random sampling, the mean and standard deviations were similar before and after sampling. However, the method pulled the least amount of samples from the largest section (rating 4). This may not represent the dataset the best, as most people gave a score of “Very High” for job satisfaction.

```
## 
##          1          2          3          4 
## 0.07612457 0.07142857 0.07466063 0.05446623 
##  Initial data  Mean =  2.728571  SD =  1.102846
##  Systematic Sampling  Mean =  2.61  SD =  1.090779
```

Stratified Sampling

```
## 
##   1   2   3   4 
## 284 287 453 446
```

```
## 
##  1  2  3  4 
## 19 20 31 30
```

```
## [1] 100
```

```
## Stratum 1 
## 
## Population total and number of selected units: 284 19 
## Stratum 2 
## 
## Population total and number of selected units: 287 20 
## Stratum 3 
## 
## Population total and number of selected units: 453 31 
## Stratum 4 
## 
## Population total and number of selected units: 446 30 
## Number of strata  4 
## Total number of selected units 100
```

	EnvironmentSatisfaction	ID_unit	Prob	Stratum
	<int>	<int>	<dbl>	<int>
19	1	19	0.06690141	1
37	1	37	0.06690141	1
46	1	46	0.06690141	1
72	1	72	0.06690141	1
77	1	77	0.06690141	1
85	1	85	0.06690141	1
87	1	87	0.06690141	1
88	1	88	0.06690141	1
121	1	121	0.06690141	1
122	1	122	0.06690141	1

1-10 of 100 rows Previous **1** 2 3 4 5 6 ... 10 Next

Frequencies of Environment Satisfaction from Sample

```
## 
##  1  2  3  4 
## 19 20 31 30 
```

Frequencies of Environment Satisfaction from Dataset

```
## 
##   1   2   3   4 
## 284 287 453 446 
```

Proportion of Stratified Sampling

Compared to systematic sampling, the stratified sampling has more even proportions from the dataset.

```
## 
##          1          2          3          4 
## 0.06690141 0.06968641 0.06843267 0.06726457 
```

Compared to the other methods of sampling, stratified sample is the most similar to the dataset mean and standard deviation.

```
##  Initial data  Mean =  2.721769  SD =  1.093082
##  Systematic Sampling  Mean =  2.72  SD =  1.09249
```