HW 5

Md Easin Hasan

I obtained a data set from Kaggle which is on human resource analytics. This dataset concerns a big company that wants to understand why some of their best and most experienced employees are leaving prematurely.

Step 1: Inspect your data, assess it for completeness, good formatting, and any errors. You should be confident after this step that the data is in good form and ready for visualization.

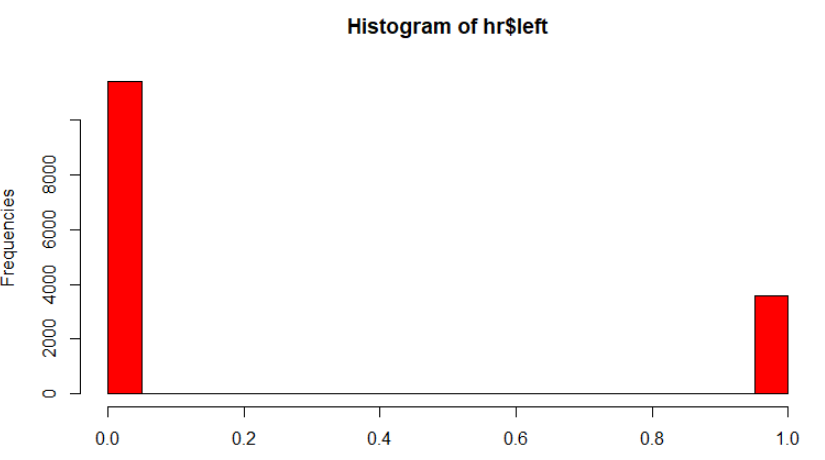
The dimension of this dataset is 14999 by 10 so this a complete dataset with good formatting. We did not find any error on this dataset. This dataset has no missing values. The variable names of this dataset are: "satisfaction\_level", "last\_evaluation", "number\_project", "average\_montly\_hours" "time\_spend\_company", "Work\_accident", "left", "promotion\_last\_5years", "sales", "salary". We also figured out which variables has character instead of a numeric value.

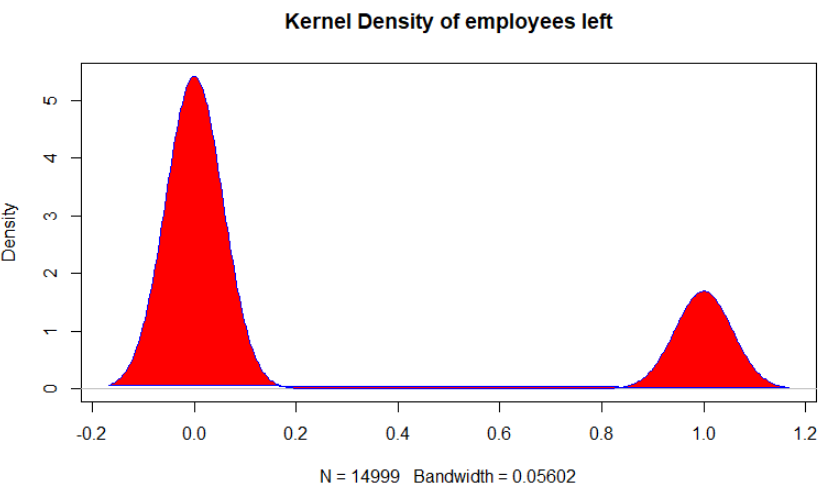
Step 2: Use some low level data visualizations (histograms, boxplots, density plots) to explore data features. Consider some data wrangling that may be needed. For example, if you notice dispartiy of scale across variables you plan to visualization, you might considering rescaling the data or transforming certain variables. If the data are skewed or exhibit extreme points, you should consider some transformations as well.

> (table(hr$left)/(length(hr$left))\*100)

0 1

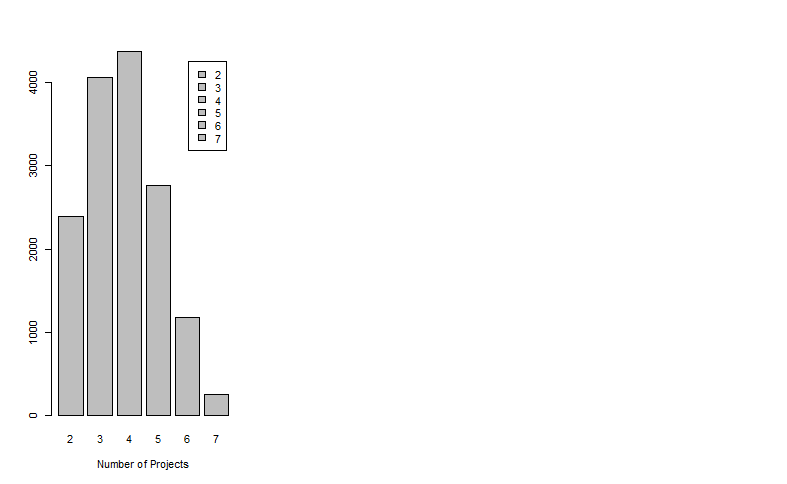
76.19175 23.80825





**Interpretation:**

We can see that almost 23.80% employee left the job.

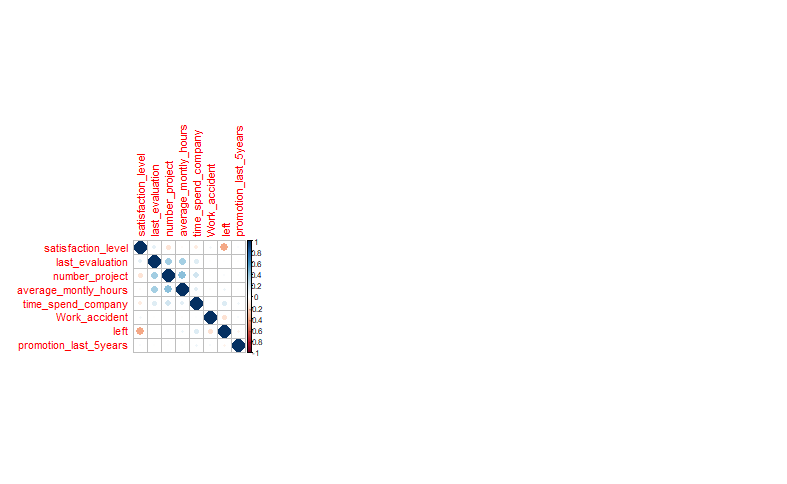


From this plot we can see how many employees working on the how many projects.

Step 3: Review the design principles we addressed in this course. Make a game plan for how you will follow. Note principles that are more or less applicable to this specific visualization. Note features of the data that should be accounted for in the visualization (outliers, unusual patterns, etc….).

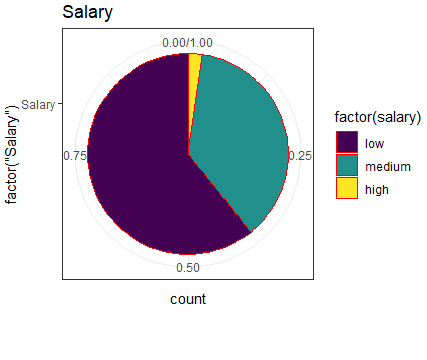
I will follow the Tufte Principle 3 of Maximize data density and the size of the data matrix, within reason. I can use small multiple graphs or parallel sequencing skill for side-by-side graph which is the Tufte principle of 4. I would like to apply the Tufte principle of 5 for providing the user with an overview and details on demand. Finally, I can edit, revision, and test against different design by grouping with colors and using color to separate graphs.

Step 4: Decide on a first attempt of a visualization. This should be a direct approach (eg not yet sophisticated) and should address the message you need to promote with the visualization. Begin with a basic plot and then think about how to message using layering. For example, if you want to show that people with high cholesterol and high BMI are more prone to heart disease, you might begin with a scatterplot (basic) then think about how you will add in the information about heart diseases by layering. Fully consider all of the features of the data you need to communicate at this stage and think about how you will include all of this information.



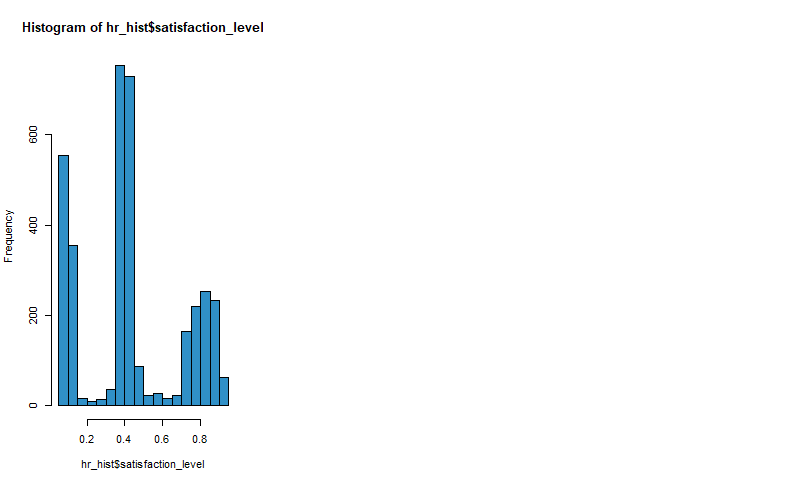
**Interpretation:**

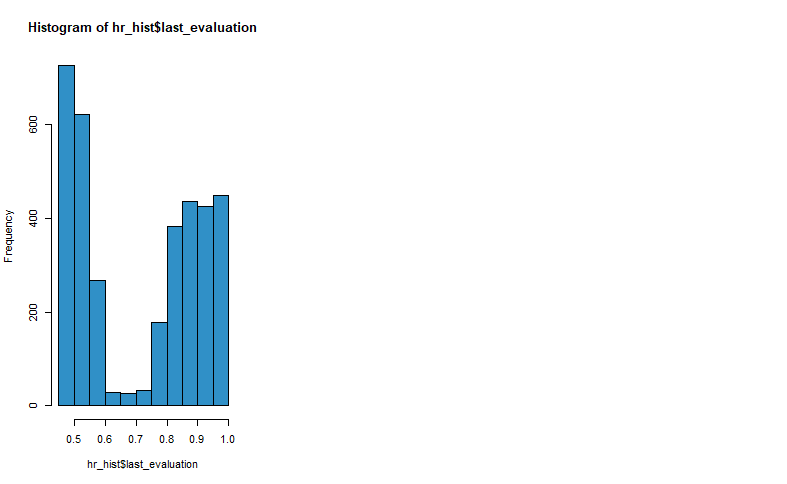
From this correlation plot we can see the correlation between the variable in this dataset.

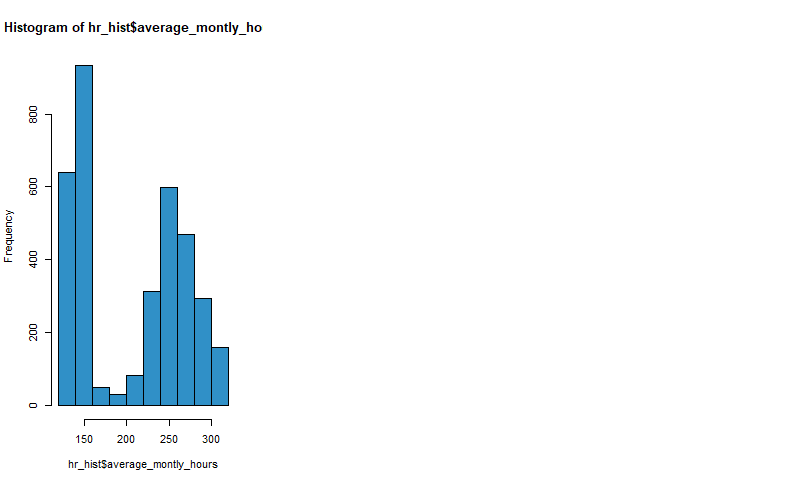


**Interpretation:**

We can observe that the people who left the job among them only 1% has high amount of salary, 62% has low amount of salary, 37% has medium amount of salary.



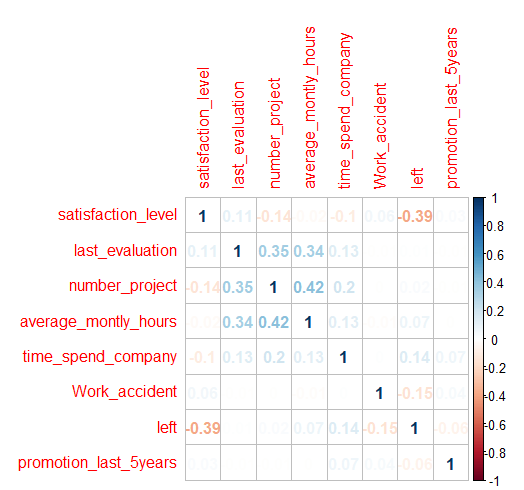




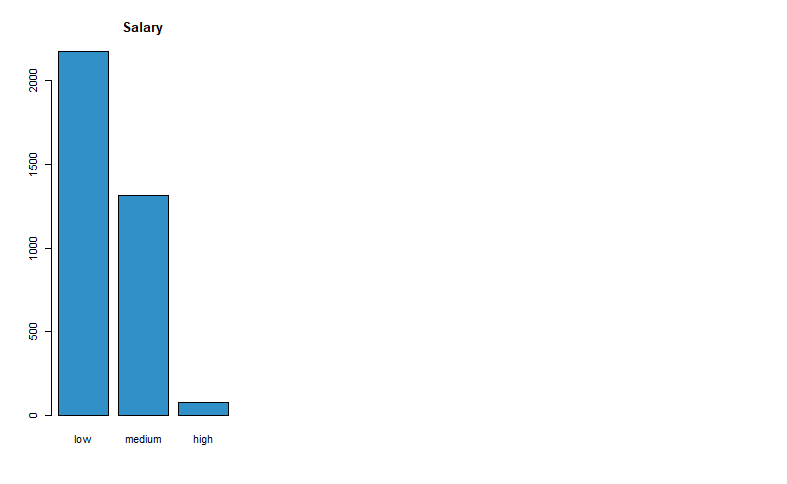
**Interpretation:**

From the graph, we can see we do not want to retain everybody since some people don’t work well as we can see from their evaluation, but clearly there are also many good workers that leave.

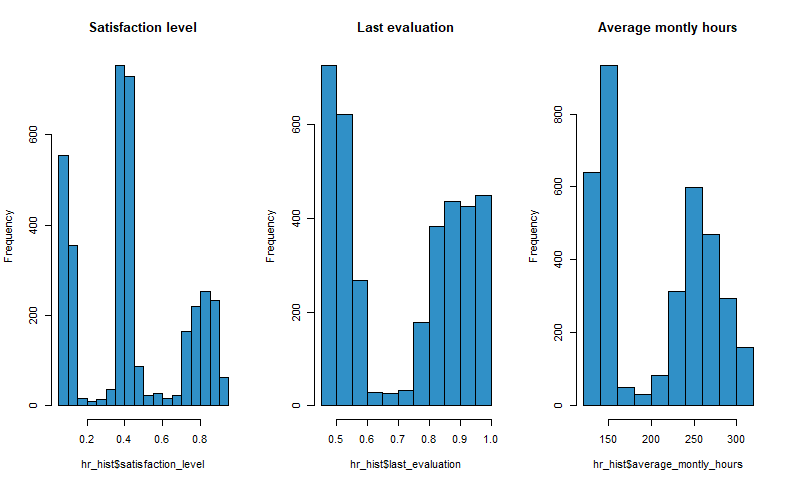
Step 5: Redesign you visualization by layering and editing the first attempt. Continually review the design prinicples so that you can make sure these attempts at redesign are improving (and not just changing) the data visualization.



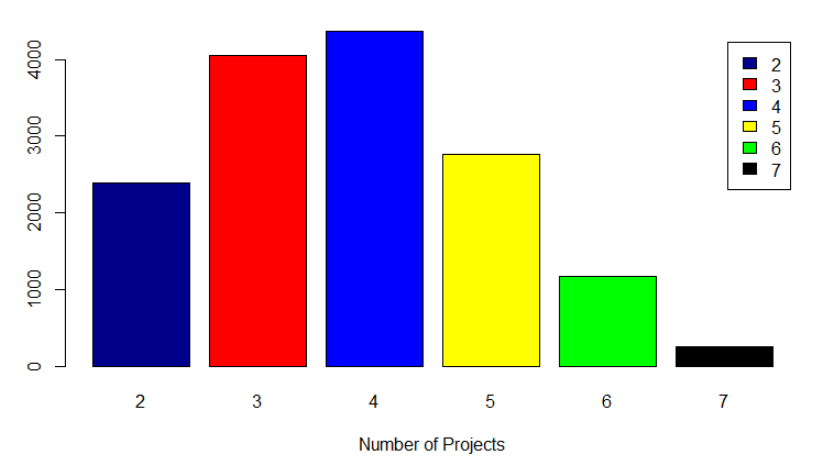
This is the redesign of the above correlation plot; I believe it is more visible since we can easily see the correlation between variables.



Instead of the pie chart of salary category I would like to keep the above bar plot since we can very easily visualize that the low salaried employees left the job compare to medium and high salary.



Instead of 3 different graphs above here I visualized all graphs side-by-side graph according to the Tufte principle of 4. In addition, added the main title so that we can easily visualize the variables of the dataset.



We have redesigned the bar plot of the number of employees working on the number of projects. Here after adding different colors, the bar plot is more visible compare to the first one.

Step 6: Document the data visualization on your portfolio and write out a COMPLETE description of the insights a person should gain from the data visualization. Check yourself at this stage to be sure that a person really can gain these insights from your visualization!

We can figure out what percentage of employees left the job. We can see there are only few employees working on 7 projects, and highest number of employees working on 4 projects. We can observe that the people who left the job they had low amount of salary. From the graph, we can see we do not want to retain everybody since some people do not work well as we can see from their evaluation, but clearly there are also many good workers that leave. Finally, the employee who leaves their satisfaction level was low and worked on many projects and did not get promotion in the last 5 years.