**Problem Statement:**

Company ABC has over 1,300 employees located in four offices around the globe. ABC would like to understand what factors, if any, correlate with employee performance. ABC would like to you to use email communication data as a proxy for employee collaboration and worked hours and skills to understand performance.

ABC has data on its employees contained within the Employee\_Roster\_Data.xlsx file, employee email data contained within Email\_Data.txt, employee worked hours on each client within hours.csv and employee skills data within skills.xlsx (see each file for explanation of variables).

● Each row within Employee\_Roster\_Data describes attributes about an individualemployee

● Each row within Email\_Data describes an individual email sent by a person (either an employee or person not employed at ABC).

● Each row within hours describes the hours spent by an employee in admin tasks and on each different client. Also includes the target total hours.

● Each row within skills describes a skill level for a given employee.

You will see within Employee\_Roster\_Data that each employee has a unique Email\_ID which corresponds to the from\_ID and to\_ID within Email\_Data. If an ID within Email\_Data does not appear within the Email\_ID column of Employee\_Roster\_Data, it is because that individual person is not an employee at ABC (e.g. a vendor sending an email to an employee, or an employee sending an email to a customer).

**Assumptions:**

As ABC try to understand if there is a correlation with employee performance I tried to check if email, worked hours and skills data correlated with it.

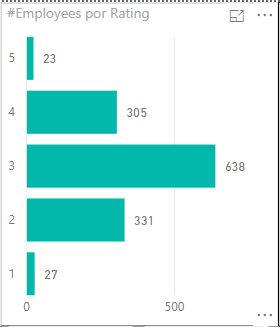
I used Employee\_Roster\_Data Rating field for performance, as it was described as “Employee's performance rating. 1 = lowest performance; 5 = highest performance”; Utilization, Client Time and Admin Time indicators for worked hours; Attribute Level from skills; and count of email sent from Email\_Data for employee collaboration.

As there was no detail on the period in which Rating field was measured, I assumed that it corresponded to the same period of time of hour’s data. Email data has no dates, so I assumed that it corresponded to the same period of time in which Rating was measured.

**Description of ABC Employees**

A total of 1324 employees.

The rating (performance evaluation) is measured with discrete values between 1 (less performant) and 5 (more performant). The average rating among employees is 2.97, with the following distribution.



Average Client Time by Employee is 71,86%

Average Admin Time by Employee is 15,31%

Average Utilization by Employee is 84,85%

Average Email Sent by Employee is 1.63K

**Findings:**

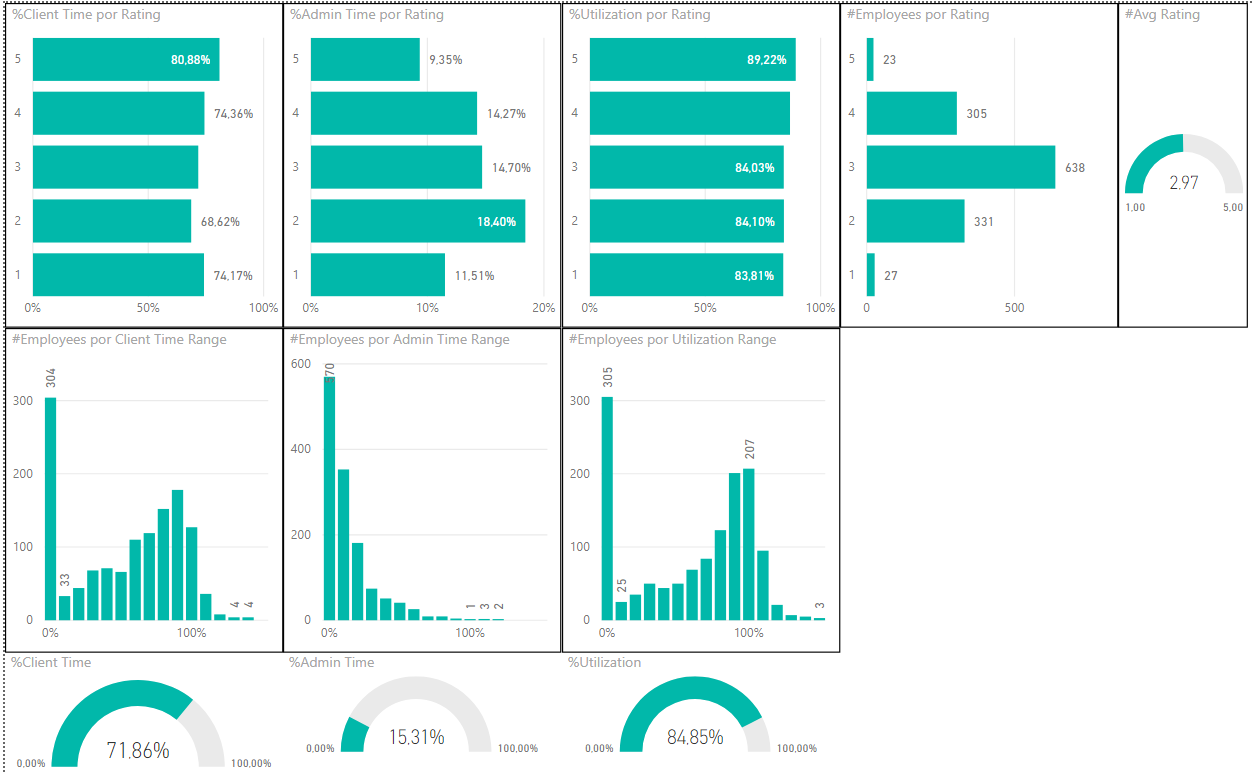
Time Indicators vs Rating

My first approach to check correlation was to use scatter charts comparing hour’s indicators vs rating.

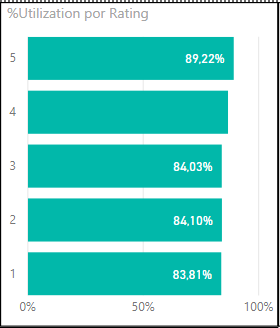


At first glance, there is no clear evidence of correlation between time indicators and rating. Elements in each class of rating are scattered, and when filtering them distributions of time indicators are similar.

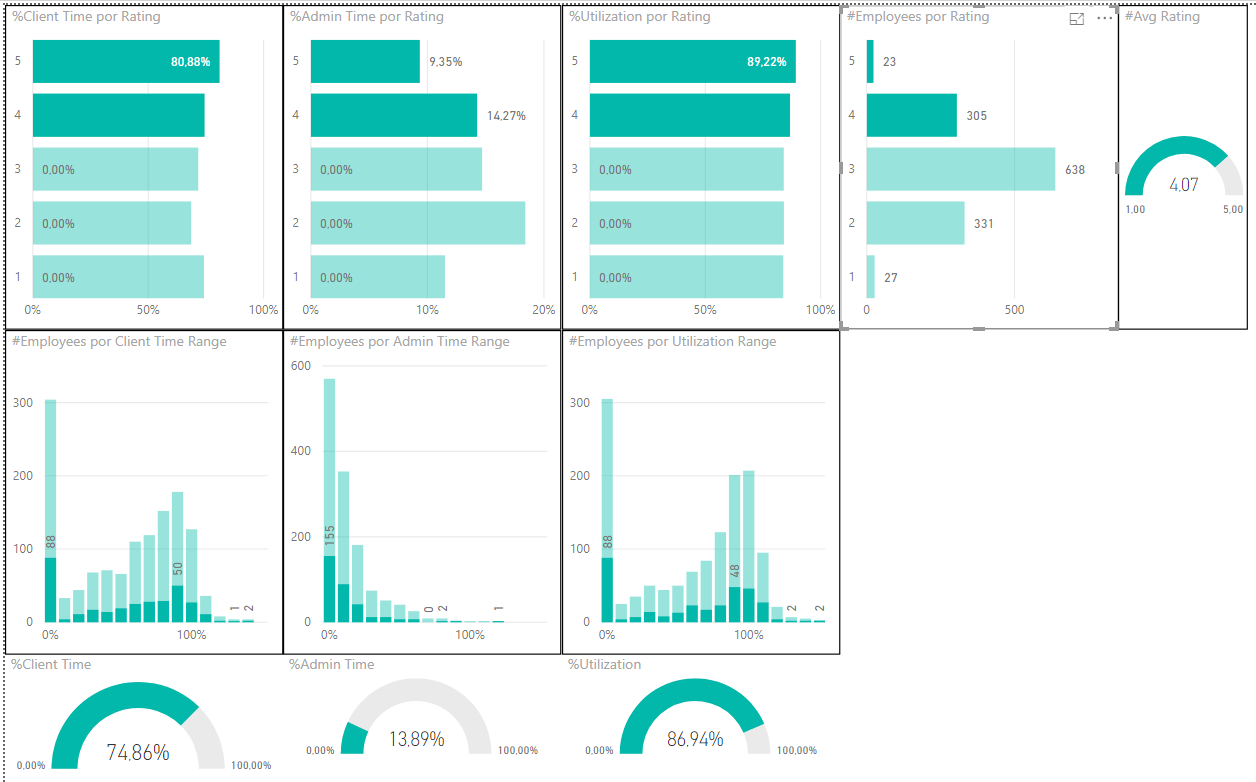
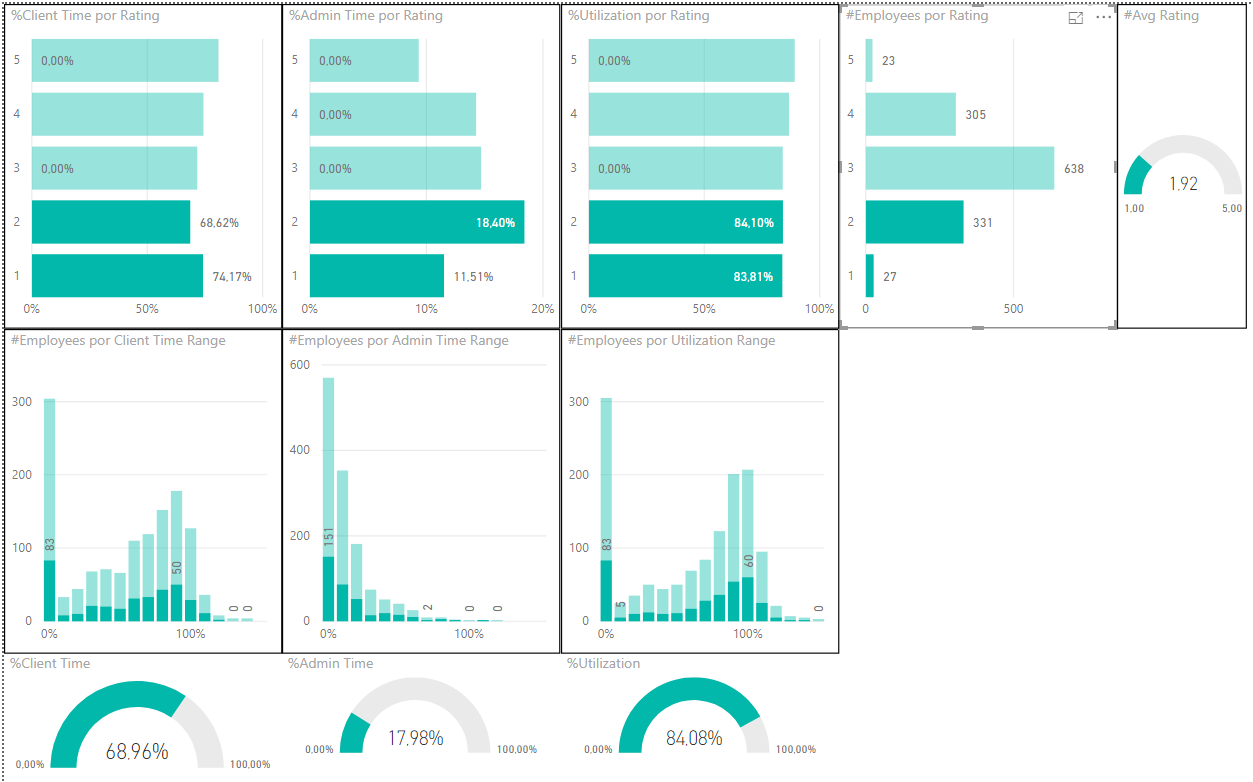
So I tried to analyze if average of time indicators was moving accordingly Rating change.



It is not clear to have a linear correlation between Client Time and Rating, and between Admin Time and Rating. But we can see average of utilization moving up accordingly Rating rise.



Going back to correlation between Client Time and Rating, and between Admin Time and Rating we can see that if we remove the extremes, there's a trend. And as the extremes have a very low frequency, if we analyze the values above and below the average rating

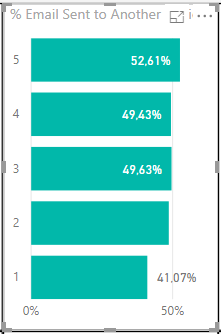
 

Employees with rating above average has higher Client time and lower Admin time, while employees with rating below average has lower Client time and higuer Admin time.

Email collaboration vs Rating

Taking into account my previous experience with scatter plots, in the case of emails I first looked at the averages, seeing that there was no obvious correlation. I tried to check if collaboration between distinct regions, Offices and Departments has an impact on Rating.

The only finding is that the top performers seem to send more mails than the average to regions other than their own, and the bottom performers send most of their mails to their own region. But the difference against the average is so small that I don't think it's right to talk about a trend.



Hours vs Email

Going to the scatter charts correlation between Hours indicators and Email is not evident.



We can see there is no clear trend and average email sent is similar in distinct beans of Client time, Admin time and Utilization.



Recommendations

Since all emails in the EmailData file had a corresponding id in the employee master, we only scan internal emails. To complete the analysis it would be interesting to include external emails (maybe client related).

Since the time information has a date field, it would be interesting if the mails also have a date field, to link the time range between both data sources.

Notes

This was my first PowerBI dashboard, as I usually work with Qlikview. Having learned the basics of PowerBI these days, I tried my best, although I apologize if it's a little poor and messy.

I tried to solve what was necessary to address the problem, but I found myself with many questions such as where I should solve the data modeling, if in SQL or with PowerQuery (in Qlikview I can do all the data processing from end to end), DAX expressions, Interactions, Segmentations, etc.