

Chase Switzer

Tyler Craig

Logan Oh

Sam Moreira

Avery Robbins

I. Analysis Overview

Questions addressed in the analysis:

- 1) Is there a pay gap based on gender or race for similar positions and experience levels?
- 2) How are we doing on employee retention?
- 3) Which survey is a better indicator of employee longevity, the Engagement Survey or the Employee Satisfaction 2 survey?
- 4) If we put employees into groups based on their months of service, do we see any correlation between how long they've been here and the results of the latest satisfaction survey (or the engagement survey)? Can we identify what the members of each group have in common?
- 5) Are we doing a good job on diversity overall? If not, which of our recruiting sources is having the biggest benefit? Where should we be focusing future recruitment efforts?

Here is the data that we used in this analysis:

<https://raw.githubusercontent.com/byui-cse/cse450-course/master/data/hr.csv>

Here is the data dictionary for the data (document that describes the columns in the dataset):

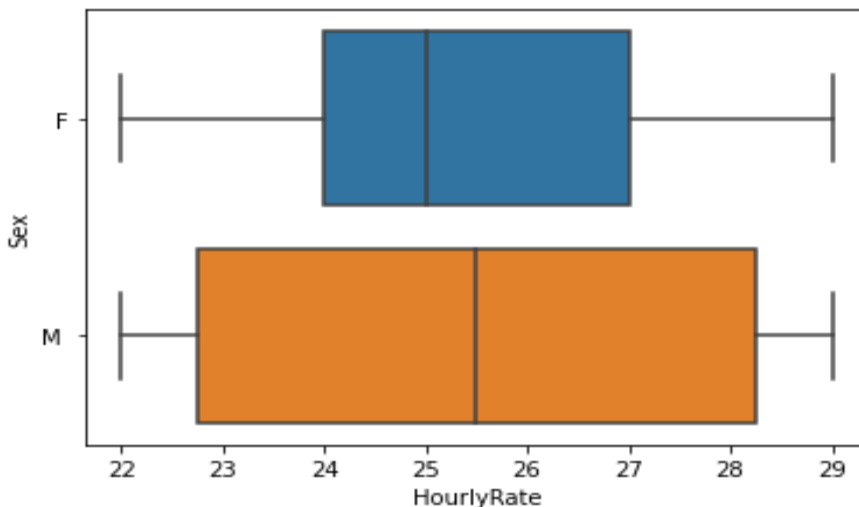
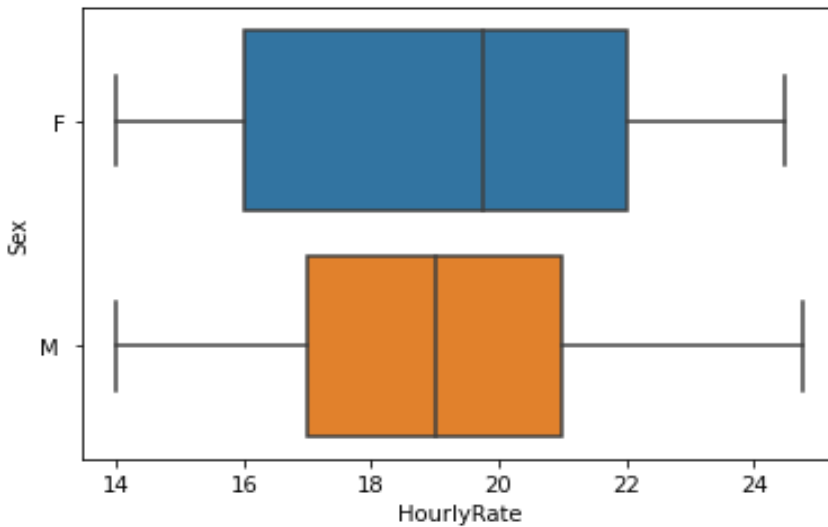
<https://byui-cse.github.io/cse450-course/module-02/hr-dictionary.txt>

II. Equitable Pay Analysis (Chase)

The question I am answering is this: Is there a pay gap based on gender or race for similar positions and experience levels? In order to answer this question, I filtered employee data into four main features: Sex, race description, pay-rate, and position.

Unfortunately, many job positions have scarce amounts of people, so I wasn't able to test every job position. The jobs that had the largest pool of races, Production Technician I (136) and Production Technician II (57), will be tested here for the best results. Here is the data for pay

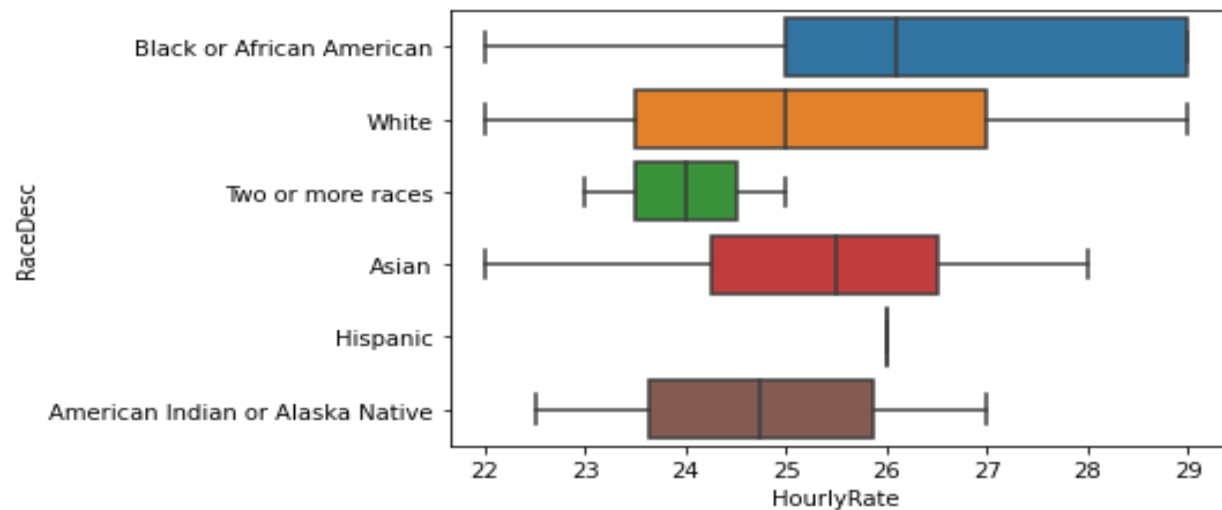
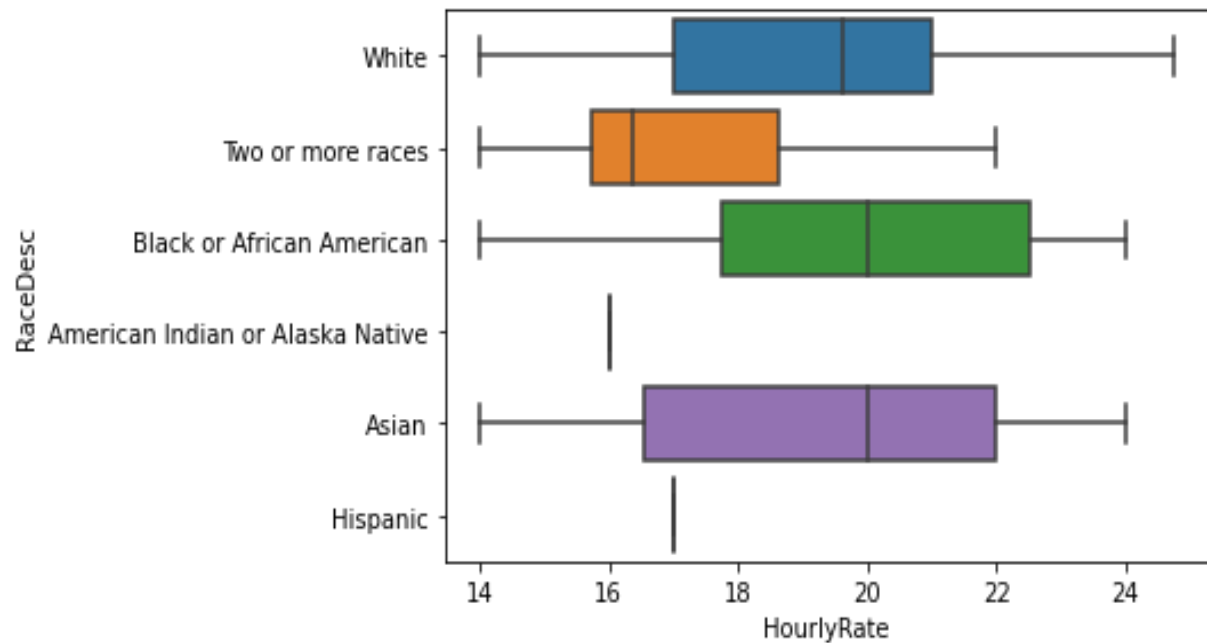
rates between males and females working as Production Technicians I and II:



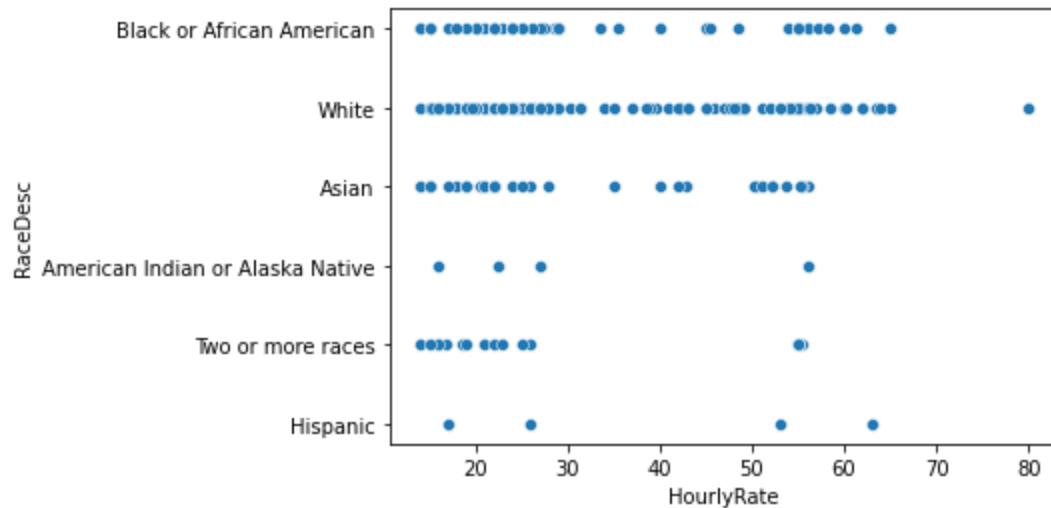
The first boxplot shows that the median hourly pay for female production technicians (I) is higher than male production technicians (I). Female's pay ranges also have a bigger range (\$16-\$22) compared to males (\$17-\$21).

The second boxplot shows the opposite. Male productions technicians (II) have a higher median hourly pay. Their hourly pay range (\$23-\$28) is also bigger than female's (\$24-\$27).

Below are two boxplots showing differences in pay for Production Technicians I (the first graph) and Productions Technicians II (the second graph) based on race:



These two graphs are consistent in one regard; employees with two or more races have lower pay compared to everyone else. Here is another graph with hourly rates of all employees (not just production technicians) that provides more evidence:

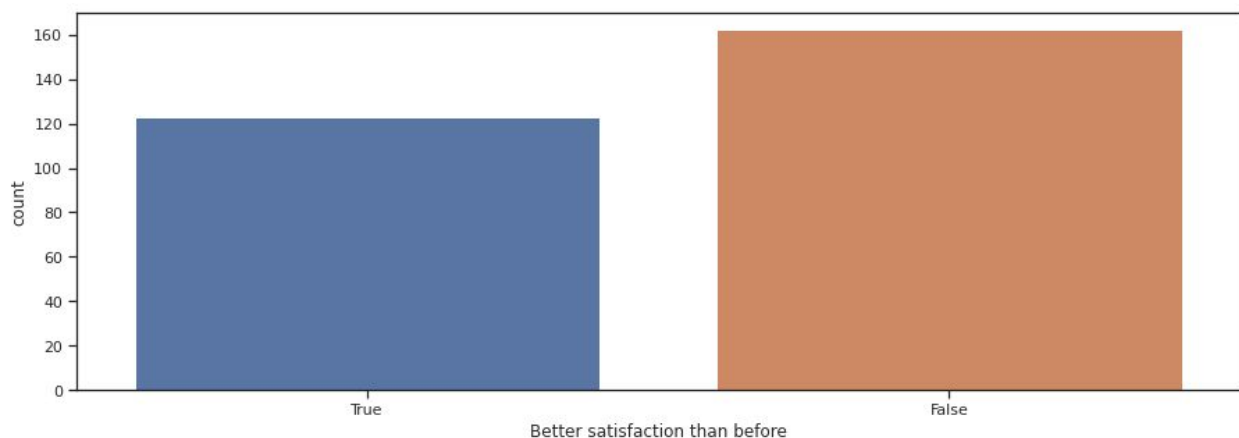


As we can see, a large majority of employees with two or more races are concentrated in a lower pay range relative to all other races (except for American Indians and Alaska Natives).

III. Employee Longevity Clusters (Logan)

I thought it would be a good idea to check employee satisfaction to see how the company is doing with employee retention. There are two columns related to employee satisfaction in the data set, EmpSatisfaction and EmpSatisfaction2. EmpSatisfaction is the employee satisfaction score that was reported several years ago, and EmpSatisfaction2 is the employee satisfaction score that was reported by Cecil last month. Before comparing the two employee satisfaction columns, employees who would start in the future were removed.

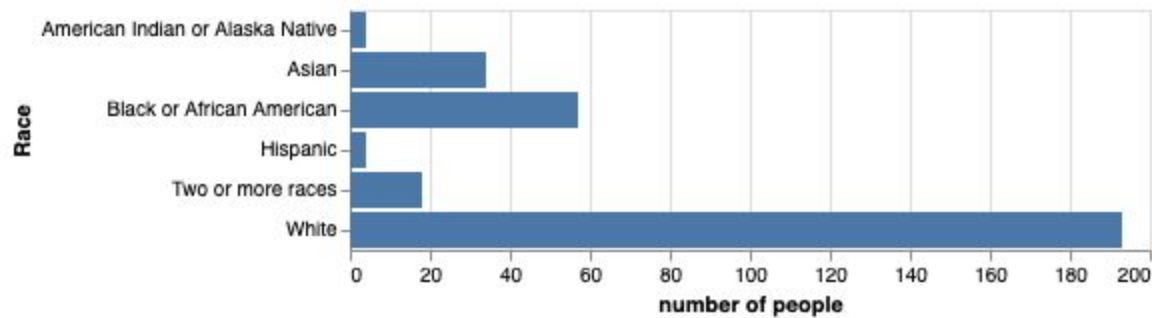
As you can see from the graph below, it can be seen that when comparing past and recent employee satisfaction, there are more cases that recently scored lower than those who did not.



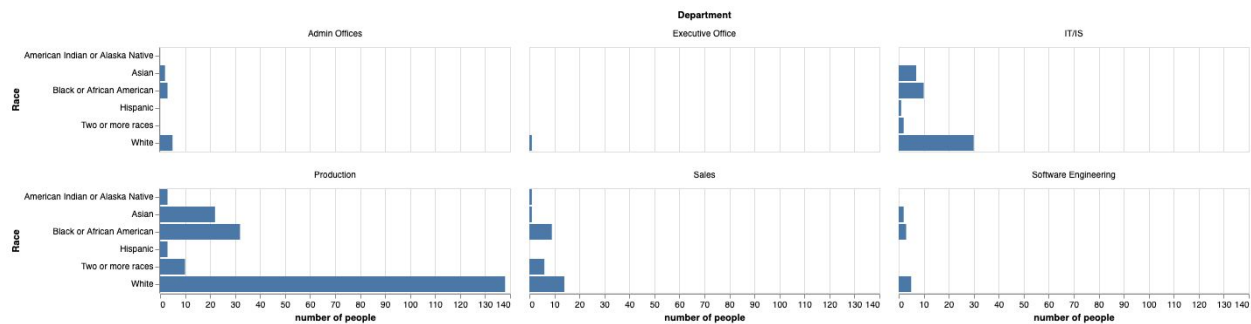
IV. Diversity (Tyler)

The most important question to the board right now is diversity. Are we doing a good job on diversity overall?

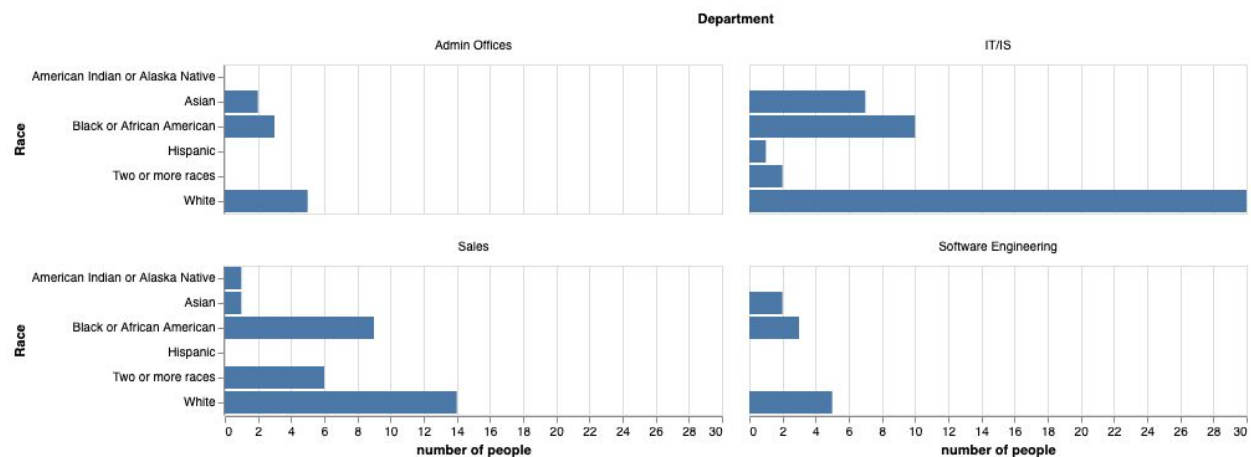
Here we see there are a lot more whites than any other race but there are a mix of other races.



Next let's compare this same thing within each department.



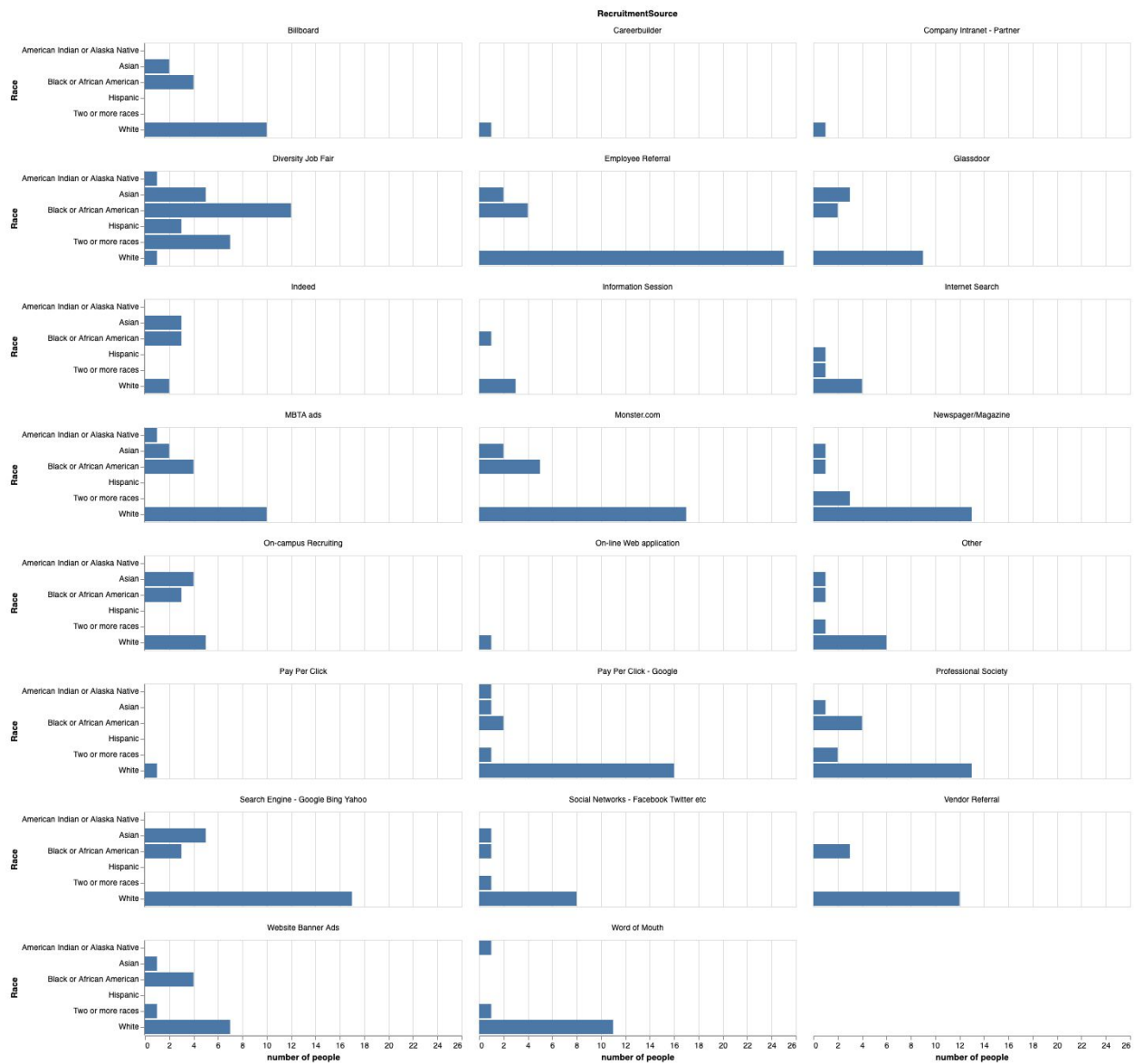
It seems like the most amount of people work in the production department and it looks like the distribution of races is nearly the same there as the overall distribution in the company. For more clarity let's look at all the other departments without the production department present.



From the looks of this it seems like the IT/IS department and the production department can be more diverse. So based off of the graphs it looks like we can use an improvement in diversity in these specific areas.

If not, which of our recruiting sources is having the biggest benefit and where should we be focusing future recruitment efforts?

Let's look at where the most minorities are being recruited from



I know it is hard to see here but from the looks of this it looks like the best recruiting source for diversity is the "Diversity Job Fair", "on campus recruiting" and "MTBA ads" so I would be

focusing our recruitment efforts there. I hope this helps analyze the overall diversity of the company and sets forth some actionable steps.

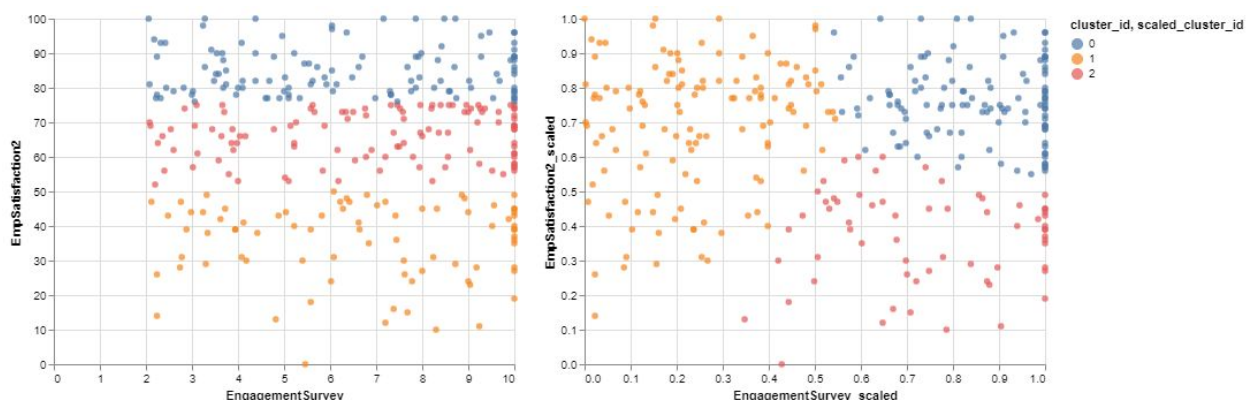
V. Employee Satisfaction and Engagement Survey Analysis (Sam and Avery)

To look more closely at the comparison of the Employee Satisfaction and Engagement Survey I had to pull in the data offered by the HR department to be able to visualize any patterns in order for me to draw up a reasonable conclusion.

As I pulled in the data it was visible that there was quite a big difference with the results of the two surveys. From the clustering technique that I used I saw that it was clustering things favoring the Employee Satisfaction and that was because the scales were different. The Employee Satisfaction scale was ranged from 0-100 and the Engagement Survey was ranged from 0-10. This was a problem because I couldn't visualize the data correctly.

This is why the data had to be normalized or scaled in a certain way that they could both be comparable. I used the Min and Max Scaler from the sk learn library. After scaling everything correctly it turned the scales from 0-1 for both of the data sets. Having turned the data sets to be on the same scale it was obvious that the clustering method changed and that the clustering looked a lot more accurate. Here is the comparison of the two graphs after scaling was done.

(Sam)



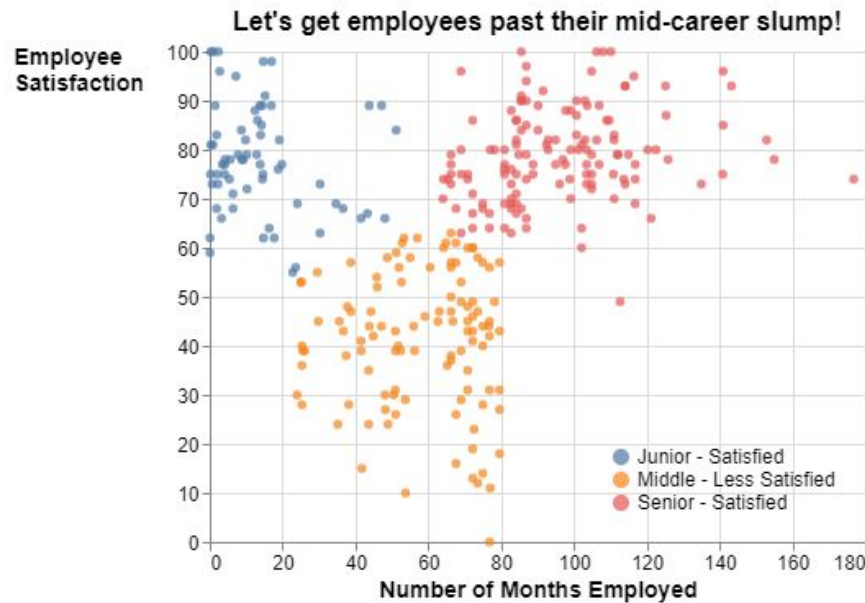
(^ Sam's Graph ^)

(Avery)

Cecil asked the following: *If we put employees into groups based on their months of service, do we see any correlation between how long they've been here and the results of the latest satisfaction survey I did (or the engagement study done by William's brother)?*

During our exploration of the data, we used a k-means clustering algorithm to cluster the employees into three different groups. The algorithm performed fairly well, and we later identified these groups through visualization, and labeled them accordingly: "Junior - Satisfied",

“Middle - Less Satisfied”, and “Senior - Satisfied.” The prefix of the description describes their time at the company, while the suffix describes their level of satisfaction.

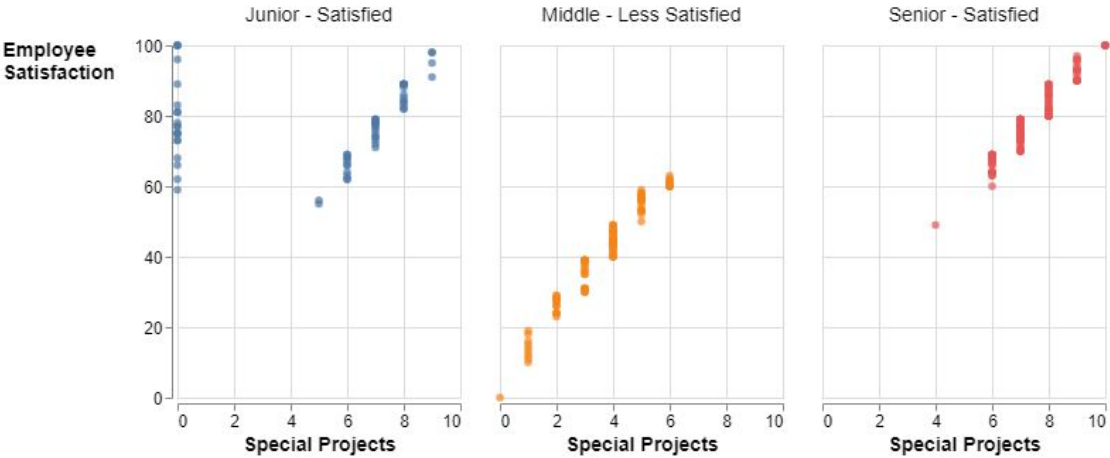


Another question that was posed: *Can we identify what the members of each group have in common?*

There is, of course, some individual deviations from their groups, but a valuable insight can be obtained from these groupings. New employees seem to be fairly satisfied overall, as well as senior employees. The employees in the middle, however, appear to be less satisfied.

The most important insight that we found is that Junior and Senior employees alike were involved in a relatively large number of special projects over the last 6 months. On the other hand, those employees that fall in between the new and the seasoned are taking on less special projects. We recommend that management give the employees in the middle group the opportunity to take on more responsibility. Feeling like a valuable member of the team and making worthwhile contributions to the company can go a long way.

Our employees respond well towards inclusion in special projects.



The typical junior employee took on an average of almost 5 special projects in the last 6 months, while their middle counterparts took on only about 3.6 (senior employees averaged 7.4 in the last 6 months). We certainly do not want to overload our employees, but instead give them plenty of responsibility and opportunity to learn and grow in their respective roles.

The last brief point to be made here: we should not be using William’s brother’s company. As is clear from the plot below, and many others that we viewed in our exploratory data analysis, the Engagement Survey provides little to no insight into employee satisfaction over time or with respect to other important variables.



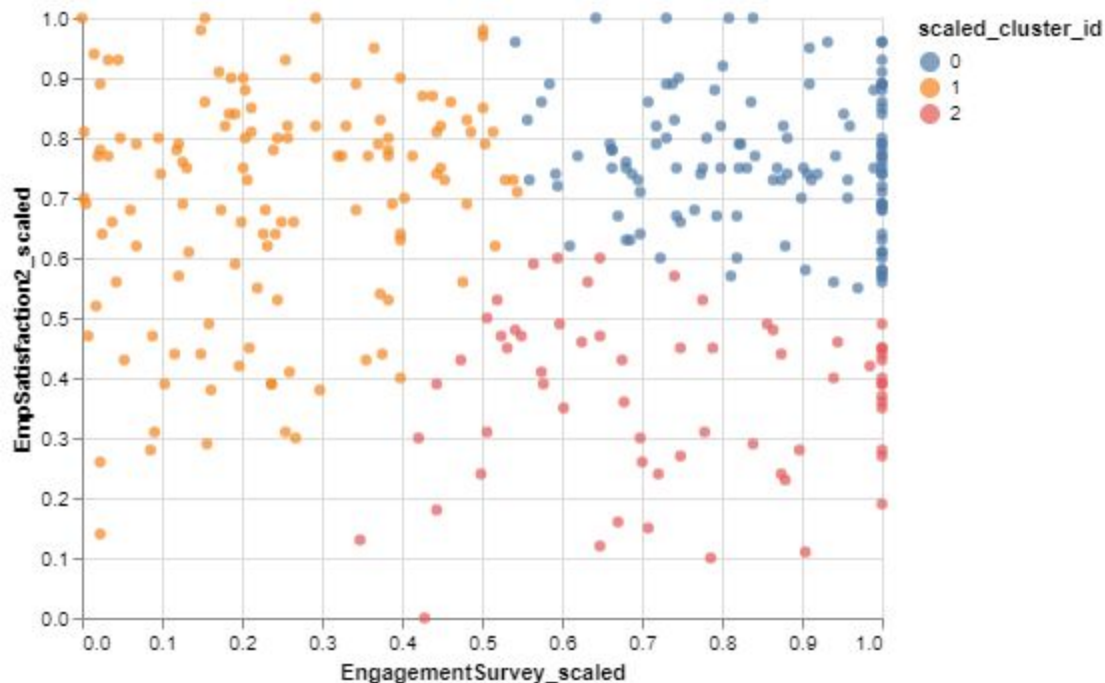
VI. Conclusion and Caveats

Employee Satisfaction and Engagement Conclusion

One of the things I noticed is that the differences in the surveys of the people's scores is strange. As you can see from the graphs the scaled clustering shows that people were not very accurate with their scores when it comes to taking one survey or the other.

What I am concluding from this data exploration is that some people, even a lot of the people in 2 clusters almost scored the polar opposites in both of the surveys. The orange and red clustering shows that those people either scored really high on one survey and low on the other or really low on one survey and high on the other. The thing is I cannot really pinpoint on what this says about either survey. Here is the final graph with the clusters as referred in my conclusion:

In addition, as mentioned previously, we should make sure that our employees take on the appropriate quantity and quality of special projects so that they can contribute and grow. We should also continue with Cecil's survey and drop the Employee Evaluators Inc survey.



(Sam)

In terms of diversity there can be some improvements made in the IT and Production sectors of the company and the best place to find the most diverse recruits are the "Diversity Job Fair", "on campus recruiting" and "MTBA ads".

VII. Python Notebooks

Below are Github Gist links to the notebooks we used during this case study:

<https://colab.research.google.com/gist/Samuelzinh0/2eebbda11678e763b1c3521c8c9c12e0/sam-m-module02.ipynb> (Sam Moreira)

<https://colab.research.google.com/gist/averyrobbins1/cb78425c18931872c5be8d320423ca80/eda-avery.ipynb> (Avery Robbins)

https://colab.research.google.com/gist/LoganOh/108e9f44ae64d978aaa6219395338e74/logan_oh_module-02.ipynb (Logan Oh)

https://colab.research.google.com/gist/cswitzer/43b6f7dfe66943c716b9b1a7294f679c/chase_s.ipynb?authuser=1 (Chase Switzer)

<https://gist.github.com/tylercraigc4/28ff3ffa331b0ed032808b491a43f0b> (Tyler Craig)