2022

SMM635 — Mid-Term Project

GROUP 6

LINH NGUYEN

SOUMYA OGOTI

WENXU TIAN

APARNA VISWANATHAN

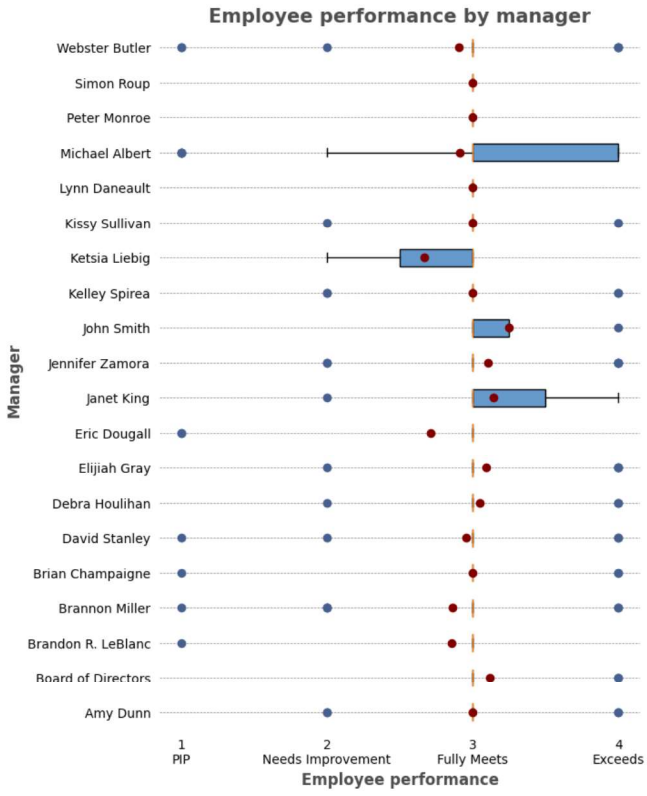
AKSHAY ARORA

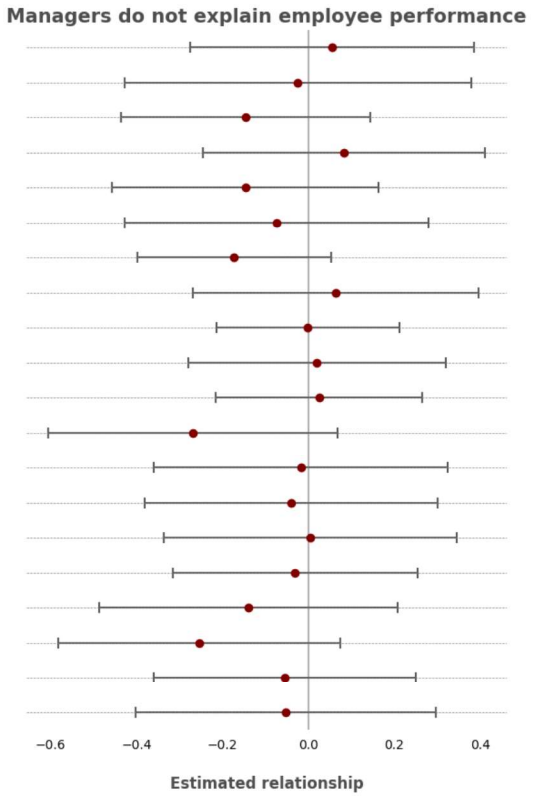
SIDDHARTH GAUR

SMM635 Mid-term Project – Group 6

Visualization #1

How do employees' performance scores change within and across managers?





What are the key design features of your visualization? (MAX 200 WORDS)

There are two graphs in the figure sharing the same vertical axis which is the managers' names.

The first demonstrates the employee performance score distribution by manager in a box and whisker plot. In this chart, the medians and means of employee performance scores associated with each manager are represented by orange vertical markers and red dots, respectively. The blue boxes indicate the interquartile range ("IQR") of the distribution. These are connected to the two lines ("whiskers") which extend to the outermost data points that are 1.5 IQR away from the boxes. The blue dots are data points that are past the end of the whiskers. The horizontal axis shows the employee performance scores and their descriptions.

The second graph is a dumbbell plot which displays the results of an ordered logistic regression model which aims to explain performance scores by managers by regressing performance scores against a set of variables provided in the dataset, including manager names. The red dots are the coefficients, which are placed on the black bars that signify their 95% confidence intervals. The negative and positive regions of the plot are separated by a grey vertical line at zero.

Why did you choose the above-mentioned design features? (MAX 200 WORDS)

Since both charts report different metrics about the same managers, sharing the Y-axis would make them cohesive and reduce the cluttering introduced by a second Y-axis.

In the first chart, although a box plot cannot exhibit the details of a data distribution like a histogram, it was selected due to its compact representation with only lines, boxes and dots, and its ability to display the distribution of numerical values especially when a comparison among multiple groups is required. Specifically, it is observable that the means and medians of performance scores within each manager can be compared with their outliers standing out to see which manager had employees with higher performance scores. We can also see if the scores are condensed at a value or spread out across 4 levels and whether the scores are skewed.

In the second chart, although reporting the p-values is enough to show the statistical insignificance of the estimated coefficients, showing the extent of the uncertainty introduces another layer of information which guide further analyses. Therefore, a dumbbell plot was chosen. The vertical line at zero is used to emphasise the point that if contained by the confidence interval, the model coefficients are not statistically significant.

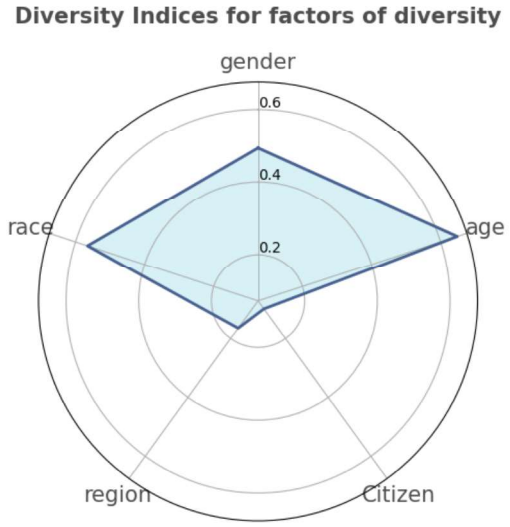
What is the main insight of the chart? (MAX 100 WORDS)

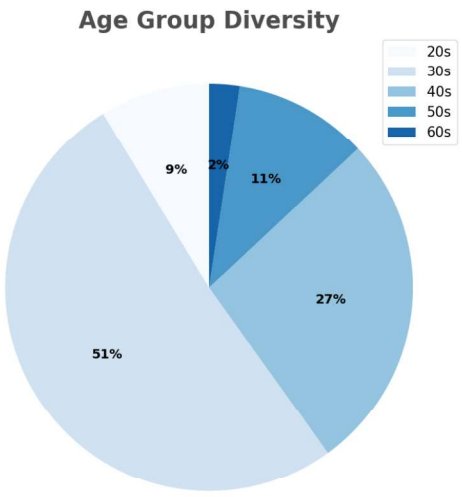
The box plots not showing up for most managers signifies that the performance scores are heavily concentrated on the median, which is 3. On the other hand, there are slight differences in the average performance, for example, John Smith has an average score closer to 4. Statistically, employee performance is not explainable by the manager factor since the zero vertical line is well contained in the

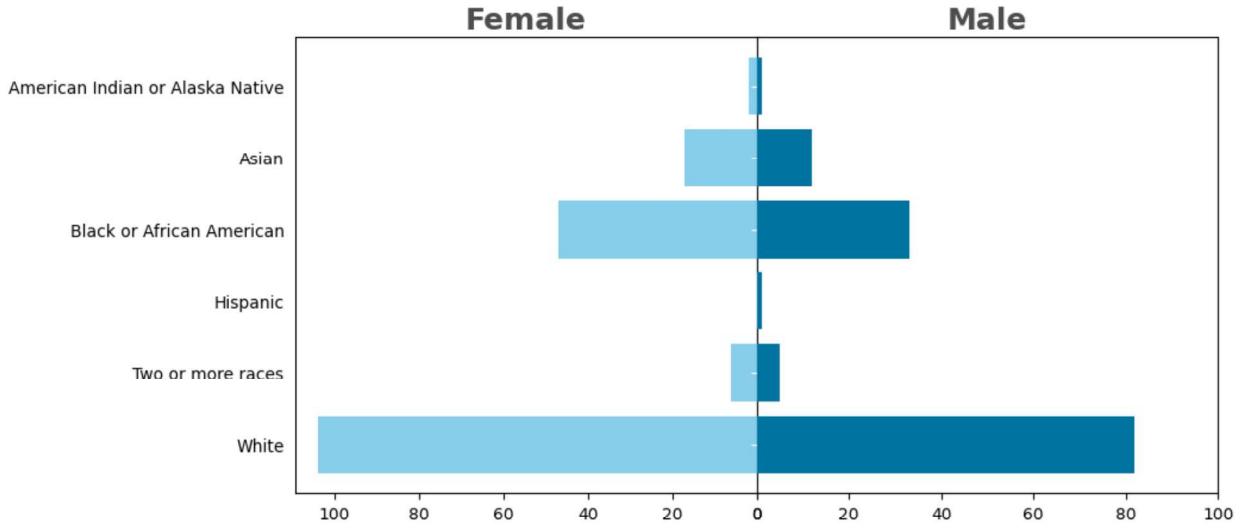
confidence intervals. However, for some managers (e.g., Ketsia Liebig), the relationship may become significant at a more generous significance level since the zero line lies towards the end of their respective confidence intervals.

Visualization #2

What is the overall diversity profile of the organization?







What are the key design features of your visualization? (MAX 200 WORDS)

A combination of charts has been chosen to depict the overall diversity profile of the organization. There are multiple factors for diversity in the dataset. Gender, Race, Age, Citizenship, and Hispanic/Latino origins are some of the relevant factors in the workforce diversity and inclusion scenario. Blau's index is calculated (with a correction mentioned by Biemann and Kearney (2010) to fix the bias in the index) for the 5 diversity factors mentioned above. The first visualization is to compare Blau's indices for these factors. An overall inspection of the composition of the chart by the occupied radar plot implies the extent of diversity. The x-axis denotes the factors, and the y-axis denotes the value of the index. The second plot is a bidirectional bar graph to portray the disparity in the combined effect of gender and race. The horizontal axis depicts the number of female employees of that race to the left, and males to the right. The third plot is a pie chart to show the composition of employees by age group.

Reference:

Biemann, T. and Kearney, E., 2010. Size does matter: How varying group sizes in a sample affect the most common measures of group diversity. *Organizational research methods*, *13*(3), pp.582-599.

Why did you choose the above-mentioned design features? (MAX 200 WORDS)

According to Harrison and Klein (2007), Blau's index is "the most commonly used measure to capture such qualitative distinctions" so it has been selected as a diversity measure. For relatively small group sizes, Blau's index strongly underestimates the variety in groups, whereas the bias becomes smaller with increasing group sizes. A correction mentioned by Harrison and Klein (2007) is used to calculate Blau's indices, which are displayed in the radar graph. This chart has been chosen to represent the overall diversity and observe the value of indices as a comparison to the maximum value of Blau's index, which is equal to 1. The inner polygon is filled so the overall diversity of the organisation can be evaluated: the more coverage, the more diverse the organisation is across all factors.

Williams and O'Reilly (1998) stated that race, gender, and age are the concentrated relational features of "social categorization perspectives" in diversity research. The two remaining plots were introduced to further assess these factors. The pie chart was selected to help quickly identify the composition of age groups, with the percentage values annotated for the exact information. Finally, the bidirectional graph is used to compare the race distribution for female and male employees.

Reference:

Harrison, D.A. and Klein, K.J., 2007. What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of management review*, *32*(4), pp.1199-1228.

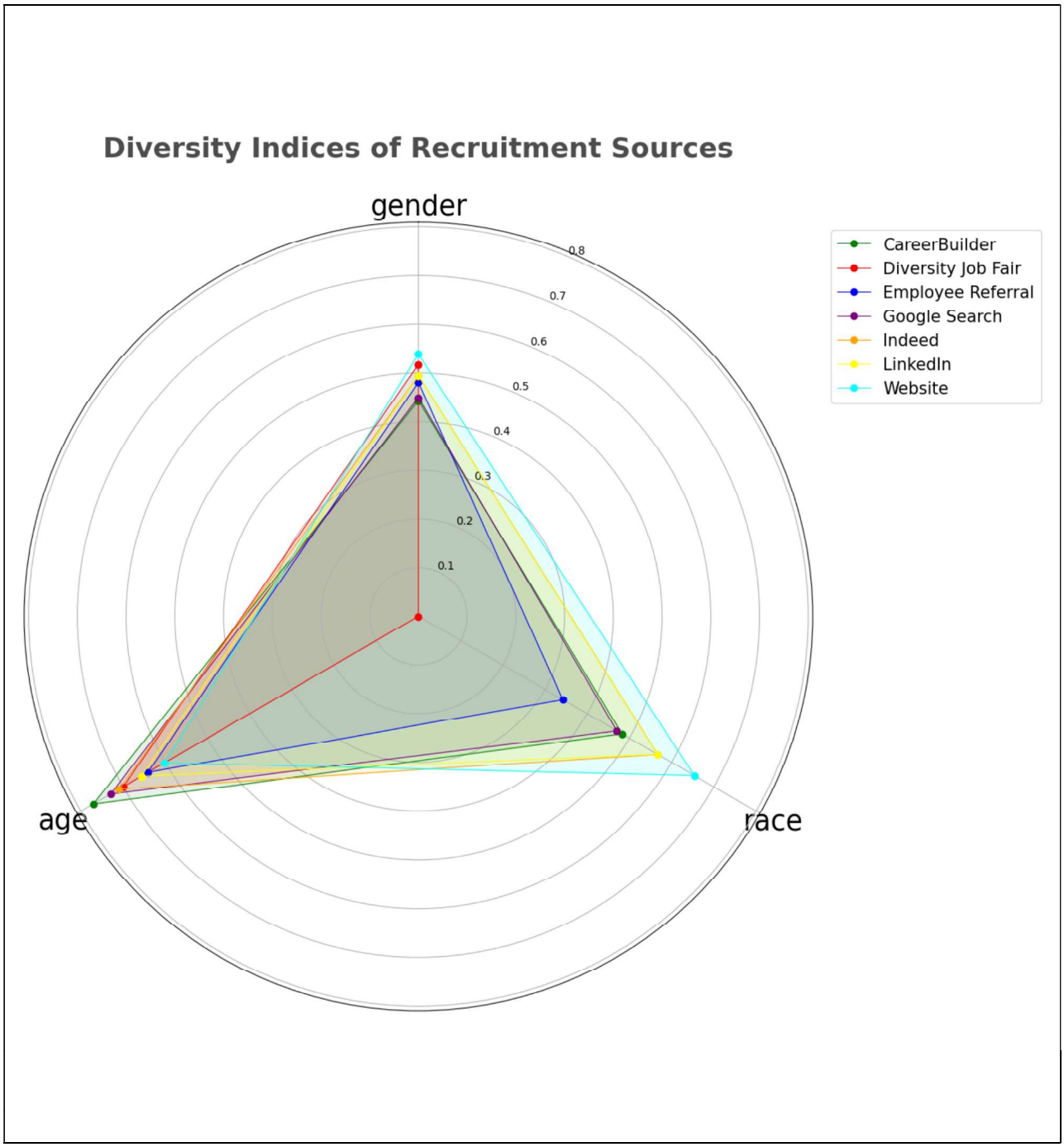
Williams, K.Y. and O'Reilly III, C.A., 1998. Demography and. *Research in organizational behavior*, *20*, pp.77-140.

What is the main insight of the chart? (MAX 100 WORDS)

The meaning of diversity does not only refer to equality, hence, it is reasonable to compare Blau's indices of completely different factors. The radar chart shows the organization's diversity presented significantly in race, age and gender, whereas, nominally in Citizenship. Citizenship is the least diverse factor with around 95% of employees being US Citizens. The bar graph implies that the employee's race is predominantly white, followed by African Americans/Blacks. It is demonstrated in the pie chart that more than half of the current employees are in their thirties and a quarter of them is in their forties.

Visualization #3

What are our best recruiting sources if we want to ensure a diverse organization?



What are the key design features of your visualization? (MAX 200 WORDS)

This figure clearly clarifies the diverse performance in multiple dimensions across different recruiting sources. This figure is a radar chart, including three dimensions which are gender, age and race. The diversity indices of different dimensions are calculated via the bias-corrected formula of Blau's index, which is better than the common one because it can reduce the bias of the diversity index when the sample size is small (Biemann and Kearney, 2010). In each dimension, different points represent the diversity index of different recruiting sources. For example, in the dimension of age, the top three recruiting sources are CareerBuilder, Google Search and Indeed. For each recruiting source, the points in the three dimensions make up a triangle. The area of the triangle indicates the overall performance of diversity in various recruiting pathways.

Reference:

Biemann, T. and Kearney, E., 2010. Size does matter: How varying group sizes in a sample affect the most common measures of group diversity. *Organizational research methods*, *13*(3), pp.582-599.

Why did you choose the above-mentioned design features? (MAX 200 WORDS)

These design features improve the efficiency and clarity of the plot. From the perspective of parameter selection, the age, gender and race included in the dataset are three of the multiple factors that indicate the diversity of an organization (Yadav and Lenka, 2020). Therefore, these three parameters are chosen to depict the chart. From the perspective of plot type selection, the main advantage of the radar chart is that it is capable of carrying large and high-dimensional datasets, which allows the readers to make comparisons quickly. For example, the chart shown here contains 21 data points (including 3 dimensions and 7 indexes in each of the dimension). On the one hand, the reader can easily identify the top and the least recruiting resource in any of the dimensions (age, gender or race). On the other hand, the reader can quickly access the overall performance of different recruiting sources by comparing the shape of the shade from different recruiting sources (i.e. colors).

Reference:

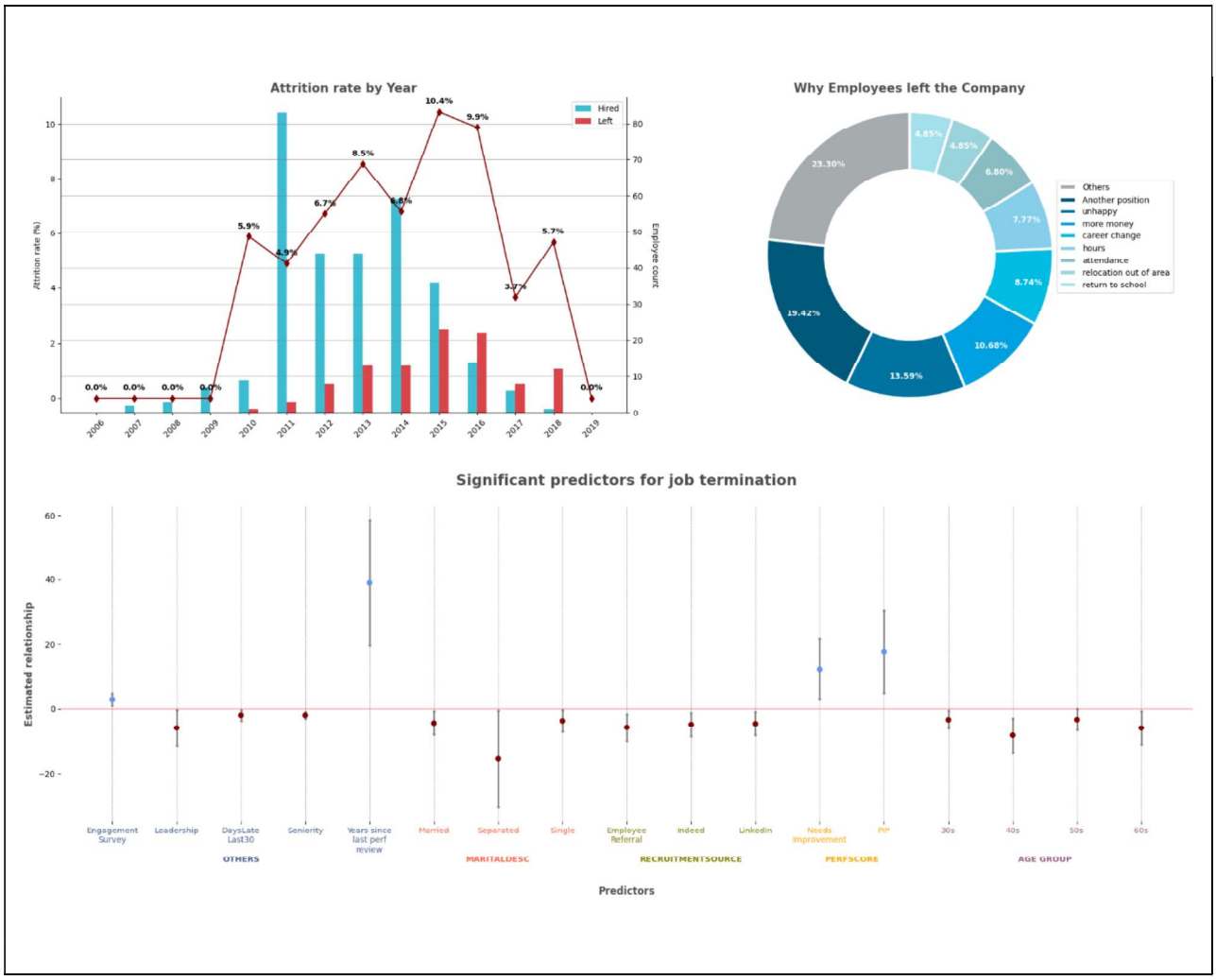
Yadav, S. and Lenka, U., 2020. Workforce diversity: from a literature review to future research agenda. *Journal of Indian Business Research*, *12*(4), pp.577-603.

What is the main insight of the chart? (MAX 100 WORDS)

The source Website shows the best diversity performance in terms of the gender and race. However, it shows the worst diversity performance in terms of age. The source CareerBuilder gives the best diversity performance in terms of age. Considering the result from question 2, we suggest using the source Website as the first recruiting source to improve the gender diversity and race diversity of the organization, while utilizing the source CareerBuilder as the second recruiting source to keep the current age diversity level of the organization.

Visualization #4

What are the most frequent factors for job termination?



What are the key design features of your visualization? (MAX 200 WORDS)

The first (top-left) shows the attrition rate by year as a red line with diamond-shaped markers and values annotated in black text. It also shows the number of employees left and hired each year as vertical bars, which can be estimated by referring to the horizontal dashed lines in combination with the secondary Y axis on the right.

The second (top-right) demonstrates the proportions of reasons why employees left the Company in a doughnut chart. The legend on the right provides an explanation for each colour of the doughnut.

Finally, the third chart displays the estimated relationship between factors that are considered significant in explaining job termination. The horizontal line at zero separates the plot into 2 regions: the upper contains factors that were estimated to have a positive relationship with job termination, while the reverse is true for the lower. The blue and red dots show the mean of the estimated coefficients if the relationship is positive and negative, respectively. They are placed on top of the grey bars representing the confidence intervals of the estimation. The factors are grouped into categories, annotated by the capital texts under the main plots and colour-coded respectively.

Why did you choose the above-mentioned design features? (MAX 200 WORDS)

In the first chart, as attrition rate is the most important information, it is emphasised with diamond markers and precise annotation of values. The bar plot further supports this by showing the number of employees left and hired. For example, the high attrition rate in 2010 could be explained by the low headcounts, while the reverse is true for 2018 which had significantly more terminations but a similar attrition rate due to high headcounts.

In the second chart, the doughnut chart was chosen to break down the termination reasons using lines instead of surfaces (as in pie charts), which facilitates comparison across categories (Cairo, 2012). The percentages are annotated to show the exact values if the reader is interested. The “Others” category is coloured in a different colour scheme since it is a derived category.

Finally, like the first question, showing the confidence intervals introduces another layer of information which guide further analyses. Although the plot spanning horizontally makes inferring the exact values difficult, the plot's intention is to demonstrate the scale of the estimated relationship relative to zero. Therefore, no horizontal gridlines were added. Finally, variables of the same groups are colour coded as they are intended to be compared together.

Reference:

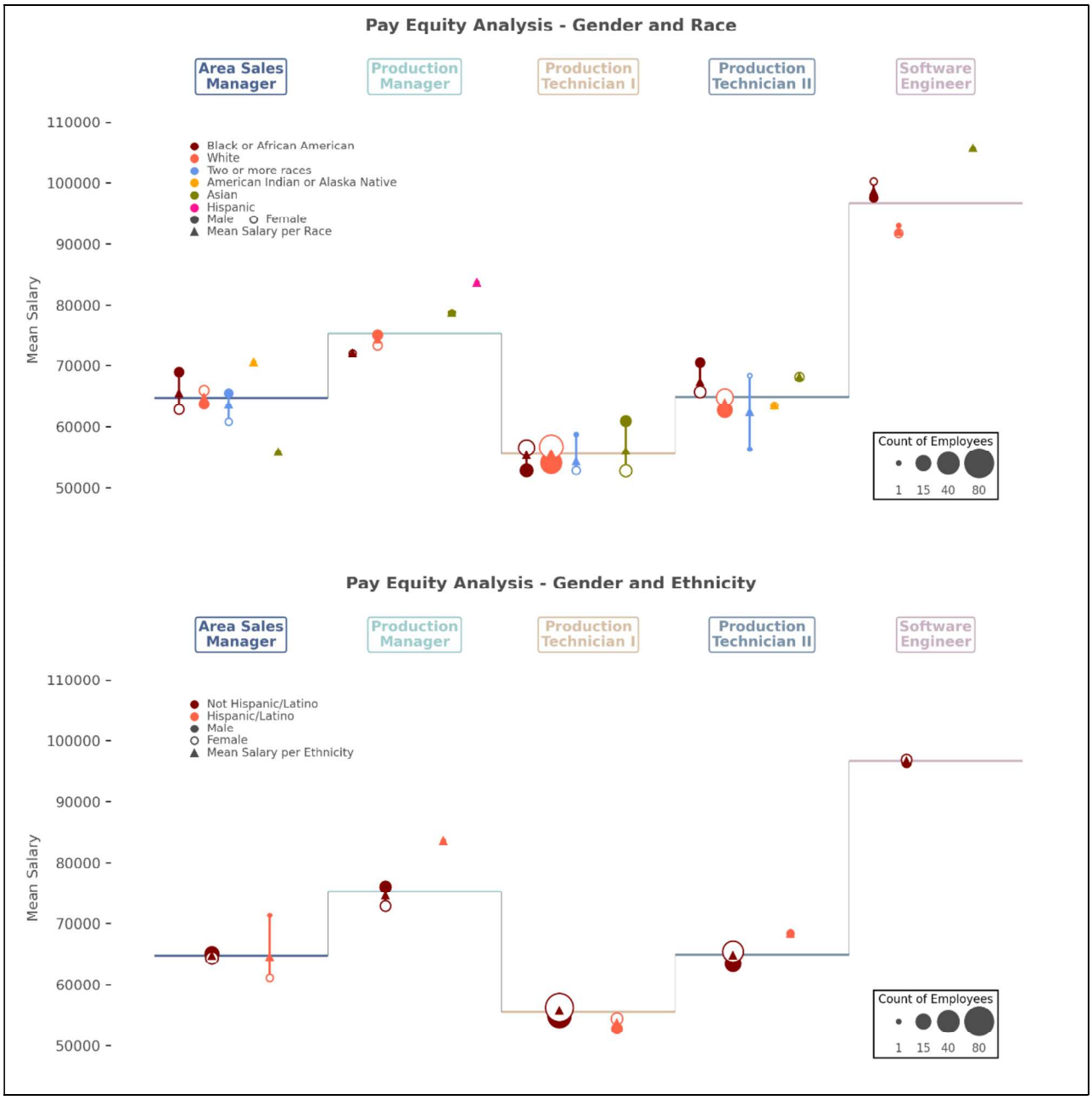
Cairo, A., 2012. *The Functional Art: An introduction to information graphics and visualization*. New Riders, p. 40.

What is the main insight of the chart? (MAX 100 WORDS)

Although the attrition rate does not indicate any clear trend, the significant drop in hiring is accompanied by an increase in terminations. This means the changing nature of the Company is one of the major factors in job termination. Furthermore, the major reasons which account for approximately 50% of terminations are “Another position”, “Unhappy”, “More money”, and “Career Change”. Finally, the model suggests that time since the last performance review is a significant factor in explaining terminations, followed by low performance. The remaining variables are barely significant and have lesser associations with termination, as indicated by the dumbbells.

Visualization #5

Are there areas of the company where pay is not equitable?



What are the key design features of your visualization? (MAX 200 WORDS)

A customised lollipop chart has been chosen here to show if there is any pay inequality with respect to gender, race and ethnicity as these are often the factors resulting in bias. One visualisation is to check the combined effect of gender and race on salary. Another one is to check the combined effect of gender and ethnicity on salary. For the two charts, the horizontal lines depict the mean salaries of the job positions with more than ten employees. In the first chart, for every position, multiple lollipops are plotted on horizontal lines, with one colour for each race. The filled and unfilled tips show the mean salary for men and women of that race respectively. The size of the tip is scaled according to the number of employees in that sub-category. The mean salary of each race is marked by a triangle for reference. Lollipops are drawn from the mean of each race to mean of the gender-specific sub-category. For cases in which there are few data points, mean salary of race overlaps with the tip and stick of the lollipop is short showing these are outliers. The second chart is drawn similarly with race replaced by ethnicity.

Why did you choose the above-mentioned design features? (MAX 200 WORDS)

The lollipop chart has been picked in place of grouped bar chart to avoid cluttering and the Moiré effect. This chart is apt here to compare one numerical variable (salary) with two categorical variables (race/ethnicity, gender). It is easy to observe the effect of interactions between multiple factors – Salary with race/ethnicity and gender concisely with less data ink. Mean salary has been considered to capture the central tendency of each sub-category. In order to draw meaningful conclusions, only job positions having more than ten employees are shown in the chart. A detailed regression analysis was performed on all factors that could contribute to pay inequity. There was no evidence of an adjusted pay gap after accounting for all factors. However, gender, race and ethnicity were chosen for our visualisations as these are the areas in which disparity in wages is typically observed (Rho, (Rho, 2021)). To guide the observer towards sub-categories with a sizeable number of employees, the tips have been scaled according to the number of employees.

Reference:

Rho, D. (2021) "The Gender Policy Report", *Status of Women & Girls in Minnesota*, 24 February. Available at: https://genderpolicyreport.umn.edu/what-causes-the-wage-gap/
(Accessed 13 November 2022).

What is the main insight of the chart? (MAX 100 WORDS)

For each job position, the mean salary for each race lies close to the mean salary of the position (no bias). When comparing salary with race and gender, there is no specific pattern that indicates a bias towards one gender. Also, for each race, the deviation from the mean caused due to gender is small (no bias). The trend is similar also for ethnicity and gender (no bias). Additionally, regression analysis with all relevant factors on salary did not show any pay inequity.