Project

STAE04/DABN19: Data Visualization

Dirk Baars

2021-10-28

1 Introduction

The HR Dataset was designed by Drs. Rich Huebner and Carla Patalano to accompany a case study designed for graduate HR students studying HR metrics, measurement, and analytics. The first version of the data stems from 2017, and the data used for this project is the fourth version, which stems from 2020. The dataset is synthetic, created specifically to go along with the case study. The data revolves around a fictitious company and the core data set contains names, DOBs, age, gender, marital status, date of hire, reasons for termination, department, whether they are active or terminated, position title, pay rate, manager name, and performance score. The original data set contains 36 variables and 311 observations. The dataset can be obtained here.

The data set contains an interesting variable, namely the absence of work. Absence of work has possible severe effects, such as inadequate staffing, which in turn may lead to increased employee stress levels (SHRM) Therefore it is interesting to look at the variables that might influence on the absence of work. The question for this research is:

"What variables have effect on absence of work?"

In this assignment, we will mostly focus on the visualization of this research question, where we will dive deeper into the variables of Salary, Recruitment Source, Work Performance and Sex. Descriptions of the variables can be found in the table below:

Table 1: *Description of HR dataset.*

Name	Type	Measurement	Description
Salary	Integer	$ U.S. Dollars	The person's yearly salary
Absences	Integer	number of times Absent	The number of times the employee was absent from work.
RecruitmentSource	Character	7 different platforms, with an 8th 'others' option	The name of the recruitment source where the employee was recruited from
PerformanceScore	Character	(Fully Meets, Partially Meets, PIP, Exceeds)	Performance Score text/category
Sex	Character	M = Male, F = Female	Sex

Some additions have been made before using the data for the analyses.PerformanceScore is not ordered in the data, where it should be clearly ordered. It ranges from (Performance Improvement Plan (PIP), Partially Meets, Fully Meets, and Exceeds. Whereby PIP means 'helping employees who are not meeting job performance goals'. Ordering the variabls is done by mutating the variable and relevel with fct_relevel(), and then ordering the variable with the as.ordered() function.

2 Analyses

First, let's examine the data. We examine the data of absenteeism against the work performance in a boxplot. A boxplot plot gives information about the distribution of the data showing the outliers, the median and the quartiles, which is in this case, gives more in-depth information than e.g. a violin plot. A sequential color scheme is added to the boxplot because the variable of Performance Levels is of ordinal origin. This has been done by adding scale_fill_ordinal().

In the boxplot (Figure 1), one outlier is noticed, in the PIP level. This employee has had 20 absence days. We keep the outlier in the data because we do not want to lose valuable data from the set. The employees who need improvement have the highest median of absence. The employees who are working with PIP have the lowest amount of absenteeism.

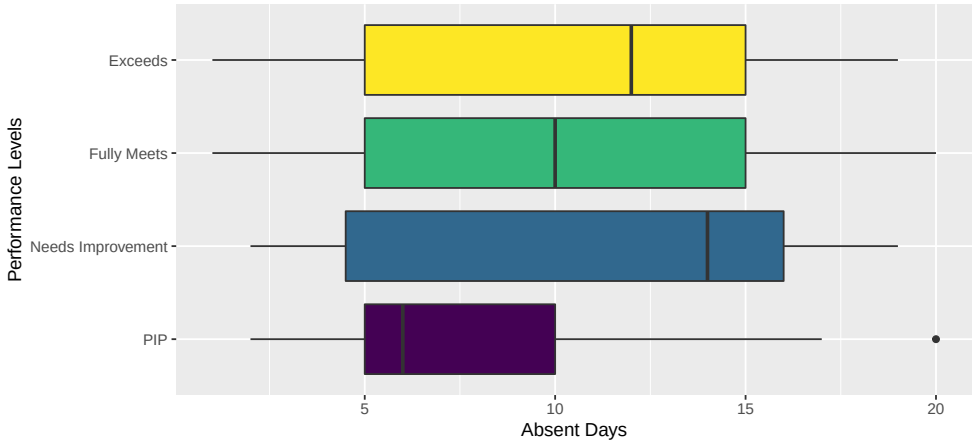

Figure 1: *Boxplot of the amount of absence days and the performance levels of the employees.*

To further examine the relationship between work performance and absenteeism, a scatterplot is created with geom_point between the variables 'Absences' and 'PerformanceScore'. Hereby the points have been made transparent with the function 'alpha' and the value points are 'jittered' to prevent overlap. A line between the data points has been added with the help of 'geom_line()', whereby a "gam" smoother is used because there are $< 1{,}000$ observations. The line of geom_smooth has a shaded area, which resembles a 95% confidence interval.

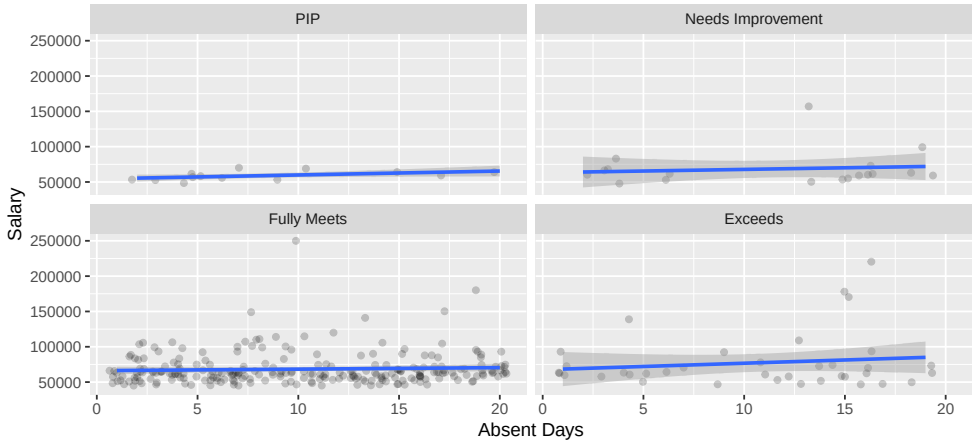

Figure 2: *For every performance level, the amount of salary and days of absence have a positive relation.*

In the scatterplot (Figure 2) it can be noticed that there are a few employees that have higher salary than the others. E.g. in the performance pool that 'exceeds' one can see that there are approximately 6 variables outside of the confidence interval, whereby the biggest earner has a salary of approximately 225000 Dollars per year.

Furthermore, there tends to be a positive relationship between the salary level and the absenteeism of work. It shows that for every performance level, the more you

earn, the higher the number of absent days.

Thirdly, one could check if there if the number of absence days might be related to the ways of Recruitment. This is done by creating a density plot. The days of absence can be seen on the x-axis, whilst the density plot is 'filled' with the source of recruitment. Also, a qualitative palette is chosen, because the sources of recruitment do not signal differences in magnitude, this has been done by adding the function: scale_fill_discrete(). Also, a geom_rug has been added, showing the marginal distribution for days of absence.

```
## Warning: Groups with fewer than two data points have been dropped.
```

```
## Warning: Removed 1 rows containing missing values (position_stack).
```

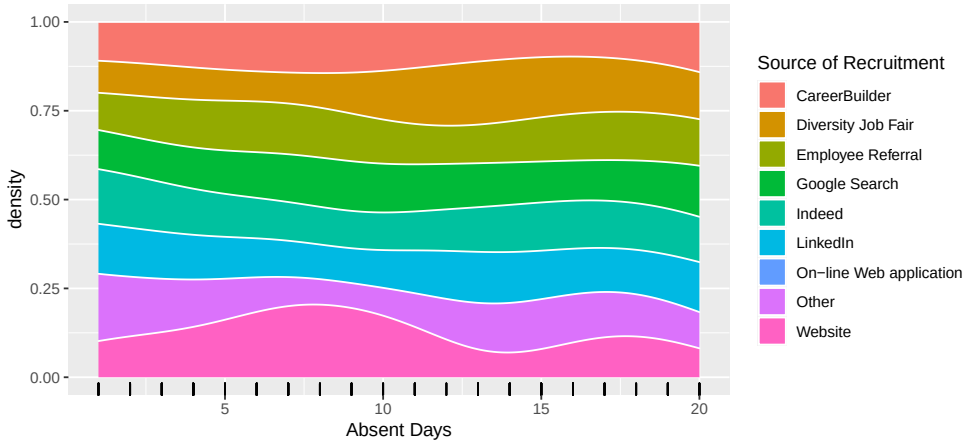

Figure 3: *The Source of Recruitment against the amount of absence in days, a density plot.*

In the density plot (Figure 3), one can see that for a low amount of absence days, employees had a high density to be recruited from Indeed, and other ways of recruitment. When employees had a high amount of absence days, they tend to be sourced from Google Search and Online Web applications.

Lastly, we can check if sex has an influence on the days of absence. This has been done by creating a mosaic graph. The main benefit of a mosaic plot is that we get to see more of the data then a regular bar chart. E.g. with a bar chart, we need to separate visualizations to provide information on the overall size of groups, whereas with the mosaic plot we can make do with only a single plot. To use the days of absence in a mosaic table, the variable is cut into three levels via the function cut_interval(), where they are ordinal in level. Therefore, we make use of the 'scale_fill_ordinal()' function, because the levels of absence are ordered.

The mosaic table can be found in figure 4. In the mosaic table, it can be seen that females have a higher amount of low absence and high absence than men, whilst men have a higher amount of average absence in the job. No clear conclusions can be drawn from this mosaic table.

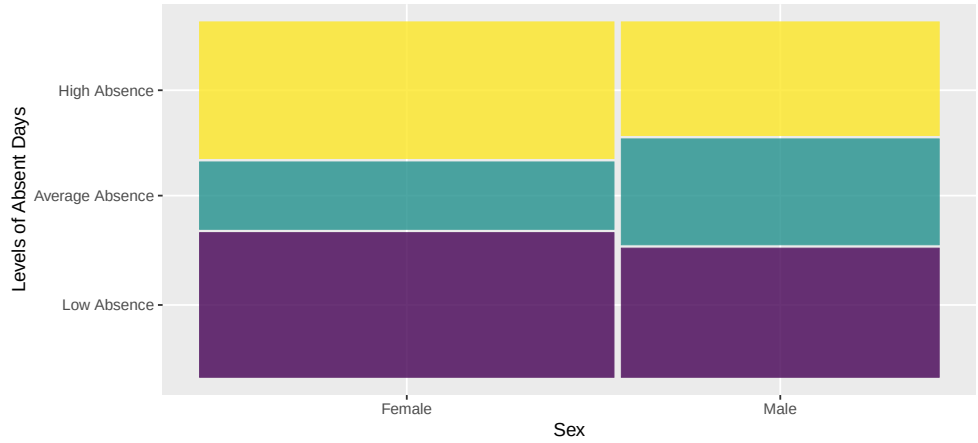

Figure 4: *The levels of absence against salary and sex, a mosaic graph.*

3 Conclusion

In general, one can argue that the employees who need to improve their level of performance have the highest median of absent days. Also for every level of performance, the more salary one earns, the higher the number of absence days. For the sourcing platform counts, employees with the highest amount of absence days are sourced via Google Search and Online Web application, and the employees with the lowest amount of absent days are sourced via Indeed and other ways of recruitment. Female employees also tend to have a higher amount of both high levels of absenteeism, as well as low levels, whilst male employees have a higher average level of absenteeism. Further research can look into these relationships to see if absent days can be influenced, which can possibly lead to lower stress levels in the work place.