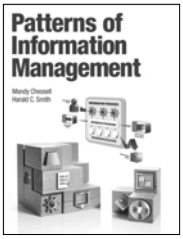
# Getting Started with Data Science

## Making Sense of Data with Analytics

Murtaza Haider

# Related Books of Interest




## Patterns of Information Management

By Mandy Chessell and Harald C. Smith

ISBN: 978-0-13-315550-1

Use Best Practice Patterns to Understand and Architect Manageable, Efficient Information Supply Chains That Help You Leverage All Your Data and Knowledge

Building on the analogy of a supply chain, Mandy Chessell and Harald Smith explain how information can be transformed, enriched, reconciled, redistributed, and utilized in even the most complex environments. Through a realistic, end-to-end case study, they help you blend overlapping information management, SOA, and BPM technologies that are often viewed as competitive.

Using this book's patterns, you can integrate all levels of your architecture—from holistic, enterprise, system-level views down to low-level design elements. You can fully address key non-functional requirements such as the amount, quality, and pace of incoming data. Above all, you can create an IT landscape that is coherent, interconnected, efficient, effective, and manageable.


## The Business of IT

### How to Improve Service and Lower Costs

By Robert Ryan and Tim Raducha-Grace

ISBN: 978-0-13-700061-6

Drive More Business Value from IT…and Bridge the Gap Between IT and Business Leadership

IT organizations have achieved outstanding technological maturity, but many have been slower to adopt world-class business practices. This book provides IT and business executives with methods to achieve greater business discipline throughout IT, collaborate more effectively, sharpen focus on the customer, and drive greater value from IT investment. Drawing on their experience consulting with leading IT organizations, Robert Ryan and Tim Raducha-Grace help IT leaders make sense of alternative ways to improve IT service and lower cost, including ITIL, IT financial management, balanced scorecards, and business cases. You'll learn how to choose the best approaches to improve IT business practices for your environment and use these practices to improve service quality, reduce costs, and drive top-line revenue growth.

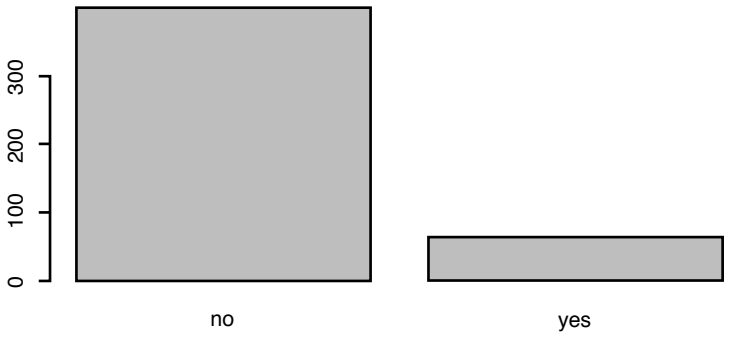



**Figure 5.7** Graphical depiction of the minority status

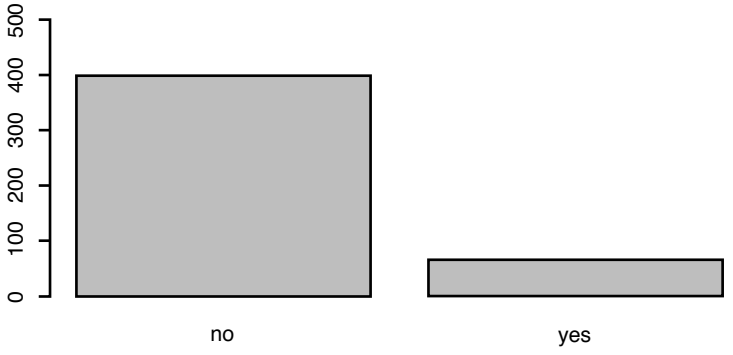



**Figure 5.8** Corrected ordinate for the graphical depiction of the minority status

Notice now that the bar labeled "no" now falls within the value plotted on the y-axis. However, we still do not know by just looking at the figure what is being depicted as no or yes on the x-axis.

Figure 5.9 addresses the two concerns I have highlighted about Figure 5.7.

```
plot(TeachingRatings$minority,ylim=c(0,500),
    xlab="minority status", ylab="number of courses")
```

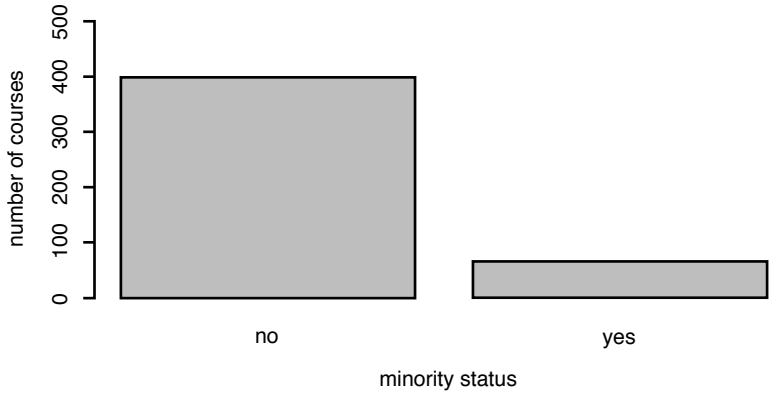



**Figure 5.9** Appropriately labelled depiction of the minority status

Notice that the label for x-axis adequately informs that we present a breakdown for the minority status of the instructors. At the same time, the values depicted on the y-axis refer to the number of courses being taught by the instructors. We can deduce by reviewing Figure 5.9 that most courses are taught by non-minority instructors.

Notice that the bars presented in Figure 5.9 are in the shape of columns. I can present the same information by rotating the bars by 90 degrees. I illustrate this by presenting the breakdown of courses taught by gender (see Figure 5.10).

```
plot(TeachingRatings$gender,horiz=TRUE, xlim=c(0,300),
     ylab="gender", xlab="number of courses")
```

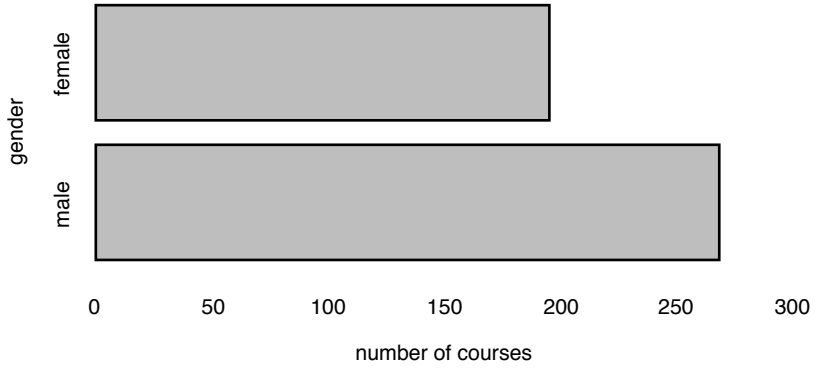



**Figure 5.10** Breakdown of courses by gender

Notice that the two bars are plotted horizontally showing that the male instructors teach more courses than the female instructors do. We can conveniently infer this from Figure 5.10 because of the illustrative labels for the x-axis, which represents the number of courses taught, and y-axis, which presents the breakdown of gender for male and female instructors.

Graphics are even more powerful when we present multiple variables in one image. For instance, Figure 5.11 presents a cross tabulation between two categorical variables, namely gender and tenure. In a tabular format, a cross-tabulation between two variables is presented as a two-by-two matrix.

```
xtab<-table(x$tenure,x$gender)
barplot(xtab, ylim=c(0,300),   legend=rownames(xtab),
        xlab= "gender", ylab= "number of courses")
```

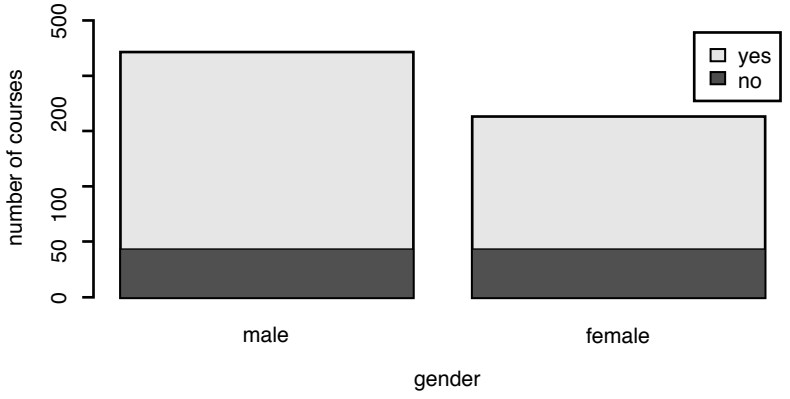



**Figure 5.11** Breakdown of courses by gender and tenure

Figure 5.11 presents the same two-by-two matrix in a graphical format. We can tell from the x-axis that the two bars represent the gender and we can tell from the y-axis that the height of the bars represents the number of courses. We also note that each bar has two distinct colors. From the legend placed in the top-right corner, we see that the darker shade represents "no" and the lighter shade represents "yes". Knowing that we have plotted gender and tenure, we can infer that the yes and no refer to the tenure status of the instructors. Thus, by reviewing the information in Figure 5.11, we can conclude that tenured faculty members teach the overwhelming majority of courses. We infer this by looking at the larger light gray shaded parts of the bars representing male and female instructors.

There are two obvious shortcomings in Figure 5.11. First is that the legend in the top-right corner refers to the tenure status as "yes" and "no". This information is automatically generated from the data labels for the minority variable. Because the labels identified tenured professors

as "yes" and non-tenured professors as "no", the plotting command automatically generates the same labels for the corresponding data in the figure. While these labels are useful, they are certainly not illustrative. The information value of this figure will improve if we were to replace "yes" with "tenured" and "no" with "non-tenured" in the legend.

The other limitation of this graph is the way it represents the breakdown between male and female instructors for their respective tenure status. A quick review of Figure 5.11 suggests that the number of courses taught by non-tenured instructors is almost the same for both male and female instructors. That is, the dark-shaded areas for the two bars representing male and female instructors are of almost equal height. This equivalency is misleading.

Even though untenured male and female instructors taught a similar absolute number of courses, relatively speaking, untenured female instructors taught a larger proportion of courses taught by female instructors than the same for their male counterparts. I address these shortcomings by plotting the respective percentage of courses across the two dimensions of gender and tenure in Figure 5.12.

```
x<-TeachingRatings
plot(x$gender, x$tenure, main="Share of courses taught", ylab="tenure
status", xlab="gender")
```

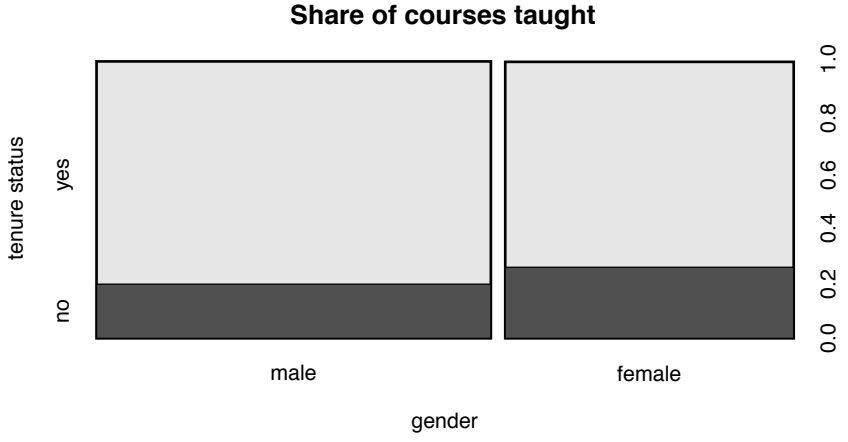



**Figure 5.12** Proportionate breakdown of courses by gender and tenure

Note that the x-axis in Figure 5.12 represents gender and y-axis represents the tenure status of instructors. Also, note that width of the bars is proportionate to the courses taught by male and female instructors, respectively. Given that male instructors taught more courses than the female

instructors, the width of the bar representing male instructors is proportionately wider than that of the bar representing female instructors. At the same time, the darker shaded portion of the bars representing non-tenured instructors for both male and female instructors illustrates that untenured female instructors taught a larger proportion of the courses taught by the female instructors than the untenured males did as a fraction of the courses taught by male instructors.

Histograms are widely used to determine the distribution of continuous variables. Histograms can help you determine outliers and the central tendency in a data set. Figure 5.13 presents the histogram for the normalized beauty scores. The figure reveals that most instructors received a score between –1 and 0, and only a small number of instructors received beauty scores in the range of 1 and 2. The y-axis presents the percentage of observations that fell in each category for the beauty score.

```
histogram(x$beauty)
```

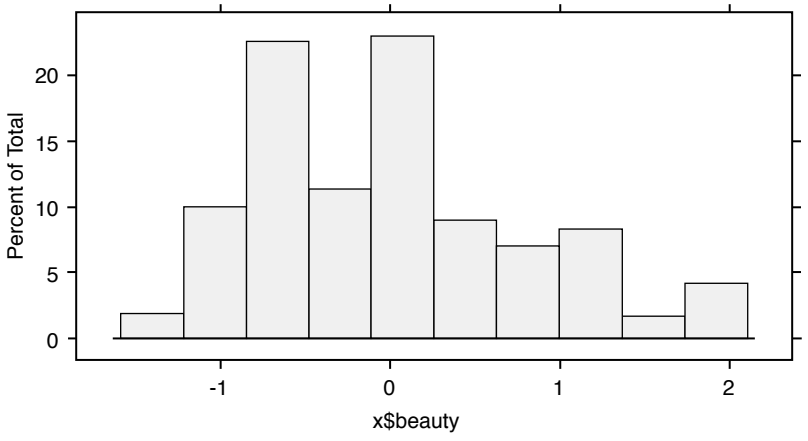



**Figure 5.13** Histogram of beauty score

I can identify at least two limitations in this graph. First, the bars are lightly shaded and they may not print well on paper. Thus, a darker shade of gray will be helpful. At the same time, the label for x-axis, which is automatically generated, is not very illustrative. This can be addressed by replacing it with a more descriptive label.

```
histogram(x$beauty, nint=15,
          xlab="normalized beauty score", col=c("dark grey"))
```

Figure 5.14 addresses the aforementioned limitations.

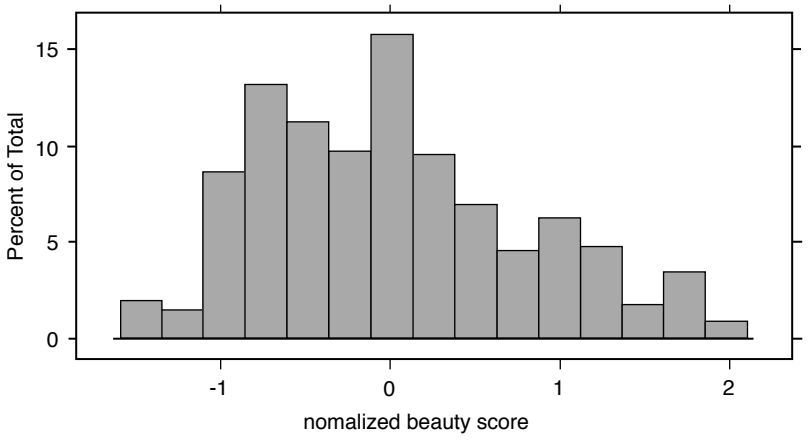



**Figure 5.14** Fine-tuned histogram of beauty score

Notice that the x-axis label appropriately identifies the data as normalized beauty score whereas the y-axis continues to represent the percentage of observations corresponding to each segment of the beauty score. You will also notice that I have added additional bars to segment the beauty score compared to the ones shown in Figure 5.13. By selecting a larger number of bars, I have obtained a better distribution of data than the one in Figure 5.13.

I present a slightly different histogram for beauty score in Figure 5.15. Notice that compared to Figure 5.14, the histogram in Figure 5.15 is narrower and taller. At the same time, the information presented in Figure 5.15 appears to have a more 'normally' distributed shape than the one in Figure 5.14.

```
histogram(x$beauty, nint=15, aspect=2,
          xlab="normalized beauty score", col=c("dark grey"))
```

Figure 5.15 differs in its layout because of a different aspect ratio. Figure 5.14 presents the same information in a rectangular format where the width is greater than the height of the figure. Because Figure 5.14 is wider, we visually deduce a different conclusion than what we are able to deduce from Figure 5.15. Notice that other than aspect ratio, everything else is the same between the two figures including the color and the number of the bars. Why is it that we see different trends in the two figures?

The answer to this question perhaps lies in how we process visual signals. Because the rectangular depiction in Figure 5.15 is characterized by taller bars, we visualize a more central tendency in the normalized beauty score in Figure 5.15 than we do in Figure 5.14.