Topic 2: Data Visualization

Instructor: Daniel L. Pimentel-Alarcón

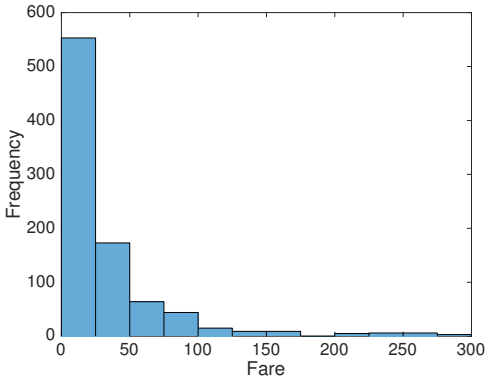
2.1 Introduction

Recall that we want to obtain insights from data. To this end, it is often convenient to visualize data. There are several ways to do this, which we will exemplify using the following dataset, containing the next information about 887 passengers aboard the Titanic ship: 1) whether they survived or not (1 = survived, 0 = deceased), 2) passenger class, 3) gender (0 = male, 1 = female), 4) age, 5) number of siblings/spouses aboard, 6) number of parents/children aboard, and 7) fare. Here is a table containing a small subset of the dataset:

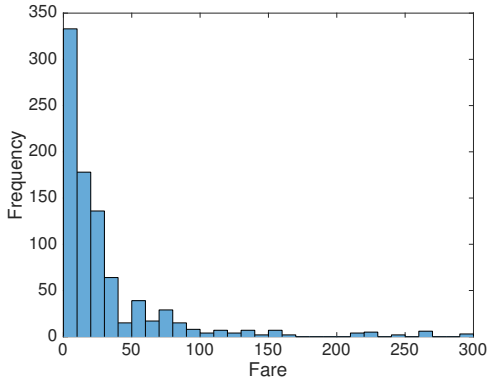
	Passenger 1	Passenger 2	Passenger 3	$\cdots$	Passenger 887
Survived	**0**	**1**	**1**	$\cdots$	**0**
Passenger Class	3	1	3	$\cdots$	3
Gender	0	1	1	$\cdots$	0
Age	22	38	26	$\cdots$	32
Siblings/Spouses	1	1	0	$\cdots$	0
Parents/Children	0	0	0	$\cdots$	0
Fare	7.25	71.2833	7.925	$\cdots$	7.75

2.2 Histograms

One of the most useful data visualization tools is the histogram. It shows in a plot the distribution of a variable, that is, the frequency with which each value appears. For example, here is a histogram showing the distribution of the *fare* variable.

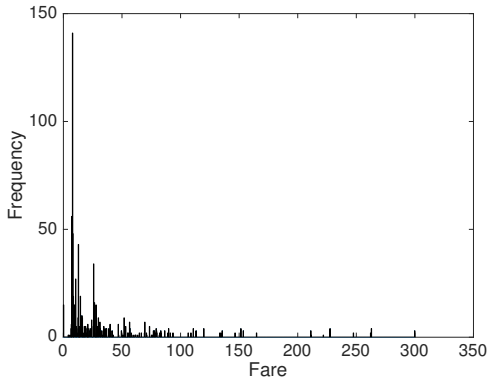


This histogram is telling me that approximately 550 people paid less than $25, that about 180 people paid between $25 and $50, and then fewer and fewer people paid more and more money, up to the point where only a handful paid more than $250. Notice that the histogram is divided by bins, that is, values are grouped in ranges. It is sometimes useful to decide how many bins to use. In the histogram above, I used 12 bins. Here is another histogram of the same data, but with 30 bins instead:

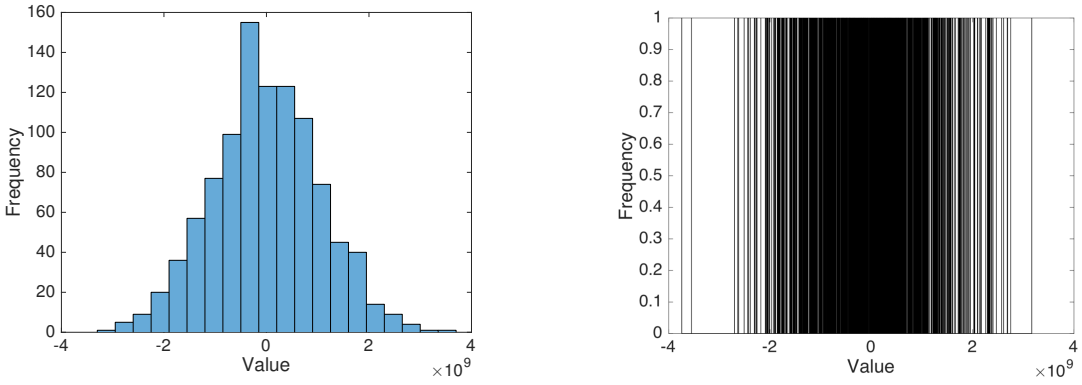


In contrast with the previous histogram, this one provides more detail. It's telling me that approximately 330 people paid less than $10, and approximately 175 paid between $10 and $20. Notice (adding bins) that both histograms agree that about 730 people paid less than $50.

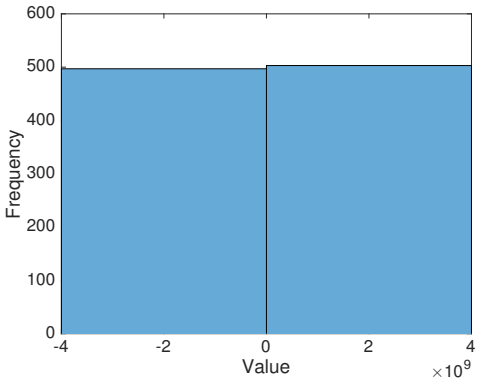
One could feel tempted to use more and more bins. However, using too many bins could backfire! Here is a histogram of the same data with 900 bins:



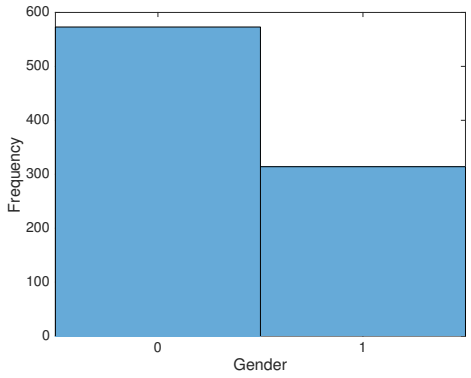
Here you might think it doesn't really look thaaaat bad. However, depending on the data, things can even get much worse. For example, here are two histograms of another dataset using 20 and 1 million bins:



Notice that the histogram on the left shows that the distribution of this data has a gaussian shape. In contrast, the histogram on the right has sooo many bins, that each value falls in a different bin, so that in the end the histogram provides no discernible insight about the data distribution. Of course, a similar problem arises if we use too few bins. Here is a histogram of the same data using only two bins:



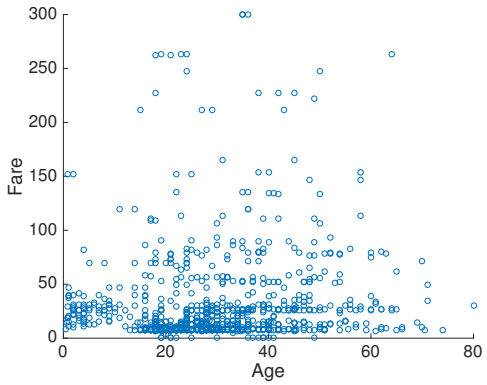
In this case, since we used too few bins, it is also impossible to recognize the distribution of this data as gaussian. Whenever using histograms, it is always important to choose the right number of bins. As a rule of thumb people often use $\sqrt{N}$ bins, where N is the number of data points. However, depending on the dataset this may or may not work, and other options might be better, typically found by trying different values, or by knowledge about the data itself. For example, since *gender* in our Titanic dataset only takes two values (0 for male nad 1 for female), a histogram of this variable can only take two bins:



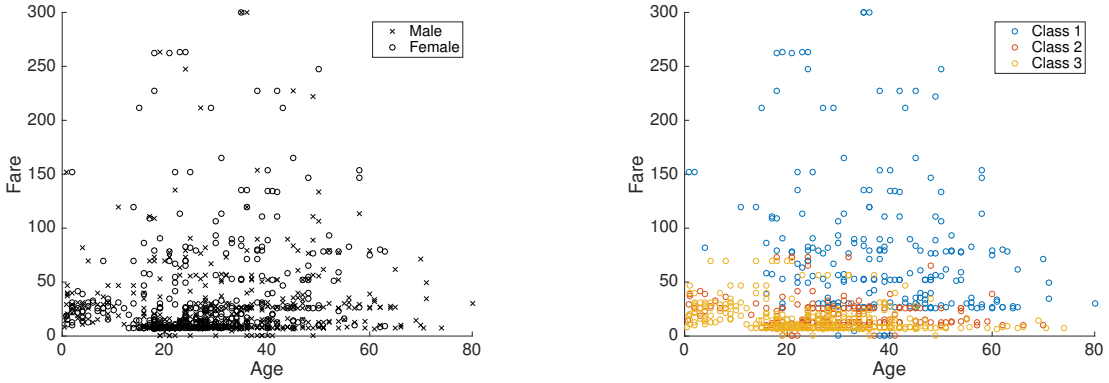
This histogram shows us that there were about twice as many men as women on board the Titanic.

2.3 Scatter Plots

Another useful visualization tool are scatter plots, which basically show the location of each sample in a plane where coordinates are given by two variables. For example, here is a scatter plot of *age* and *fare*:



Scatter plots allow to plot a third categorical variable using symbols or colors. For example, the next scatter plots shows *age* and *fare* classified by *gender* (left) and *passenger glass* (right):

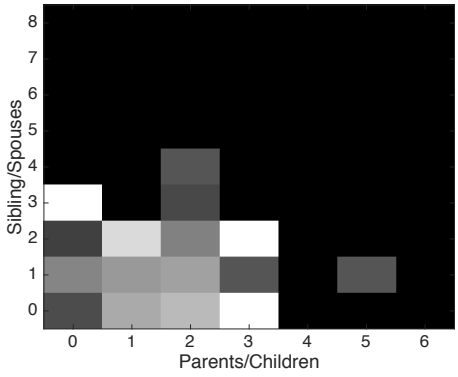


The scatter plot on the left *suggests* that the three most expensive tickets were bought by a couple of around 35 years old, and their friend or brother. The scatter plot on the right shows that some first-class and second-class passengers actually paid less than some third-class passengers! 😱

We could simultaneously use symbols *and* colors to produce a scatter plot showing *age* and *gender* simultaneously classified by *gender* and *passenger class*. What would be the advantages and disadvantages of this?

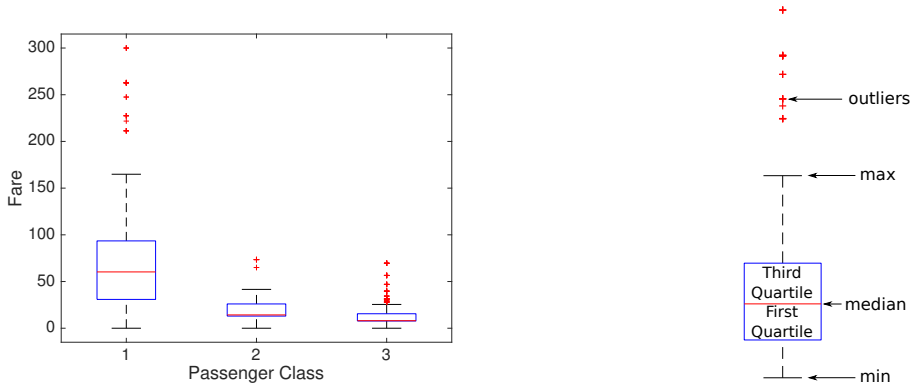
2.4 Heat Maps

Heat maps (aka tile plots) are essentially images showing the value of one variable (color-coded) as a function of two others. For example, here is a heat map indicating the survival rate (the lighter the higher) as a function of the number of *siblings/spouses* and the number of *parents/children*.



2.5 Box Plots

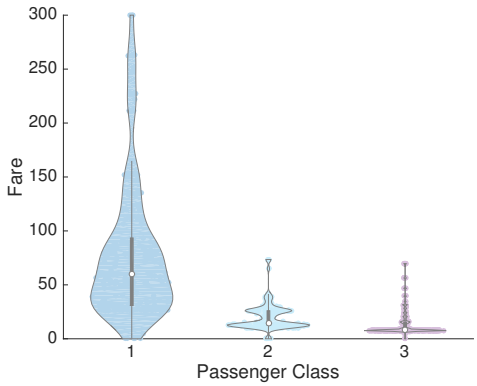
Box plots are useful to visualize a summary of the data distribution. For example, here is a box plot displaying *fare* ranges for each *passenger class*:



The low and top markers show the minimum and maximum values in the dataset. The boxes show the first and third quartiles (containing 25% and 75% of the data). The middle line shows the median, and outliers (extreme values that are considered to be not representative of the data) are show individually as crosses.

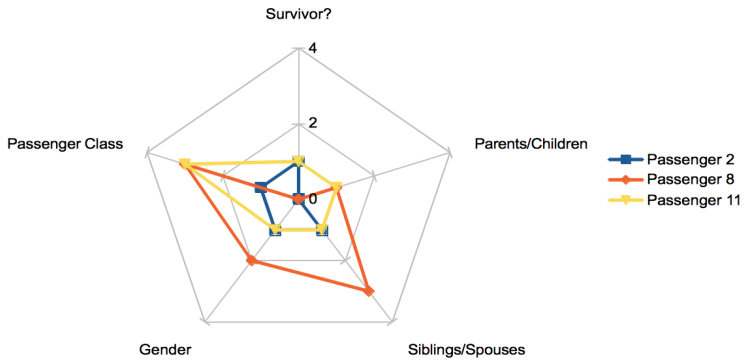
2.6 Violin Plots

Violin plots are like box plots on steroids. Instead of simply showing the minimum, the maximum, a box enclosing 50% of the data, the median, and the outliers (like box plots), violin plots show an interpolated of histogram describing the entire distribution of the data. Here is a violin plot displaying *fare* distributions for each *passenger class*:



2.7 Spider Plots

Spider plots (aka radar or net plots or charts) are useful tools to represent multivariate samples in a two-dimensional chart. Spider plots are formed by several axis starting from the origin (one axis for each variable/feature). In these plots, each multivariate sample is depicted as a figure formed by connecting its features' values in these axes. Here is a spider plot displaying five features of three random passengers:



2.8 Normalization

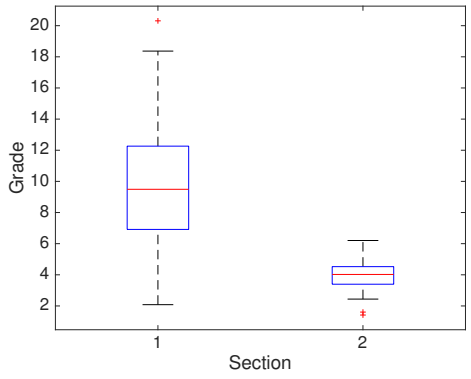
Would you rather get a 4 with professor A, or a 6 with professor B? Of course, it depends on which scale is each professor using. If both professors are using a scale of 1-10, of course I'd prefer the 6. However, if professor A is using a 1-5 scale, and professor B is using a 1-10 scale, then I'd much rather get the 4, which amounts for an 8 in a 1-10 scale. The process of adjusting values in different scales to a common scale is called *normalization*. One common practical strategy is to transform all values into a 0-1 scale. This can be easily done with the following formula:

$$\text{normalized value} = \frac{\text{original value} - \min(\text{original values})}{\max(\text{original values}) - \min(\text{original values})} \tag{2.1}$$

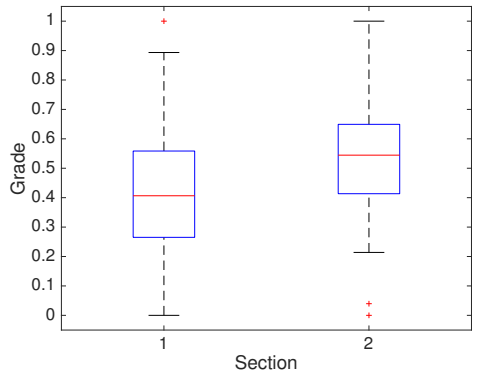
Applying this formula to our grades example, we would get that the normalized grades are

$$\frac{6-0}{10-0} = 0.6 \qquad \text{and} \qquad \frac{4-0}{5-0} = 0.8.$$

Normalization is often a useful step before visualization. It allows more direct and fair comparisons. For example, suppose I give you a box plot summarizing the grades in two sections of CS 4780/6780:



At first glance it might look like that Section 1 did much better than Section 2. However, this could be an artifact of the scale that was used in each section. To obtain a more meaningful comparison, we should normalize data first. Here is a box plot summarizing the normalized grades in the same two sections:

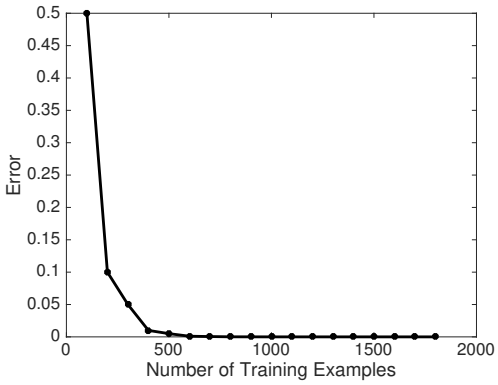


Once we normalize, we can see that contrary to what initially appeared, Section 2 did in fact slightly better than Section 1.

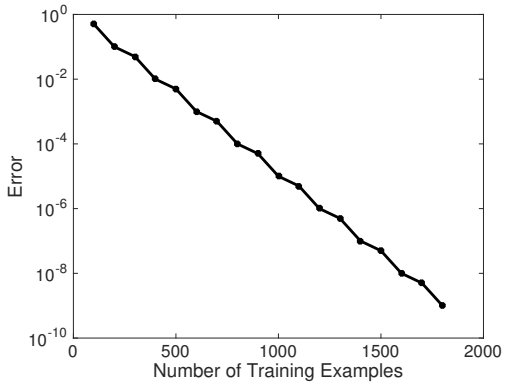
Of course, equation (2.1) is not the only way to normalize data. Other options include adjusting data to other ranges, for example $[-1, 1]$, or 0-100 (you might be familiar with this one; its simply percentiles!).

2.9 Log-scale

Presenting data according to a logarithmic scale is a useful trick to allow a larger range of values in skewed distributions, or ranges that include multiple orders of magnitude. For example, here is a simple plot showing the error rate of a machine learning algorithm as a function of the amount of available training data:



By looking at this plot, it might appear like the error stays steady at zero after approximately 600 samples. However, we can plot the same results in a logarithmic scale (meaning the y-axis is simply transformed according to a log operator):



This shows that even after 600 samples, the algorithm's error keeps decreasing by approximately two orders of magnitude for each 500 samples! In words, this means that the algorithm fails 100 times fewer with each 500 samples. This might seem overly ambitious in some applications, like recommending a Netflix movie, but it could be crucial in critical applications, like self-driving vehicles or automatic medical diagnosing.

2.10 Conclusions

These notes show several data visualization tools that might be useful throughout this course. Of course, there are many other ways to visualize data, some more sophisticated than others, and some more adequate to specific datasets. Additional examples range from basic pie and bar charts, to 3D plots and networked graphs. In general, it is up to the data scientist to decide what is the best way to present results/data in order to convey a message, or the insights they have obtained from data. As a rule of thumb, the simpler the better.