STAT2201

Analysis of Engineering & Scientific Data

Unit 5

Slava Vaisman

The University of Queensland
School of Mathematics and Physics

Descriptive Statistics

- An important aspect of dealing with statistical data analysis is the data summarization and organization.
- Given a statistical data, we would like to estimate some characteristics of a population such as numbers, tables and graphs.
- Descriptive Statistics and data visualization (graphics) are used to achieve this task.

Data types

We start with the discussion of possible data types. From the bird'seye view, the variables can be either *Continuous quantitative* or *Discrete quantitative*.

1. The *continuous quantitative* data represents values in a continuous range. Some examples are height, width, length, temperature, humidity, an object's volume, area, and price.
2. The *discrete qualitative* data (also called *factor* or *categorical*) represents values in (a small) discrete range. For example, the number of family members, gender (male or female), or a count of some object. It is common to further distinguish factors by the following sub-types.
   - *Nominal* factors are variables that represent groups without order, such as males and females.
   - *Ordinal* factors are variable that represent groups with a certain order. For example, consider an *age group* of an individual, say $(0-5)$, $(5-18)$, $(18-45)$, $(45-67)$ and $(67-180)$. These groups can be ordered since the ages can be ordered.

Data configurations (1)

In general, where are many possible data configurations. Each configuration will consist from variables discussed in the previous slide. Below, we discuss a few major configuration types.

- A single sample configuration consists of m scalars: $\mathcal{D} = \{x_1, x_2, \ldots, x_m\}$. Each $x \in \mathcal{D}$ can (for example) correspond to a subject's survival time in a medical experiment.

- Time series (single sample): $\mathcal{D} = \{x_{t_1}, x_{t_2}, \ldots, x_{t_m}\}$, where $t_1 < t_2 < \cdots < t_m$. Such configuration can represent a process such stock price, sensor measurements, etc.

Data configurations (2)

- Two (or more) sets of samples:

$$\mathcal{D} = \left\{ \left\{x_1^1, \ldots, x_{m_1}^1\right\}, \left\{x_1^2, \ldots, x_{m_2}^2\right\}, \ldots, \left\{x_1^k, \ldots, x_{m_k}^k\right\} \right\}.$$

  In this case, $\mathcal{D}$ can represent the survival time of $m_1, \ldots, m_k$ subjects in a medical experiment, in which, k different treatments are tested.

- Data tuples: $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$. Tuples can show a (one-dimensional) correspondence between x and y. In other words, such data represents a possibly unknown function $y = f(x)$. For example, x can stand for height and y for the weight of a subject.

- Tuples can be easily generalized to vectors of observations (of length n). Specifically, the data takes the form:

$$\mathcal{D} = \{(x_1^1, \ldots, x_n^1), \ldots, (x_1^m, \ldots, x_n^m)\}$$

Data tables

- While several additional data representation types such as graphs, are possible, we will consider the classical statistical data representation via a table.
- Under this setting, the table rows and columns represent *independent* observations and the observed measurements (also called *variables, predictors, or features*), respectively. The table rows are (usually) identified by unique identification number (Id).

Table: Data representation by the table

Id	variable 1	variable 2	$\cdots$	variable i	$\cdots$	variable n
1	.	.	.	.	.	.
2	.	.	.	.	.	.
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	.	.	.	.	.	.

Cars

```
mpg cylinders displacement horsepower weight acceleration
model_year origin car_name
-----------------------------------------------------------
0 18 8 307 130 3504 12 70 US chevrolet chevelle malibu
1 15 8 350 165 3693 11.5 70 US buick skylark 320
2 18 8 318 150 3436 11 70 US plymouth satellite
3 16 8 304 150 3433 12 70 US amc rebel sst
4 17 8 302 140 3449 10.5 70 US ford torino
.
.
.
.
.
```

Data summarization

A *statistic* is a numerical quantity that is computed from a sample $x_1, \ldots, x_m$. We present some very common and useful statistics, while distinguishing between factors and quantitative variables.

	count	freq
US	254	0.625616
Japan	79	0.194581
Europe	73	0.179803

We can also study a correlation between the two factor variables using the so called contingency table, which shows the distribution of one variable in rows and another in columns.

cylinders	3.0	4.0	5.0	6.0	8.0
origin					
US	0	72	0	74	108
Europe	0	66	3	4	0
Japan	4	69	0	6	0

Describing quantitative data

A *summary statistics* is a great tool for an exploration of a continuous variable. Specifically, given a data vector of numbers $\boldsymbol{x} = (x_1, \ldots, x_n)$, we calculate the following descriptors.

- The *sample mean* of the data is given by:

$$\overline{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

- The *median* $\tilde{\boldsymbol{x}}$ is essentially the "middle" of the data. To obtain the median, order the data such that $x_1 \leq x_2 \leq \cdots \leq x_n$. If n is odd, $\tilde{\boldsymbol{x}} = x_{\frac{n+1}{2}}$, (the value at position $\frac{n+1}{2}$ in the ordered sequence). If, on the other hand, n is even, then one can take values at positions $\frac{n}{2}$ or $\frac{n+1}{2}$ can be used. In practice, an average between these values is taken.

Describing quantitative data

- The data *range* is calculated via

$$\text{range} = \max_{1 \le i \le n} x_i - \min_{1 \le i \le n} x_i.$$

- The *order statistics*. First, sort the data to obtain $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$, and observe the following.
  1. The minimum: $x_{(1)}$.
  2. The maximum: $x_{(n)}$.
  3. The median:

$$\begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+1}{2}\right)}\right) & \text{if } n \text{ is even.} \end{cases}$$

Describing quantitative data

- The *p-quantile* of $\boldsymbol{x}$ for $0 < p < 1$ is a value z such that a fraction p of $\boldsymbol{x}$ is less than or equal to z and a fraction $1 - p$ of $\boldsymbol{x}$ is greater than or equal to z. Note that 0.5-quantile is actually the sample median.
  A different name of the p-quantile is the $100 \cdot p$ percentile. In descriptive statistics, it is common to report the 25, 50, and 75 percentiles. These are also called the first, second, and third quartiles.
- It is common to report the *sample variance* (spread) of the data. The variance measures the deviation of the data from its mean. The sample variance is given by:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{\boldsymbol{x}})^2,$$

  where $\overline{\boldsymbol{x}}$ is the sample mean. The square root of the sample variance $s = \sqrt{s^2}$ is called the *sample standard deviation*.

Describing quantitative data

- Finally, the sample correlation coefficient r_{xy} is an estimate for the correlation coefficient, ρ, and is given by:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

	mpg	displacement	horsepower	weight	acceleration
count	398.0	406.0	400.0	406.0	406.0
mean	23.5	194.7	105.0	2979.4	15.5
std	7.8	104.9	38.7	847.0	2.8
min	9.0	68.0	46.0	1613.0	8.0
25%	17.5	105.0	75.7	2226.5	13.7
50%	23.0	151.0	95.0	2822.5	15.5
75%	29.0	302.0	130.0	3618.2	17.1
max	46.6	455.0	230.0	5140.0	24.8

The quantile of a probability distribution

- Given $\alpha \in [0, 1]$, what is x such that $\mathbb{P}(X \leq x) = \alpha$?
- By definition:

$$\mathbb{P}(X \leq x) = F(x) = \int_{-\infty}^{x} f(u)\mathrm{du} = \alpha.$$

- Conclusion: to find the quantile, solve the equation for x.

Visualization

When we start a data analysis, it is often beneficial to look at our data. To do so, we need to introduce an appropriate graphical representation for qualitative and quantitative variables. This will allow us to obtain the necessary intuition. Specifically, we will observe the following.

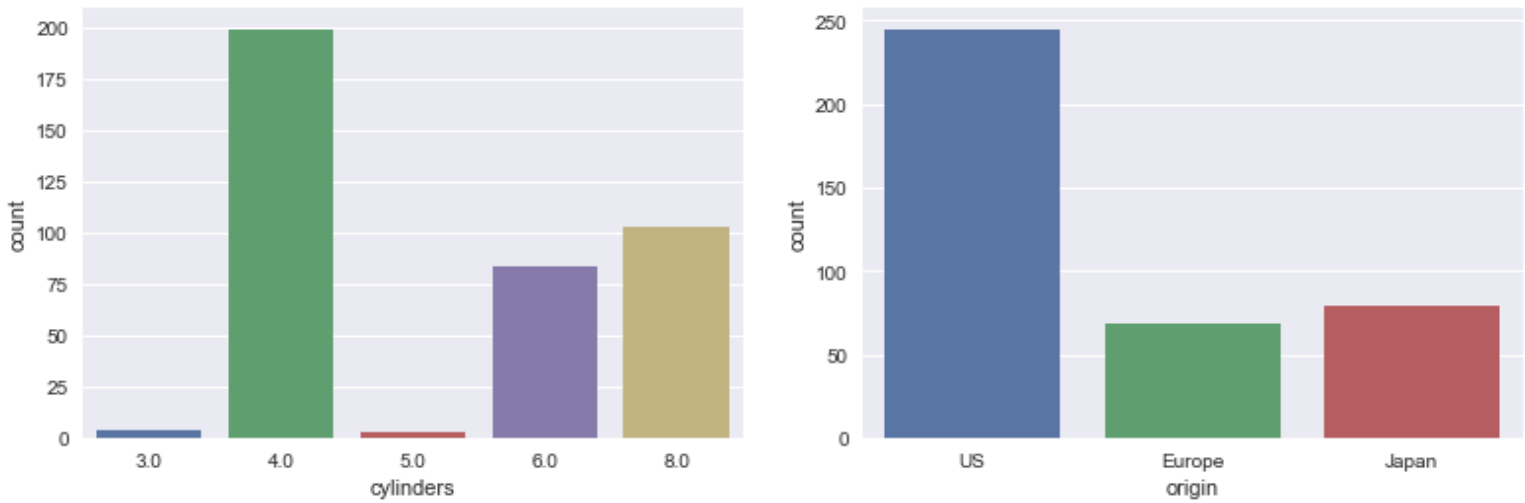
1. For each variable, identify the most common values, and the amount of variability.

2. Diagnose unusual observations.

3. Explore trends in the data.

Single factor variable

The main graph type for a factor variable is a *bar* chart, which counts the categories of the corresponding factor variable. An alternative that is frequently used in industry, is the well-known *pie* chart.

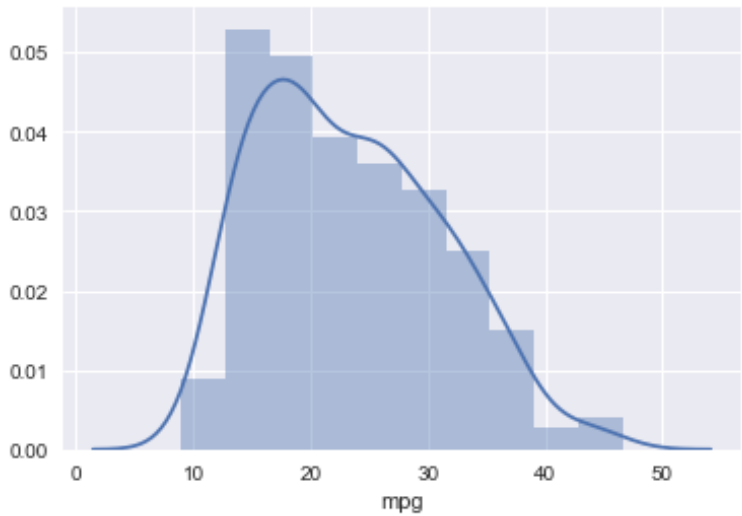


(a) Bar chart cylinders (b) Bar chart origin

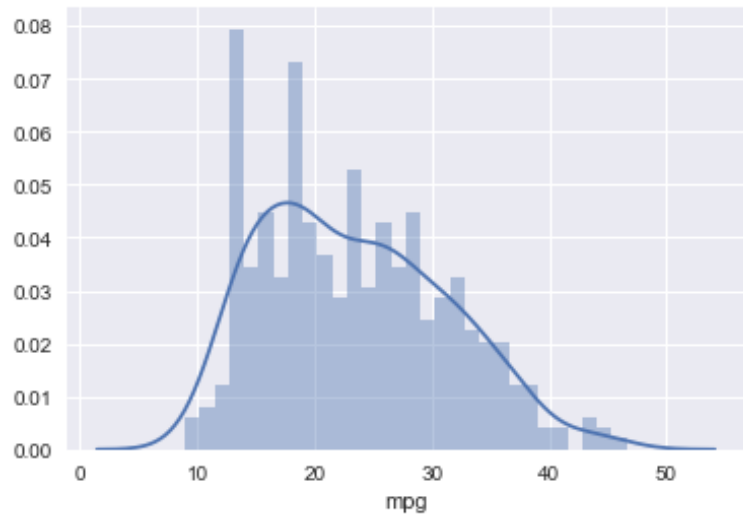
Figure: Bar charts for categorical variables.

Single continuous variable

Similar to the bar plots, a *histogram* is considered as the main graphical tool for visualization of the distribution of a quantitative variable. The main idea is to divide the range of a continuous variable into a predefined number of bins on the x-axis, and plot the associated frequencies for each bin on the y-axis.



(a) Histogram with 10 bins

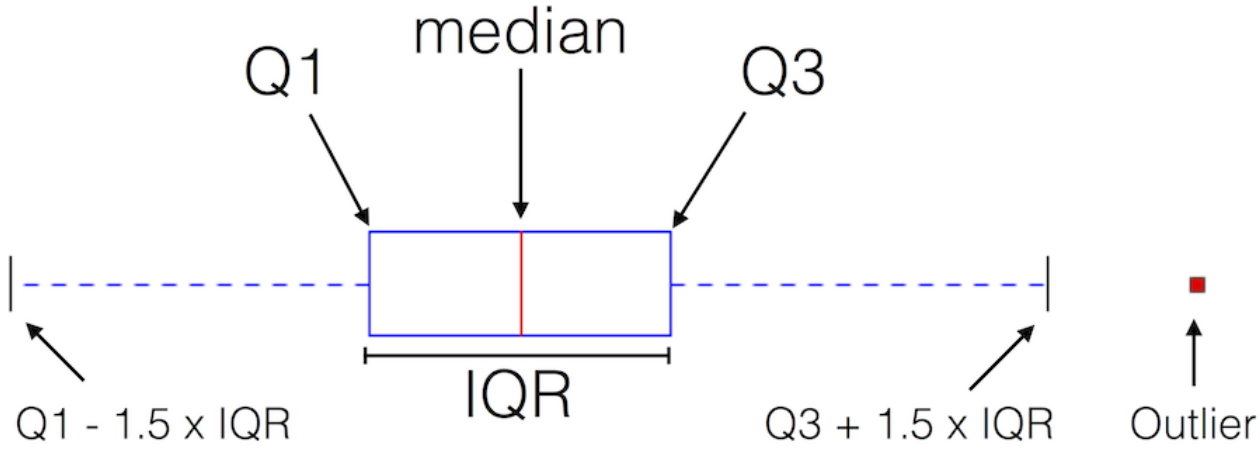


(b) Histogram with 30 bins

Note that both the histogram and the KDE are not unique, since they depend on the number of bins.

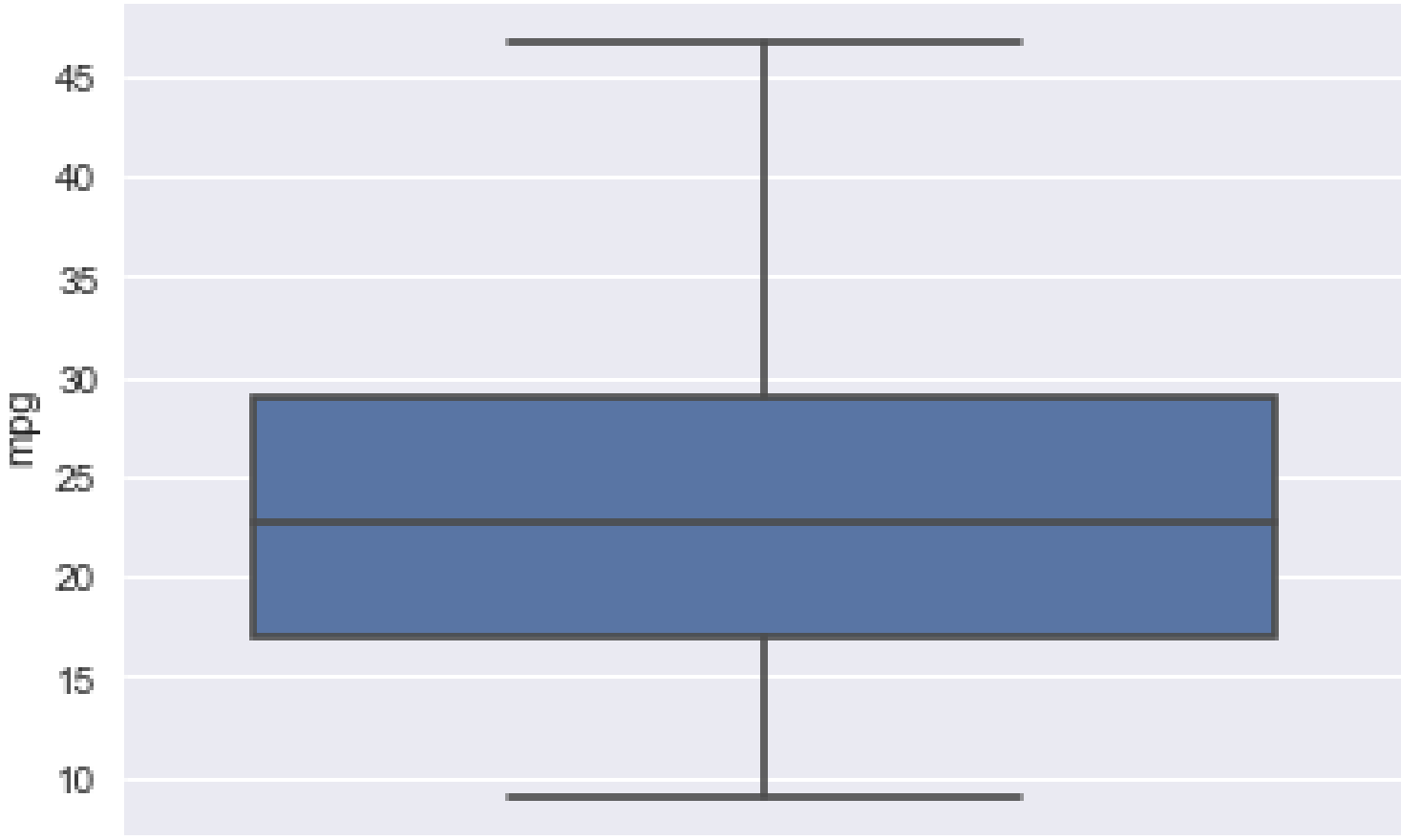
The box plot

The box plot is a graphical display that simultaneously describes several important features of a data set such as *centre*, *spread*, *departure from symmetry*, and *identification of unusual observations or outliers*.



Specifically, this plot is drawn using the values of these three quartiles. The data points that are represented as small circles (drawn outside the box) are outliers that lay further than 1.5 times the interquartile range IQR, where $\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1$.

The box plot



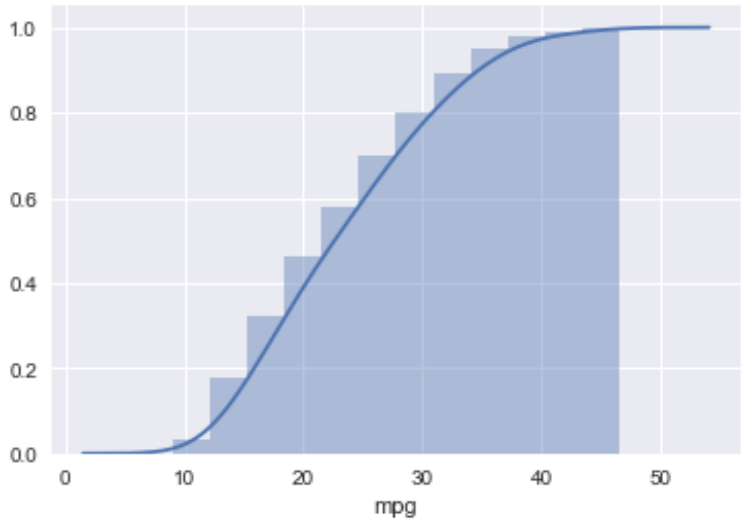
(c) Box plot for miles per gallon.

The Cumulative frequency plot

Of interest is the Cumulative frequency plot. The height of each bar shows the probability (y axis) of x been less or equal to the mpg (x axis). The Empirical Cumulative Distribution Function (ECDF) can be constructed via

$$\hat{F}(x) = \frac{1}{m} \sum_{i=1}^{m} 1_{\{x_i \leq x\}},$$

where $1_{\{.\}}$ is the indicator function.



(d) Box plot for miles per gallon.

Bi-variate scatter plots

To visualize two continuous variables, we use the so-called *scatter plot*, which can be drawn by making two axes, one for each variable, and then adding the values of the variables as the coordinates of the corresponding points.

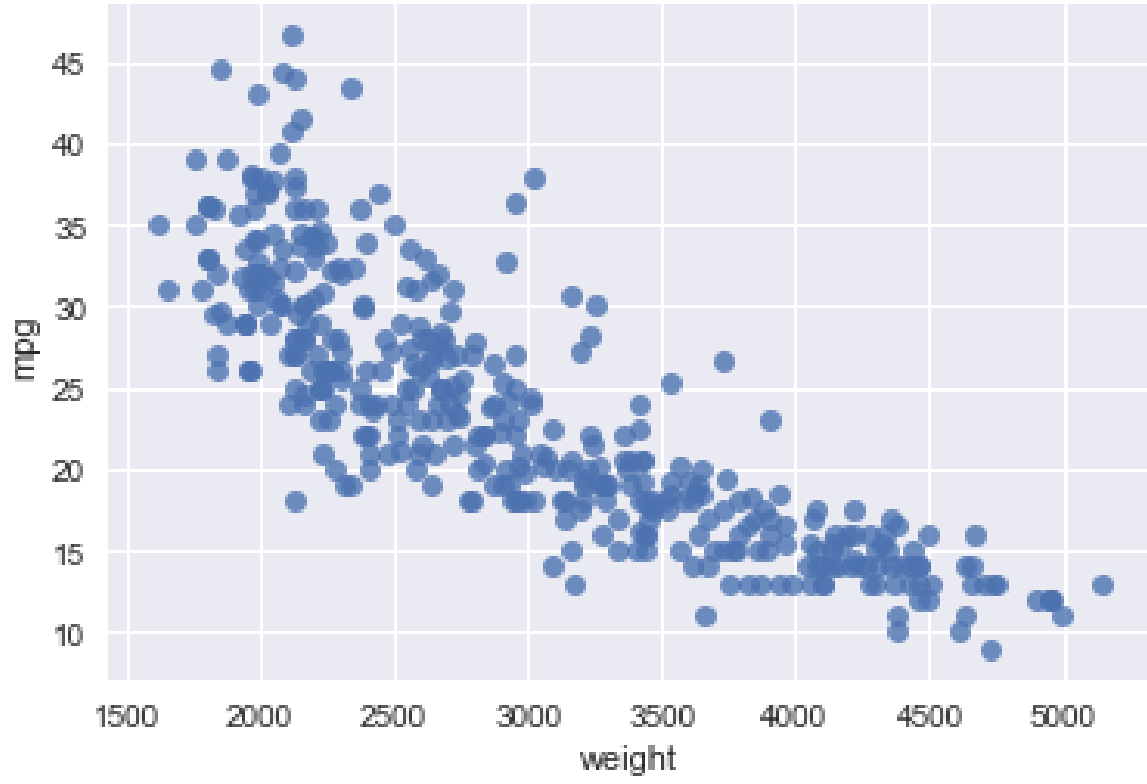


Figure: Scatter plot of two continuous variables.

Mixing variable types

In this cases, we will generally use a categorized box plot as shown in the Figure.

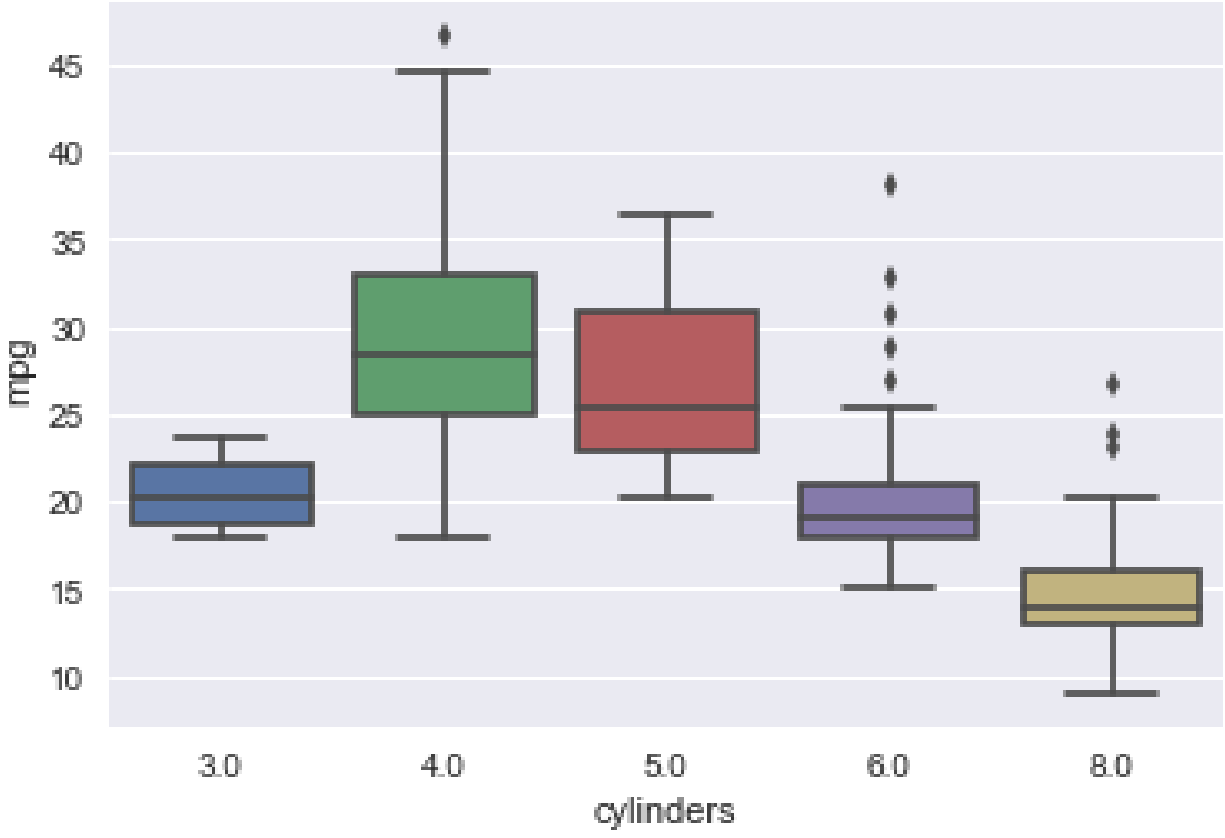


Figure: Box plot by category.