

Exploring and Visualizing Data.

Rick White

November 18, 2015

Outline

- ▶ Types of variables
- ▶ Plotting and summarizing a single variable
- ▶ Categorical variables
- ▶ Numeric variables
- ▶ More than one variable
- ▶ two or more categorical variables
- ▶ one categorical and one numeric variable
- ▶ two numeric variables
- ▶ More than two variables when at least one is numeric

Types of Data

- ▶ Categorical Data
- ▶ Nominal scale: Label with no logical order
- ▶ Ordinal scale: Label but the data can be sorted
- ▶ Numeric data
- ▶ Interval scale: has an arbitrarily defined zero.
- ▶ Ratio scale: has a meaningful zero
- ▶ Categorical and Numeric data are treated very differently
- ▶ Usually distinctions between nominal/ordinal or interval/ratio will not make much difference

Plotting in general

The purpose of graphics is not to be pretty or complicated but to convey information in a clear and unbiased way.

Graphics need not be fancy or complex to be effective. In fact the simplest graphic is usually the most effective.

Adding extra dazzle to a graphic can be distracting and make the graphic confusing and hard to interpret.

Every feature on a graphic should have a clear purpose. If it doesn't have a purpose, remove it.

Pie charts are an example of a graphic that contain added features that otherwise convey no information (3d effects, for example).

Plotting and summarizing one variable

With a single variable the usual goal is to examine the distribution and summarize some features of the data. Rarely is comparing a feature to a specific value the goal.

A feature could be the middle of the data, the dispersion of the data, the most common value or the frequency of an occurrence.

How we summarize and plot categorical variables is very different from how we summarize and plot numeric variables.

Categorical Data

Nominal Data, sometimes referred to as qualitative data, is a name or label but otherwise we cannot talk about relative order other than equality. Do not confuse qualitative data with qualitative research.

Ordinal data is quantitative data because it can be ordered and allow logical comparisons between observations (i.e., small, medium, large). Likert scales are ordinal variables not numeric variables.

We cannot use arithmetic on categorical variables. Small + medium does not make sense.

Categorical variables usually have many subjects with the same observed value for the variable. The frequency of occurrence is what interests us.

Tabulating data

We tabulate categorical data by counting the occurrence of each unique value. Presenting the data as a percentage is usually more informative but it does hide the sample size.

```
## Titanic:
```

```
## Number of passengers by Class
```

```
## Class
```

```
## 1st 2nd 3rd Crew Sum
```

```
## 325 285 706 885 2201
```

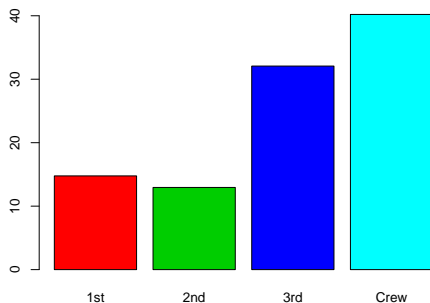
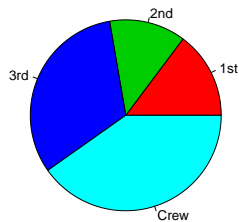
```
## Percentage of passengers by Class
```

```
## Class
```

```
## 1st 2nd 3rd Crew Sum
```

```
## 14.8 12.9 32.1 40.2 100.0
```

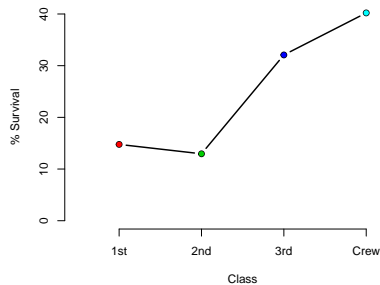
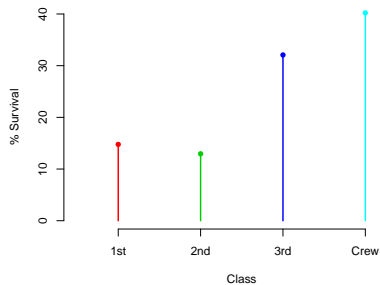
We can plot the data using either a pie chart or a bar chart.



Which graphic do you find more informative?

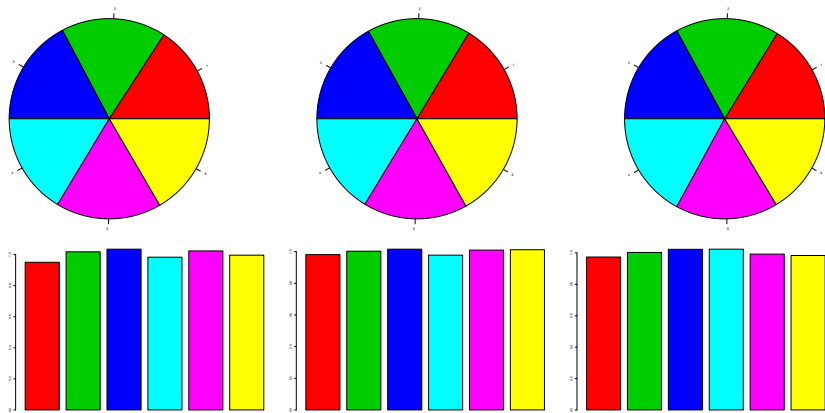
The eye is good at judging linear measures and bad at judging relative areas.

Dot chart or line chart



These use less ink but display the same information as a bar chart. The line chart allows several lines to be displayed simultaneously which can provide better group comparisons than a bar chart or dot chart.

Pie Charts vs Bar Charts:



See Wikipedia, [here](#) or [here](#) for more information.

Numeric Variables

A numeric variable gives real numbers that can be interpreted directly. Ten kilometers is 5 kilometers further than 5 kilometers. Ten Celsius is 5 degrees warmer than 5 Celsius.

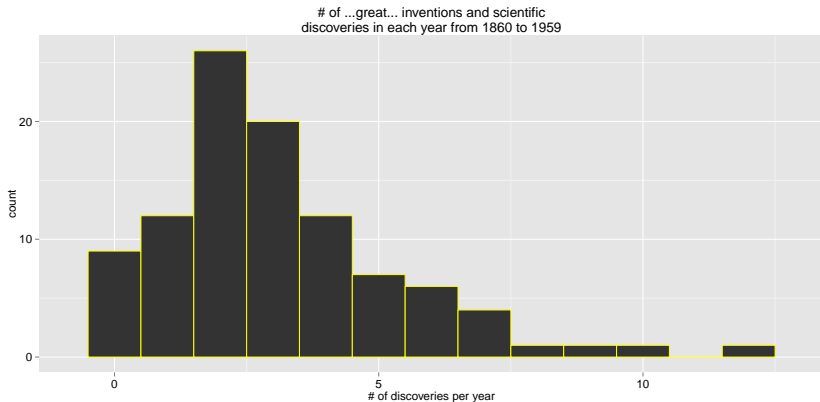
They can be discrete or continuous (number of people a room, number of correct answer on an exam, a person's height, time to complete a task).

Ratio data has a meaningful zero which allows relative values to be interpreted but ratio's do not make sense for interval data. Ten kilometers is twice as far as 5 kilometers but 10 Celsius is not twice as warm as 5 Celsius.

Assigning numbers to an ordinal scale does not make it numeric. (Likert scales are not numeric).

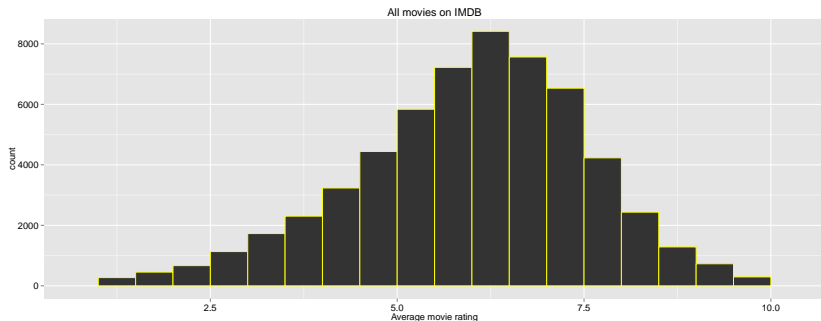
Plotting a numeric variable

For discrete numeric variables with relatively few values use a bar chart as we did for categorical variables.



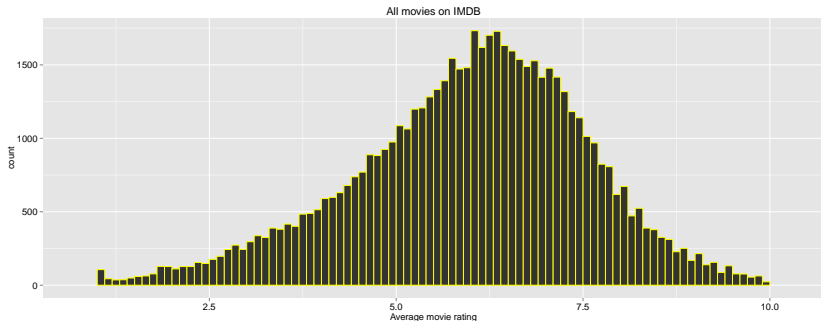
Histograms

If we have too many categories or we have continuous data we can create small intervals and count the number of observation contained in each interval.



Most software will choose a bin width automatically but you can change it if you like. Here I chose a bin width of 0.5.

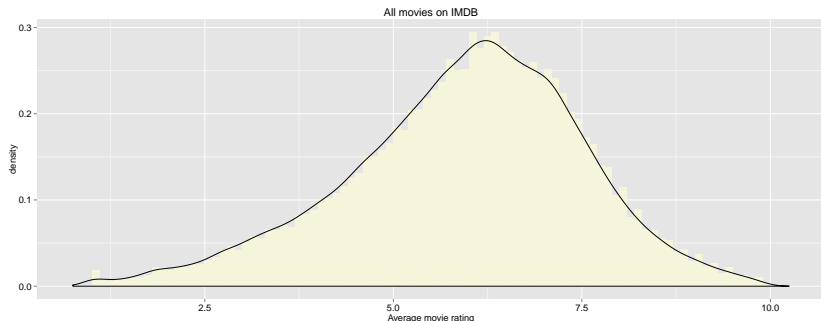
The choice of bin width can be very important.



With a large sample size ($n=58788$ for IMDB movies), we can choose a relatively small bin width. The graphic is not as “smooth”.

Density Plots instead of histograms

Put a smooth bump over each observation. Add up the bumps to get the final graphic.



Like a histogram there are parameters that can be adjusted to control the smoothness of the plot but most software provides reasonable defaults.

Summarizing a numeric variable

Unlike a categorical variable where all the information is captured in a table, we summarize numeric data into single values that describe a certain feature of the data.

We begin by finding the central tendency or “middle” of the data.

The most common measure is the sample mean denoted by \bar{x} , computed by summing the data then dividing by the sample size ($\sum x_i/n$).

Another measure is the median. The median \tilde{x} is any value for which at least 50% of the data are $\leq \tilde{x}$ and 50% of the data are $\geq \tilde{x}$.

Example

```
## Data is  $X = \{ 1 2 12 17 18 \}$ 
```

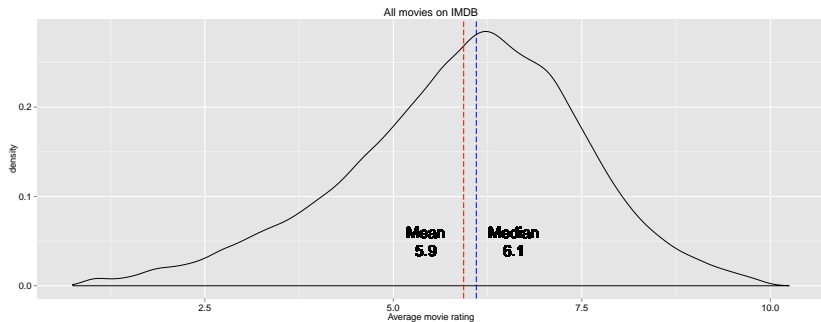
```
## Sum of  $X = 50$ 
```

```
## Mean of  $X = 10$ 
```

```
## Median of  $X = 12$ 
```

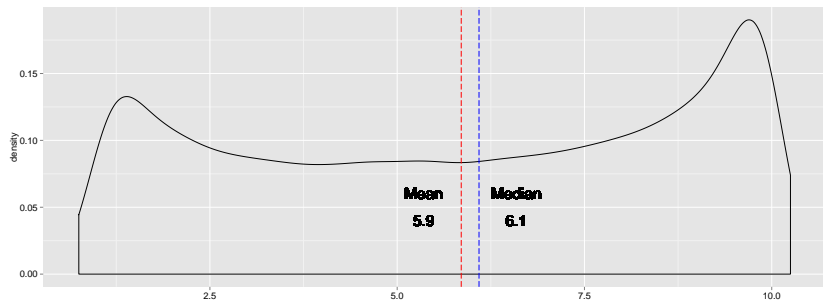
Note if the sample size is odd, the median is determined by a specific observation in the data. If the sample size is even, the median can be any value between the two middle values of the data. The common choice is the average of the two middle values.

Add mean and median of the IMDB data with vertical lines.



Should the mean be to the left of the median or do you think this just occurred by chance?

The mode is not a good measure of central location.



Variation

After the central location we usually describe the spread of the data.

The most common measure is the standard deviation denoted by

$$s = \sqrt{\sum(x_i - \bar{x})^2 / (n - 1)}$$

Other measures include the inter-quartile range (IQR) and the median absolute deviation (MAD).

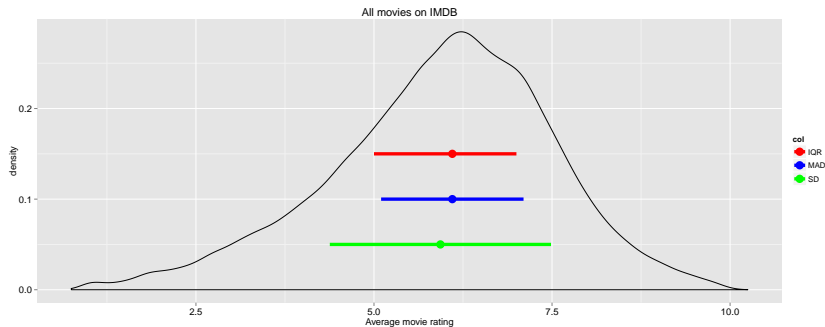
The quartiles are similar to the median except they split the data into quarters instead of halves.

The median is the second quartile. The IQR is the difference between the 3rd and 1st quartile.

The MAD is $\widetilde{|x_i - \tilde{x}|}$. It is usually scaled to make the estimate equal to the SD for a normal distribution.

Summary statistics for the average rating in the IMDB data.

##	N	Mean	Median	SD	IQR	MAD
##	58788	5.93	6.1	1.55	2	1



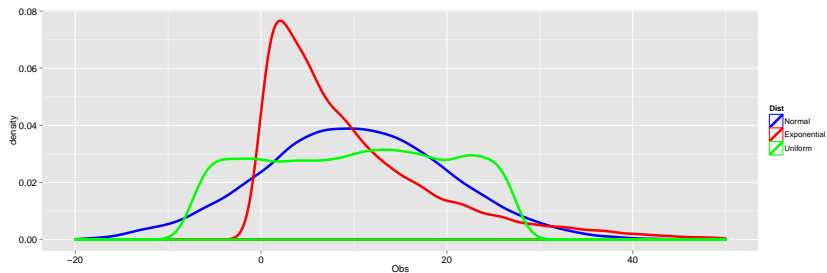
Other numerical summaries

Other features used to describe numeric variables is the skewness, which measures the asymmetry in the data, and the kurtosis, which measures the “peakedness” of the data.

The mean, standard deviation, skewness and kurtosis are the first four moments of the data. We do not have names for higher order moments and we typically never consider them.

The median, quartiles, percentiles are collectively known as quantiles. Quantiles are used to describe the empirical distribution of the data which provides a much more detailed picture of the data than the moments.

These distributions have the same mean and sd.



```
##          Dist      mean      sd
## 1      Normal 10.09330  9.988894
## 2 Exponential 10.14665 10.063010
## 3      Uniform 10.15216  9.943505
```

We can distinguish the distributions by their skewness and kurtosis.

```
##           Dist mean    sd    skew    kurt
## 1      Normal 10.1  9.99  0.0285 -0.0466
## 2 Exponential 10.1 10.06  2.0170  6.7805
## 3      Uniform 10.2  9.94 -0.0349 -1.1783
```

Skewness and kurtosis are difficult to interpret.

Lets try some quantiles instead.

```
##           10%  25%  50%  75%  90%
## Normal      -2.69 3.35  9.98 16.86 23.00
## Exponential  1.09 2.97  7.08 14.06 23.32
## Uniform     -3.78 1.50 10.48 18.59 23.82
```


More than one variable

We are usually interested in looking at relationships in the data. How we examine the relationship might depend on the type of data we are considering and the roles they play.

- ▶ Are some variables considered responses while others considered predictors?
- ▶ Do we want to see the form of the relationship between the variables or just determine if some features are different for various subgroups.

Data should come in a rectangular format. Each row is a sampled unit and each column is a variable measured on that unit. If there are repeated measures the data could be presented in a wide or narrow format.

More than one categorical variable

We cross tabulate the occurrences.

```
## The fate of the passengers of the Titanic
```

```
##           Sex           Male           Female
##           Survived    No Yes    No Yes
## Class
## 1st           118   62           4 141
## 2nd           154   25           13 93
## 3rd           422   88          106 90
## Crew          670  192            3 20
```

Here we have the passengers split by Class, Sex and Survival, 3 categorical variables. Sex and Survival have 2 levels while Class has 4 levels. Class is a nominal variable while Sex and Survival are binary variables.

We still want to talk about percentage but what is the appropriate denominator to use?

It depends on what you want to compare.

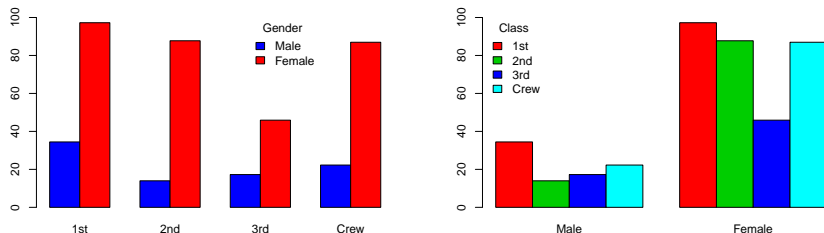
Let go back to the Titanic example. It makes sense that we want to draw conclusions about survival rates. Therefore we should compute our percentage of survival within Class and Sex.

```
## Survival percentage of Titanic passengers
```

```
##           Class
## Sex       1st  2nd  3rd  Crew
##  Male   34.4 14.0 17.3 22.3
##  Female 97.2 87.7 45.9 87.0
```

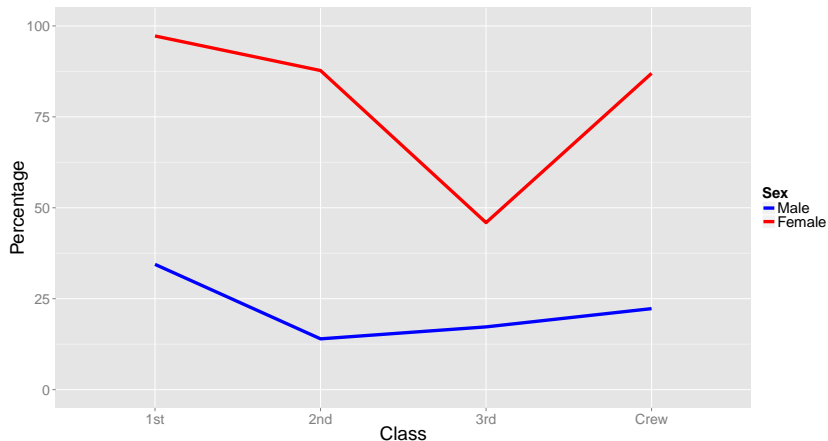
We can use a bar chart to display the data for more than one categorical variable and distinguish the extra groups by colour.

Titanic: Percentage Passenger Survival



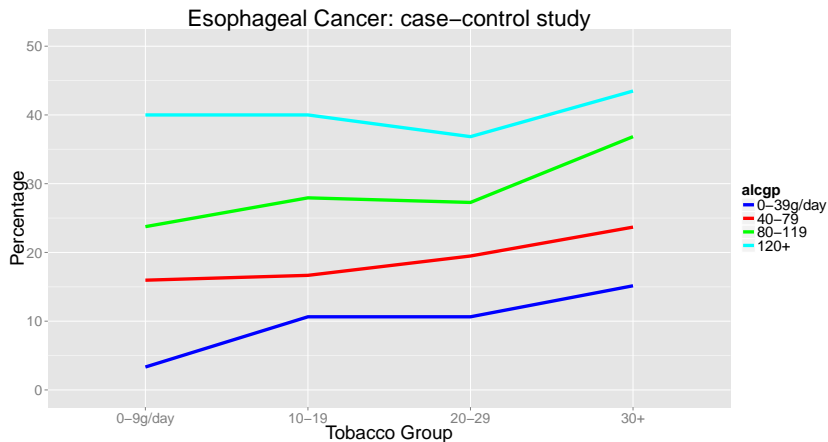
This could get very busy if we had more categories.

We can use a line chart instead of a bar chart

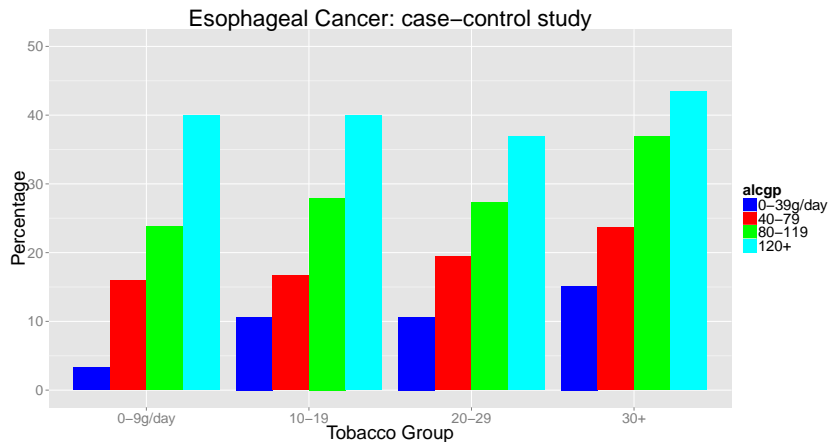


This type of plot can be handy if the number of categories becomes large.

Cases and controls with esophageal cancer in 4 alcohol consumption groups by 4 tobacco consumption groups.

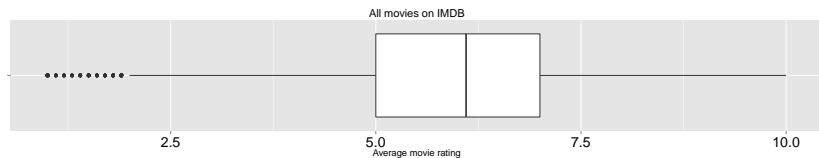


Compare the previous line charts to a bar chart.



Which do you prefer?

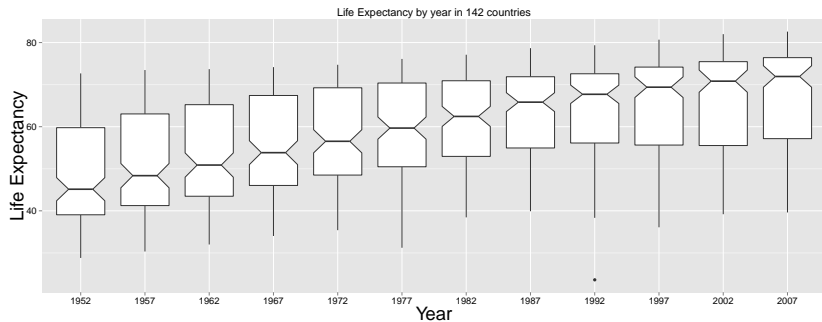
One Numeric Variable and one categorical variable



You can plot a numeric variable using a Boxplot. This is based on summary statistics rather than all the data but can be useful when comparing data between several groups.

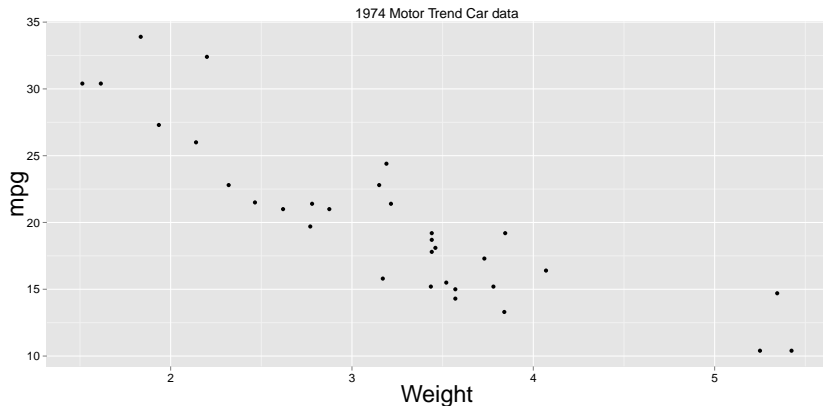
The central line is the median. The box covers the lower to upper quartile. The whiskers extend to an “acceptable” limit. Points outside these limits are possible outliers.

Comparing a numeric variable across the levels of a categorical variable.



We can quickly inspect the difference in location (median), dispersion (IQR) and skewness. The “notch” is a 95% confidence interval for the median.

Scatterplots for two numeric variable



Select one variable for the y axis and one for the x axis. This defines a grid of possible values for x and y data pairs. Put a point on the grid for each data pair where the x value and y value cross.

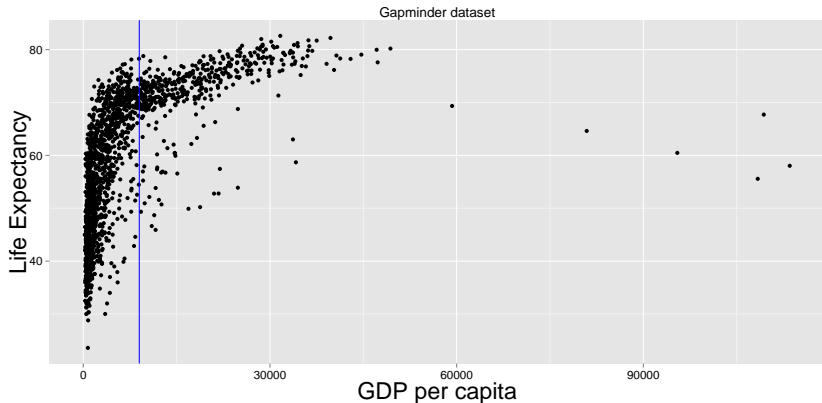
We quantify the relationship between 2 continuous variables by the correlation.

The usual correlation coefficient is the Pearson product-moment correlation r . It quantifies the linear relationship between 2 variables. $(-1 \leq r \leq 1)$ is a unitless quantity where 0 means no relationship and ± 1 is a perfect linear relationship.

Other measures of correlation are Spearman's rank correlation (r_s) and Kendall's τ . Both measure how likely one variable will increase with another without requiring a linear relationship.

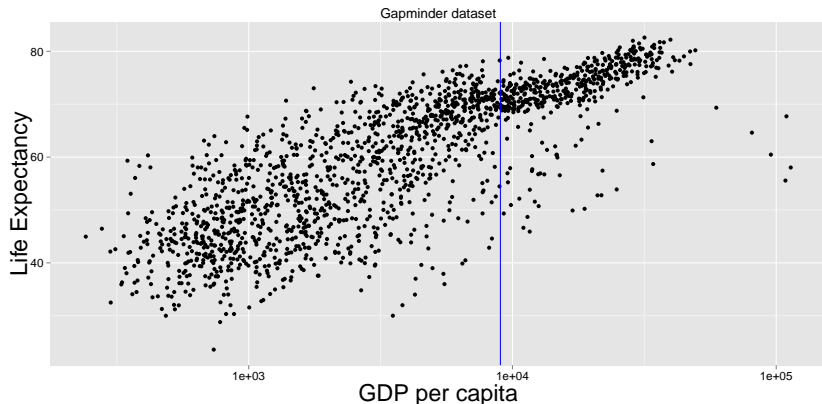
In the previous example, the correlation between mpg and weight by each measure is $r = -0.87$, $r_s = -0.89$ and $\tau = -0.73$.

Gapminder data from Ted Talk.



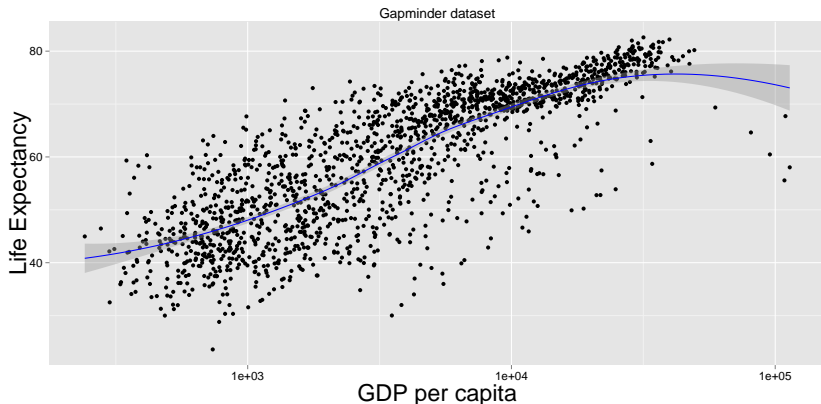
There is a definite relationship between life expectancy and GDP/capita but it is nonlinear. For this data $r = 0.58$, $r_s = 0.83$ and $\tau = 0.64$

Express the GDP/capita on a logarithmic scale.



The relationship looks more linear and we have a better resolution of the data at the lower values of GDP/capita. Now $r = 0.81$, while $r_s = 0.83$ and $\tau = 0.64$ do not change.

Add a curve that approximates the relationship.



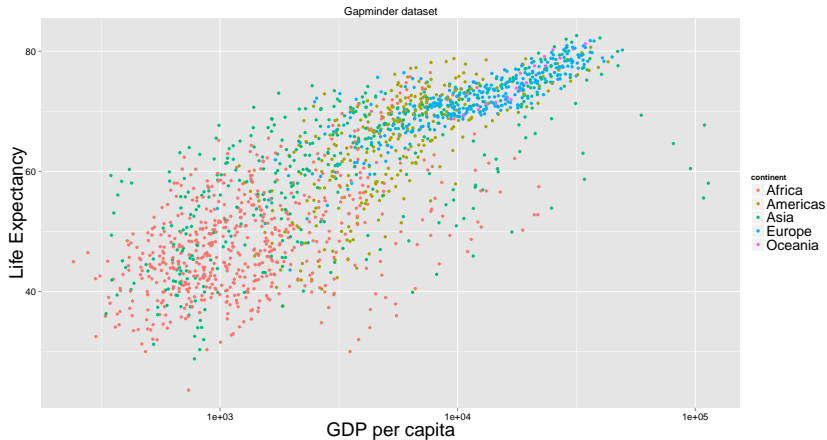
We add a smooth curve estimated from the data. A linear fit seems reasonable except at extremely large values of GDP/capita.

More than 2 variables when at least one is numeric.

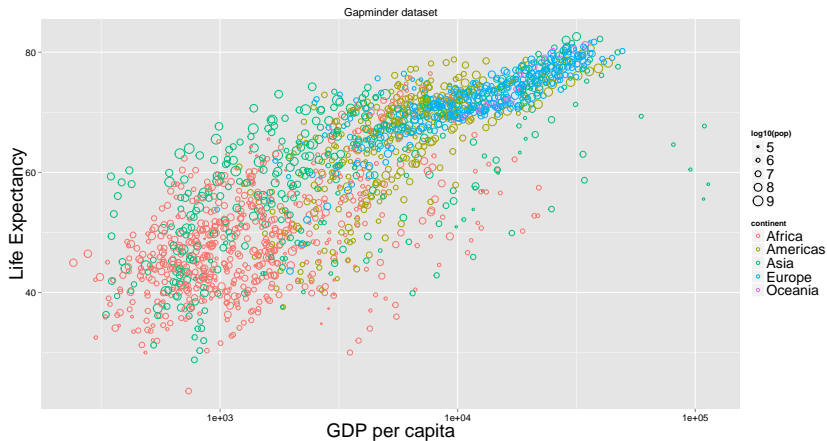
Each variable in a graphic uses one dimension of space but most graphic devices are still only two dimensional. Previously with 3 categorical variables we used colour to represent the third dimension. This is commonly done in graphics. Depending on the information we need to add, we can use colour, size or symbol to represent the additional data. Using all three we can have up to 5 dimensions represented on the plot.

When designing such a plot, you need to make sure the graphic does not become too cluttered and hard to interpret. Different symbols are useful for a categorical variable with only a few categories. Size might be used to represent a positive numeric variable. Colour is the most versatile.

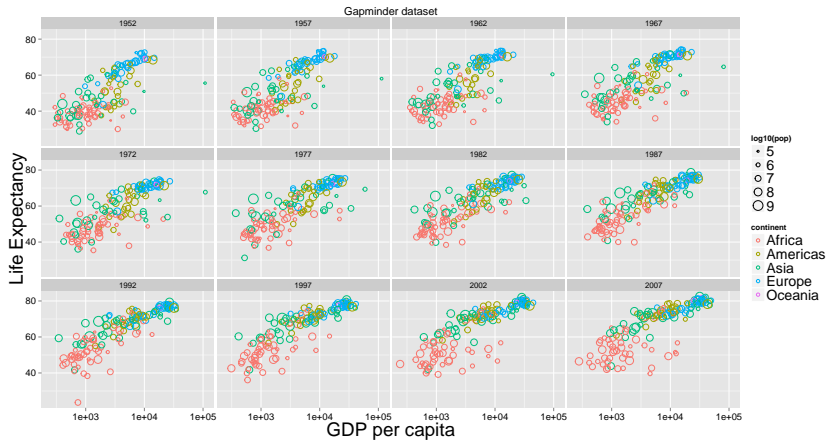
Let's add continent information to the previous graphic.



Let's add population information to the previous graphic.

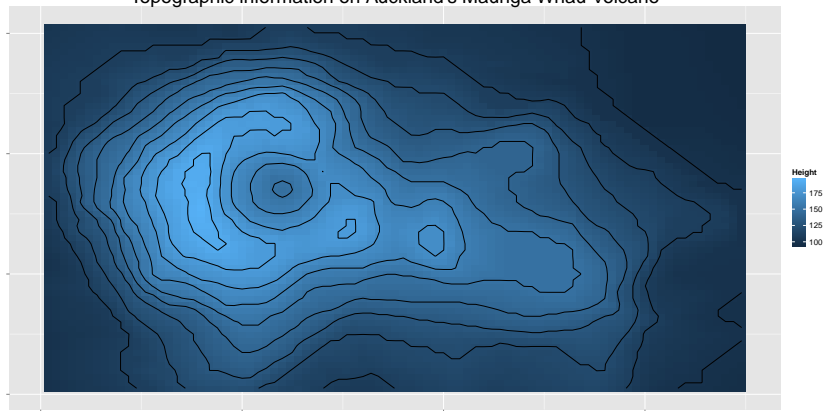


One final detail is year.



Contour Plots

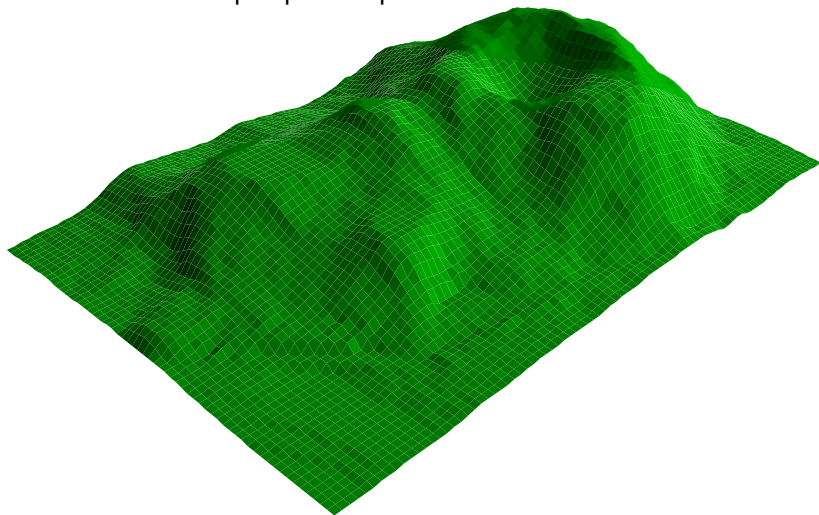
Topographic information on Auckland's Maunga Whau Volcano



If you have 2 dimensions that form a grid for a third numeric variable you can use a contour plot.

Perspective plots

We can also use a perspective plot.

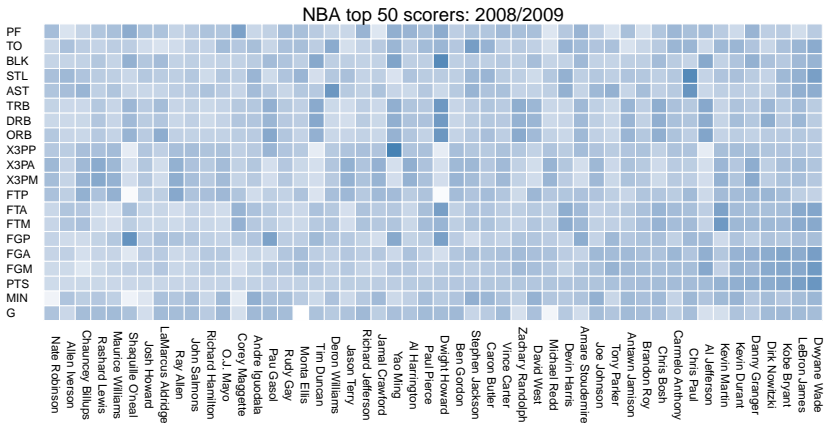


The challenge is picking a good viewing point.

Heatmaps

If you have 2 discrete dimensions and you wish to display numerical information for each combination of the categorical variables you can use a heatmap.

Using Name as id variables



Questions?

www.stat.ubc.ca/SCARL

STAT 551 - Stat grad students taking this course offer free statistical advice. Fall semester every academic year.

SOS Program - An hour of free consulting to UBC graduate students. Funded by the Provost and VP Research.

Short Term Consulting Service - Advice from Stat grad students. Fee-for-service on small projects (less than 15 hours).

Hourly Projects - SCARL professional staff. Fee-for-service consulting.

The End