ITEC874 — Big Data Technologies

Week 10 Lecture 1: Visualising Big Data

Diego Mollá

ITEC874 2019H2

Abstract

It's said that a picture is worth a thousand words. Images can quickly convey complex information, especially for non-technical audiences, but you need to get it right. This lecture will focus on methods and techniques for visualising Big Data. We will cover the typical uses of data visualisation, the most common techniques, and address common problems when displaying information visually.

Update October 11, 2019

Contents

Reading

- Wickham, H. 2009. ggplot2: Elegant graphics for data analysis.

1 Data Visualisation

1.1 Why Visualising Data?

Why Visualising Data?

- It's said that a picture is worth a thousand words.
- Images can quickly convey complex information.
- But you need to get it right.

https://en.wikipedia.org/wiki/A_picture_is_worth_a_thousand_words

Uses of Data Visualisation

Uses of Data Visualisation

- Exploratory Analysis for data analytics.
- Presentation of results from data analytics.
- *Visual Analytics*: As a substitute for data analytics.

Visual Analytics: Wikipedia Definition
Visual analytics is "the science of analytical reasoning facilitated by interactive visual interfaces." It can attack certain problems whose size, complexity, and need for closely coupled human and machine analysis may make them otherwise intractable.

Visualisation in Data Mining Projects

Visualisation can be integrated in several of the steps in a Data Mining Project:

1. Develop an understanding of the purpose of the data mining exercise.
2. Obtain the data set, e.g. by sampling.
3. *Explore, clean, and preprocess the data.*
4. Reduce and partition the data.
5. Determine the data mining task and technique.
6. Iterative implementation and parameter tuning.
7. *Assess the results; compare models.*
8. Deploy the best model.
9. *Evaluate or Monitor Results.*
10. Start all over again!

Uses of Visual Analytics

Visual Analytics is particularly useful for *descriptive* and *diagnostic* analytics.

Descriptive Analytics
Analyse past and present data with the aim to understand it.

Diagnostic Analytics
Analyse past and present data to determine what happened and why.

Predictive Analytics
Use models based on past data to predict the future. The deliverables are usually a predictive forecast.

Prescriptive Analytics
Use models to specify what actions should be taken. This is the most valuable kind of analysis.

Visual Analytics is not the same as Image Analytics!

Visual Analytics
Use visualisation as a tool to analyse data.

Image Analytics
Analyse image data.

Use Cases of Image Analytics

1. Identify bags at airports.
2. Analyse social media images for missing persons.
3. Real-time vehicle damage assessment.
4. Detect pneumonia from chest x-rays.

2 Common Building Blocks for Data Visualisation

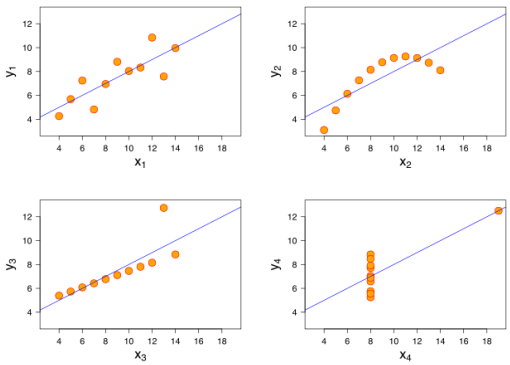
2.1 Visualisation for Summarising Data

Common Summary Statistics
Some of the most common metrics are:

- *Average* or *mean*: $\bar{x} = \frac{\sum x_i}{n}$
- *Median*: The value in the middle.
- *Minimum*, *Maximum*, *Range*.
- *Standard deviation*: $s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$
- *Counts*, *Percentages*
- *Quartiles*: Q1, Q2, Q3, Q4

The Problem with Summary Statistics
They may not suffice to describe the data

Anscombe's Quartet



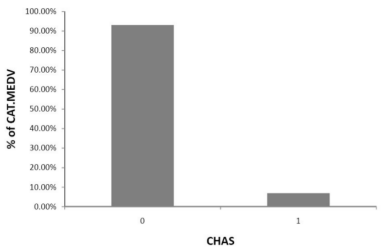
Mean of x: 9

Mean of y: 11
Variance of y: 4.125
Correlation between x and y: 0.816

From Wikipedia (*https://en.wikipedia.org/wiki/Anscombe's_quartet*):
Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."

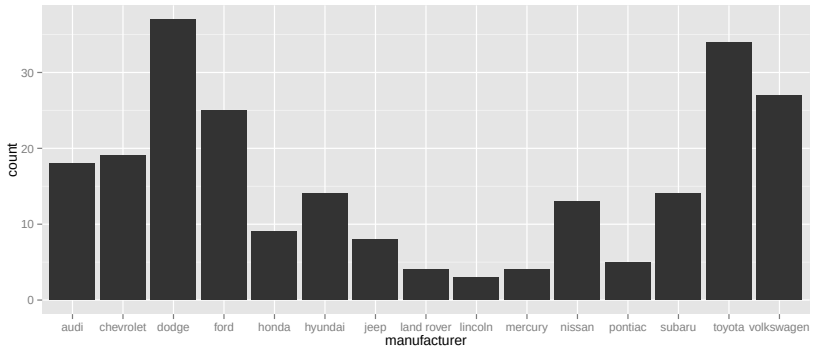
Bar Chart

- Can be used for categorical variables.
  - Numerical variables need to be *binned* to intervals.
- Each bar represents the counts of the value of a variable.
- In this example, 95% of tracts do not border Charles River.



This bar chart uses the Boston Housing data set (*https://www.kaggle.com/c/boston-housing*) and plots the percentage of CAT MEDV by CHAS, where CAT MEDV represents the median value of properties in a suburb, and CHAS has the value 1 if a suburb borders the Charles River.

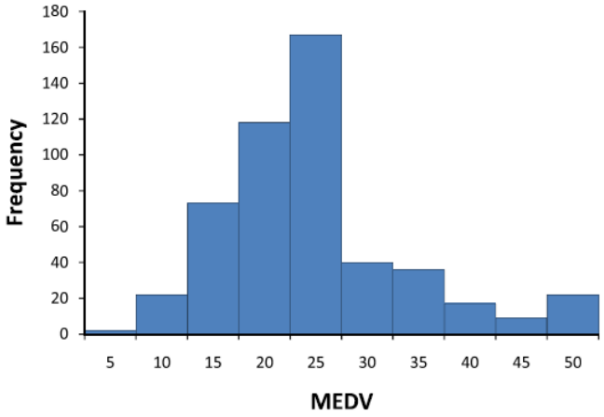
Bar Charts of Categorical Data



This bar chart uses the Auto-MPG data set (*https://ggplot2.tidyverse.org/reference/mpg.html*) and plots the counts of numbers of samples of each car manufacturer.

Histogram

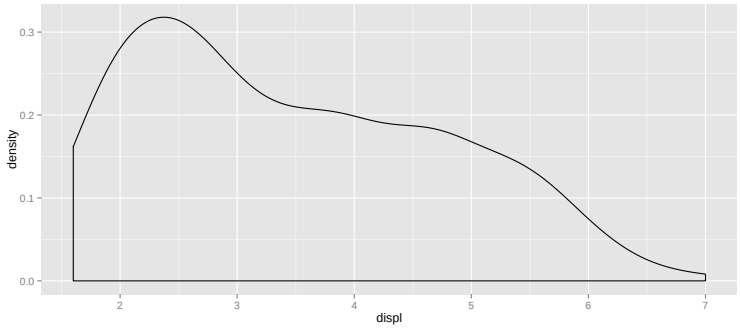
- A bar chart that shows the *distribution* of the values of a *numerical* variable.
- The values are *binned* and then counted.



This histogram uses the Boston Housing data set (*https://www.kaggle.com/c/boston-housing*) and plots the distribution of values of the MEDV variable (median value of houses in a suburb).

Density Plots of Continous Data

- A density plot approximates a histogram using a continuous line.



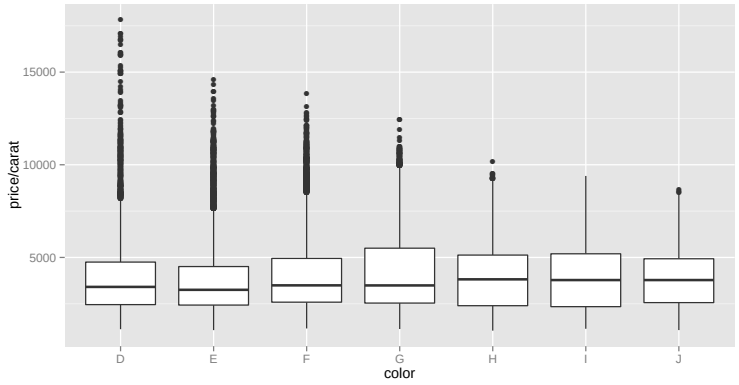
This density plot uses the Auto-MPG data set (*https://ggplot2.tidyverse.org/reference/mpg.html*) and plots the density of the variable "displacement".

Boxplot

- Also called "quartile plot" or "box and whiskers".
- The box delimits the data between Q1 and Q3, which is called the *interquartile range* (IQR).
- The line in the box is the median.
- The whiskers delimit the maximum allowed; several possibilities, including:
  - Absolute maximum and minimum.
  - 1.5 IQR of the lower and upper quartiles.
    - $Q3 + 1.5 \times (Q3-Q1)$
    - $Q1 - 1.5 \times (Q3-Q1)$
- Any values outside the whiskers are usually plotted as circles.

Side-by-side Boxplots

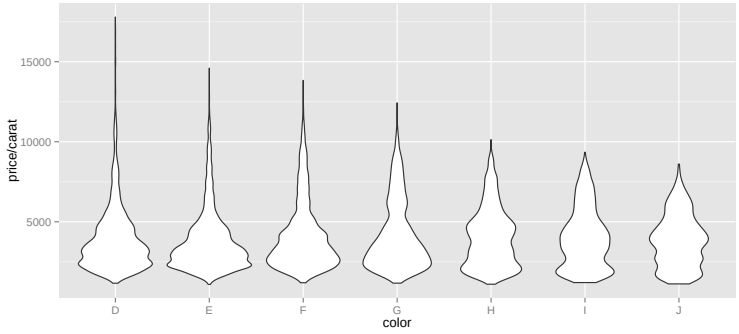
Side-by-side boxplots are useful for comparing subgroups.



These side-by-side boxplots use the diamonds data set (*https://ggplot2.tidyverse.org/reference/diamonds.html*) and plots the distributions of price/carat in each of 7 colors.

Violin Plots

Violin plots incorporate a density plot of each subgroup.

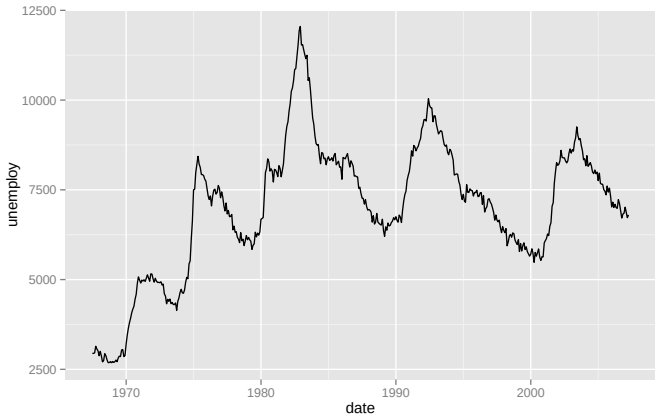


These side-by-side violinplots use the diamonds data set (*https://ggplot2.tidyverse.org/reference/diamonds.html*) and plots the distributions of price/carat in each of 7 colors.

2.2 Visualisation for Assisting Prediction

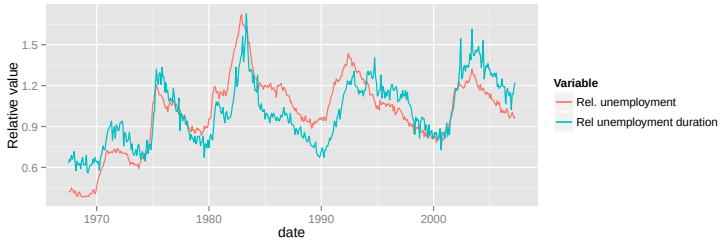
Line Graph for Time Series

Useful to display the values of a variable in sequence.



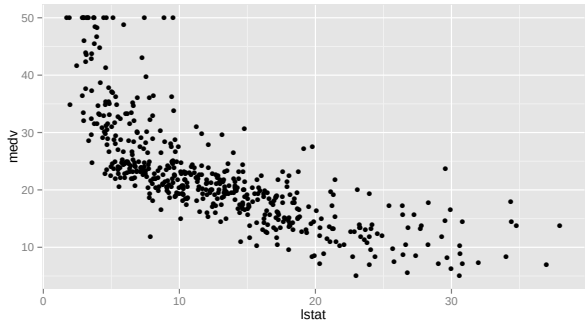
This lineplot uses the US economic time series dataset (*https://ggplot2.tidyverse.org/reference/economics.html*) and plots the evolution of the number of unemployed.

Multiple Lines on a Single Plot



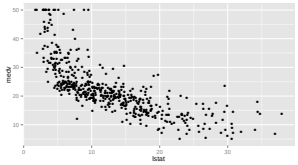
Scatterplots

Display the relationship between two numerical variables.



This scatterplot uses the Boston dataset and plots the relationship between the variables 'lstat' (percentage of lower status of the population) and 'medv' (the median house value). It shows a clear relationship between these two variables: as the percentage of lower status of the population increases, the median house value decreases. The plot shows a strong correlation between these two variables, but the correlation is not linear.

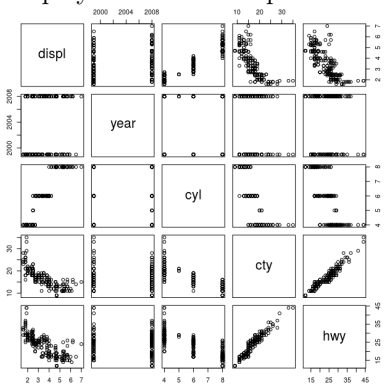
What can we learn from this plot?



- There is a strong correlation between 'lstat' and 'medv'.
- The correlation is negative: as 'lstat' increases, 'medv' decreases.
- The correlation is not linear.
- If we wish to predict the value of 'medv', then 'lstat' looks a good predictor.

Matrix Plots

Display the relationship between pairs of variables.



What can we deduce from this plot?

- 'cty' and 'hwy' are strongly correlated.
- 'displ' and 'cty' are strongly correlated.
- 'displ' and 'hwy' are strongly correlated.

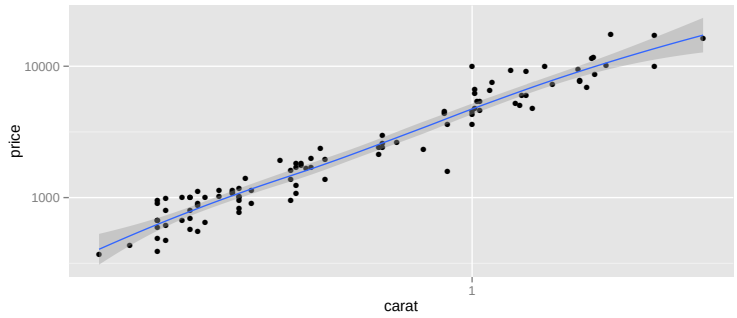
This matrix plot uses the Auto-MPG data set (*https://ggplot2.tidyverse.org/reference/mpg.html*) and shows strong correlations in these variables:

displ: displacement.

cty: city miles per gallon.

hwy: highway miles per gallon.

Adding a Smoother to a Plot



This smoother also shows the 95% confidence intervals.

Keyword Extraction and Word Clouds

- *Keyword extraction*: Extract the most important words in a document or collection of documents.
- *Word cloud*: a graphical interface that displays words according to their importance.

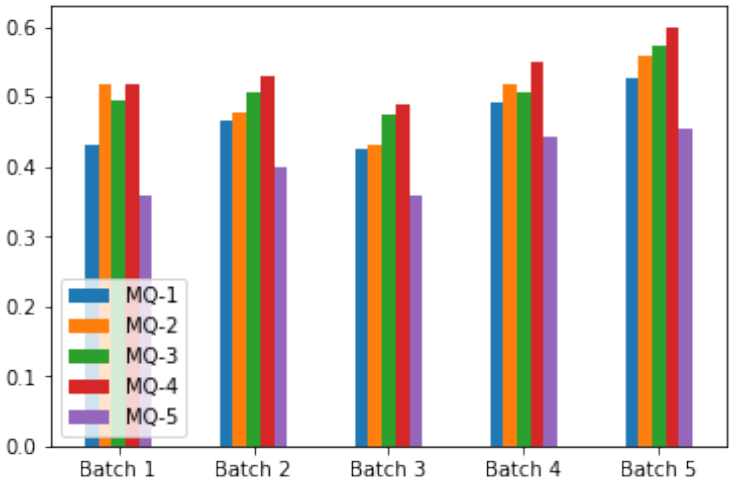
How to Select and Score words?

- Remove stop words.
- Select words by frequency.
- Use tf.idf
- ...

2.3 Visualisation for Evaluation

Bar Chart

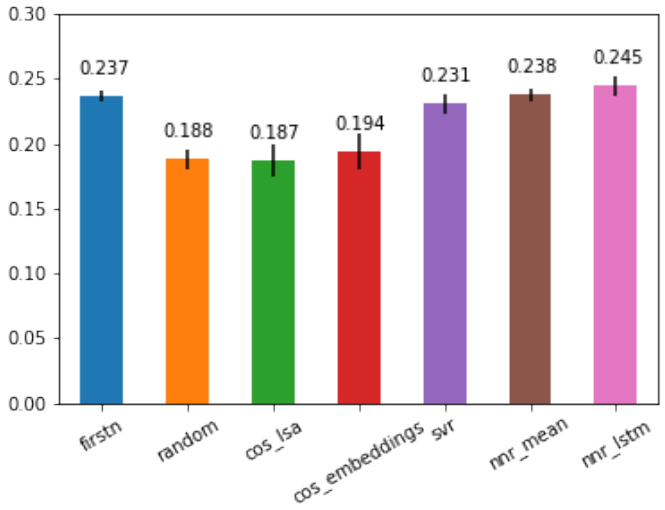
A bar chart can compare the results of several experiments.



This bar chart shows the results of 5 different approaches (MQ-1 to MQ-5) on 5 different sets of data (Batch 1 to Batch 5). The results show that all algorithms perform comparatively similarly in all batches. This gives some confidence that these results can be trusted. For example, we observe that MQ-5 performs worst, and MQ-4 performs best.

Bar Chart with Error Bars

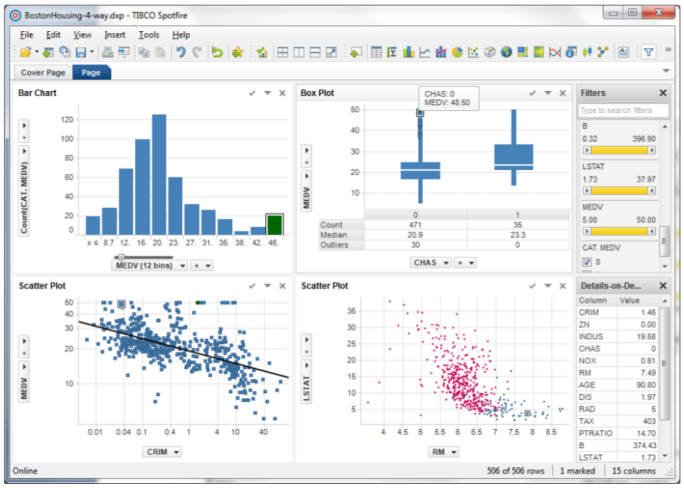
Error bars usually indicate the 95% confidence intervals.



This bar chart with error bars compares the performance of seven different algorithms. Each algorithm was run 10 times on different data, and the error bars indicate the standard deviation on the 10 runs.

Linked Plots

- Same record is highlighted in each plot.
- Useful for error analysis and for data exploration.



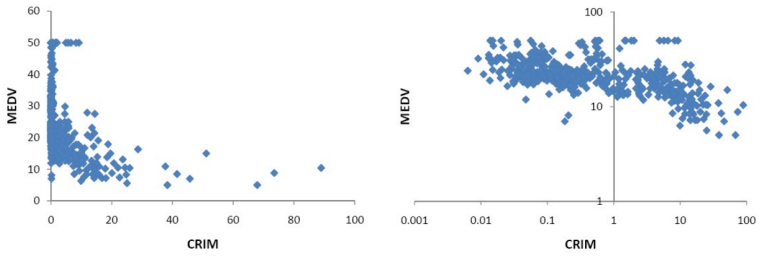
Example: SAS Visual Analytics

https://www.sas.com/en_au/software/visual-analytics/demo/banking-and-risk-insights/sample-report.html

3 Enhancing Visualisation

Rescaling to Log Scale

- Usually distributions that are symmetrical are more useful for analytics.
- Re-scaling is a possible approach to transform a distribution into a more symmetrical distribution.

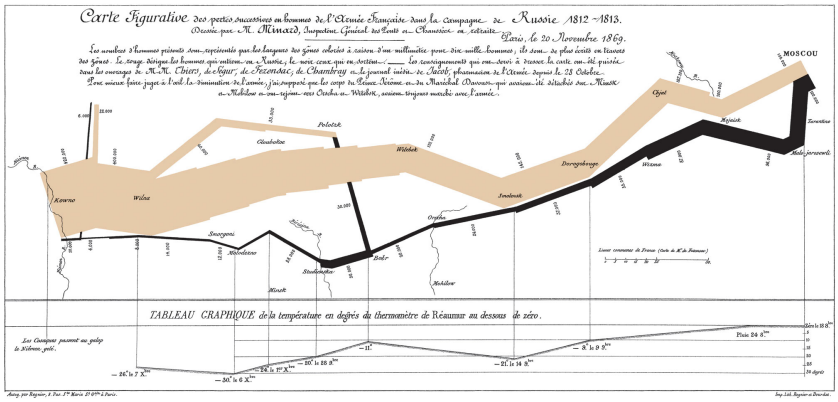


Before After

3.1 Visualising Large Dimension Spaces

Visualising Large Dimension Spaces — Motivational Example

Charles Minard's Famous Plot

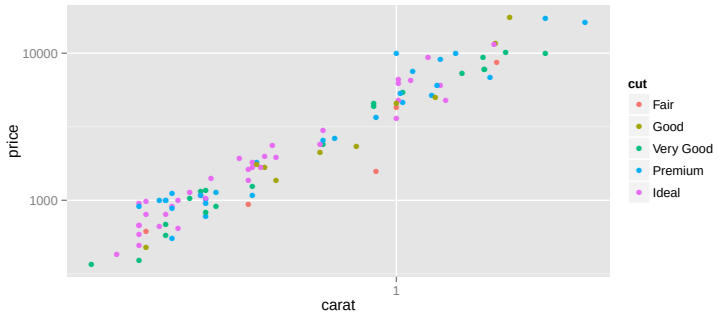


https://en.wikipedia.org/wiki/Charles_Joseph_Minard

Charles Minard's map of Napoleon's disastrous Russian campaign of 1812. The graphic is notable for its representation in two dimensions of six types of data: the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates.

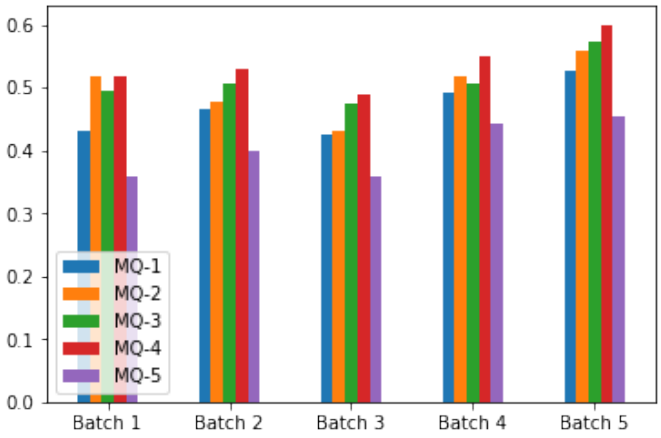
Introducing Colours

- Plots can show two dimensions only.
- A third dimension can be visualised by adding colours.



Grouped Bar Chart

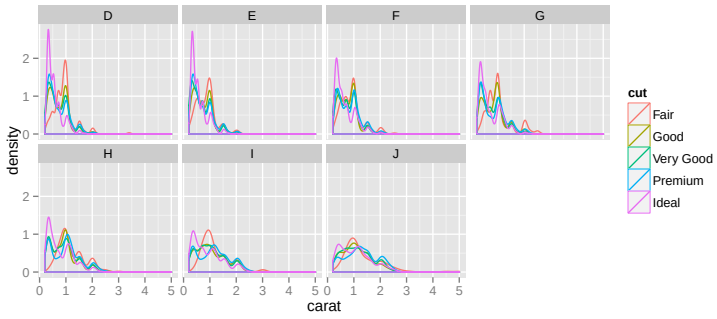
A grouped bar chart can represent the values of two independent variables.



This bar chart shows the evaluation results of five systems (MQ-1 to MQ-5) on 5 different data sets (batch 1 to batch 5).

Faceting

- Faceting is an alternative to colours.
- The data are divided into subsets or *facets*.



This plot shows density plots of each type of cut in each of the diamond colours.

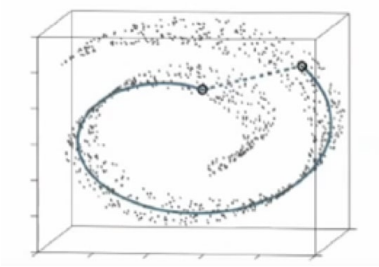
Dimension Reduction

- Sometimes you need to plot a scatterplot with more than two dimensions.
- Dimension reduction techniques will map n dimensions to $m < n$ dimensions. For visualisation, $m = 2$.
- Principal Components Analysis (PCA) projects m dimensions to n so that the resulting projection has the highest dispersion possible.
  - Think of what would be the best angle to photograph an object.
  - Latent Semantic Analysis (LSA) and Singular Value Decomposition (SVD) are based on the same principles.
- Other approaches such as t-SNE perform a non-linear mapping.
  - t-SNE maps n dimensions to $m < n$ dimensions so that small distances between points is preserved as much as possible.

t-SNE vs. PCA
The "Swiss roll" example

Mapping to 1 dimension

- PCA would map to the x coordinate.
- t-SNE would map following the solid line.



https://www.kdnuggets.com/2018/08/introduction-t-sne-python.html

3.2 Visualising Many Samples

Data Aggregation

We may change the level of aggregation:

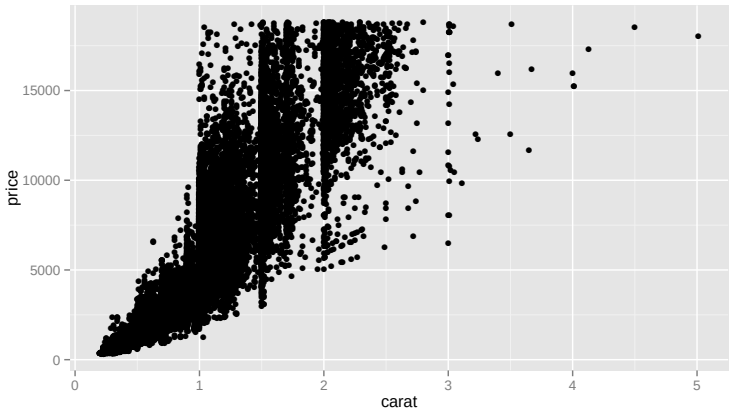
- Change the time scale (e.g. weekly, monthly, yearly).
- Group by region.
- Bin the values.

Scaling Up to Large Datasets

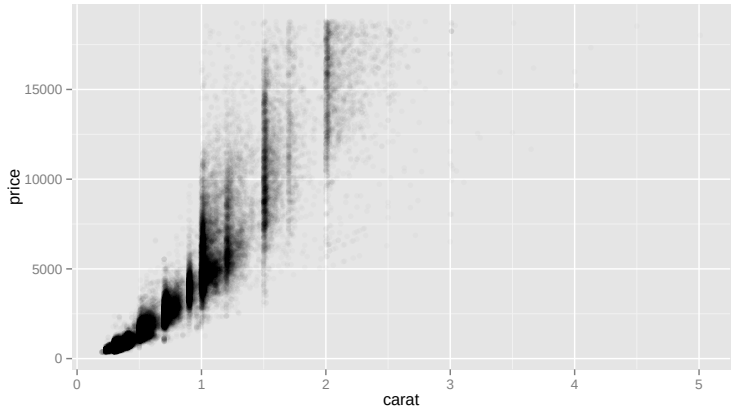
Aside from applying aggregation we may try:

- Sampling.
- Reducing the marker size (the circles in a scatterplot).
- Using more transparent marker colours.
- Using density plots.

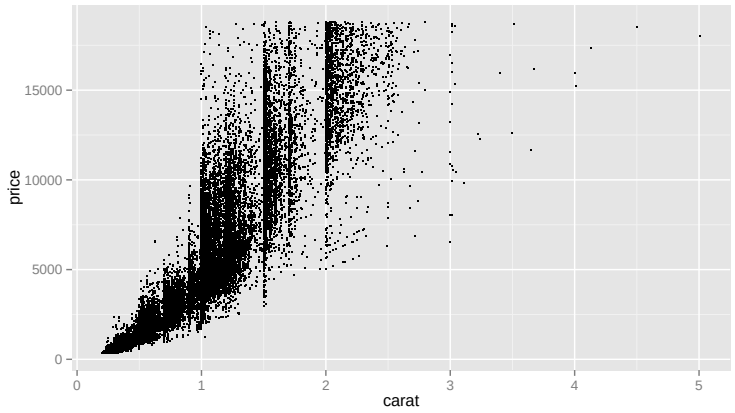
Original Plot



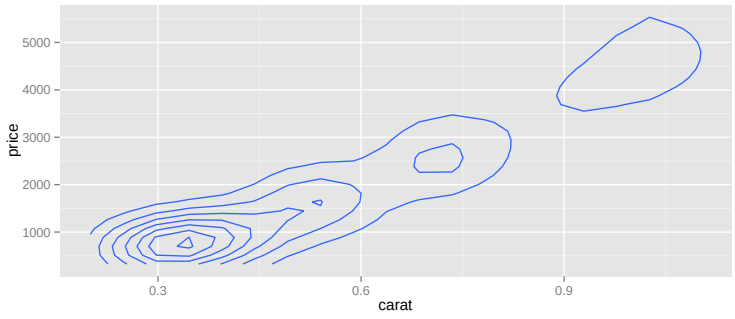
Using Transparency

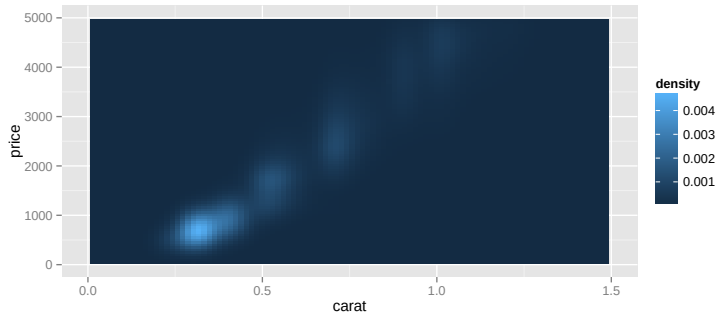


Using Smaller Marker Size

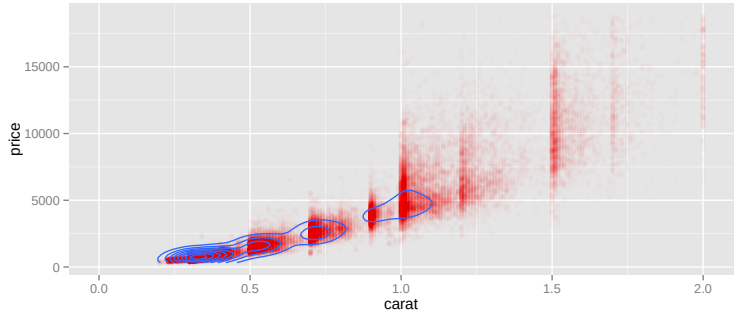


Using Density Plots





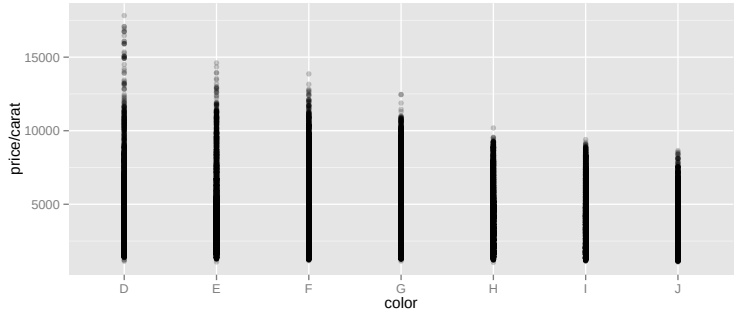
Overlaying Several Approaches



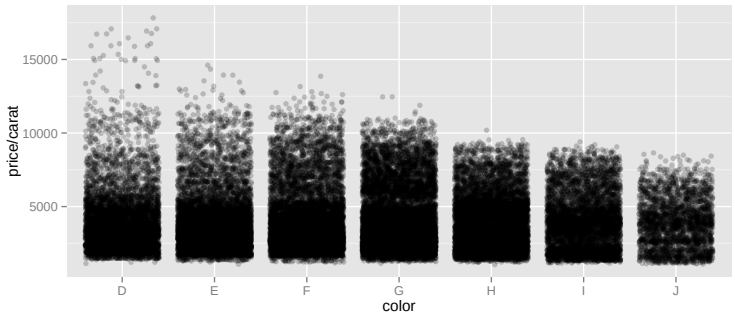
Jittering for Categorical Variables

- Categorical variables will have instances of the same value.
- When we plot these values, all will be in the same place.
- By introducing jittering, we add random noise to the position of the value in the plot.

Without Jittering



With Jittering



3.3 Common Mistakes with Data Visualization

Using the Wrong Plot

- *Most common mistake*: use a line graph when you should use a bar chart.
  - If plotting values that change in sequence, use a line graph.
  - If plotting values of multiple variables, use a bar chart.
- *Second most common mistake*: use pie charts. Usually, bar charts will give more precise information.

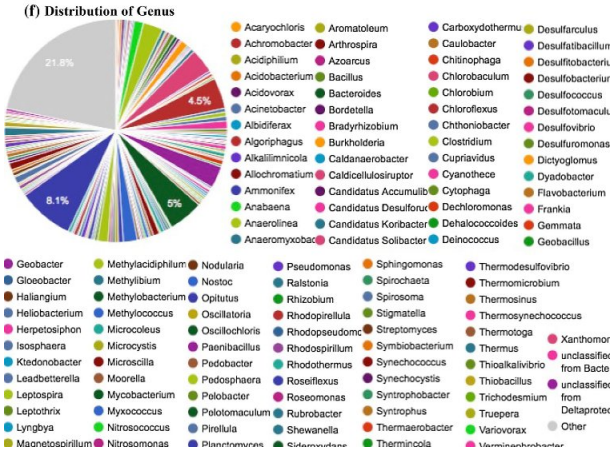
Always think

1. What information do I want to convey?
2. What is the best way to convey the information?

Multiple plots with different scales

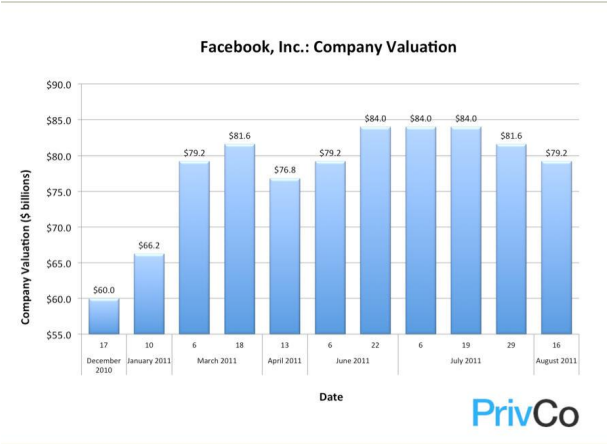
- When using multiple plots, make sure that the plots are comparable.
- If two plots use different scales, comparison is more difficult.
- The plots would be worthless or misleading.

Using too much information



https://www.kdnuggets.com/2017/10/5-common-mistakes-bad-data-visualization.html

What's Wrong with This Graph?



https://www.forbes.com/sites/naomirobbins/2011/11/17/whats-wrong-with-this-graph/

Take-home Messages

- Uses of data visualisation.
- Visual analytics.
- Types of data visualisation.
- What data visualisation for what application?
- Visualising large volumes of data.
- Common mistakes with data visualisation.

What's Next

Week 11

- Stream Processing.