

ITEC874 — Big Data Technologies

Week 9 Lecture 1: Visualising Big Data

Diego Mollá

Department of Computer Science
Macquarie University

COMP348 2018H1

Programme

- 1 Data Visualisation
- 2 Common Building Blocks for Data Visualisation
- 3 Enhancing Visualisation

Reading

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

Tentative Outline of ITEC874 — Weeks 7 to 12

- ⑦ Analysing Big Data
- ⑧ Analysing Unstructured Data
- ⑨ Visualising Big Data
- ⑩ Analysing Streaming Data
- ⑪ Big Data and Society
- ⑫ Industry Talk: Amazon

Programme

1 Data Visualisation

- Why Visualising Data?

2 Common Building Blocks for Data Visualisation

3 Enhancing Visualisation

Programme

1 Data Visualisation

- Why Visualising Data?

2 Common Building Blocks for Data Visualisation

3 Enhancing Visualisation

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.

One Look Is Worth A Thousand Words--

One look at our line of Republic, Firestone, Miller and United States tires can tell you more than a hundred personal letters or advertisements.

WE WILL PROVE THEIR VALUE BEFORE YOU INVEST ONE DOLLAR IN THEM.

Ever consider buying Supplies from a catalog?

What's the use! Call and see what you are buying. One look at our display of automobile and motorcycle accessories will convince you of the fact.

THAT WE HAVE EVERYTHING FOR THE AUTO

Piqua Auto Supply House
133 N. Main St.—Piqua, O.

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

<https://www.zencos.com/blog/5-applications-of-image-analytics-with-SAS-Viya/>

- 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.

Programme

1 Data Visualisation

2 Common Building Blocks for Data Visualisation

- Visualisation for Summarising Data
- Visualisation for Assisting Prediction
- Visualisation for Evaluation

3 Enhancing Visualisation

Programme

1 Data Visualisation

2 Common Building Blocks for Data Visualisation

- Visualisation for Summarising Data
- Visualisation for Assisting Prediction
- Visualisation for Evaluation

3 Enhancing Visualisation

Common Summary Statistics

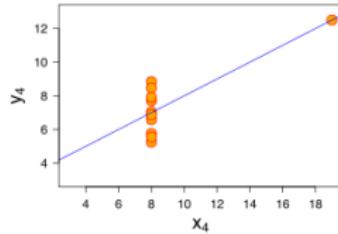
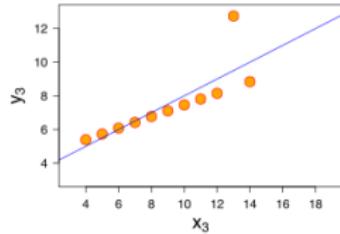
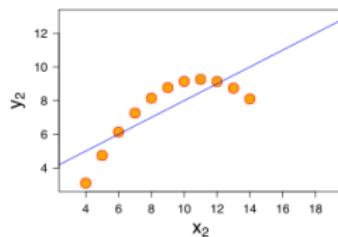
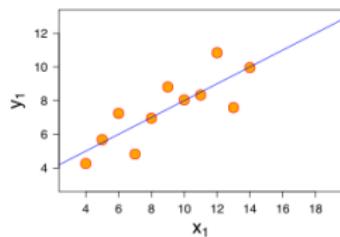
Some of the most common metrics are:

- 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

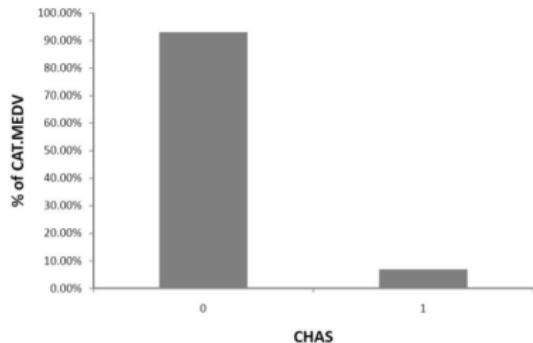
Variance of y : 4.125

Correlation between x and y : 0.816

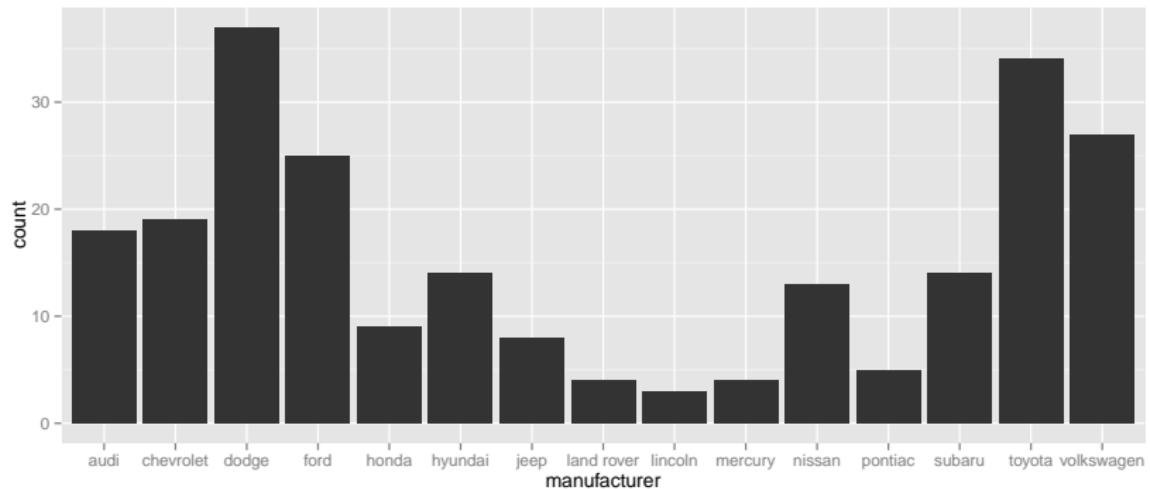
([https://en.wikipedia.org/wiki/Anscombe's_quartet](https://en.wikipedia.org/wiki/Anscombe%27s_quartet))

Bar Chart

- Can be used for numerical or categorical variables.
- Each bar represents the counts of the value of a variable.
- In this example, 95% of tracts do not border Charles River.

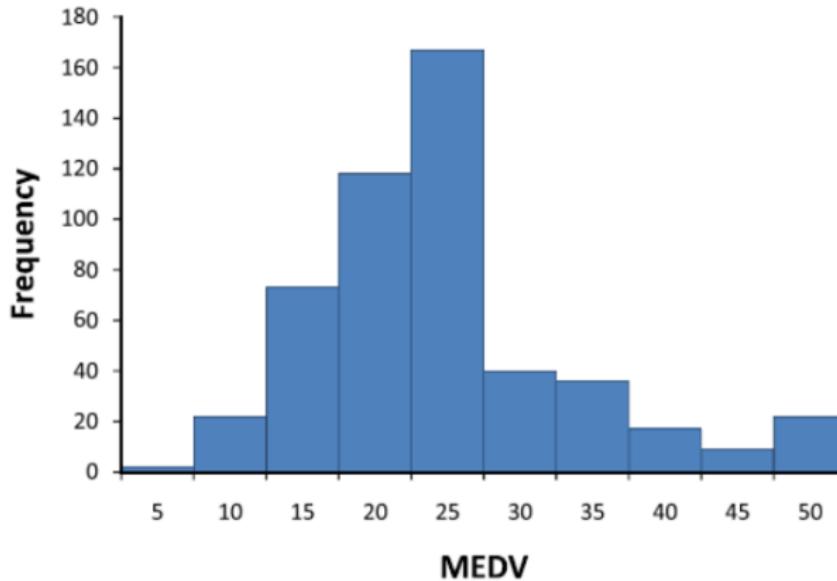


Bar Charts of Categorical Data



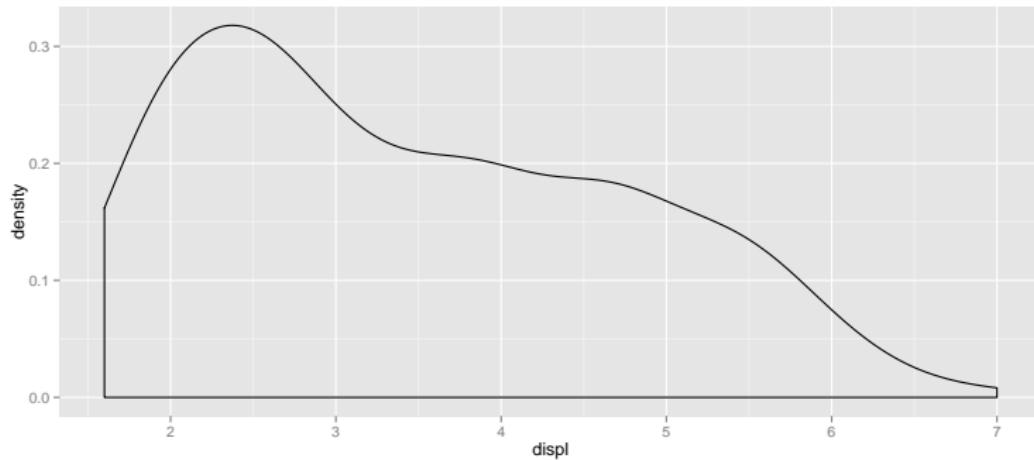
Histogram

A bar chart that shows the distribution of the values of a variable.



Density Plots of Continuous Data

- Density plots approximate histograms using a continuous line.
- Useful when the variable is numerical.

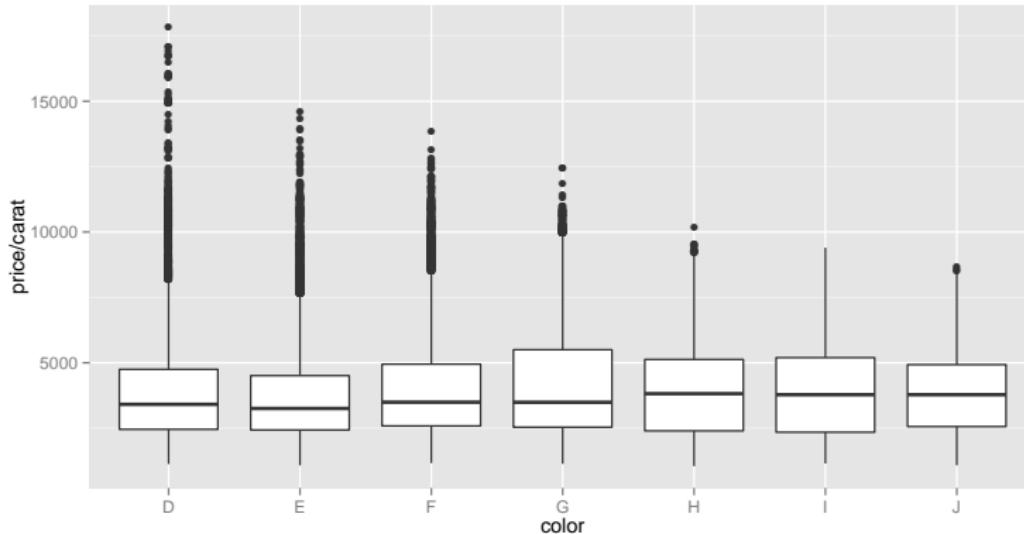


Boxplot

- 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.
 - $Q3 + 1.5 \times (Q3 - Q1)$
 - $Q1 - 1.5 \times (Q3 - Q1)$
- Anything outside the whiskers is usually plotted as circles.

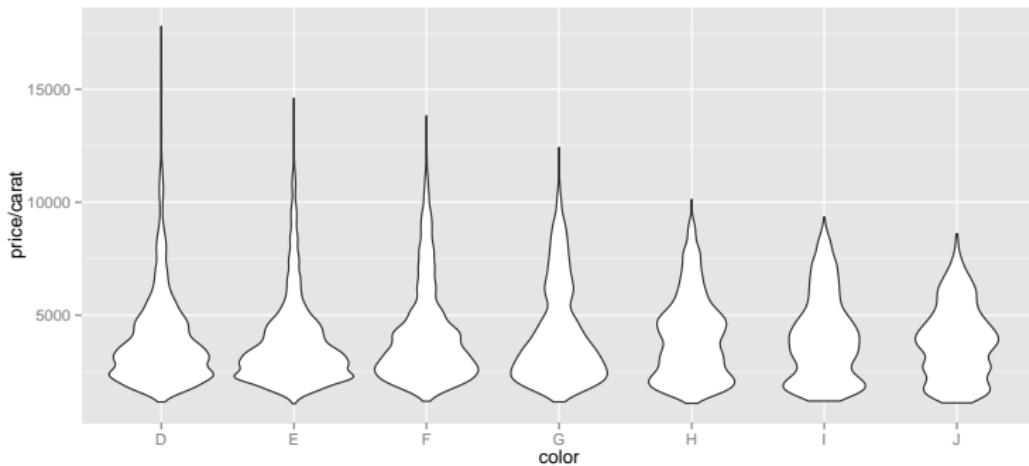
Side-by-side Boxplots

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



Violin Plots

Violin plots incorporate a density plot of each subgroup.



Programme

1 Data Visualisation

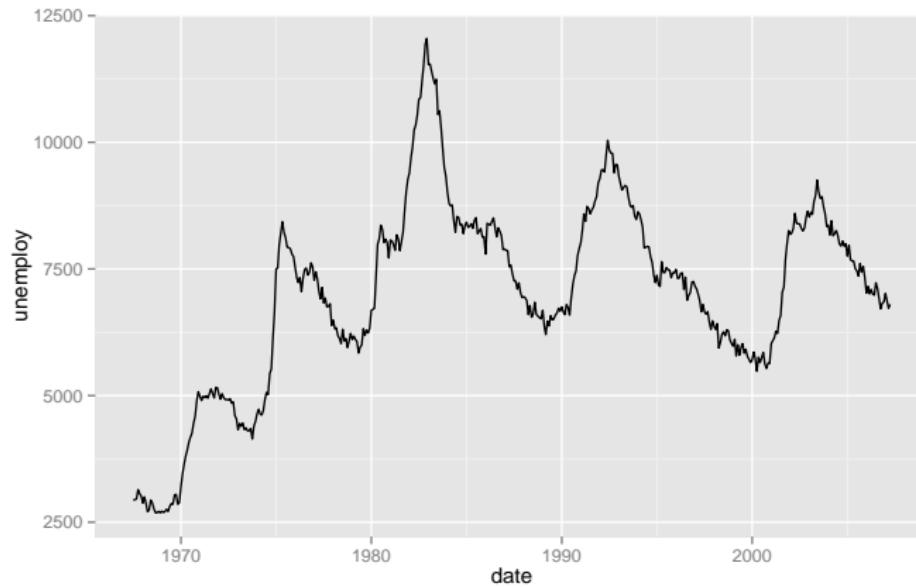
2 Common Building Blocks for Data Visualisation

- Visualisation for Summarising Data
- Visualisation for Assisting Prediction
- Visualisation for Evaluation

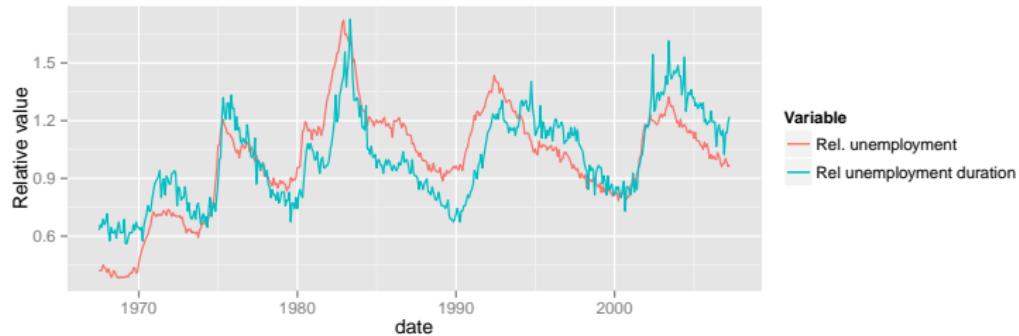
3 Enhancing Visualisation

Line Graph for Time Series

Useful to display the values of a variable in sequence.

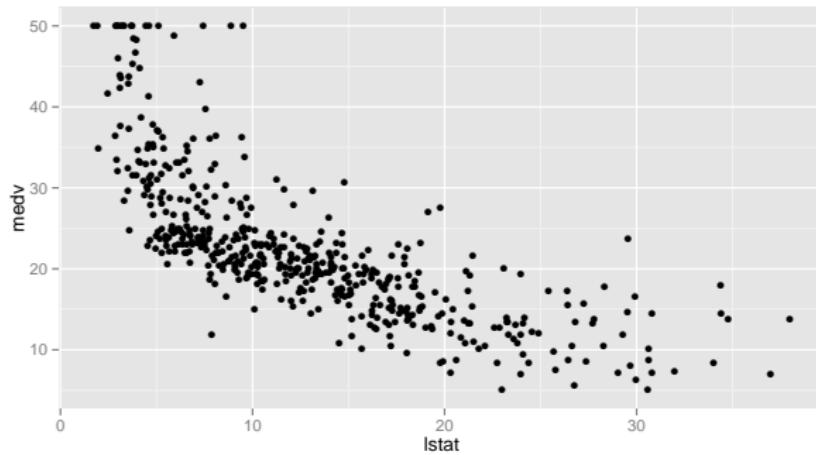


Multiple Lines on a Single Plot

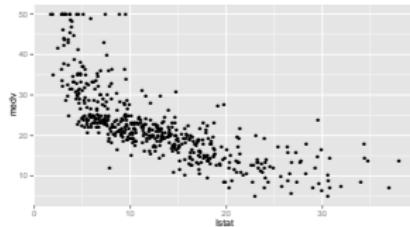


Scatterplots

Display relationship between two numerical variables.



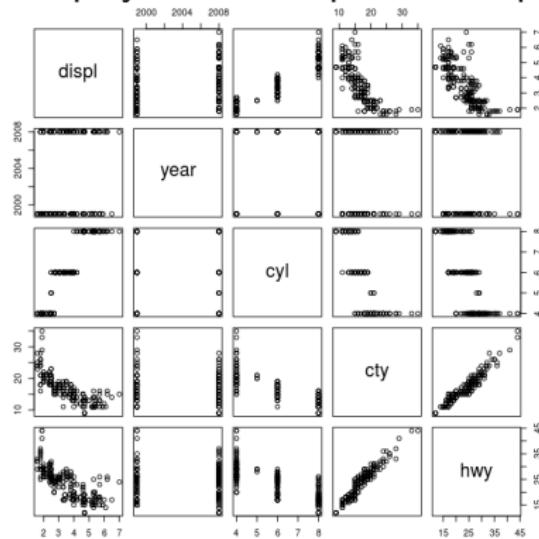
What can we learn from this plot?



- 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

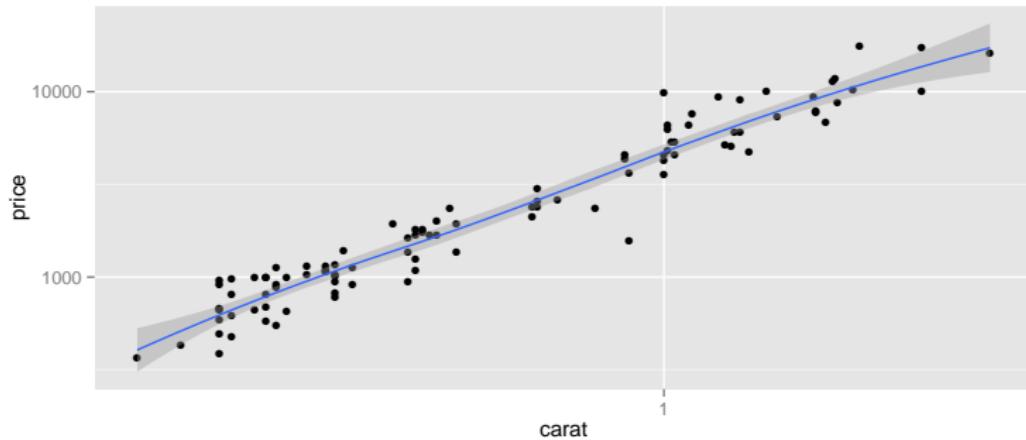
Display relationship between pairs of variables.



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.

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
- ...



Programme

1 Data Visualisation

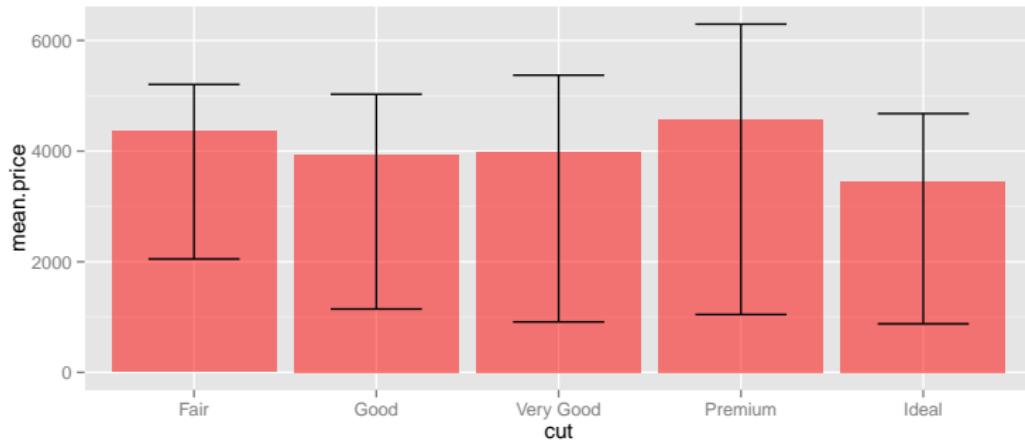
2 Common Building Blocks for Data Visualisation

- Visualisation for Summarising Data
- Visualisation for Assisting Prediction
- Visualisation for Evaluation

3 Enhancing Visualisation

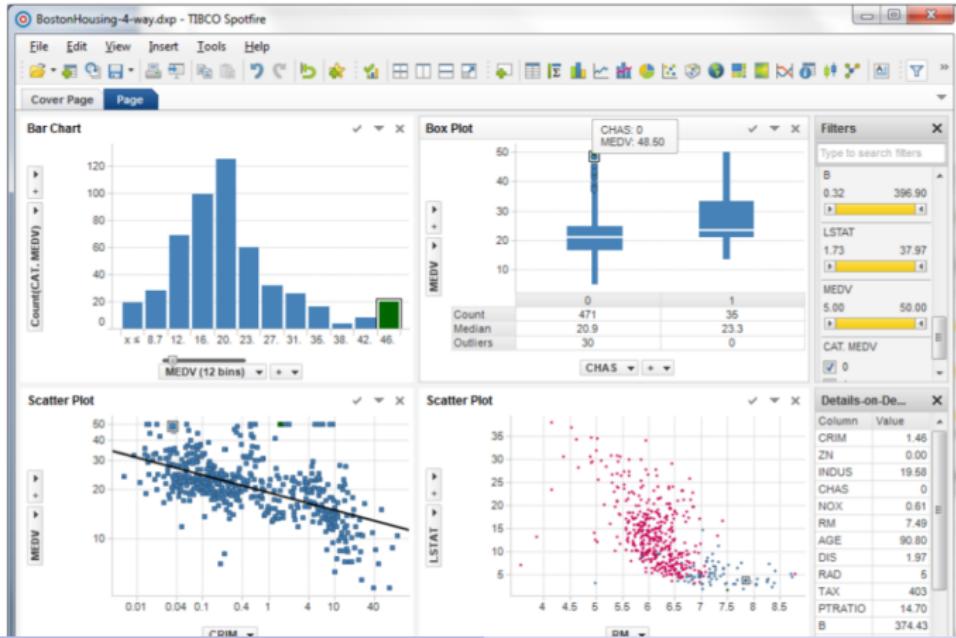
Bar Plot with Error Bars

Error bars usually indicate the 95% confidence intervals.

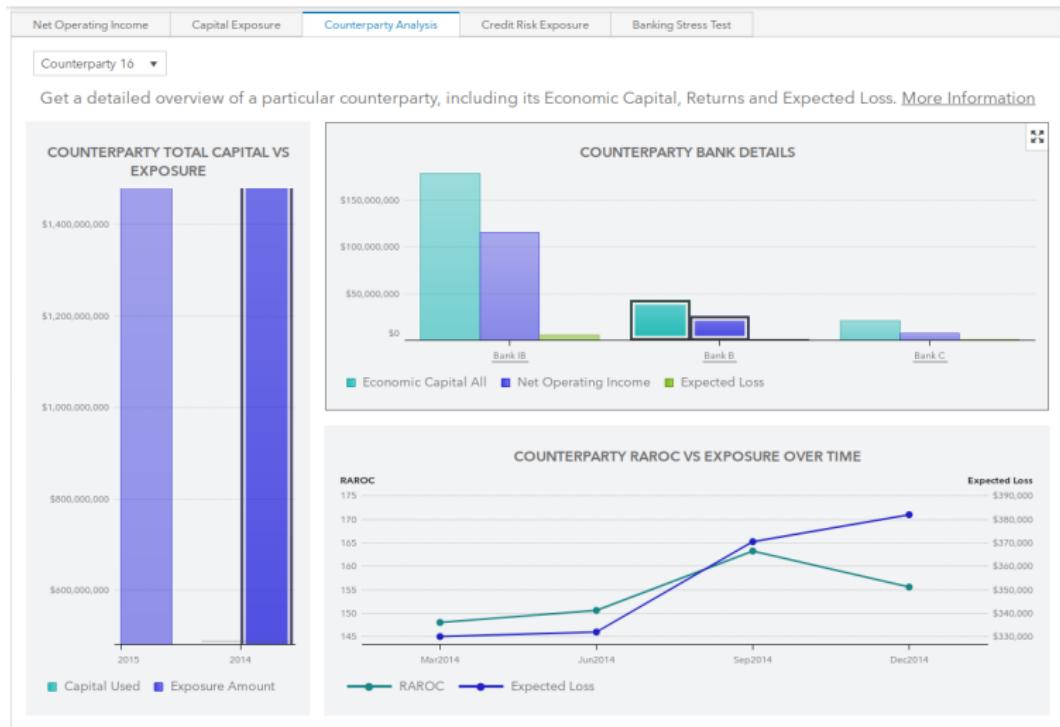


Linked Plots

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



Example: SAS Visual Analytics



Video: What-If Tool

What If...

you could inspect a machine learning model,
with no coding required?

Building effective machine learning systems means asking a lot of questions. It's not enough to train a model and walk away. Instead, good practitioners act as detectives, probing to understand their model better.

But answering these kinds of questions isn't easy. Probing "what if" scenarios often means writing custom, one-off code to analyze a specific model. Not only is this process inefficient, it makes it hard for non-programmers to participate in the process of shaping and improving machine learning models. For us, making it easier for a broad set of people to examine, evaluate, and debug machine learning systems is a key concern.

That's why we built the What-If Tool. Built into the open-source TensorBoard web application - a standard part of the TensorFlow platform - the tool allows users to analyze an machine learning model



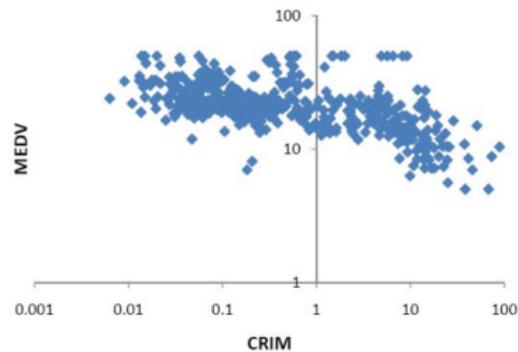
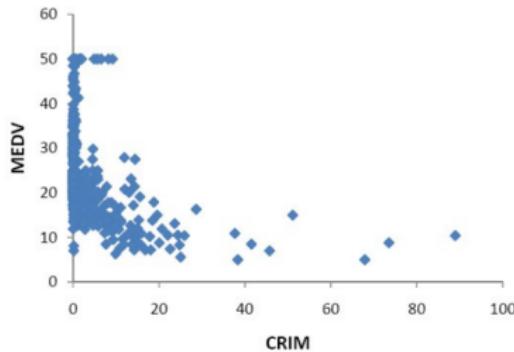
<https://pair-code.github.io/what-if-tool/>

Programme

- 1 Data Visualisation
- 2 Common Building Blocks for Data Visualisation
- 3 Enhancing Visualisation
 - Visualising Large Dimension Spaces
 - Visualising Many Samples
 - Common Mistakes with Data Visualization

Rescaling to Log Scale

Allows to expand on the smaller numbers and compact on the large numbers.

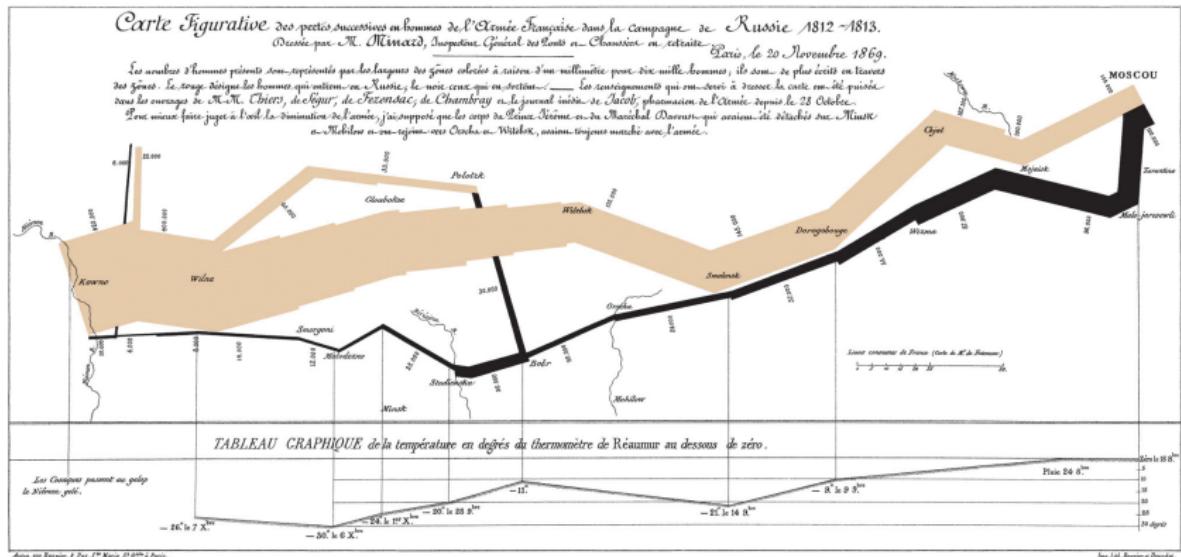


Programme

- 1 Data Visualisation
- 2 Common Building Blocks for Data Visualisation
- 3 Enhancing Visualisation
 - Visualising Large Dimension Spaces
 - Visualising Many Samples
 - Common Mistakes with Data Visualization

Motivational Example

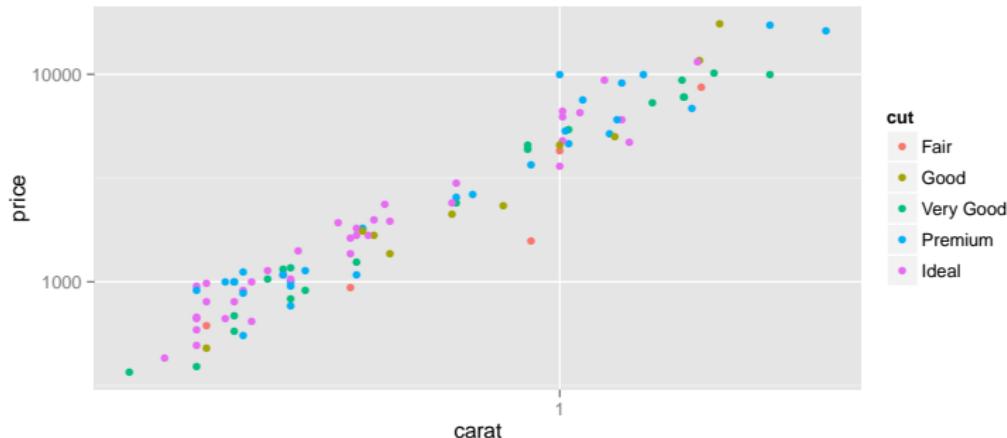
Charles Minard's Famous Plot



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

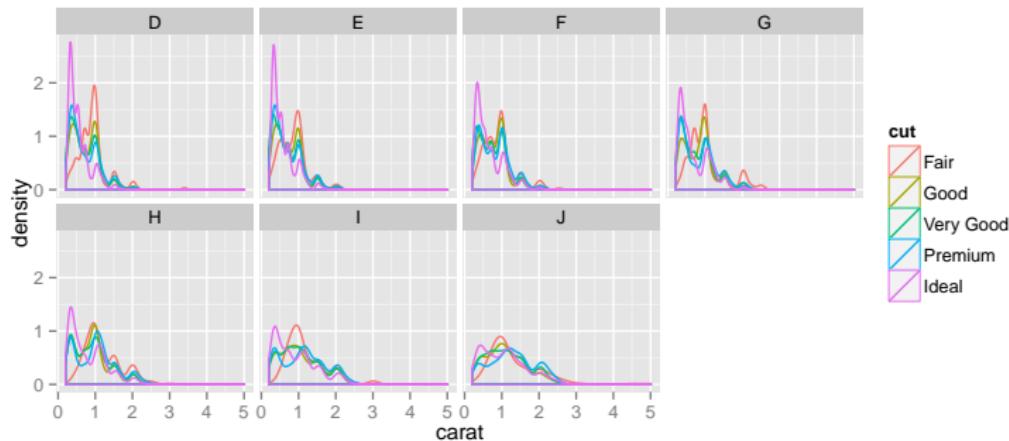
Introducing Colours

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



Faceting

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



Dimension Reduction

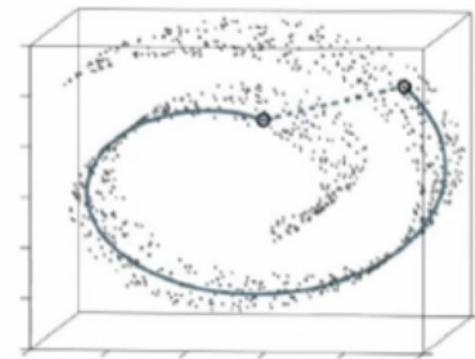
- Sometimes you need to plot a scatterplot with more than two dimensions.
- Dimension reduction techniques will map n dimensions to $m < n$ dimensions. For visualisation, $m = 2$.
- Principal Components Analysis (PCA) projects m dimensions to n so that the resulting projection has the highest dispersion 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>

Programme

- 1 Data Visualisation
- 2 Common Building Blocks for Data Visualisation
- 3 Enhancing Visualisation
 - Visualising Large Dimension Spaces
 - Visualising Many Samples
 - Common Mistakes with Data Visualization

Data Aggregation

We may change the level of aggregation:

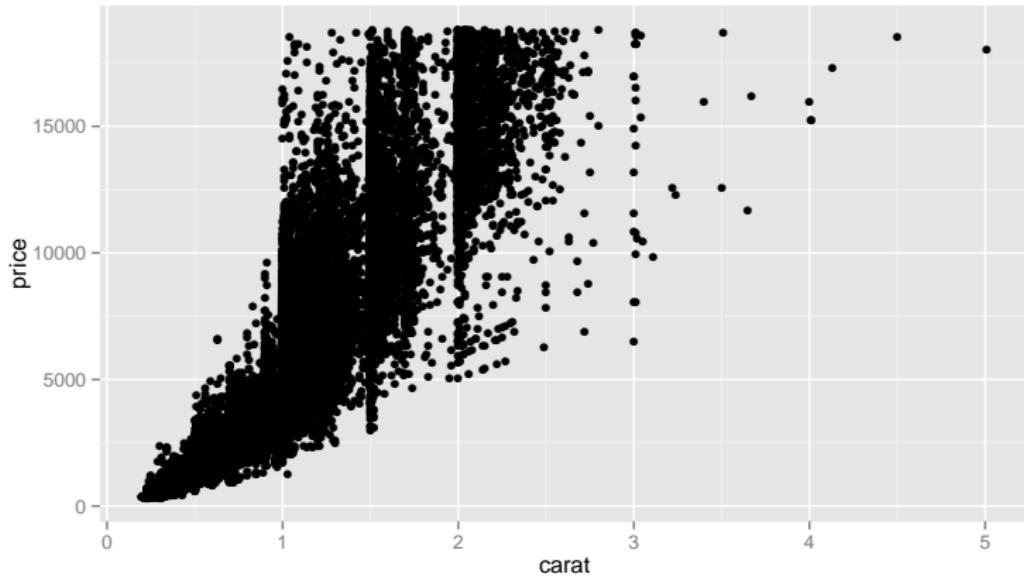
- Change the time scale (e.g. weekly, monthly, yearly).
- Group by region.
- Bin the values.

Scaling Up to Large Datasets

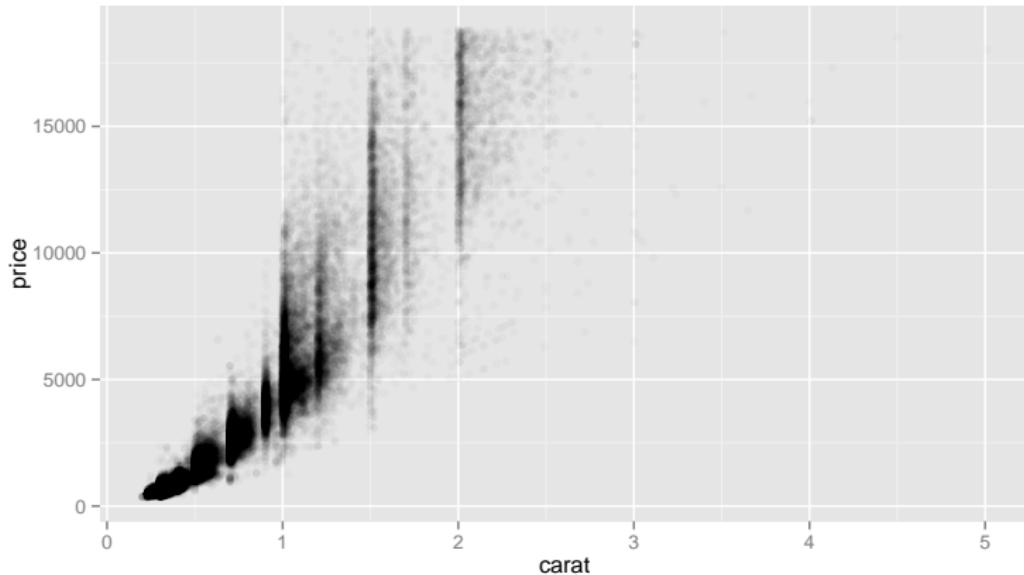
Aside from applying aggregation we may try:

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

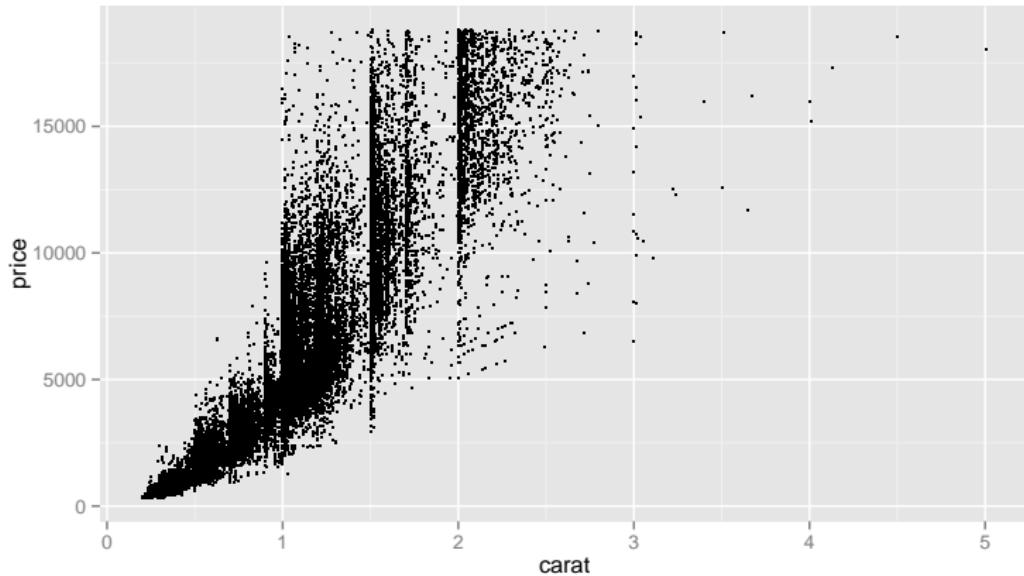
Original Plot



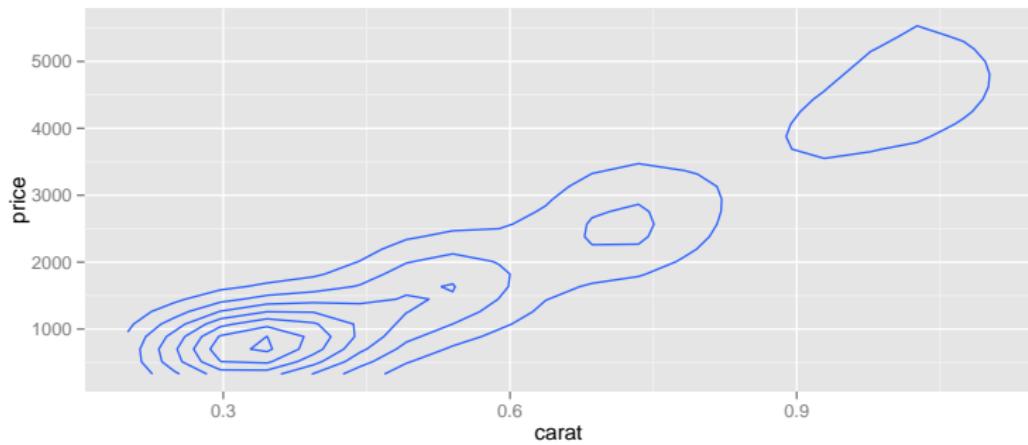
Using Transparency



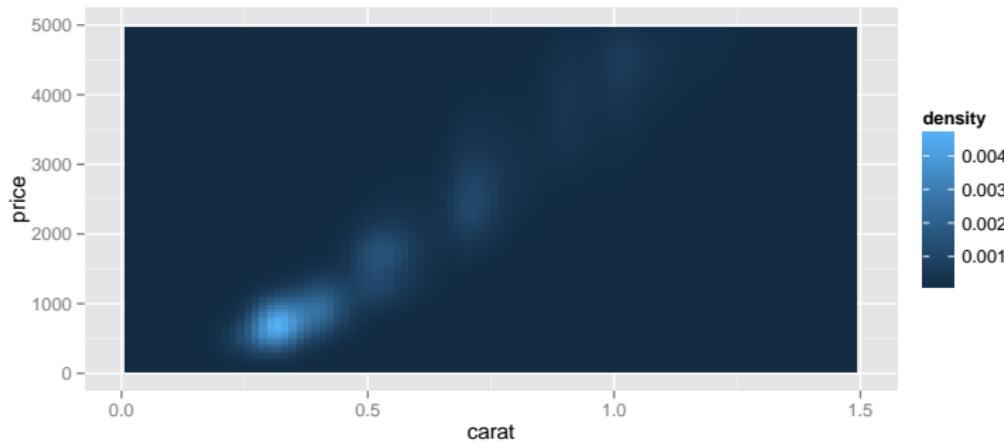
Using Smaller Marker Size



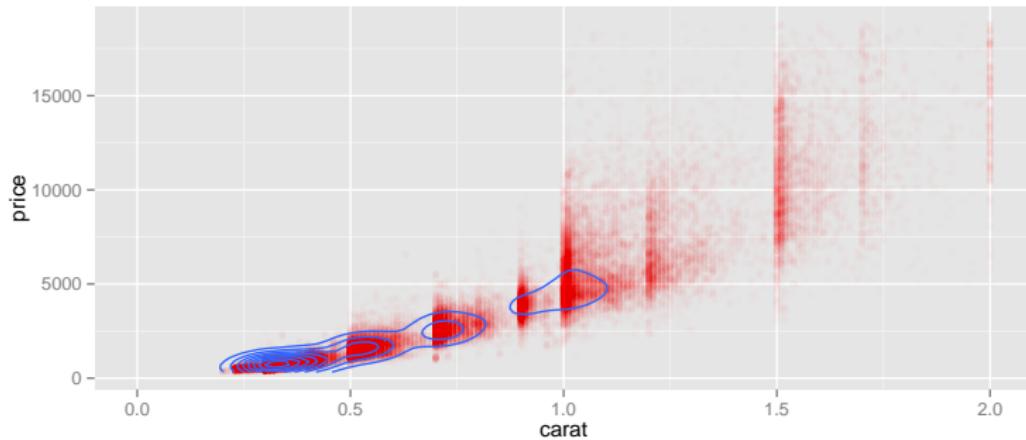
Using Density Plots I



Using Density Plots II



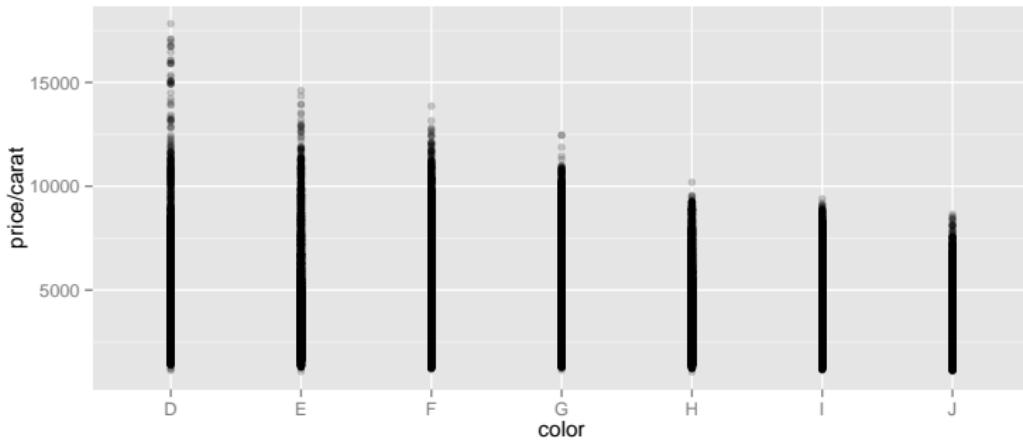
Overlaying Several Approaches



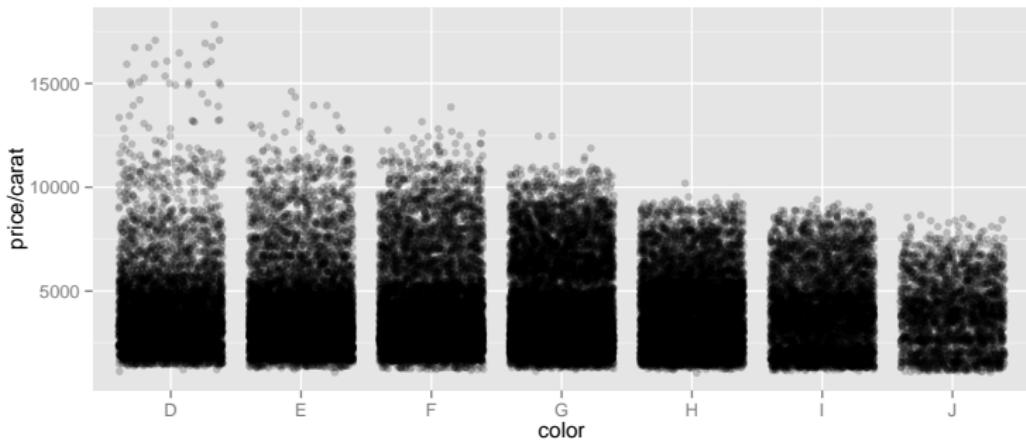
Jittering for Categorical Variables

- 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



Programme

- 1 Data Visualisation
- 2 Common Building Blocks for Data Visualisation
- 3 Enhancing Visualisation
 - Visualising Large Dimension Spaces
 - Visualising Many Samples
 - 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 a pie chart. Usually, a bar chart will be better.

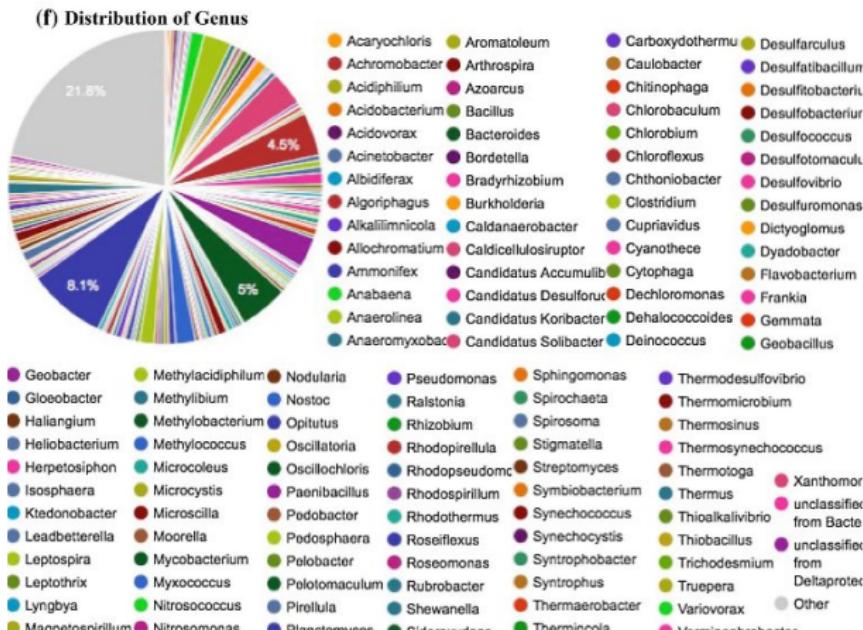
Always think

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

Multiple plots with different scales

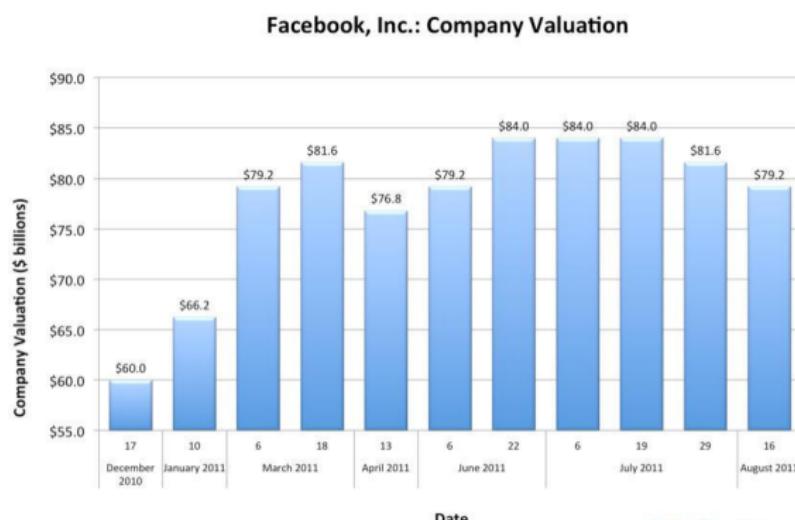
- When using multiple plots, make sure that the plots are comparable.
- If two plots use different scales, comparison is more difficult.
- The plots would be worthless or misleading.

Using too much information



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

What's Wrong with This Graph?



PrivCo

<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 10

- Stream Processing.