Notes on Creating a data-driven organization

Chapter 5:

- Types of analysis
    - There are 6 types of statistical analysis, ordered from simple to complex:
        - Descriptive
        - Exploratory
        - Inferential
        - Predictive
        - Causal
        - Mechanistic
- In week 2 we discussed analytics. Analytics and analysis have 1 thing in common: Predictive.
    - The above mentioned analysis types can drive multiple levels of analytics. Exploratory data analysis can be used for ad hoc reports(analytics level 2) but also for alerts(analytics level 4)
- Levels of analytics can be seen as activities using one or more types of statistical analysis with some business rules and constraints.
- Below is a picture that shows a mapping of the different types of analytics and analysis, and shows what types of analytics can be used together with what type of analysis:

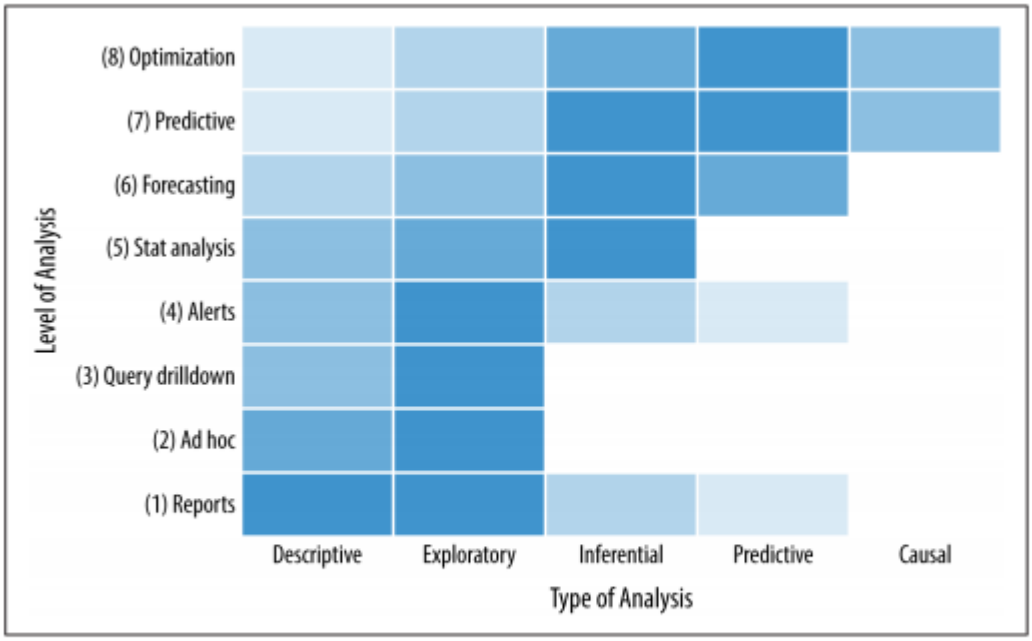

Figure 5-2. Crude mapping between level of analytics (left) and types of analysis (bottom). See detailed explanation in text.

Descriptive analysis:

- Descriptive analysis is the simplest type of analysis. It describes and summarizes a dataset quantitatively. It also characterizes the sample of data at hand and does not attempt to describe anything about the population from which it comes. It can often form the data that is displayed in dashboards.
- The simplest but one of the most important measures is Sample size; the number of data points or records in the sample.
- Location metrics include:
    - **Mean; sum of values divided by number of values**
    - **Geometric mean; mean with multiplicative effects**
    - **Harmonic mean; reciprocals of the value.**
    - **Median; 50th percentile**
    - **Mode; most frequently occurring value**
- Dispersion or central tendency measures include:
    - **Minimum; smallest value in a sample**
    - **Q1;25th percentile**
    - **Q3;75th percentile**
    - **Maximum; largest value in a sample**
    - **Interquartile range; Q3-Q1, or the central 50% of data**
    - **Range; Maximum – minimum**
    - **Stand deviation; measure of the dispersion from the arithmetic mean of a sample**
    - **Variance; measure of dispersion and is the average squared difference from the arithmetic mean and is the square of the standard deviation**
    - **Standard error; standard deviation divided by the square root of the sample size**
    - **Gini coefficient; mean of dispersion originally developed to quantify the degree of inequality in incomes in a population, but which can be used more broadly**
- Shape measures:
    - **Skew; a measure that captures the asymmetry of a distribution**
    - **Kurtosis; measure of the sharpness of the peak of a distribution**
- Descriptive analysis might involve tables of relative frequencies for different categories or contingency tables.
- **The goal of descriptive analysis is to describe the key features of the sample numerically. It should shed light on the key numbers that summarize distributions within the data.**

Exploratory Analysis:

- Descriptive analysis will only get you so far. A problem would be that you are condensing a large number of values down to a few summary numbers. Unsurprisingly, different samples with different distributions, shapes, and properties can result in the same summary statistics.

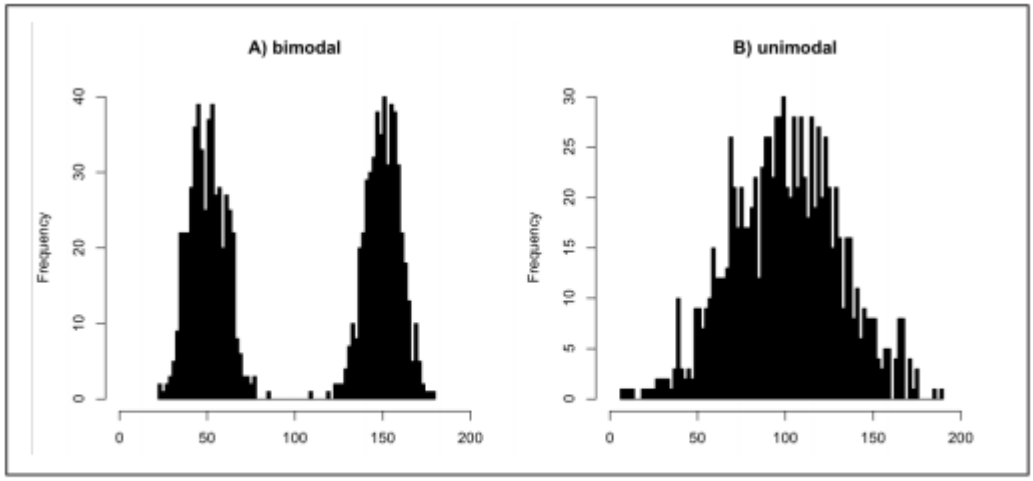

Figure 5-3. A) a bimodal distribution and b) a unimodal distribution, both of which have the same mean value of ~100.

- Another example is Anscombe's quartet:

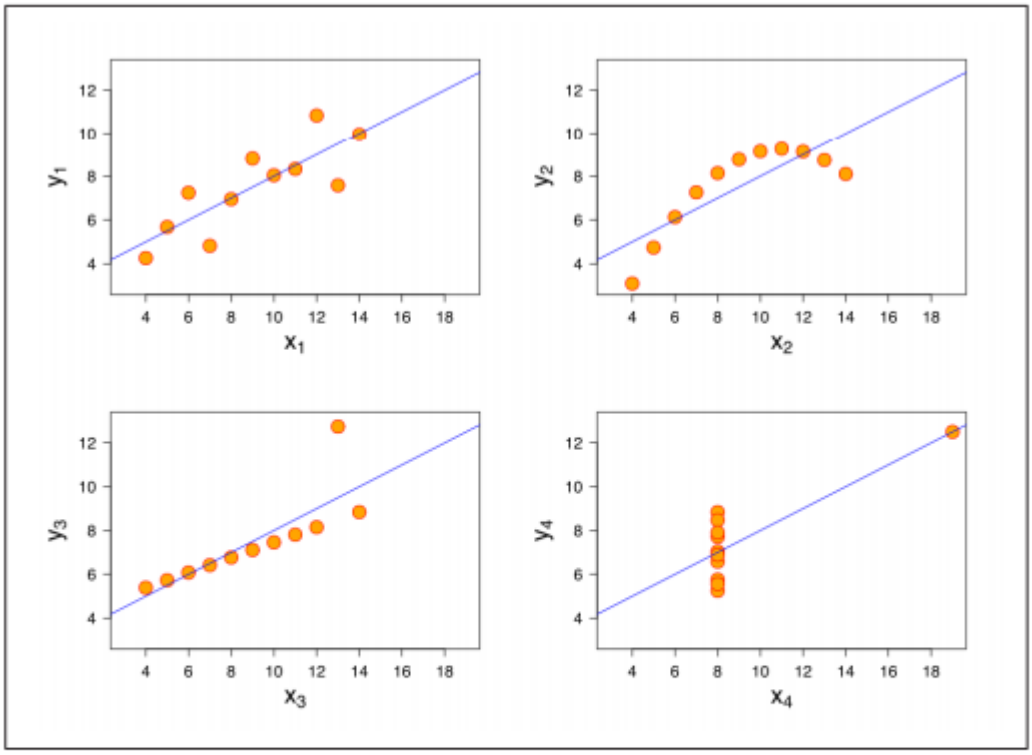

Figure 5-4. Anscombe's quartet. Each of these four samples has the same mean x, mean y, variance of x, variance of y, correlation, and regression line (at least to two decimal places). Figure from http://bit.ly/anscombes-quartet.

- The following graphs are examples of univariate data, where the data is continuous:

Stem	Leaf
1	0 3 6
2	1 6 7 8
3	5 5 6
4	1 1 5 6 9
5	0 3 6 8

Figure 5-5. A stem-and-leaf plot.

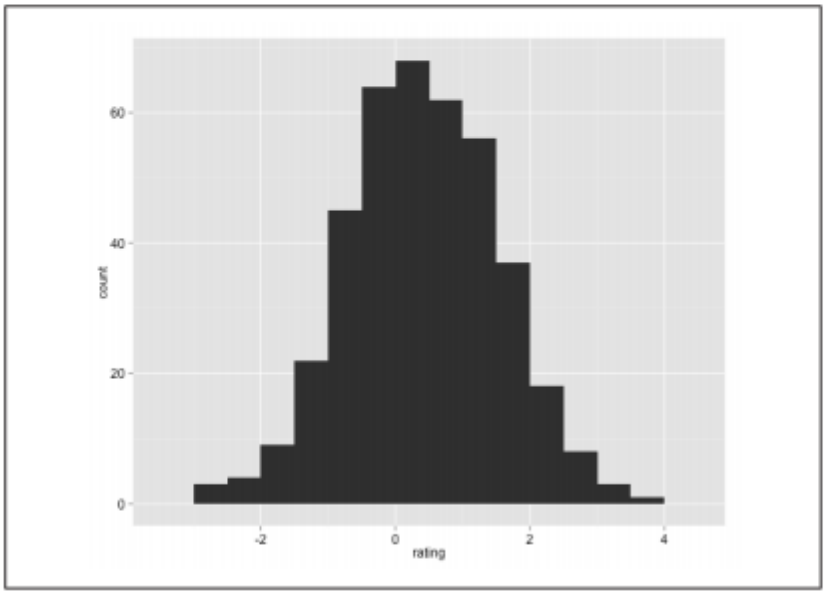

Figure 5-6. A histogram.

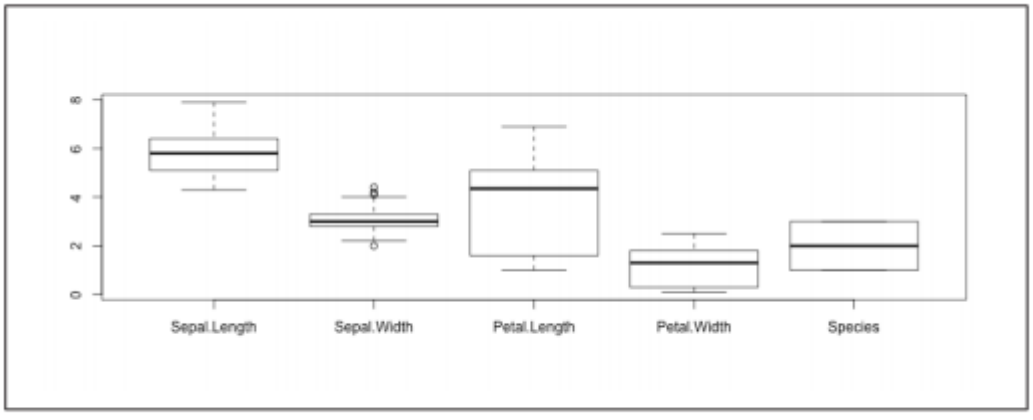

Figure 5-7. A box plot.

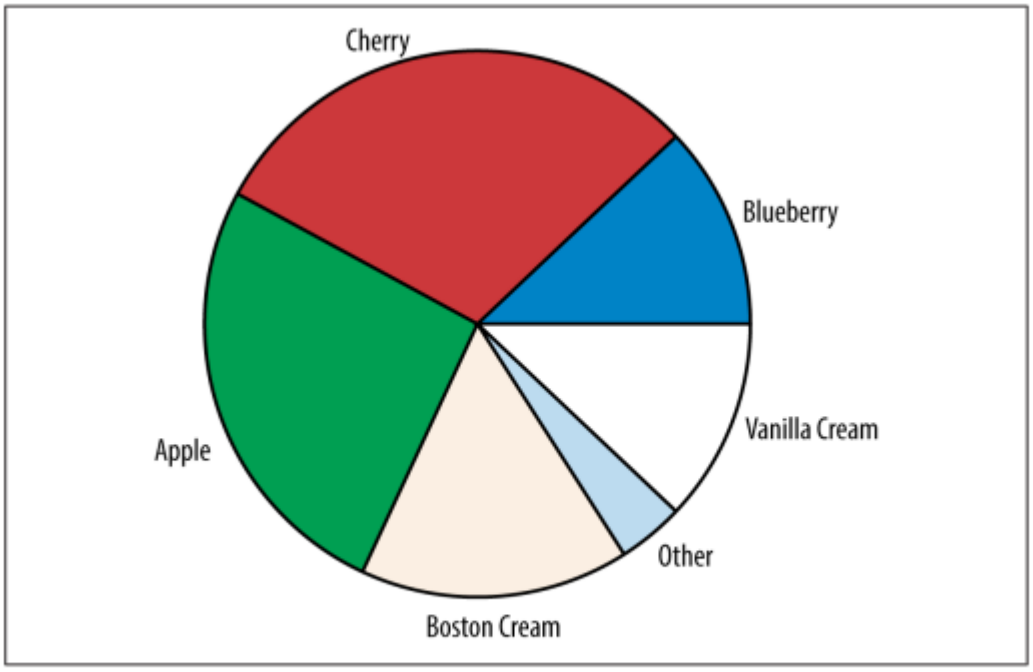

Figure 5-8. A pie or donut chart.[3]

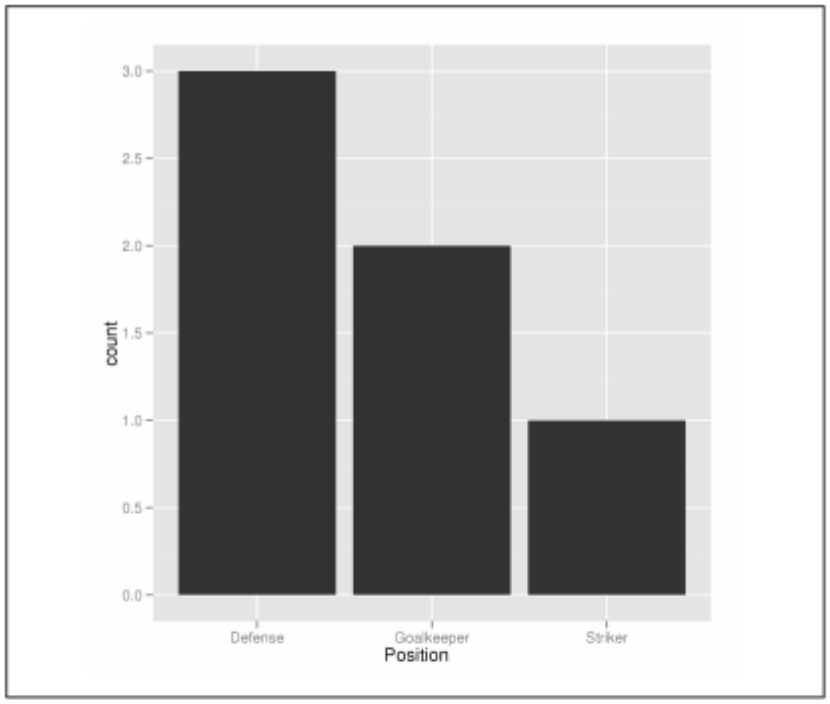

Figure 5-9. A bar chart.

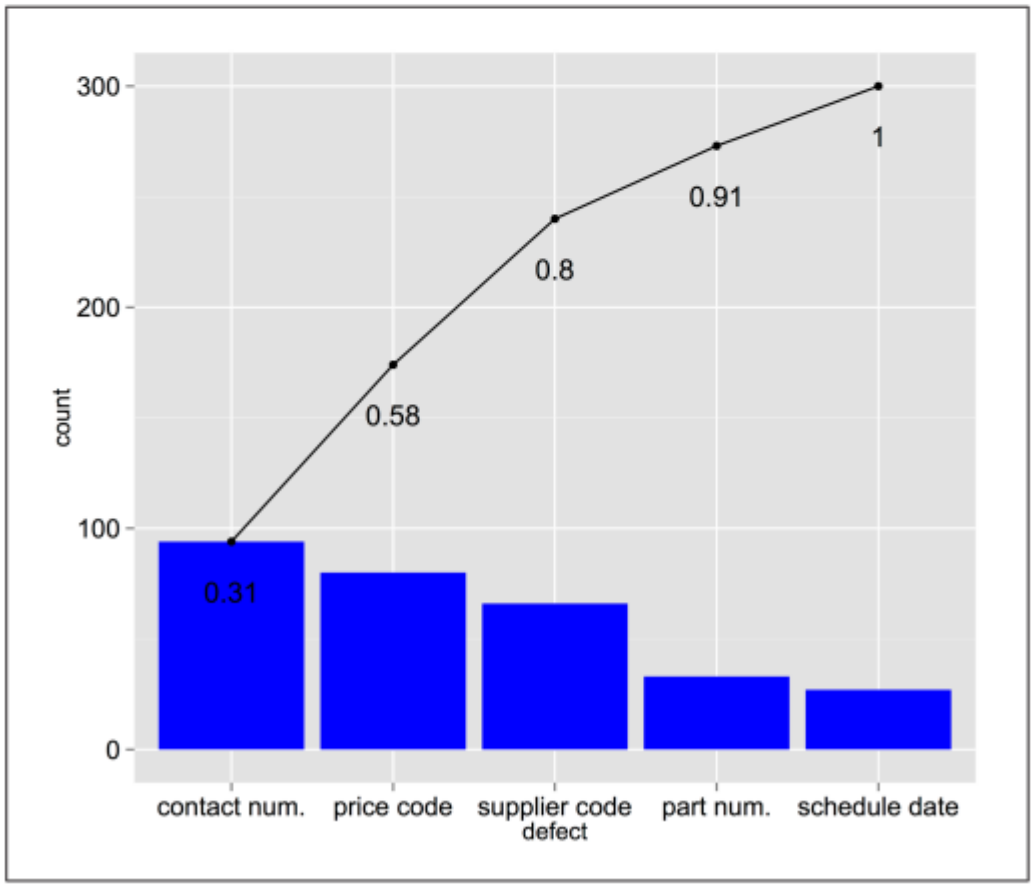

Figure 5-10. A pareto chart.

Chapter 6: Metric design

A data-driven organization needs to set out a clear strategy, that is, a direction that the business is heading, and then determine a set of top-level metrics—key performance indictors (KPIs)—to track whether the business is indeed heading in the right direction and to monitor progress and success. Responsibility for driving those toplevel KPIs flows down to divisions or business units where they may define additional KPIs specific for that business unit.

- **Metric data**
    - There are several considerations when choosing or designing a metric. In an ideal world, metrics should exhibit a number of traits.
    - **Simple**
        - Simple metrics are, by definition, simple to define, which in turn means they are:
          Simpler to convey to others: there is less chance of confusion
          Simpler to implement: they are more likely to be calculated correctly
          More likely to be comparable to other teams or organizations
        - Of course, there are many reasons why one might legitimately want to add on additional business logic and edge cases to create a more complex metric. You may need to filter out sources of bias or extremal edge cases. Or, you may need a metric that explicitly tracks a particular biased subsample, such as the subset of customer cases that cost you the most to resolve. You have to take each case on its merits, but try to avoid adding additional complexity for rare edge cases that add little to the overall business value and insight of that metric.
    - **Standardized**
        - Match standard metric definitions wherever possible. For instance, if there is a standard, well-defined metric for website bounce rate, then use it unless there is a very good reason to define and imple- ment a home-grown variant. If the retail industry uses exits to meas- ure foot traffic at stores, use that metric and not enters, even if they might be highly numerically or conceptually comparable. For instance, to track monthly active users, Facebook only includes peo- ple logged in, whereas Yelp considers both those logged in as well as guests.
        - Being standardized will generate less confusion, especially for colleagues joining your teams from other organizations.
    - **Accurate**
        - Metrics should be accurate. That is, their mean numerical value should be close to the true underlying mean value. If you compare this to archery, it is the equivalent of an arrow being on target.
        - Imagine a sharpshooter at a shooting range, firing a rifle at a distant target and using a scope to help him see. There is a constant breeze blowing the bullet off target; thus, he turns a knob on the side of the scope to adjust the (mis)alignment of the scope and the barrel—the “fudge factor”—to account for the wind. If the wind drops or picks up, however, that scope alignment is

stale and will no longer help get the bullet on the target. Conditions change, and you have to keep your models and any fudge factors up to date

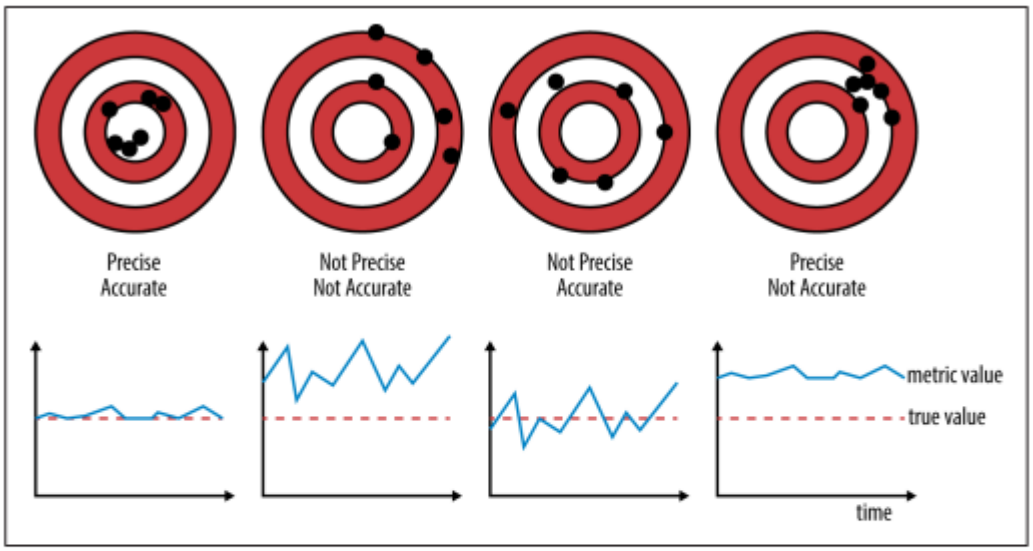

Figure 6-1. Precision (being stable or clustered) and accuracy (being on target) with illustrative example in two-dimensional data. Inaccurate metrics have a bias such that their mean differs from the true mean in a systemic way. Precision captures the variability: how much the mean value would differ if you were to repeat the experiment multiple times and collect new samples of the same size.

- **Precise**
  - Metrics should be precise. If you look at the picture above, precision means how close the arrows are to eachother(the 1st and 4th ones being the most precise)
  - One important thing when talking about precision is sample size. The bigger the sample size, the smaller the standard error becomes.
  - **Strive for accurate and precise metric, and consider the costs and benefits of larger samples.**
- **Relative vs Absolute**
  - Think through what you want to have happen in the underlying data, and choose the metric to be absolute or relative so that it will appropriately track that change.
- **Robust**
  - Use exploratory data analysis to get a feel for the data, and use those to guide the choice of appropriate robust metrics and measures.
- **Direct**
  - Where possible, instrument your systems and processes as much as possible and at the lowest level possible to try to avoid proxies. Don't always take the easy way out, and use the data that you happen to have. Focus on the data you should be collecting and using that if it would better serve your needs.

- **Key performance Indicators**

- KPI's are the suite of highest-level measures linked to the company's strategic objectives. They help define and track the direction that the business is going and help it meet its goals.
- KPI's are:
    - **Clearly defined; you don't want confusion or ambiguity about a core metric**
    - **Measurable; KPI's need to be quantifiable**
    - **Having targets; KPI's should be achieveable but at a stretch with hard work**
    - **Visible; KPI's need to be visible to at least those responsible for driving them, but ideally broader than that.**
    - **Something that reflect what the organization is trying to achieve**
- KPI's are often written in the SMART(ER) format.
- KPI's tend to cover all major areas of the business and any parts of the business that are the particular strategic focus for that period, usually that being a year.
- There seems to be a standard for the amount of KPI's a business can have. "twenty is plenty"
- Too many KPI's lead to staff having to divide their focus, and that leads to being less effective overall.

Chapter 7: storytelling

- Every dataset has a story to tell. It is a data scientists job to find this story and find out how he will translate it to something that everyone can understand.
- Take a look at this graph about Australian twitter accounts:

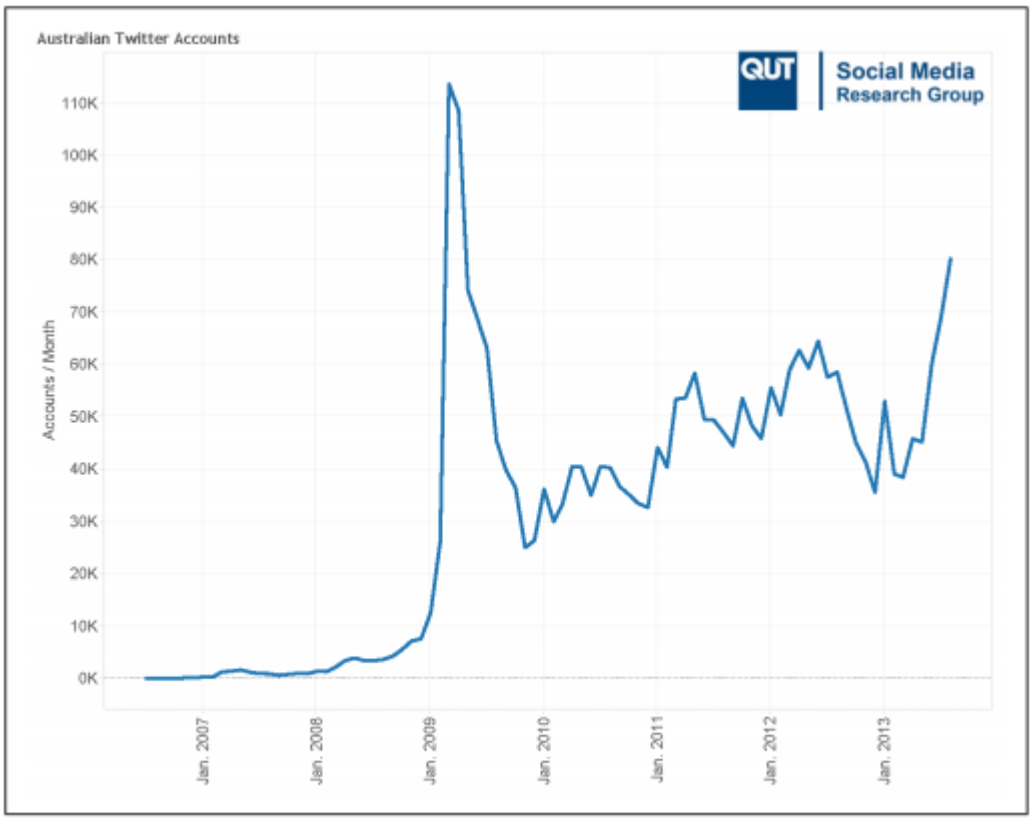

Figure 7-1. New account creation of Australian twitter accounts versus time. From http://bit.ly/aus-twitter.

You can see that somewhere in 2009 there was an enormous jump in Australian twitter accounts. Okay, what now?

This graph on its own is not really that interesting, because everyone can see the huge jump in accounts, but has no idea why.

Now take a look at the graph below:

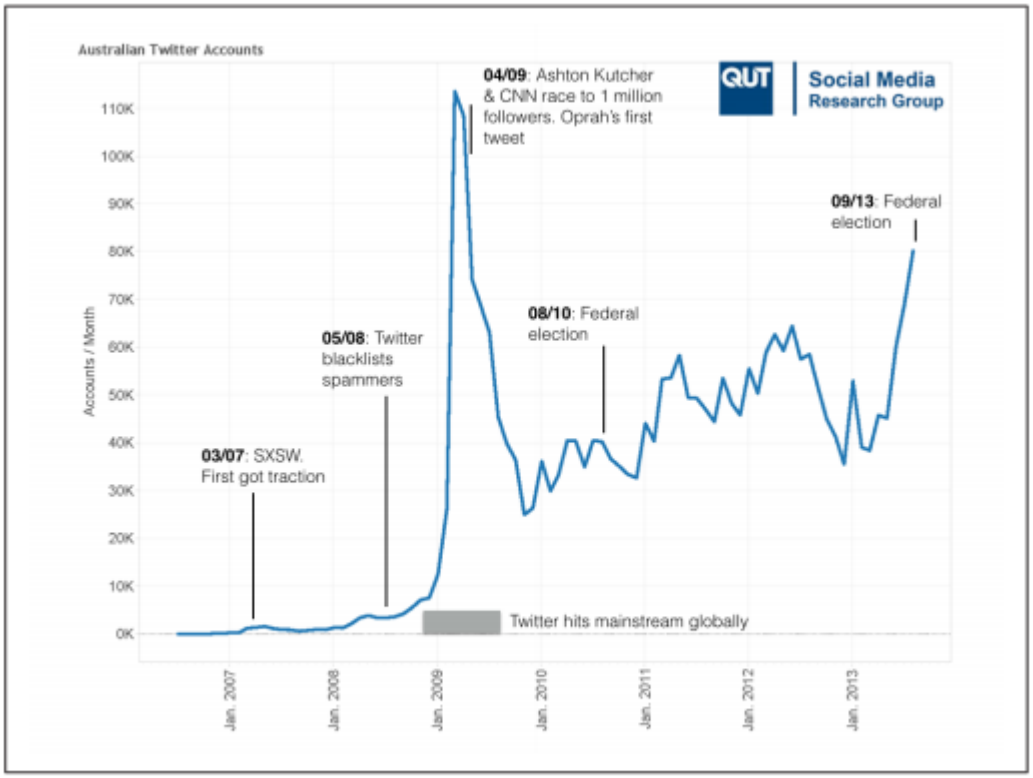

Figure 7-2. Annotated version of Figure 7-1.

This jump now becomes clear to everyone, because there is a description of the events that happened back then; there was a race to 1 million followers AND a huge celebrity made her first twee AND twitter hits mainstream globally(which in my eyes means that it was hyped back then). Other big rises involve big events too, such as the federal elections, with people browsing twitter to either spread their love for a certain party or to find information perhaps.

- **First steps**
    - Ask yourself the following questions if you want to think about the best way to present your data;
        - What are you trying to achieve?
        - Who is your audience?
        - What's your medium?
- **What are you trying to achieve**
    - What is your objective? Why are you putting this report together?
    - There must be a clear understanding of why you are presenting these data or findings
- **Who is your audience?**
    - How data literate is your audience? How technical are they?
    - These are questions you must ask yourself in order to come up with a good presentation. You will need to adapt your presentation and visualizations based on how well your audience is experienced with data science
- **What's the medium?**
    - How are you going to present your findings? Will you do a written report? A dashboard full of visualizations? A powerpoint?

- ## Data visualization
    - **Choosing a chart**
        - There is a lot of choice
        - the right chart of choice depends on your types of variables, how many variables or other factors such as the values themselves.
    - **You can categorize the charts by focusing on the following four points:**
        - Comparisons
        - Distribution
        - Relationships
        - Comparisons
    - **There has been made a data visualization checklist, which tells you about the essentials that you need to think about when visualizing data:**
        - Text
        - Arrangement
        - Color
        - Lines
        - Overall
    - **Focusing the message**
        - Your goal is to create something visual to make your message clear to other people. There are a lot of ways how you can achieve this, but the general consensus is that your visualization needs a clean, slick design. This design should be easy to understand while it still contains the message you want to make clear.