Analytics – Data Discovery
QlikView
3rd-5th September 2014

KS Gopinath Narayan, IAAS
CIA, CFE, PMP
Pr. Director (IT Audit)
Office of the CAG of India
narayanKSG@cag.gov.in
Presentation Outline

• About Data Analytics
• Data Discovery through Data Visualisation
• Various types of Graphs and Charts
• Tools for Data Discovery
  – Microsoft Excel – Pivot and PowerPivot
  – QlikView / QlikSense
  – Tableau Desktop/Public
• QlikView
About Data Analytics

• Analytics – Reduction of data to understandable findings.

• Data analytics is an analytical process by which **insights are extracted** from operational, financial, and other forms of electronic data.

• **provide the “how?” and “why?” answers to initial “what?” questions frequently found in the information initially extracted from the data.**

(KPMG-2013)
About Data Analytics

• Data Analytics driven by the Hype around Big Data Analytics
• **Big data**: “collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications”
• Massive growth in data volumes
• **Google CEO Eric Schmidt in 2010**: “There were 5 Exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days.”
• Growth Largely in unstructured data.
Data Size Primer

• KB = kilobyte, a paragraph of text
• MB= Megabyte, $10^6$ Bytes, Complete works of Shakespeare = 5MB
• GB= Gigabyte, $10^9$ Bytes, Accounts Transaction Data of one year 1-3 GB (States in India)
• TB=Terabyte, $10^{12}$ Bytes, Tweets created daily = 12+ TB, US Library of Congress=235 TB
• PB=Petabyte, $10^{15}$ Bytes, Data processed in a day by Google=24 PB
• EB=Exabyte, $10^{18}$ Bytes, Total data created in 2006 = 161 EB
Technological Changes Impacting Analytics

• Traditionally, Analytics capability through Business Intelligence (BI) Tools / Data Warehouse
  – Top down, IT modeled with reports, KPIs, slice and dice capability
• BI moving towards self-service delivery model. New tools/segments being created
• New opportunity in the form of In-Memory Analytics
• Advanced analytics capability now affordable. Powered by:
  – Increased computational power
  – Lower cost of RAM
  – 64 Bit computing
# Data Analytics Classification

<table>
<thead>
<tr>
<th>Descriptive Analytics (Business Intelligence)</th>
<th>Predictive Analytics</th>
<th>Prescriptive Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>o What and when did it happen? &lt;br&gt; o How much is impacted and how often does it happen? &lt;br&gt; o What is the problem?</td>
<td>o What is likely to happen next? &lt;br&gt; o What if these trends continue? &lt;br&gt; o What if?</td>
<td>o What is the best answer? &lt;br&gt; o What is the best outcome given uncertainty? &lt;br&gt; o What are significantly differing and better choices?</td>
</tr>
<tr>
<td>Statistics</td>
<td>Data Mining &lt;br&gt; Predictive Modeling &lt;br&gt; Machine Learning &lt;br&gt; Forecasting &lt;br&gt; Simulation</td>
<td>Constraint-based optimization &lt;br&gt; Multiobjective optimization &lt;br&gt; Global optimization</td>
</tr>
</tbody>
</table>

**Descriptive Analytics / Data Discovery**  
**Predictive Analytics**  
**Prescriptive Analytics**
CAATs and Audit - Why Data Discovery?

- Traditional CAATs - IDEA, ACL Software
- Some usage of MS Excel and MS Access and SQL
- Generalised Audit Software with transaction based analytics - Rule based or micro-level analytics
  - Data extraction and analysis entailing sorting, grouping, filtering, joining, sampling, irregularity testing-gap detection, Benford analysis
- Good for evaluating known conditions/Compliance audit
- Lack Macro level Analytics capability - Understand the big picture to identify key areas to check for non-compliance
- Need for new generation Analytics tools to supplement Generalised Audit Software - Data Discovery/Visualisation
... there are known knowns; there are things we know that we know. 
There are known unknowns; that is to say, there are things that we now know we don't know.  
But there are also unknown unknowns – there are things we do not know we don't know.

Donald Rumsfeld: Secretary of Defense
February 2002

about the lack of evidence linking the government of Iraq with the supply of weapons of mass destruction to terrorist groups
Why Visualisation?

• Visual analysis aids analytical reasoning
• leverages the incredible capabilities and bandwidth of the visual system
• takes advantage of our brains’ built-in “software” to identify patterns and communicate relationships and meaning
  – Identify trends and outliers, discover or search for interesting or specific data points in a larger field
• inspire new questions and further exploration

John Tukey - “…the picture-examining eye is the best finder we have of the wholly unanticipated”
Figure 16: Who buys the most books?
In this market segmentation analysis, the heat map reveals a new campaign idea. High-income households of people in their sixties buy children's books. Perhaps it's time for a new grandparent-oriented campaign?
Anscombe’s Quartet

• Four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.
• Each dataset consists of eleven (x,y) points.
• 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.
### Anscombe’s Quartet Data

<table>
<thead>
<tr>
<th>Anscombe’s quartet</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$y$</td>
<td>$x$</td>
<td>$y$</td>
<td>$x$</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>10.0</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
</tr>
</tbody>
</table>

For all four datasets:

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of $x$ in each case</td>
<td>9 (exact)</td>
</tr>
<tr>
<td>Variance of $x$ in each case</td>
<td>11 (exact)</td>
</tr>
<tr>
<td>Mean of $y$ in each case</td>
<td>7.50 (to 2 decimal places)</td>
</tr>
<tr>
<td>Variance of $y$ in each case</td>
<td>4.122 or 4.127 (to 3 decimal places)</td>
</tr>
<tr>
<td>Correlation between $x$ and $y$ in each case</td>
<td>0.816 (to 3 decimal places)</td>
</tr>
<tr>
<td>Linear regression line in each case</td>
<td>$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)</td>
</tr>
</tbody>
</table>
Anscombe’s Quartet - Graphed
Data Visualisation-1- John Snow’s map of London Cholera Epidemic of 1854
Figurative Map of the successive Losses in men of the French Army in the Russian Campaign 1812-1813.

Drawn up by M. M. M. C. J. de Chasseloup Laubat.

Paris, November 25, 1869.

The numbers of men present are represented by the widths of the colored zones at a rate of one millimeter for every ten thousand men. The numbers are further written across the zones. The red designates the men who entered Russia, the blue those who left it. The information which has served to draw up the map has been extracted from the works of M. J. F. Payen, of the 4th Corps, of Chasseloup Laubat, and of the unpublished story of Jacob, physician of the army since October 1812.

In order to better judge with the eye the movements of the army, I have assumed that the troops of General Murat and of Marshal Davout, who had been detached at Smolensk and Mogilev, and have, respectively, crossed the Vistula and Vistula, have always marched with the army.

GRAPHIC TABLE of the temperature in degrees of the Résamur thermometer below zero.

The Gzhatsk given the frozen Neman at avalise.
Data Visualisation-2: Charles Minard’s graph (1869) of French invasion of Russia in 1812
1. Exploratory Visualization
2. Explanatory visualization

Purpose of Visualisation - to move information from point A to point B

• In Exploratory Visualisation:
  – A: Dataset
  – B: Designer’s mind

• In Explanatory Visualisation:
  – A: Designer’s mind
  – B: Reader’s mind
Charts and Graphs in Visualisation

1. Bar Chart
2. Line Chart
3. Pie Chart
4. Map Chart
5. Scatter Plot
6. Bubble Chart
7. Histogram Chart
8. Heat Chart
9. TreeMap
10. Box-and-whisker Plot
Bar Chart

• Quick to compare information, revealing highs and lows at a glance.
• Effective with numerical data that splits nicely into different categories.

• When to use bar charts:
  – Comparing data across categories
Line Chart

• Connect individual numeric data points.
• Simple way to visualize a sequence of values.
• When to use line charts:
  – Viewing trends in data over time
Pie Chart

• To show relative proportions – or percentages – of information
• the most commonly mis-used chart type
• Not good for comparing data

• When to use pie charts:
• Showing proportions.
Map Chart

- New Chart Type
- When you have any kind of location data

- When to use maps:
- **Showing geocoded data.** Examples: Insurance claims by state, product export destinations by country etc.
Scatter Plot

• To see how different pieces of information relate
• Effective way to get a sense of trends, concentrations and outliers – identify areas to focus further investigation
• When to use scatter plots: Investigating the relationship between different variables
Bubble Chart

• A technique to accentuate data on scatter plots or maps
• Not a separate type of visualization
• varied size of circles provides meaning about the data

• When to use bubbles:
• **Showing the concentration of data along two axes.**

Examples: sales concentration by product and geography
Histogram Chart

• To see how data distributed across groups
• To understand which categorization approach makes sense

• When to use histograms:
• Understanding the distribution of your data. Examples: Number of customers by company size, student performance on an exam
Heat Chart

• Excellent for comparing data across two categories using colour
• Shows where intersection of the categories is strongest and weakest

• When to use heat maps:
  • **Showing the relationship between two factors.** Examples: product adoption across regions

*Figure 16: Who buys the most books?*

In this market segmentation analysis, the heat map reveals a new campaign idea. High-income households of people in their sixties buy children’s books. Perhaps it’s time for a new grandparent-oriented campaign?
TreeMap

• To see data at a glance and discover how the different pieces relate to the whole
• Uses a series of rectangles, nested within other rectangles, to show hierarchical data as a proportion to the whole

• When to use treemaps:
  • Showing hierarchical data as a proportion of a whole:

Figure 18: Support Cases at a Glance
This treemap shows all of a company’s support cases, broken by case type, and also priority level. You can see that Document, Feedback, Support and Maintenance make up the lion share of support cases. However, in Feedback and Support, P1 cases make up the most number of cases, whereas most other categories are dominated by relatively mid P4 cases.
**Box-and-whisker Plot**

- Shows distributions of data
- Box contains the median of the data along with the 1st and 3rd quartiles
- Whiskers typically represents data within 1.5 times the Inter-quartile Range (or Maximum/Minimum)

- When to use box-and-whisker plots:
  - **Showing the distribution of a set of data**: Examples: understanding data at a glance, seeing how data is skewed towards one end, identifying outliers
Analytics for Audit-Tool Fitment Criteria

- Strong Data Discovery and Data Visualisation capabilities
- Reasonably easy to learn - quick solution development time
- Operate from Desktop/Laptop of the Auditors
- Capable of scaling up to handle large volumes of data through a central server based model
- Reasonably priced
Data Discovery/Visualisation Tools
(Selected by IAAD)

• QlikView:
  http://www.qlikview.com/

• Tableau:
  http://www.tableausoftware.com/

• Microsoft BI:
  (Pivot/PowerPivot)
Gartner’s Magic Quadrant for Analytics - 2014

Business Intelligence & Analytics Platforms

Advanced Analytics Platforms

As of February 2014

Source: Gartner (February 2014)
In Memory Analytics Performance

• Compared performance for data stratification.
• **Stratification** on Billed Units, summation on Fixed Charge and Energy Charge
• 2 hours 10 minutes in IDEA, 3 seconds in QlikView
THANK YOU