Welcome
Think Outside the Box(plot)

Jerry Valerio
Datavangelist
Tableau
Jerry Valerio
Senior Manager, Sales Consulting
gvalerio@tableau.com

Gerard is a data engineer, data evangelist, and data strategist with customer advisory experience working for Tableau, and previously Vertica and Informatica and management consulting experience previously working for Accenture and PricewaterhouseCoopers.
Jerry Valerio

§ Foodie since girth and it shows!

- Side hustles as adjunct professor and data science bootcamp instructor.
- Sky-dived (tandem) and also zip-lined once because YOLO!
Audience

- Basic knowledge of statistics
- Interested in Tableau's statistical capabilities
  - Distribution
  - Summary
  - Modeling
Agenda
Why Visual Analysis?
# Anscombe's Quartet

Let's analyze some data ...

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th></th>
<th>II</th>
<th></th>
<th>III</th>
<th></th>
<th>IV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>10</td>
<td>8</td>
<td>8.04</td>
<td>10</td>
<td>7.46</td>
<td>8</td>
<td>6.58</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>10</td>
<td>8</td>
<td>9.14</td>
<td>13</td>
<td>8.74</td>
<td>8</td>
<td>5.76</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>8</td>
<td>8</td>
<td>6.95</td>
<td>8</td>
<td>6.77</td>
<td>8</td>
<td>7.71</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>8</td>
<td>8</td>
<td>8.14</td>
<td>8</td>
<td>7.11</td>
<td>8</td>
<td>8.84</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>13</td>
<td>9</td>
<td>7.58</td>
<td>9</td>
<td>8.11</td>
<td>8</td>
<td>8.47</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>13</td>
<td>9</td>
<td>8.77</td>
<td>9</td>
<td>9.26</td>
<td>8</td>
<td>7.04</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>9</td>
<td>9</td>
<td>8.81</td>
<td>9</td>
<td>8.84</td>
<td>8</td>
<td>7.04</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>11</td>
<td>11</td>
<td>8.33</td>
<td>11</td>
<td>8.71</td>
<td>8</td>
<td>8.47</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>14</td>
<td>9</td>
<td>9.96</td>
<td>11</td>
<td>7.81</td>
<td>8</td>
<td>7.04</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>11</td>
<td>11</td>
<td>8.26</td>
<td>12</td>
<td>8.15</td>
<td>8</td>
<td>5.25</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>6</td>
<td>14</td>
<td>6.24</td>
<td>6</td>
<td>6.08</td>
<td>8</td>
<td>5.25</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>4</td>
<td>6</td>
<td>6.13</td>
<td>6</td>
<td>5.39</td>
<td>8</td>
<td>12.5</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>4</td>
<td>12</td>
<td>4.26</td>
<td>4</td>
<td>9.13</td>
<td>8</td>
<td>5.56</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>5</td>
<td>12</td>
<td>3.1</td>
<td>4</td>
<td>6.15</td>
<td>8</td>
<td>7.91</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>12</td>
<td>5</td>
<td>5.68</td>
<td>5</td>
<td>7.26</td>
<td>8</td>
<td>6.89</td>
<td></td>
</tr>
</tbody>
</table>
## Anscombe's Quartet

Let's summarize the data ... 

<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

Let's visualize the data ...
Distribution
Histograms
Histograms show us the distribution of numerical data.
Histograms

Basic Histogram

Cumulative Histogram
Percentiles indicate the value below which a given percentage of the observed data falls.

Ex: If a school is in the 66.7\textsuperscript{th} percentile, their teacher score is better or stronger than 2/3 of compared schools.
Box Plots
Box Plots

Traditional Box Plot

Tableau Box Plot
Anatomy of the Tableau Box Plot
Anatomy of the Tableau Box Plot

Interquartile Range
Anatomy of the Tableau Box Plot
Anatomy of the Tableau Box Plot

Outliers!
Summary Statistics?
<table>
<thead>
<tr>
<th>Summary</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Count:</td>
<td>108</td>
</tr>
<tr>
<td>SUM(Sales)</td>
<td></td>
</tr>
<tr>
<td>Sum:</td>
<td>$609.206</td>
</tr>
<tr>
<td>Average:</td>
<td>$5,641</td>
</tr>
<tr>
<td>Minimum:</td>
<td>$259</td>
</tr>
<tr>
<td>Maximum:</td>
<td>$22,171</td>
</tr>
<tr>
<td>Median:</td>
<td>$4,011</td>
</tr>
<tr>
<td>Standard deviation:</td>
<td>$4,824</td>
</tr>
<tr>
<td>First quartile:</td>
<td>$2,180</td>
</tr>
<tr>
<td>Third quartile:</td>
<td>$7,647</td>
</tr>
<tr>
<td>Skewness:</td>
<td>1.51</td>
</tr>
<tr>
<td>Excess Kurtosis:</td>
<td>2.17</td>
</tr>
</tbody>
</table>
A measure of the tendency of your data to have extreme values to one side. Positive skewness means the extreme values are to the right, while negative skewness means the extreme values are to the left.
Summary Card - Kurtosis

A measure of the tendency of your data to have more extreme or outlying values than a normal distribution. A normal distribution has a kurtosis of 3.

Positive Excess Kurtosis

Negative Excess Kurtosis
Modeling
What do we mean by Modeling?

Applying mathematical functions to data in an attempt to surface hidden insights.
Classifying Data

Unsupervised Classification
Similar with respect to several attributes

- **Examples:**
  - Trend / Regression Lines
  - Forecasts
  - K-Means Clustering

Supervised Classification
Similar with respect to a target

- **Examples:**
  - Logistic Regression
  - Decision Trees
  - Neural Networks
  - Random Forest
Trend Lines / Regression Lines
Trend Lines

Options
- Linear
- Exponential
- Logarithmic
- Polynomial
- Power

\[ y = mx + b \]
Exponential

Options
- Linear
- Exponential
- Logarithmic
- Polynomial
- Power

\[ y = b_2 \times e^{b_1 \times x} \]
Logarithmic

Options

- Linear
- Exponential
- Logarithmic
- Polynomial
- Power

Y = b0 + b1 * ln(X)
Polynomial and Power

Options
- Linear
- Exponential
- Logarithmic
- Polynomial
- Power

Symbolically:

\[ y = b_0 + (b_1 \times x) + (b_2 \times x^2) \ldots \]

\[ \ln(Y) = \ln(b_0) + b_1 \times \ln(X) \]
Find a line that minimizes the squared values of these distances - the distance from the actual value in the data set to the trend line.
Trend Models: Describing the Formula

A linear trend model is computed for sum of Power output (MW) given sum of Wind Speed (m/s). The model may be significant at \( p < 0.05 \).

**Model formula:**

\[
(\text{Wind Speed (m/s)} + \text{intercept})
\]

- **Number of modeled observations:** 365
- **Number of filtered observations:** 0
- **Model degrees of freedom:** 2
- **Residual degrees of freedom (DF):** 363
- **SSE (sum squared error):** 1076
- **MSE (mean squared error):** 2.96419
- **R-Squared:** 0.956448
- **Standard error:** 1.72158
- **p-value (significance):** < 0.0001

### Individual trend lines

<table>
<thead>
<tr>
<th>Term</th>
<th>Value</th>
<th>StdErr</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wind Speed (m/s)</td>
<td>2.24418</td>
<td>0.025133</td>
<td>89.2853</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>intercept</td>
<td>-6.45329</td>
<td>0.20413</td>
<td>-31.6136</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>
Trend Models: Evaluating Model Fit

A linear trend model is computed for sum of Power output (MW) given sum of Wind Speed (m/s). The model may be significant at $p \leq 0.05$.

Model formula: $\text{(Wind Speed (m/s) + intercept)}$

- Number of modeled observations: 365
- Number of filtered observations: 0
- Model degrees of freedom: 2
- Residual degrees of freedom (DF): 363
- SSE (sum squared error): 1075
- MSE (mean squared error): 2.96419
- R-Squared: 0.936448
- Standard error: 1.72158
- p-value (significance): < 0.0001

Individual trend lines:

<table>
<thead>
<tr>
<th>Power output (MW)</th>
<th>Wind Speed (m/s)</th>
<th>p-value</th>
<th>DF</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>&lt; 0.0001</td>
<td>363</td>
<td>Wind Speed (m/s)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>intercept</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

p-values:

- Wind Speed (m/s): < 0.0001
- Intercept: 0.0001

Copy to Clipboard
The R-Squared value shows the ratio of variance in the data, as explained by the model, to the total variance in the data. The P-value reports the probability that the equation of the line was a result of random chance. The smaller the p-value, the more significant the model is. A p-value of 0.05 or less is often considered sufficient.
Forecasting
Forecast Requirements

- At least:
  - One Dimension
  - One Measure

- Dimension Requirements
  - Date Field
  - Integer Field

**NOTE:** Tableau requires at least five data points in the time series to estimate a trend, and enough data points for at least two seasons or one season plus five periods to estimate seasonality.
Forecasting Terms

Exponential Smoothing:
more recent values are given
greater weight

Trend
Tendency in the data to increase
or decrease over time

Seasonality
Repeating, predictable variation in
date value, such as an annual
fluctuation in temperature relative
to the season.

Granularity
The unit you choose for the date
value is known as the granularity of
the date
Forecasting Models

Multiplicative models can significantly improve forecast quality for data where the trend or seasonality is affected by the level (magnitude) of the data.
Forecast Description

- OK: less error than a naïve forecast
- GOOD: less than half as much error as a naïve forecast
- POOR: means that the forecast has more error.
Clustering
Clusters

K-means is a simple algorithm that tries to minimize the distance from a center point to all points in the same cluster.

But first, we need to make a reasonable estimate of the number of clusters in our data.

How many clusters should there be with this visualization?
Clusters

3 – simple enough. We use Calinski’s algorithm to determine “k”.

Then we use Lloyd’s algorithm to compute the distances from each center point in our three clusters to every point in our data. Assign each point to the closest center.

Repeat until points don’t change center assignments.
Clusters – Describing the results

This shows the inputs to the clusters. We see our two variables, we were not aggregated and scaling was not adjusted.
Clustering

Grouping a set of objects such that marks within each cluster are more similar to one another than they are to marks in other clusters
Clusters – Saving the results

The clustering is done. Three clusters with default names and colors. Plus, if new data comes in the data gets re-clustered and results may change.

Drag/drop the clusters pill onto the Data pane.

I’m going to rename it to Vehicle Type Clusters. Notice the icon will change.
Clustering – Fine tuning the saved group

Rename the groups to something more meaningful e.g.:

- Fuel Efficient
- Middle of the Road
- Gas Guzzlers
Clusters – Fine tuning the saved group

Now I can remove the ad-hoc “Clusters” group and replace it with my saved group.

Update the colors and I am done!
Correlation (is not Causation)
Pearson’s correlation coefficient is a measure of the strength and direction of the linear relationship between two variables.
Correlation computation

Correlation Function:

- Built into Tableau
- Uses Calculated Fields

```
CORR(EXPR1, EXPR2)

Returns the Pearson correlation coefficient of two expressions.

Example: CORR([Sales], [Profit])
```
Recap & Last Notes
Recap

Distribution

• Histograms
• Percentiles
• Box Plots
• Control Charts

Modeling

• Trend Lines
• Forecasting
• Clustering
• Correlation Coefficients
Thank You

Jerry Valerio

gvalerio@tableau.com