Insight Types Specification

Introduction

We have developed 12 different types of insights, corresponding to 12 different perspectives commonly adopted in practice. They are:

1. *Attribution*
2. *Outstanding No. 1*
3. *Outstanding Top 2*
4. *Outstanding Last*
5. *Evenness*
6. *Change Point*
7. *Outlier*
8. *Seasonality*
9. *Trend*
10. *2DClustering*
11. *Correlation*
12. *Cross-Measure Correlation*

These 12 insight types can be grouped into 3 categories according to their definitions and semantics, as depicted in Table 1.

5 insight types fall into the category of SinglePointInsight. SinglePointInsight refers to the insights with single subspace and single measure, and breakdown by a non-ordinal dimension.

4 insight types belong to the category of SingleShapeInsight, which only differs from SinglePointInsight by the use of ordinal breakdown dimension. Semantically, SingleShapeInsight refers to the insights related to time series.

3 insight types belong to the category of CompoundInsight. CompoundInsight refers to the insights with multiple subspaces or measures, which provides relatively richer semantics. Specifically, Correlation insight compares two subspaces in the insight subject; Cross-Measure-Correlation and 2DClustering compare two measures in the insight subject.

Table 1. Insight Categorization

Insight Category	*SinglePointInsight*	*SingleShapeInsight*	*CompoundInsight*
Insight types	Outstanding No. 1 Outstanding No. Last Attribution Outstanding Top 2 Evenness	Change Point Trend Seasonality Outlier	Correlation Cross-Measure-Correlation 2DClustering
#Types	5	4	3

Insight Type Specification

SinglePointInsight

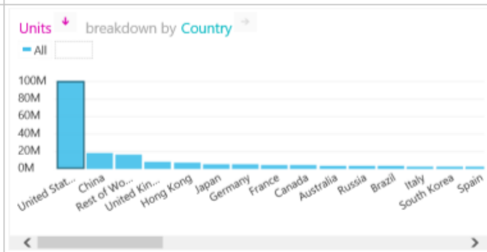
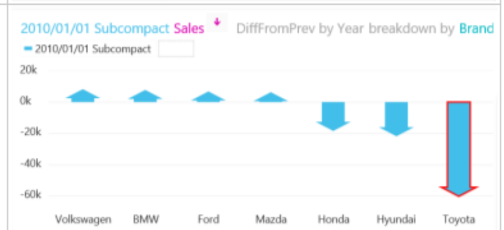
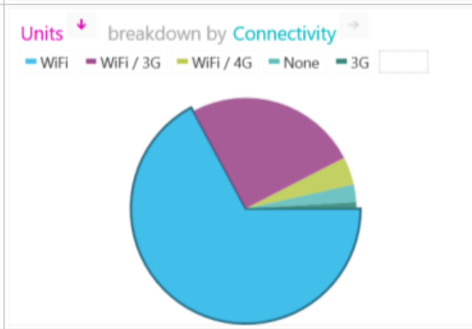
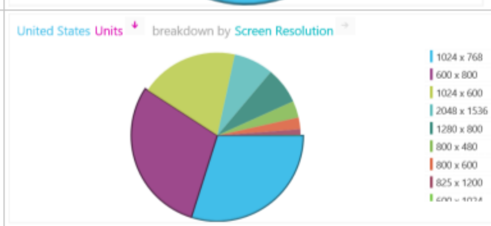
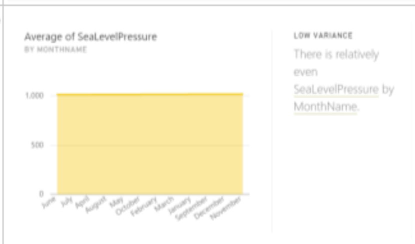
Insight type	Description	Example
Outstanding No.1	Among a comparison group with non-negative aggregation results, Outstanding No.1 shows the fact that the leading value is remarkably higher than all the remaining values. (But not as high as being dominant, which will be introduced later with Attribution)	
Outstanding No.last	Similar to Outstanding No.1, it is for negative aggregation results and shows the fact that the most negative value is remarkably lower than all the remaining values (only negative values are taken into account).	
Attribution	Among a comparison group with non-negative aggregation results, Attribution shows the fact that the leading value dominates (accounting for >= 50% market share of) the group.	
Outstanding top 2	Similar to Attribution and Outstanding No.1, among a comparison group with non-negative aggregation results, Outstanding top 2 shows the fact that the leading two values are remarkably higher than the remaining values.	
Evenness	The cases where all values of a measure for a given category are very close to each other.	

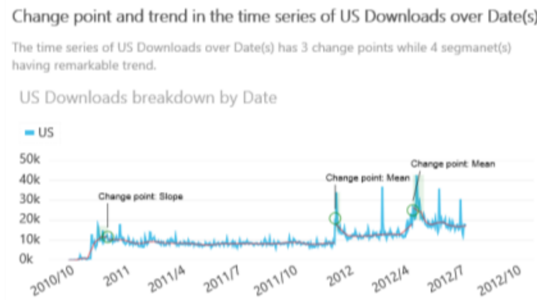
Figure 1. Description of SinglePointInsight

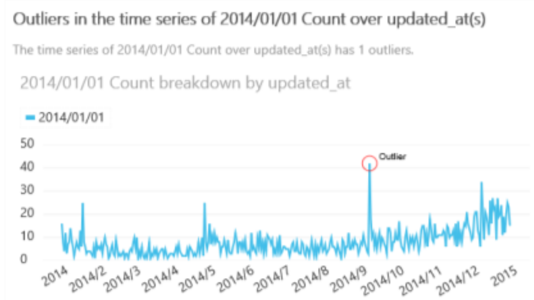
The significance calculation of SinglePointInsight shares similar logic. Take Outstanding No. 1 as an example:

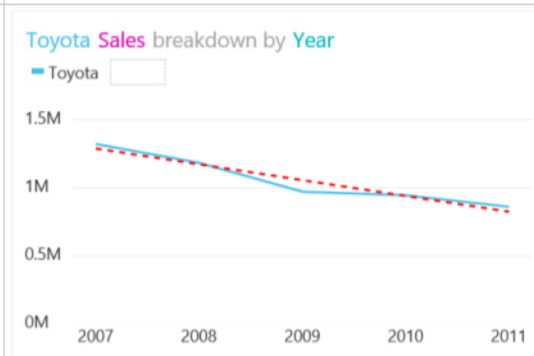
Significance of Outstanding No. 1: Given a group of non-negative numerical values $\{x\}$ and their biggest value x_{max}, the significance of x_{max} *being Outstanding No.1 of* $\{x\}$ *is d*efined based on the p-value against the null hypothesis of $\{x\}$ *obeys an ordinary long-tail distribution*. The p-value will be calculated as follows:

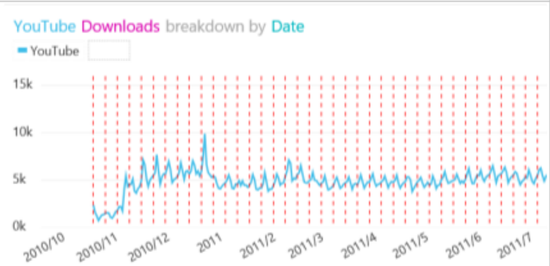
1. We sort $\{x\}$ in descending order;
2. We assume the long-tail shape obeys a power-law function. Then we conduct regression analysis for the values in $\{x\}\backslash x_{max}$ using power-law functions $\alpha \cdot i^{-\beta}$, where i is an order index and in our current implementation we fix $\beta = 0.7$ in the power-law fitting;
3. We assume the regression residuals obey a Gaussian distribution. Then we use the residuals in the preceding regression analysis to train a Gaussian model H;
4. We use the regression model to predict x_{max} and get the corresponding residual R;
5. The p-value will be calculated via $P(R|H)$.

SingleShapeInsight

<table>
<tr><th>Insight type</th><th>Description</th><th>Example</th></tr>
<tr><td>Change point</td><td>Change point of time-series signals regarding significant change of (1) mean value or (2) curve slope or (3) their combination between its preceding and successive regions</td><td>


</td></tr>
<tr><td>Outlier</td><td>Outlier of time-series signals</td><td>


</td></tr>
<tr><td>Trend</td><td>A time series has an obvious trend (increase/decrease) with a certain turbulence level (steadily/with turbulence).</td><td>


</td></tr>
<tr><td>Seasonality</td><td>A time series shows clear seasonality.</td><td>


</td></tr>
</table>

Figure 2. Description of SingleShapeInsight

Since all SingleShapeInsights are time series related insights, we can follow the standard statistical hypothesis testing procedure for time series data. Take Change Point as an example:

Significance of Change Point. A change point is typically modelled as a mean-value change point.

1. A change point candidate is evaluated against its left window of n preceding points and its right windows of n successive points, denoted as $\{X_{left}, Y_{left}\}$ and $\{X_{right}, Y_{right}\}$ respectively. The entire window surrounding the change point candidate is denoted as $\{X, Y\}$.
2. For mean-value change point
   a. $\bar{Y}_{left} = \frac{\sum y_{left}}{n}, \bar{Y}_{right} = \frac{\sum y_{right}}{n}$
   b. $\sigma_{\mu_Y} = \frac{1}{\sqrt{n}}\sigma_Y = \frac{1}{\sqrt{n}}\sqrt{\frac{\sum y^2}{2n} - \left(\frac{\sum y}{2n}\right)^2}$
   c. $k_{mean} = \frac{|\bar{Y}_{left} - \bar{Y}_{right}|}{\sigma_{\mu_Y}}$

   and we define the significance based on the p-value of k_{mean} against Gaussian distribution $N(0, 1)$.

CompoundInsight

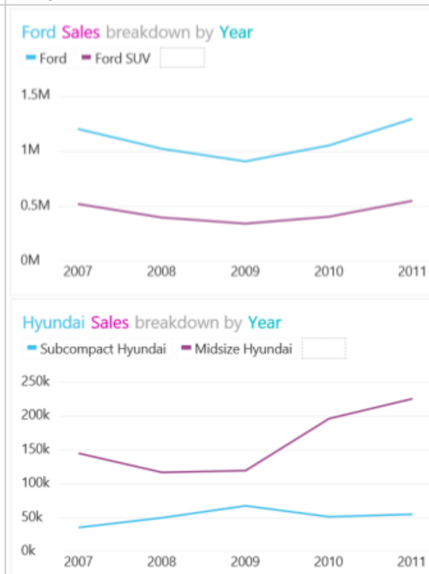
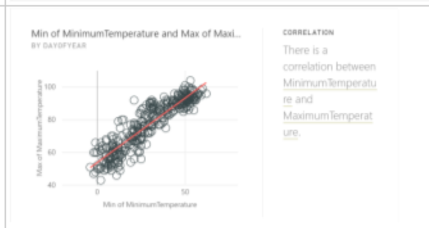
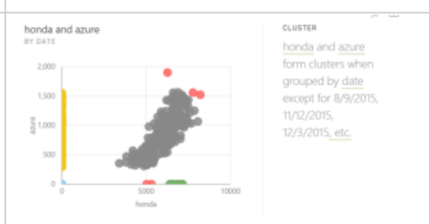
<table>
<tr><th>Insight type</th><th>Description</th><th>Example</th></tr>
<tr><td>Correlation</td><td>Two time series have remarkable positive/negative correlation.</td><td>
</td></tr>
<tr><td>Cross-measure correlation</td><td>It reports cross-measure analysis results regarding remarkable correlation between two measures.</td><td>
</td></tr>
<tr><td>Scatterplot Clustering (2DClustering)</td><td>A scatterplot is generated by: two measure breakdown by a specific dimension. Clustering on scatterplot is complementary to the Cross-measure correlation, to address the cases where data distribution over the 2-dimensional scatterplot is complicated.</td><td>
</td></tr>
</table>

Figure 3. Description of CompoundInsight

Significance of Correlation. The significance of *two time-series signals X and Y being correlated* is defined based on testing using Student's t-distribution with Pearson's correlation coefficient *r*, where *r* is defined as

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)}\sqrt{\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)}}$$

Following are the detailed steps for significance calculation

1 – specify the null and alternative hypotheses:

Null hypothesis $H_0: \rho = 0$

Alternative hypothesis $H_A: \rho \neq 0$

2 – calculate the value of test statistic

$$t = r\sqrt{\frac{n-2}{1-\ r^2}}$$

3 – use the resulting test statistic *t* to calculate the p-value, which is determined by referring to a t-distribution with n-2 degrees of freedom.

4 – the p-value is translated into significance. The lower the p-value, the higher the significance.