GSA

U.S. General Services Administration

D2D Advanced Tableau Training

April 23rd and April 25th

presented by Walter Mehra
D2D Team Member

Pre-Class Activities

- Please make sure you do the following prior to the training session:
  - Ensure that Tableau Desktop (version 10) or higher (D2D is on 10.4.1) is installed on your computer and you have a valid license (or are on Tableau granted two-week trial license);
  - Download the dataset used for this training to your computer's desktop

Objectives

- Analysis on the City Pairs Program (2018 Data)
  - LOD Calculation
  - Z-Scores
  - Pearson Correlation Coefficient (CORR)
  - Regression Analytics
  - Clustering Analytics

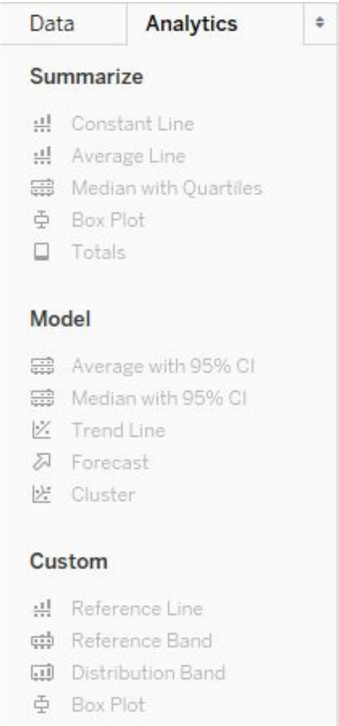
Data Terminology

Data Type	Descriptions/Details	Examples
Number (Decimal) or Number (Whole)	These are either integers or floating points. If a variable can take on any value between two specified values, it is called a continuous variable; otherwise, it is called a discrete variable	3 or 3.14159265359 (Continuous)
Date or Date and Time	Tableau recognizes dates in almost all formats, these values are typically used for time series or trend analysis	11/28/2017 or 11/28/2017 1:00 PM
Boolean	They are logical values	True or False
String	Any sequence of characters. They are enclosed within single quotes. The quote itself can be included in a string by writing it twice.	GSA, Budget Activity, Lease, NAICS
Geographic Role	*Not a datatype, but does clarify the desired outcome for the data field* Identifier for Tableau to facilitate map building included roles are Airport, Area Code, CBSA/MSA, City, Congressional District, Country / Region, County, Latitude, Longitude, NUTS Europe, State/Province, Zip Code/Postcode	KIAD (Airport), 202 (Area Code), VA (State/Province), etc.

Tableau will categorize your data fields into one of two buckets:

- *Dimension* is something you categorize with (e.g., color of a shirt)
- *Measure* is something you do math with (e.g., the number of white shirts)

Advanced Analytics Pane



Model

- Average w/ 95% CI
- Median w/ 95% CI
- Trend Line: Build a linear regression model
- Forecast: Quantitative time-series data using exponential smoothing models in Tableau Desktop. These models capture the evolving trend or seasonality of your data and extrapolate them into the future.
- Cluster: Using k-means, partition the data into k [number of] clusters using the cluster mean. Tableau automates the creation of clusters, including the number of clusters created; clustering aggregates data into groups that are most similar to each other vs. the other groups

Level of Detail Analysis

(Advanced Analytics)

- Just like basic calculations, LOD calculations allow you to compute values at the data source level and the visualization level. However, LOD calculations give you even more control on the level of granularity you want to compute. They can be performed at a more granular level (INCLUDE), a less granular level (EXCLUDE), or an entirely independent level (FIXED) with respect to the granularity of the visualization.

- LOD Expressions can be set to:
  - FIXED (not impacted by the dimensions in the Viz)
  - INCLUDE (includes dimensions within expression AND in the Viz)
  - EXCLUDE (exclude dimensions in the Viz from the calculation)

Tableau Resource on LOD Expressions

❑ Dataset(s): Sample - Superstore

Steps:

1. Create a new sheet, call it **LOD**
2. Create two new Calculated Fields:

   First Purchase Date: *{FIXED [Customer Name] : MIN([Order Date])}*

   Days Since First Purchase: *DATETRUNC('day', [Order Date])-DATETRUNC('day', [First Purchase Date])*
3. Convert **Days Since First Purchase** to a Dimension
4. Drag **Days Since First Purchase** to the Columns shelf and convert it to a Continuous field
5. Drag **Sales** to the Rows shelf
6. Change the aggregation for Sales on Rows from SUM to AVG
7. Add a quick table calculation to Sales on Rows: Running Total
8. Drag First Purchase Date to Color.
9. Click the + in the YEAR(First Purchase Date) field on Color to add the next level down in the date hierarchy: QUARTER(First Purchase Date).
10. Change the QUARTER(First Purchase Date) to affect the viz Color

Z-Score Analysis

(Analytics)

- Z-Score enables us to standardize the data and compare it among datasets

- It’s a calculation that enables you to quantify how many standard deviations from the mean the data is

- Roughly 95% of the data should have a Z-Score between -2.0 and +2.0

$$z = \frac{x - \mu}{\sigma}$$

❑ Dataset(s): FY18 - CPP Award Data

❑ Question: Can we quantify the difference of the unrestricted fare relative to the population set?

Steps:

1. Open Tableau
2. In the Connect pane, select **More**
3. Browse to the FY18 – CPPAwardData Tableau Extract file
4. Create a new sheet, title it **Z-Score Viz**
5. Create new Calculated Fields:

   AVG UNR FARE: *WINDOW_AVG(avg([Unrestricted Fare (YCA)]))*

   STDEV UNR FARE: *WINDOW_STDEV(avg([Unrestricted Fare (YCA)]))*

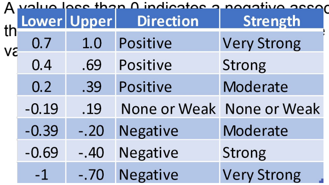
   Z-Score: *(avg([Unrestricted Fare (YCA)]) - [AVG UNR Fare]) / [STDEVP UNR Fare]*
6. Drag **Origin City Name** to the Rows shelf
7. Drag **Z-Score** to the Columns shelf
8. Sort the data in descending order
9. Calculate the Z-Score for **Destination City Name**

Question – What are the interesting values? What is driving the Z-Score?

Correlation Analysis

(Advanced Analytics)

- Identify linear relationship between a sample set of data for two variables (theoretically can include more variables but not in Tableau)
- Tableau includes Pearson Product-Moment Correlation Coefficient (or Pearson Correlation Coefficient)
- Coefficient value ranges from -1.0 to 1.0
  - A value of 0 indicates that there is no association between the two variables.
  - A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.
  - A value less than 0 indicates a negative association; that is, as th of the other va



Lower	Upper	Direction	Strength
0.7	1.0	Positive	Very Strong
0.4	.69	Positive	Strong
0.2	.39	Positive	Moderate
-0.19	.19	None or Weak	None or Weak
-0.39	-.20	Negative	Moderate
-0.69	-.40	Negative	Strong
-1	-.70	Negative	Very Strong

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

❑ Dataset(s): FY18 - CPP Award Data

❑ Question: Do Passenger Counts Affect Unrestricted Fares?

Steps:

1. Create a new sheet, call it **Correlation Analysis**
2. Create a new Calculated Field, name it **CORR Fare:Passenger**

   CORR([Passenger Count (PAX)],[Unrestricted Fare (YCA)])
3. Drag **Origin City Name** and **Origin State** to the Rows shelf
4. Drag **CORR Fare:Passenger** to the Text marks card, then drag it to the filter (select Special and Non-null values) and Show Filter

Question – What does the data tell us?

1. Create a new sheet, call if **Correlation Viz**
2. Drag **Passenger Count** to the Columns shelf
3. Drag **Unrestricted Fare** to the Rows shelf
4. Select Analysis from the menu and uncheck **Aggregate Measures**
5. Drag **City Pair** to the Labels marks card
6. Bring in the filters from our first sheet
7. Review the data by selecting a positive correlation coefficient and a negative correlation coefficient

End of Part 1

- Any questions?
- Anything you'd like covered for next time?

Linear Regression

(Advanced Analytics)

- Identify relationship between an independent and dependent variable
- Build upon correlation analysis
- Tableau enables us to build a regression model using either linear, logarithmic, exponential, power, or polynomial types
- We can leverage R^2 to validate our model (recall that we generated our correlation coefficient, r, in an earlier analysis)
- Our goal is to achieve a trend line that generalizes well, that is, we don't want our trend line to underfit or overfit the data
- Since we are not performing any cross-validation (i.e., training our model with a sample), we can generalize that we don't want an R^2 that is close to 0 (underfit) or 1 (overfit)
- Additionally, we want a p-value < 0.05 for statistical significance of rejecting our null hypothesis (i.e., there is no affect of our independent variables to our dependent)

❑ Dataset(s): FY18 - CPP Award Data with Distance

❑ Question: Does the route distance correlate well with unrestricted fares?

Steps:

1. Open Tableau and import the new dataset (FY18 - CPPAwardData with Distance)
2. Confirm in the Data Source preview pane that you have a calculated field called Distance (if it's not called that, simply rename it)
3. Create a new sheet, call it **Correlation Analysis**
4. Drag **Awarded Serv** to the Filters shelf, check N (for non-stop flights only); select Apply to Worksheets -> All Using this Data Source
5. Create a new calculated field call **CORR Distance:Fare**

   CORR([Distance],[Unrestricted Fare (YCA)])
6. Drop **CORR Distance:Fare** into the Viz (or Text marks card)

Question *– What's the value? What does the data tell us?*

1. Create a new sheet, call if **Regression Analysis**
2. Drag **Distance** to the columns shelf
3. Drag **Unrestricted Fare (YCA)** to the rows shelf
4. Switch to the Analytics Pane
5. Drag Trend Line to the Viz (select Linear)
6. Hover over our Trend Line and review the model formula, the R^2 value, and p-value

Question *– Does this model reject the null hypothesis? What else does the data tell us?*

K-Means

(Advanced Analytics)

- Cluster analysis partitions marks in the view into clusters, where the marks within each cluster are more similar to one another than they are to marks in other clusters

- Tableau uses the k-means algorithm for clustering (vs. other methods such as hierarchal, density, etc.)

- The general steps for k-means are:
  - Randomly select cluster centers
  - Assign each instance to the nearest center
  - Recalculate the new cluster centers
  - Reassign each instance to the new closest cluster center
  - The process stops either when no instances are reassigned to a different cluster or when the specified number of maximum iterations is reached.

❑ Dataset(s): FY18 - CPP Award Data with Distance

❑ Question: Can we segment routes based on distance?

Steps:

1. Create a new sheet, call it **Cohort Analysis**
2. Drag **City** Pair to the Rows shelf
3. Drag **Distance** to the Text marks card
4. Switch to the Analytics Pane and drag **Cluster** to the viz
5. Sort the data descending (by Distance)
6. Drag **Cluster** from the Color marks card into the Data Pane
7. Click on the **Cluster** in the Color marks card and select Describe Clusters

Question *– How many clusters did Tableau calculate for this analysis? Any idea why?*

Linear Regression (Advanced Analytics)

- Identify relationship between an independent and dependent variable
- Build upon correlation analysis
- Tableau enables us to build a regression model using either linear, logarithmic, exponential, power, or polynomial types
- We can leverage R^2 to validate our model (recall that we generated our correlation coefficient, r, in an earlier analysis)
- Our goal is to achieve a trend line that generalizes well, that is, we don't want our trend line to underfit or overfit the data
- Since we are not performing any cross-validation (i.e., training our model with a sample), we can generalize that we don't want an R^2 that is close to 0 (underfit) or 1 (overfit)
- Additionally, we want a p-value < 0.05 for statistical significance of rejecting our our null hypothesis (i.e., there is no affect of our independent variables to our dependent)

❑ Dataset(s): FY18 - CPP Award Data with Distance

❑ ● Question: Does creating cohort groups by distance improve our linear model?

Steps:

1. Create a new sheet, call it **Regression Cohort Analysis**
2. Drag **Distance** to the columns shelf
3. Drag **Unrestricted Fare (YCA)** to the rows shelf
4. Switch to the Analytics Pane
5. Drag Trend Line to the Viz (select Linear)
6. Drag **City Pair (cluster)** to the Filter shelf, select None then Show Filter
7. Hover over our Trend Line and review the model formula, the R^2 value, and p-value
8. Review the model results for each Cluster

Question *– Which Cluster has the best generalized model? Why?*

Outside Tableau Resources*

Official Tableau Resources

1. Online Help: http://onlinehelp.tableau.com/v9.0/pro/online/windows/en-us/help.htm
2. Quick Start Guide: http://onlinehelp.tableau.com/v9.0/pro/online/windows/en-us/help.htm
3. Training, videos, webinars, whitepapers, events: http://www.tableau.com/learn
4. Tableau User Groups -- these exist across the country. For example, here is the one for Washington, DC: http://community.tableau.com/groups/washington-dc
5. Tableau Public Gallery, lots of interesting-looking examples: https://public.tableau.com/s/gallery
6. **Whitepaper "Designing Efficient Workbooks." http://www.tableau.com/learn/whitepapers/designing-efficient-workbooks**

Unofficial Tableau-related Sites

1. http://www.dataplusscience.com/TableauReferenceGuide/
2. http://vizpainter.com/
3. http://drawingwithnumbers.artisart.org/
4. https://3danim8.wordpress.com/

GSA®