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!

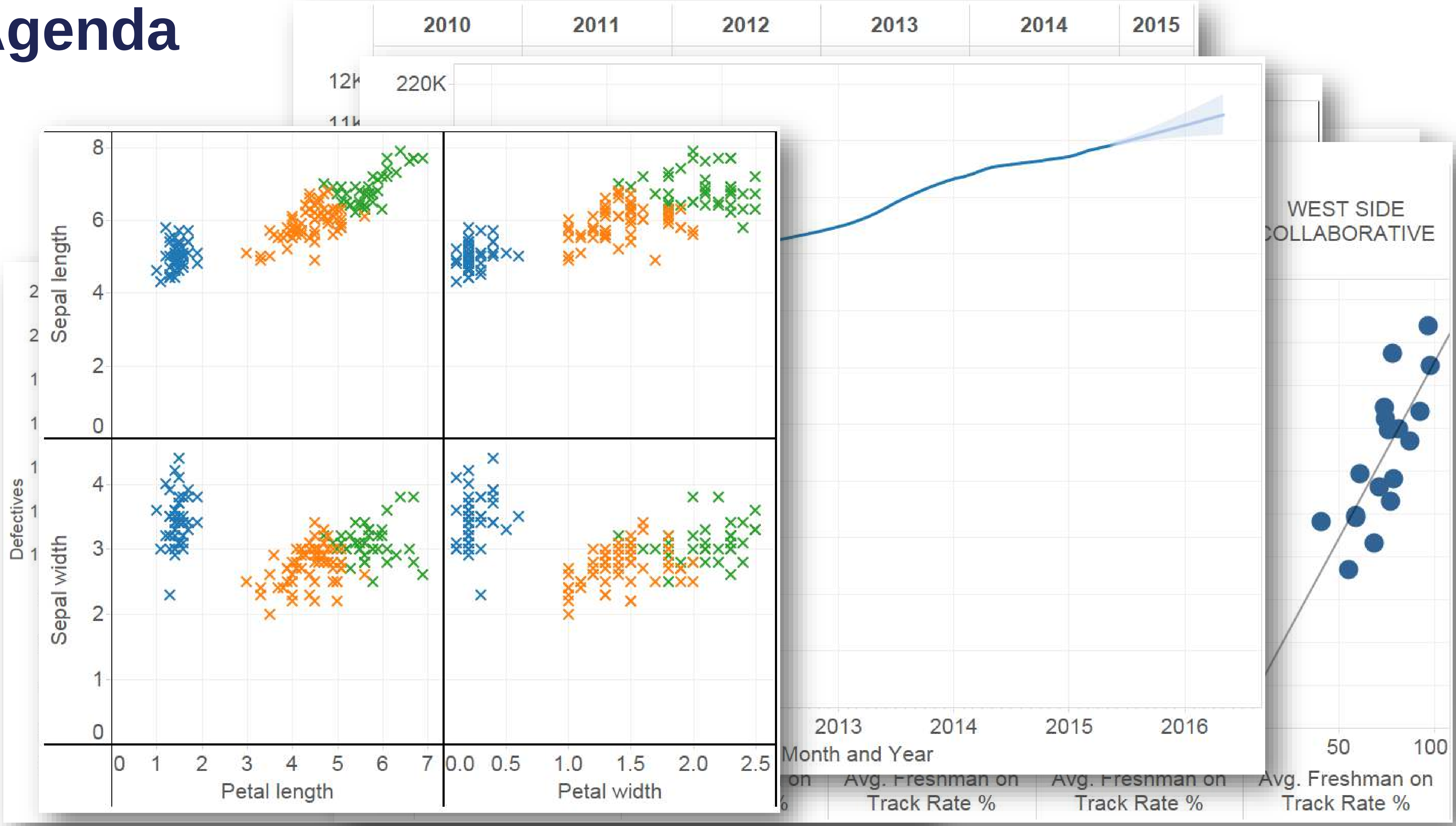
- Side hustles as adjunct professor and data science bootcamp instructor.
- Sky-dived (tandem) and also zip-lined once because YOLO!

NOVA DATA SCHOOL

Audience

- Basic knowledge of statistics
- Interested in Tableau's statistical capabilities
  - ✓ Distribution
  - ✓ Summary
  - ✓ Modeling

Agenda



2010
2011
2012
2013
2014
2015
220K
2013
2014
2015
2016
Month and Year
WEST SIDE
COLLABORATIVE
50
100
Avg. Freshman on
Track Rate %
Avg. Freshman on
Track Rate %
Avg. Freshman on
Track Rate %
Defectives
Sepal length
Sepal width
Petal length
Petal width

Why Visual Analysis?

Anscombe's Quartet

Let's analyze some data ...

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

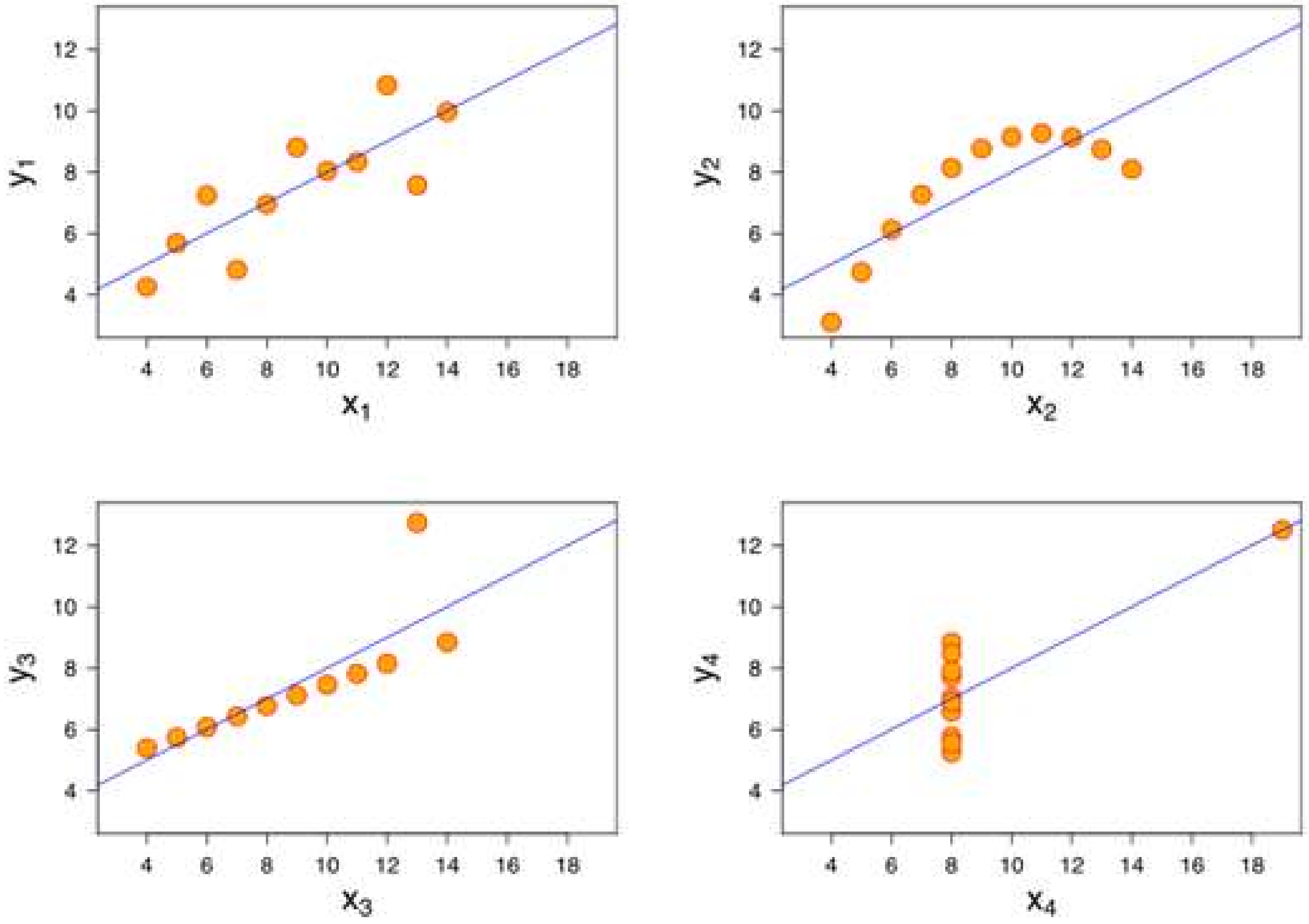
Anscombe's Quartet

Let's summarize the data ...

Property	Value
Mean of x in each case	9 (exact)
Variance of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	y = 3.00 + 0.500x (to 2 and 3 decimal places, respectively)

Anscombe's Quartet

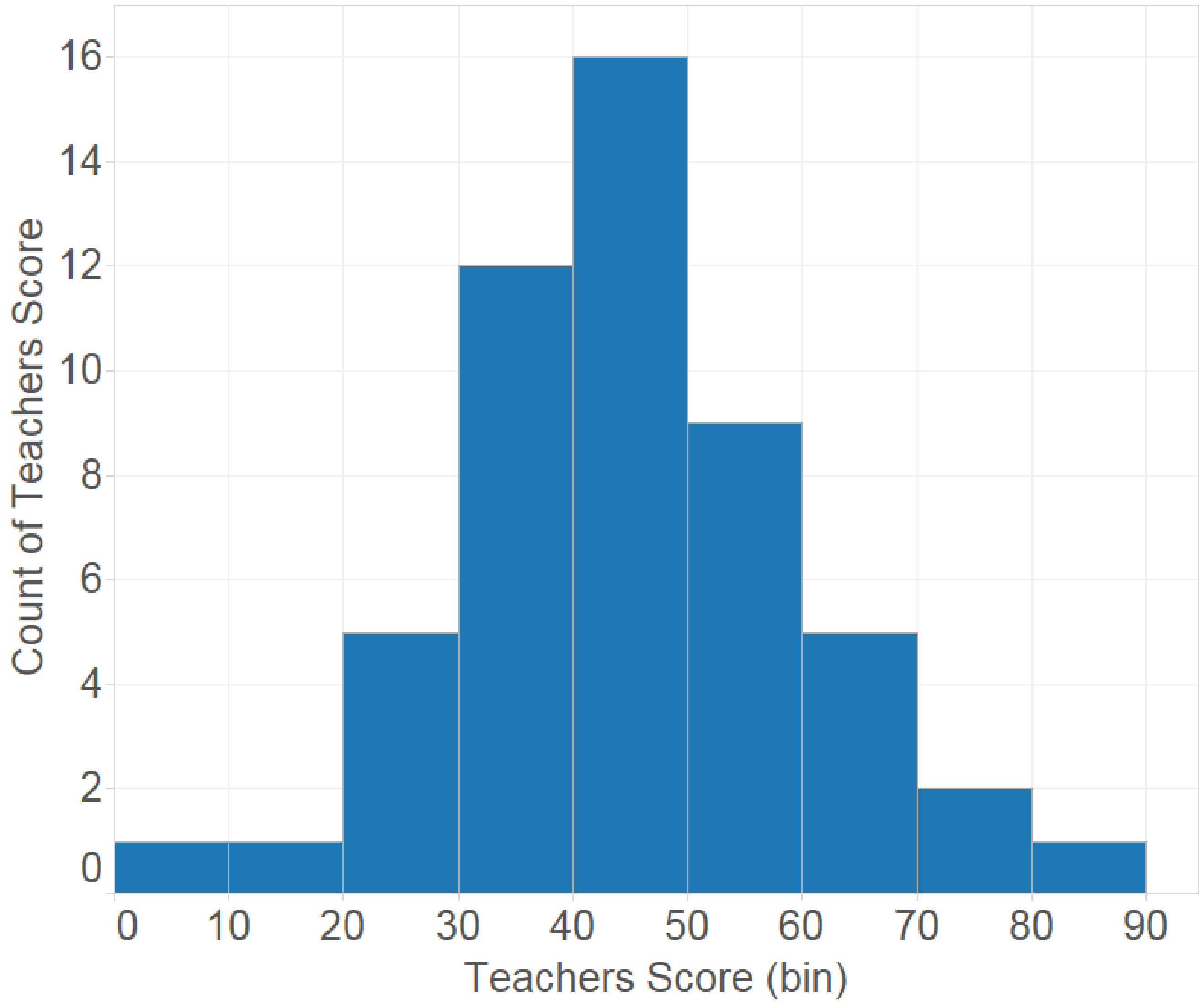
Let's visualize the data ...



Distribution

Histograms

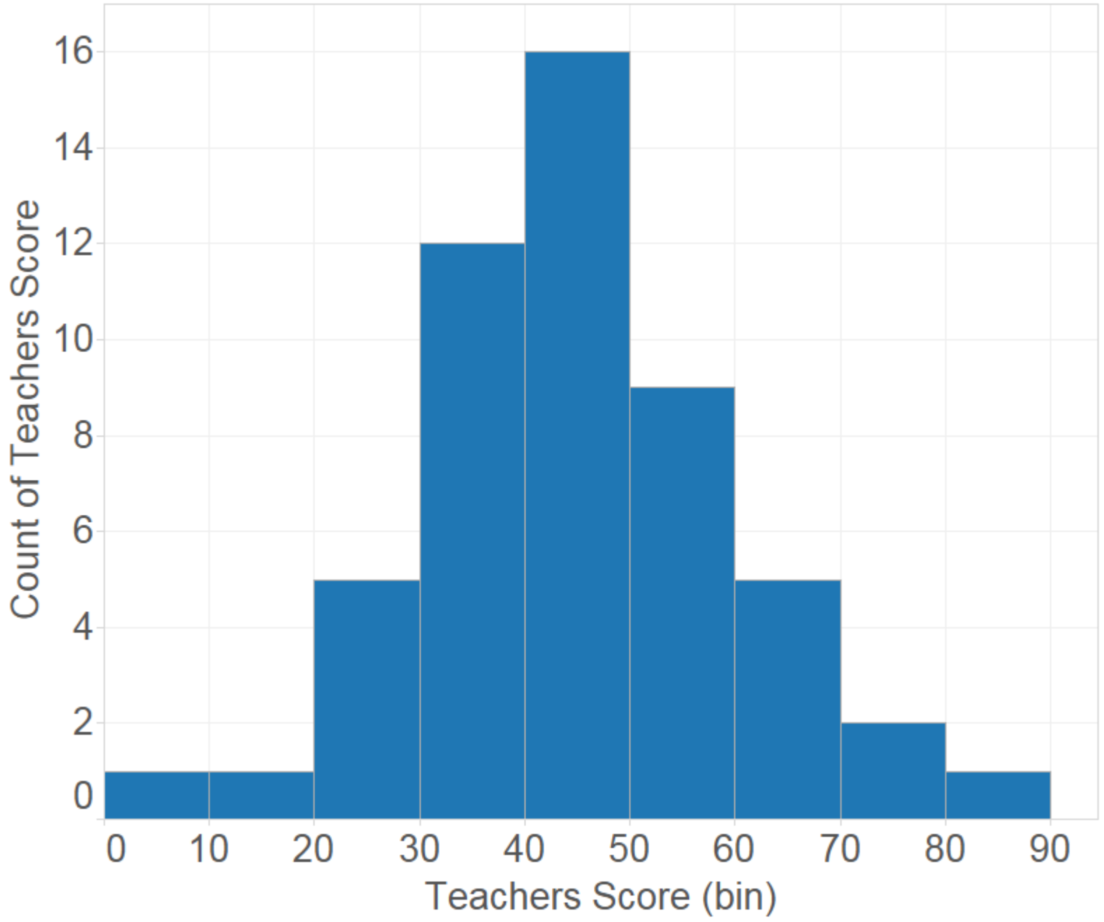
Histograms



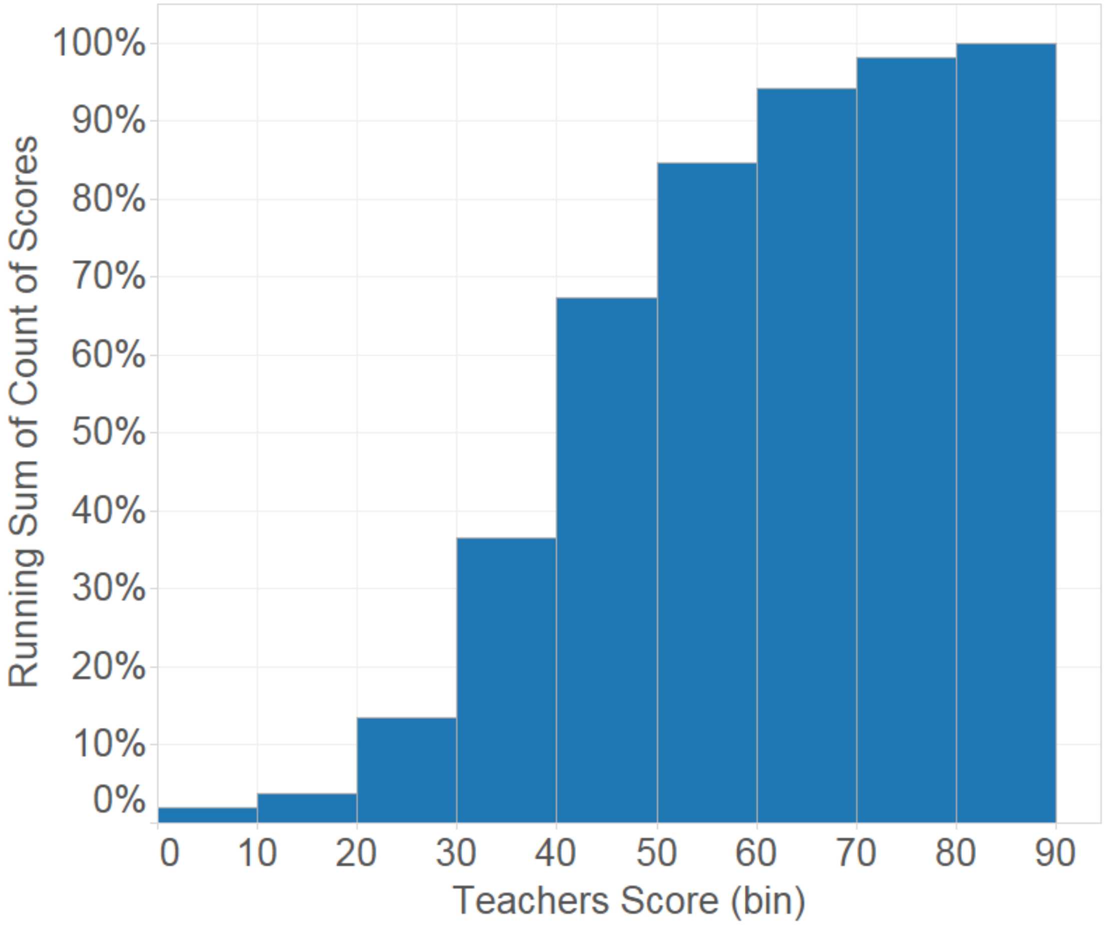
Histograms show us the distribution of numerical data

Histograms

Basic Histogram

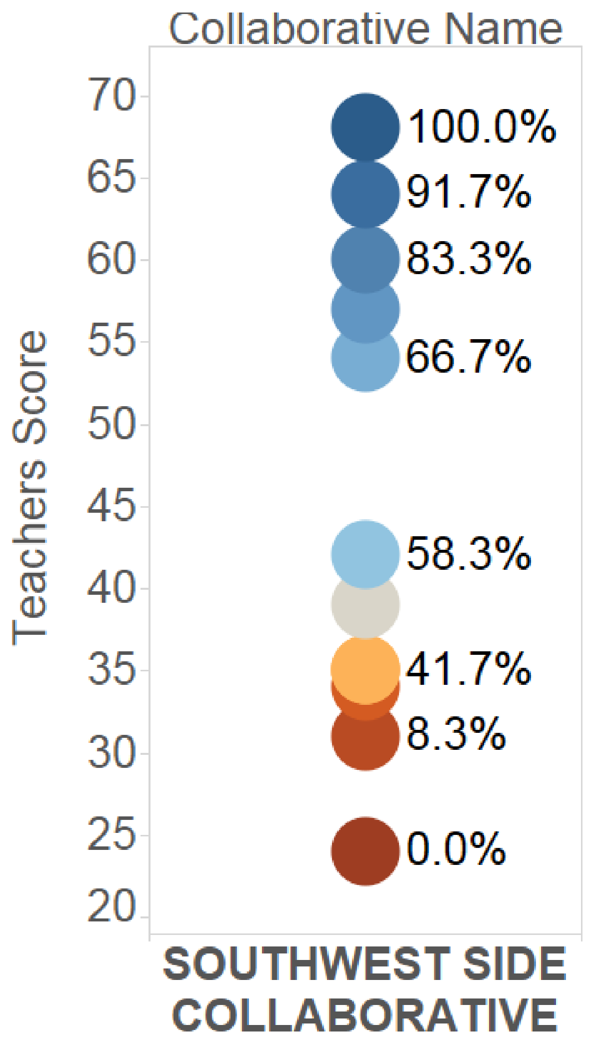


Cumulative Histogram



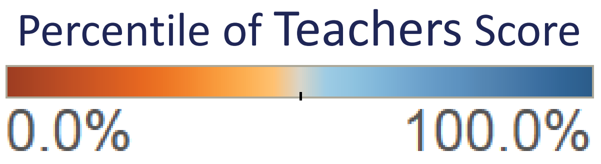
Percentiles

Percentiles



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

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



Box Plots

Box Plots

Traditional Box Plot

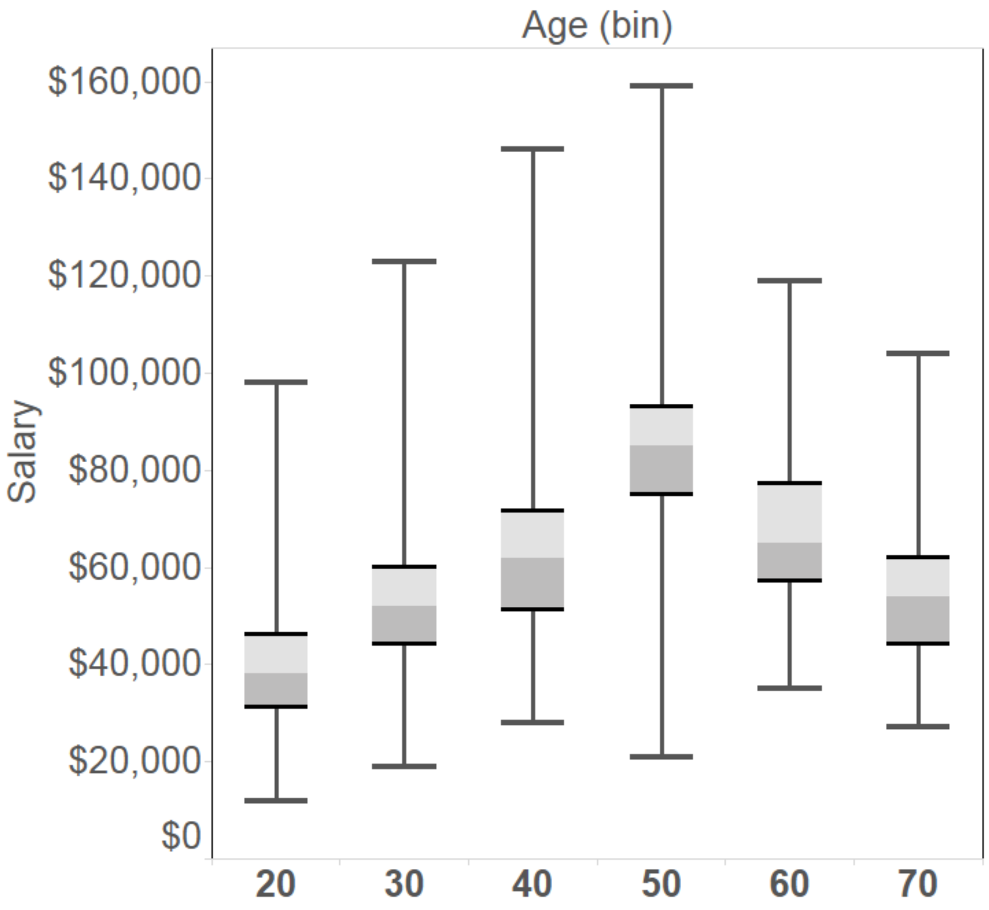
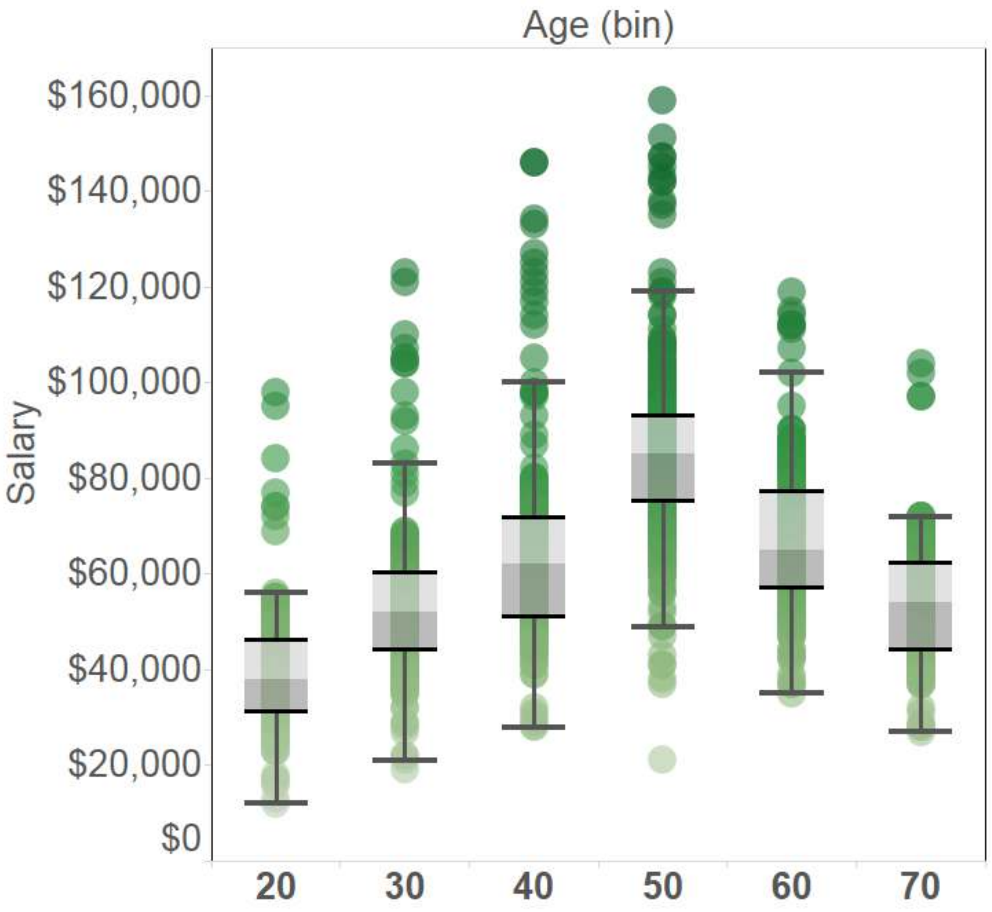
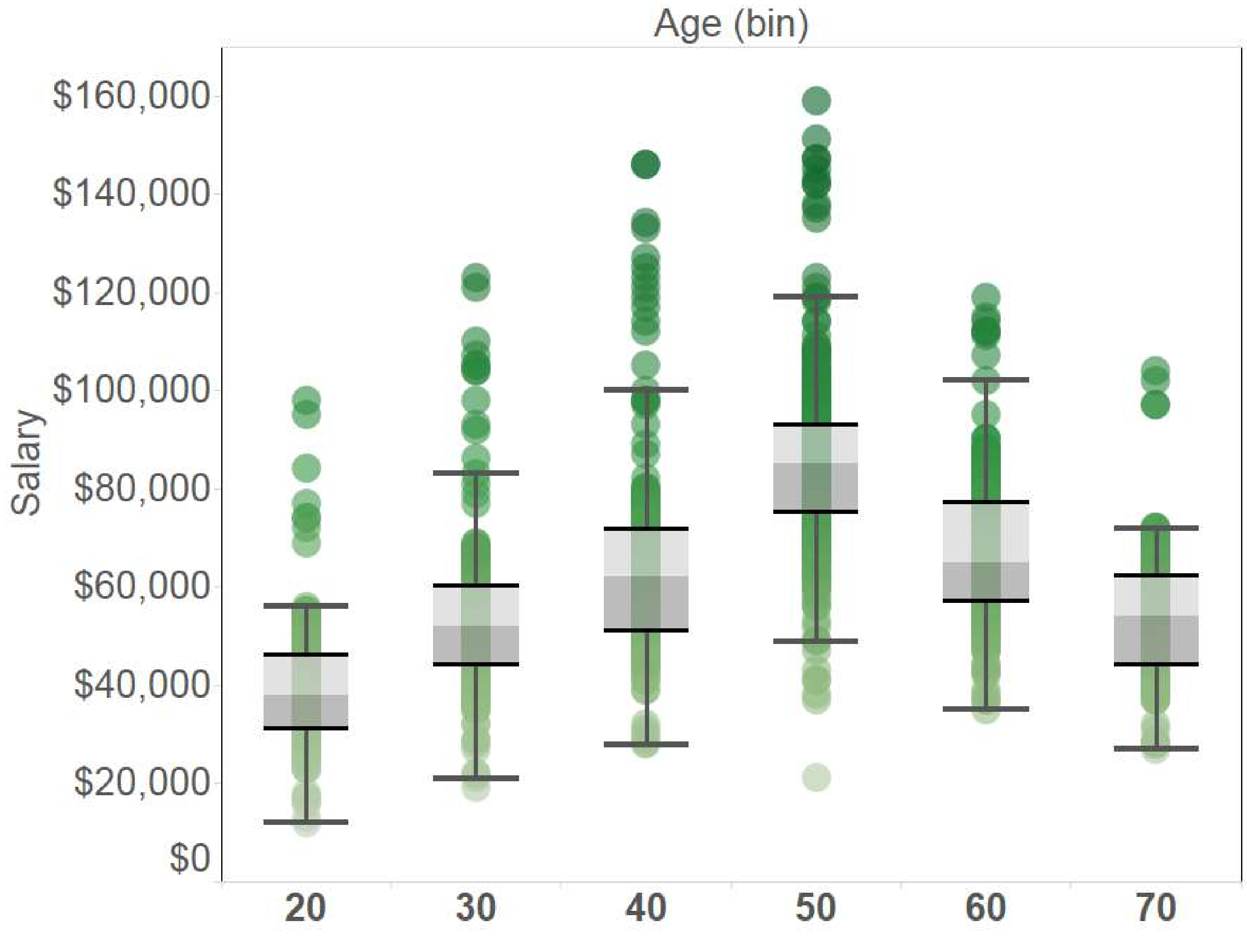


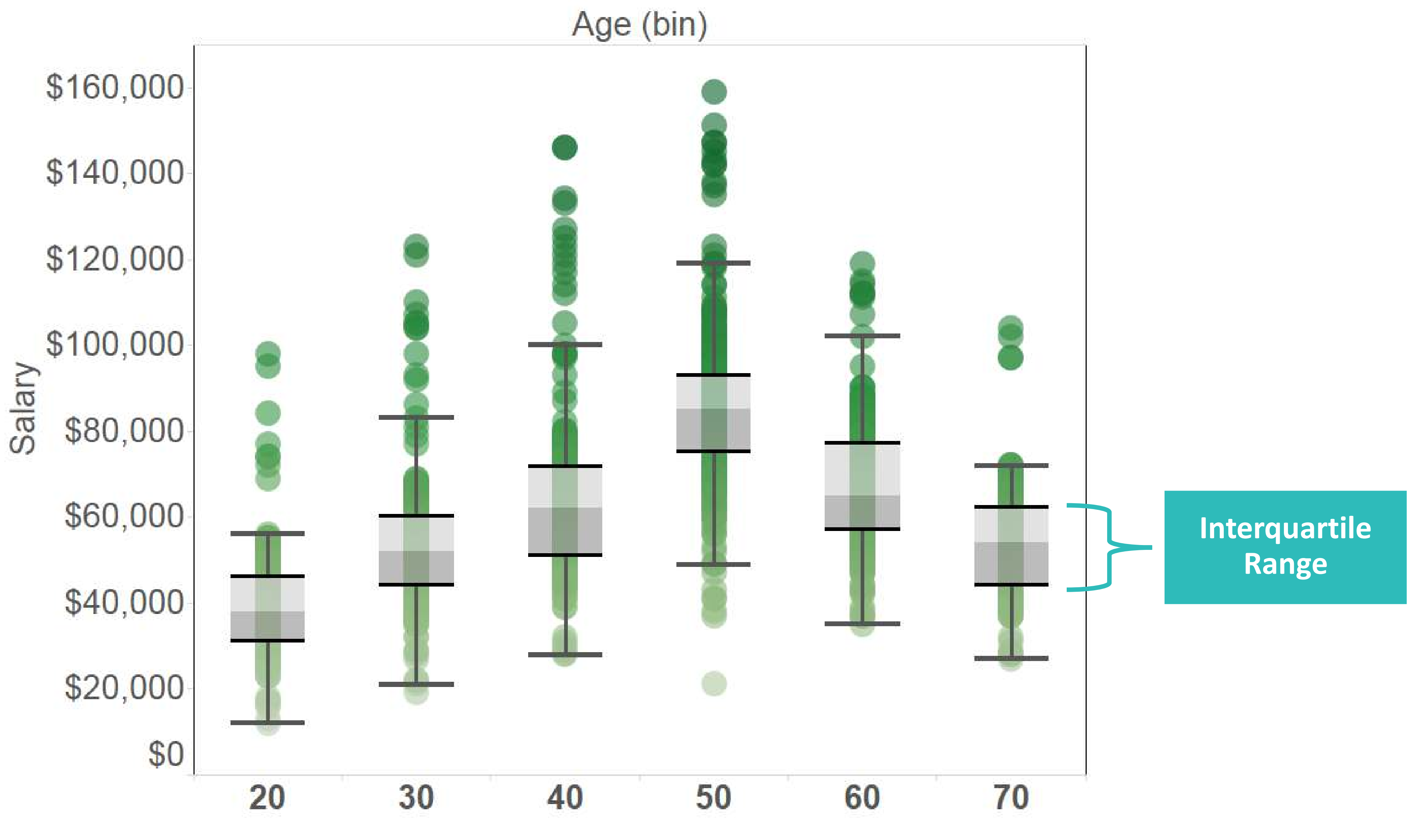
Tableau Box Plot



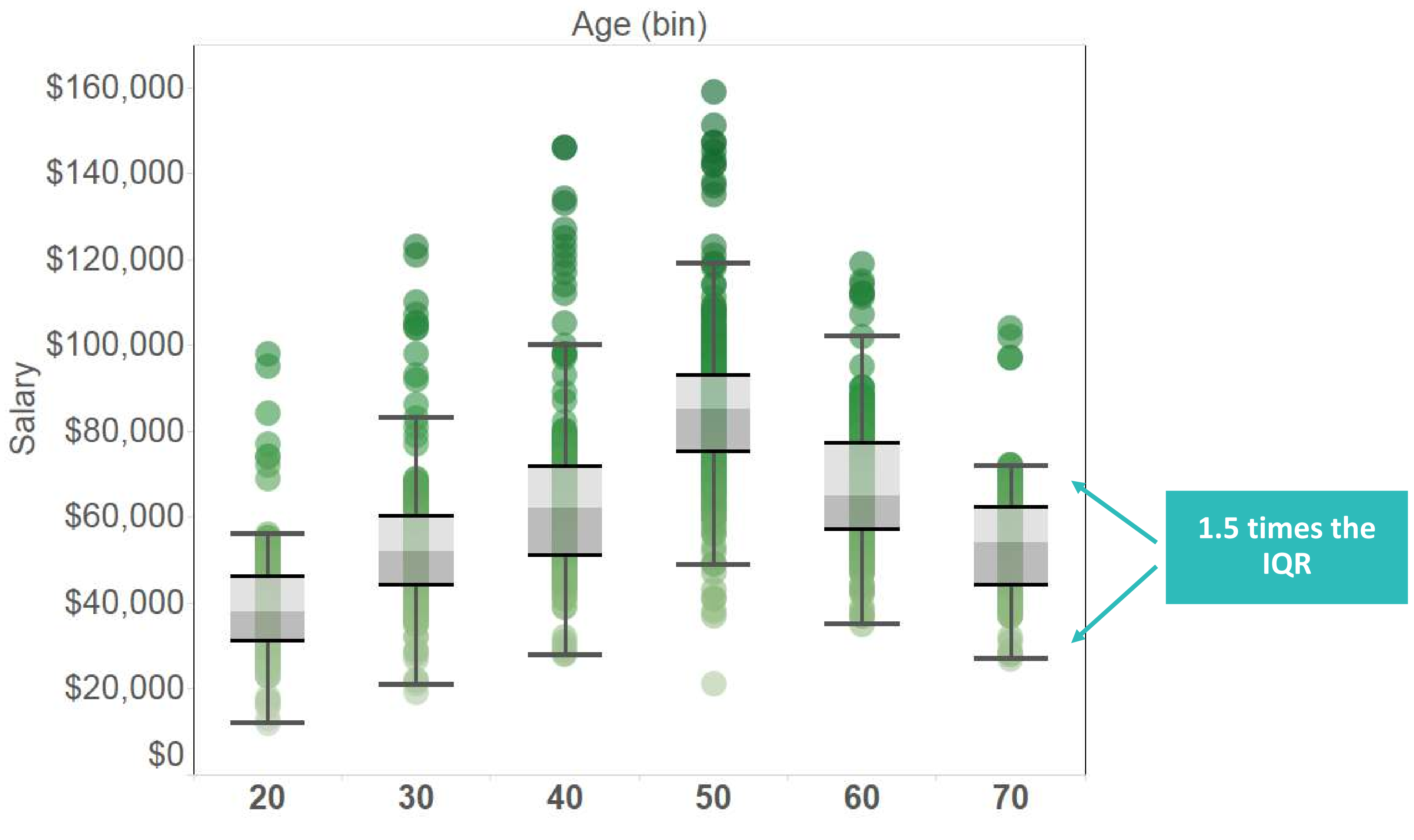
Anatomy of the Tableau Box Plot



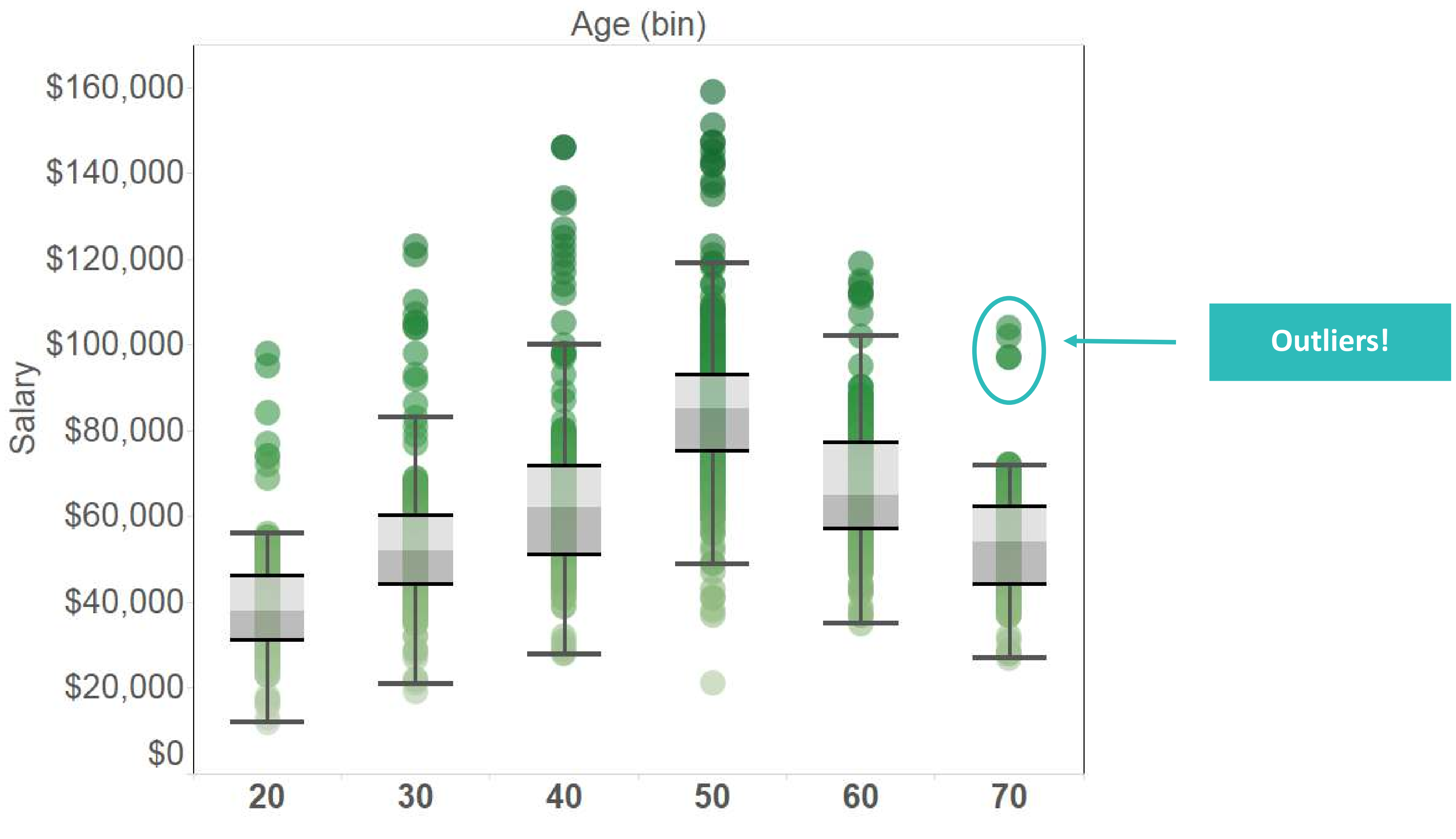
Anatomy of the Tableau Box Plot



Anatomy of the Tableau Box Plot



Anatomy of the Tableau Box Plot



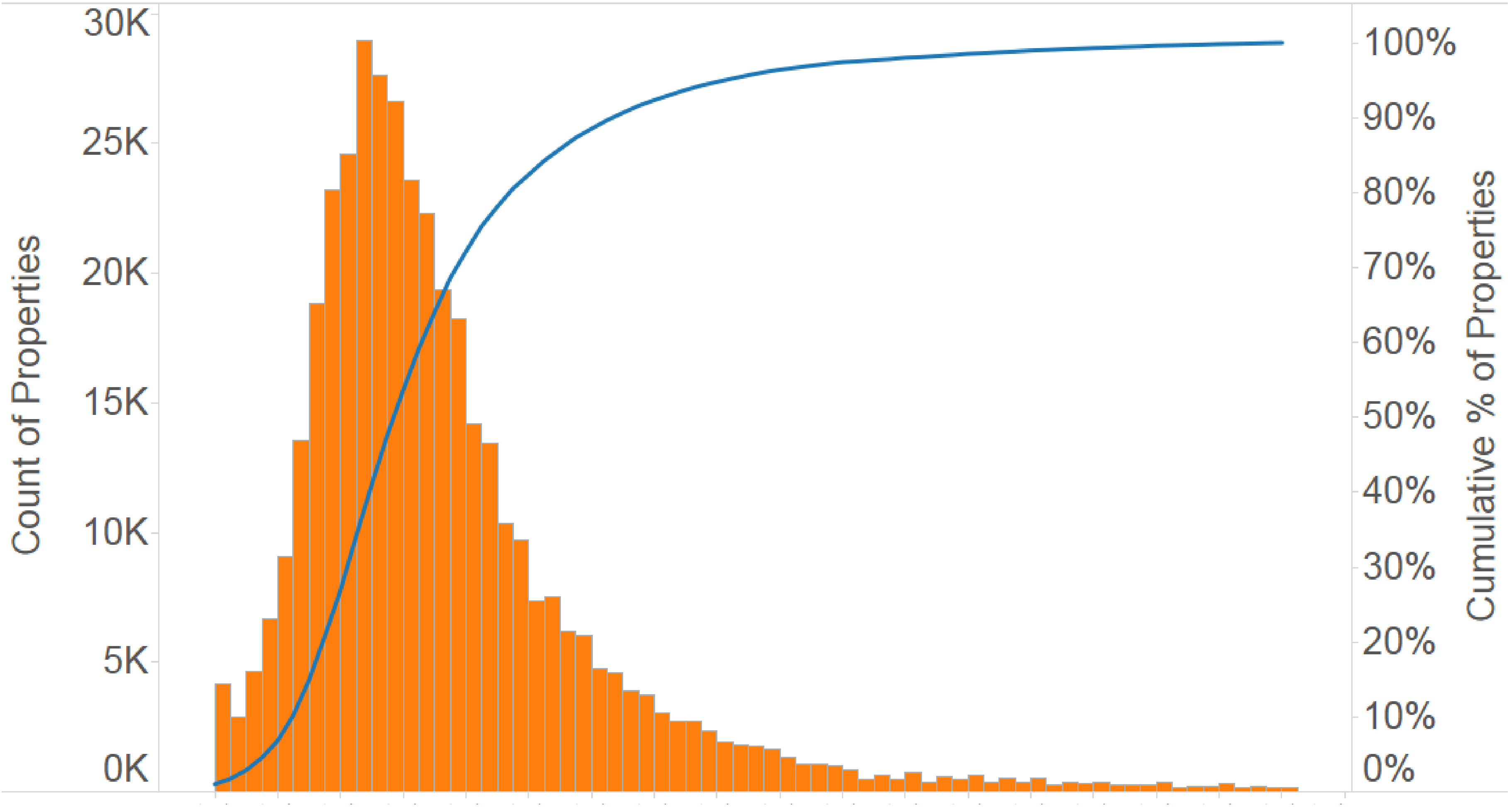
Summary Statistics?

Summary Card

Summary

Count:	108
SUM(Sales)	
Sum:	$609,206
Average:	$5,641
Minimum:	$259
Maximum:	$22,171
Median:	$4,011
Standard deviation:	$4,824
First quartile:	$2,180
Third quartile:	$7,647
Skewness:	1.51
Excess Kurtosis:	2.17

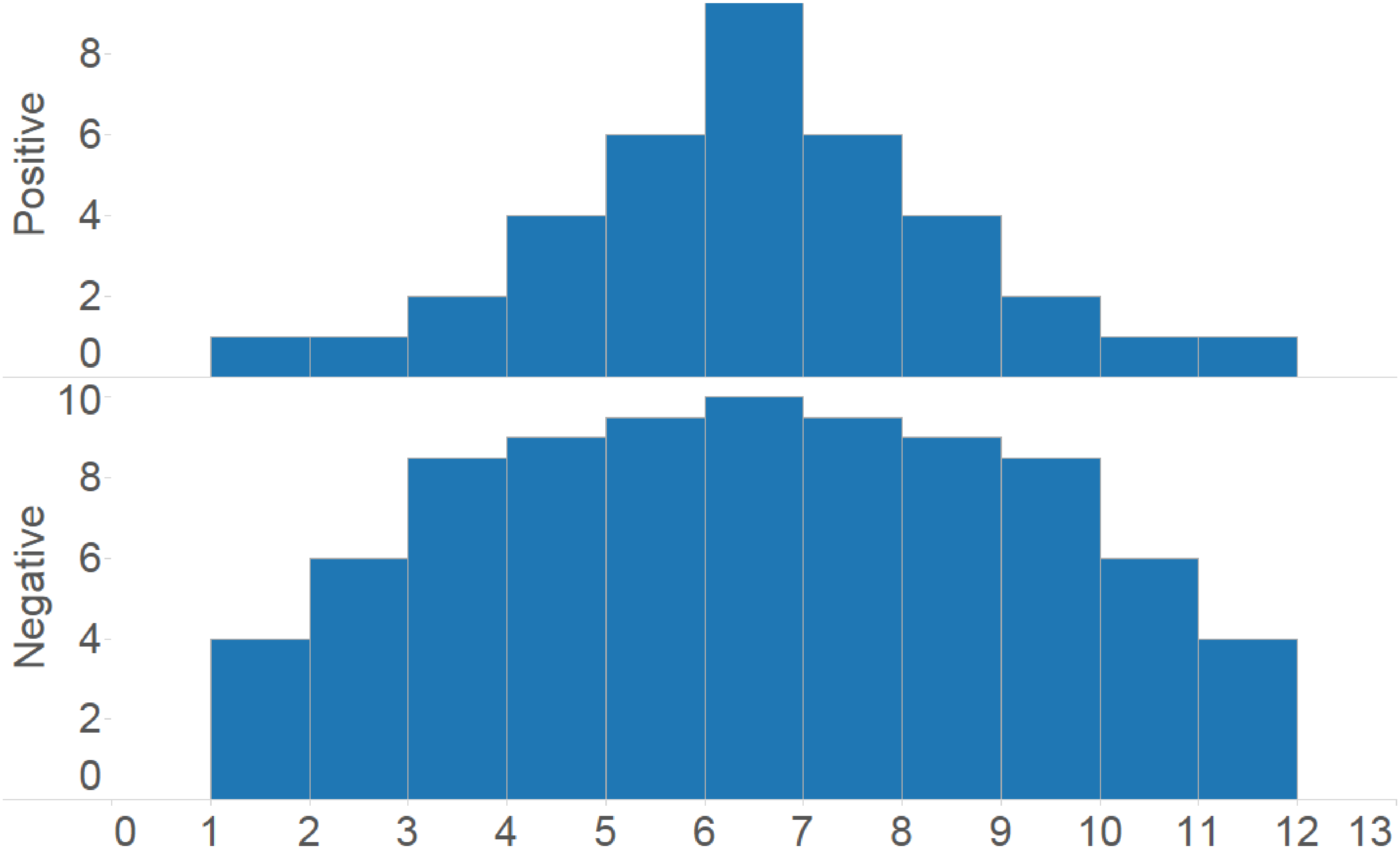
Summary Card - Skewness



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

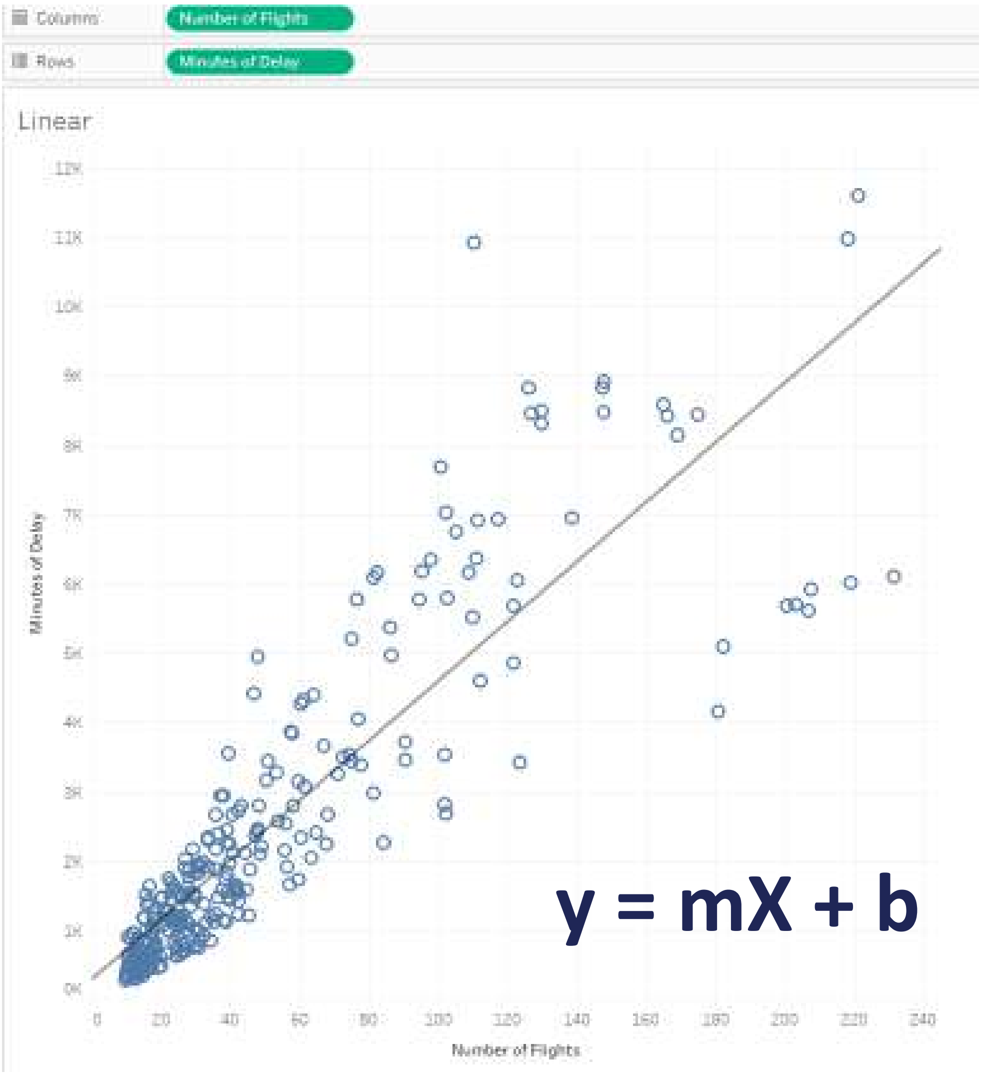
Supervised Classification

Similar with respect to a target

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

Trend Lines / Regression Lines

Trend Lines

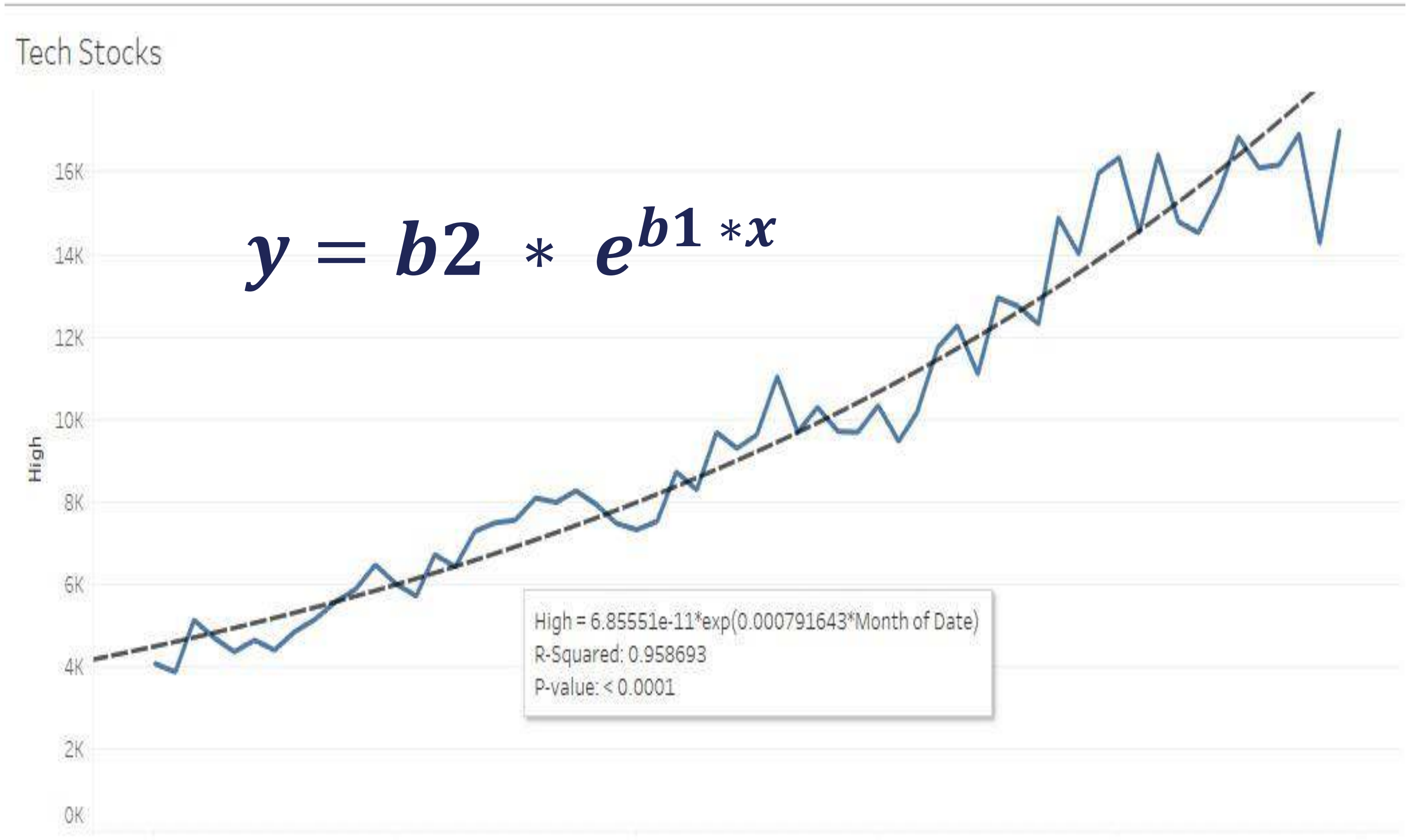


$$y = mX + b$$

Options

- **Linear**
- Exponential
- Logarithmic
- Polynomial
- Power

Exponential

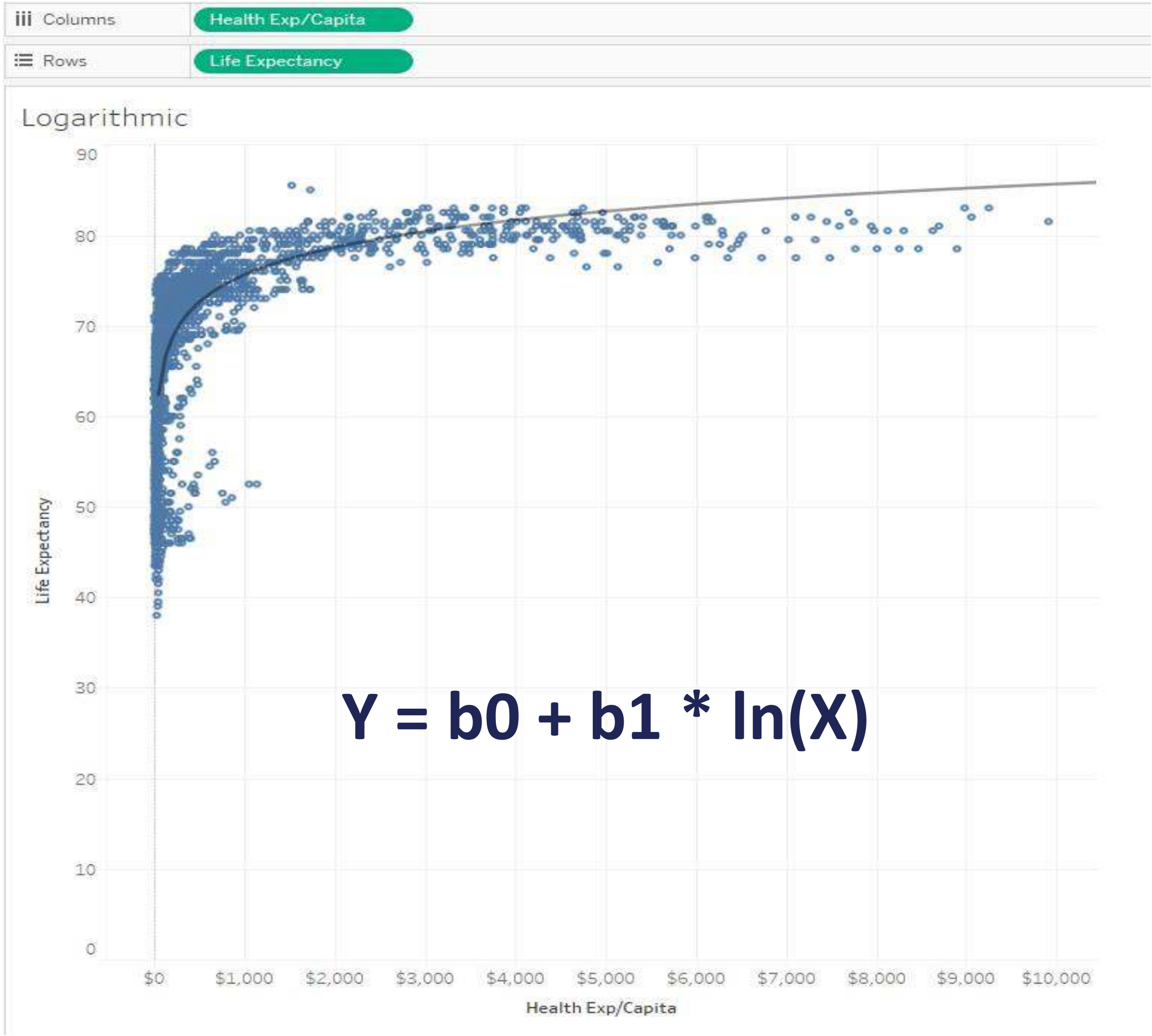


$$y = b2 \ * \ e^{b1 \, * x}$$

Options

- Linear
- **Exponential**
- Logarithmic
- Polynomial
- Power

Logarithmic

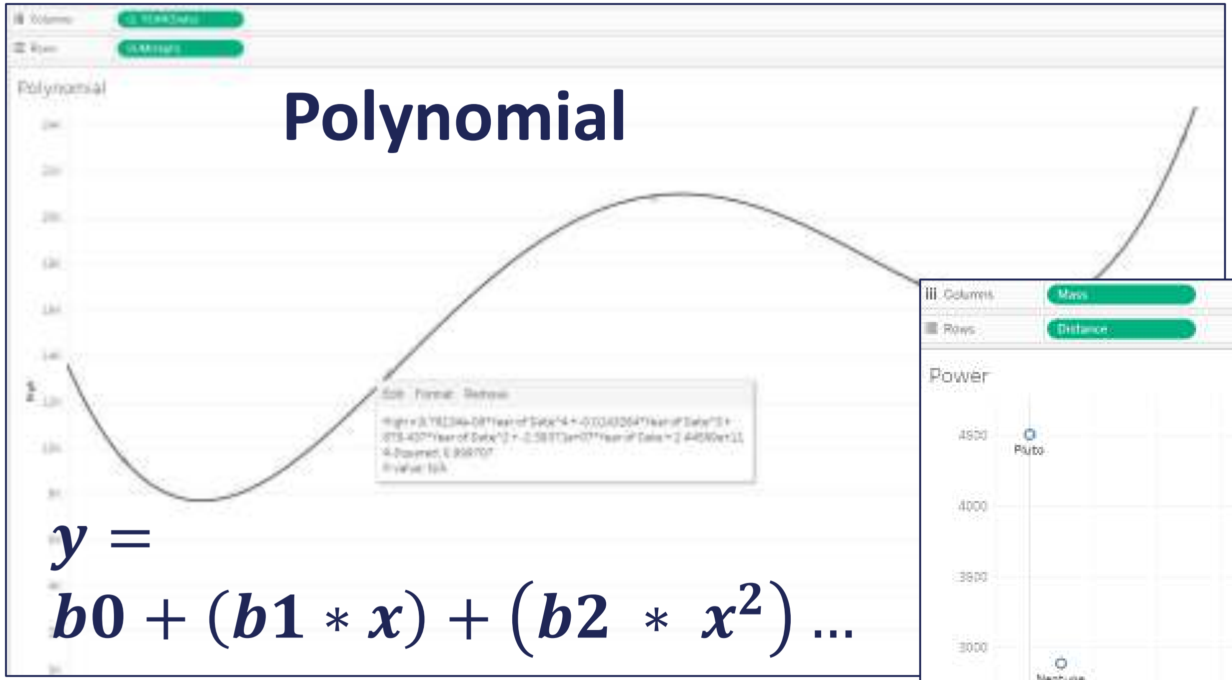


$$Y = b0 + b1 * \ln(X)$$

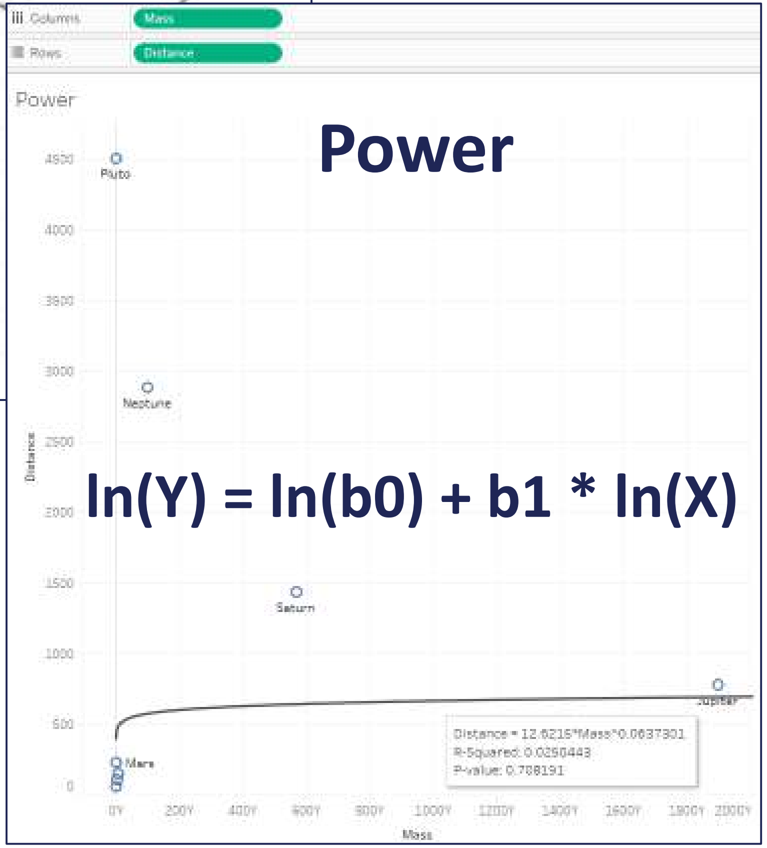
Options

- Linear
- Exponential
- **Logarithmic**
- Polynomial
- Power

Polynomial and Power



$$y = b0 + (b1 * x) + \left(b2\ *\ x^2\right) \ldots$$

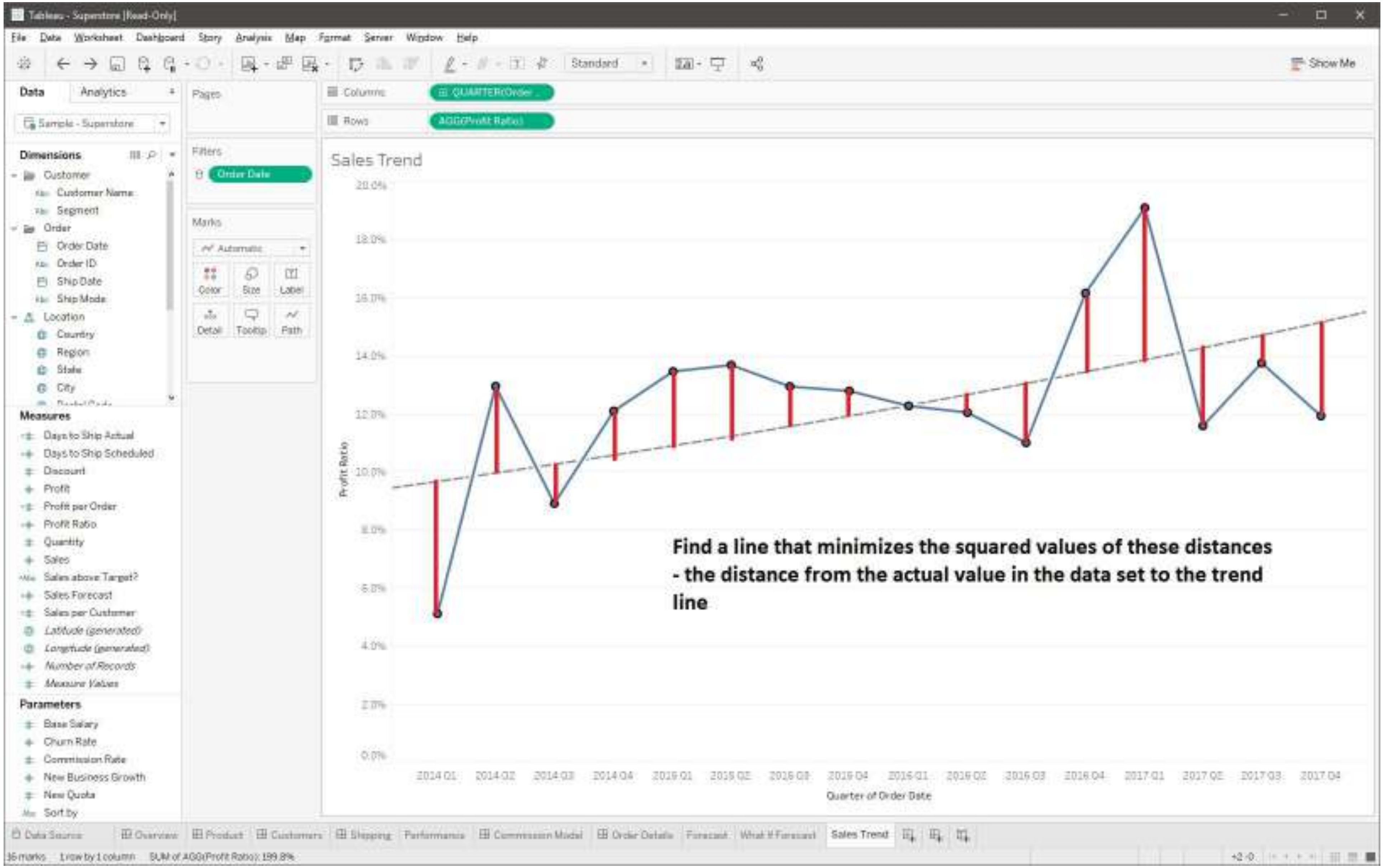


ln(Y) = ln(b0) + b1 * ln(X)

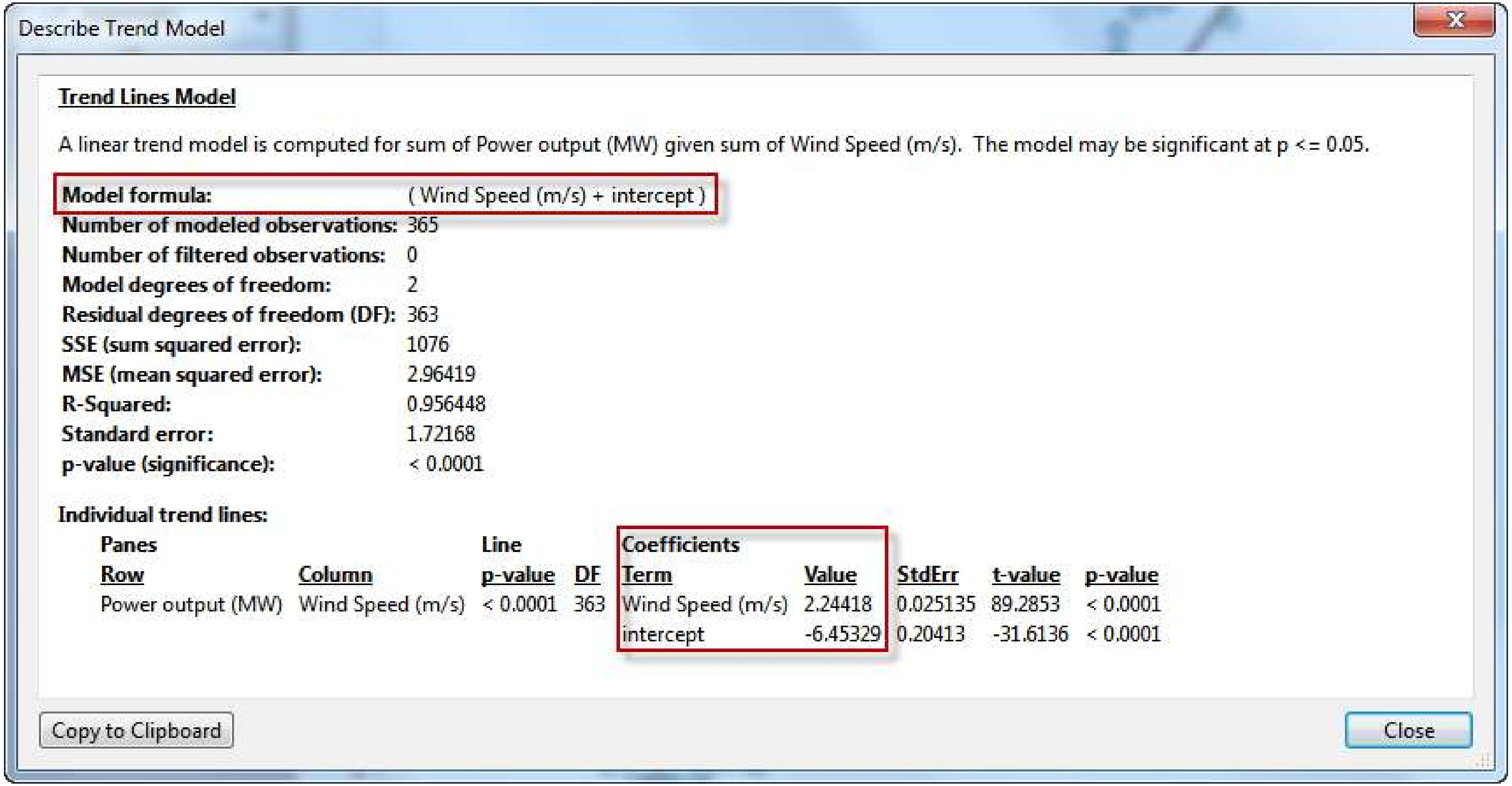
Options

- Linear
- Exponential
- Logarithmic
- **Polynomial**
- **Power**

Trend Lines (Overview)



Trend Models: Describing the Formula



Describe Trend Model

Trend Lines Model

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:	(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):	1076
MSE (mean squared error):	2.96419
R-Squared:	0.956448
Standard error:	1.72168
p-value (significance):	< 0.0001

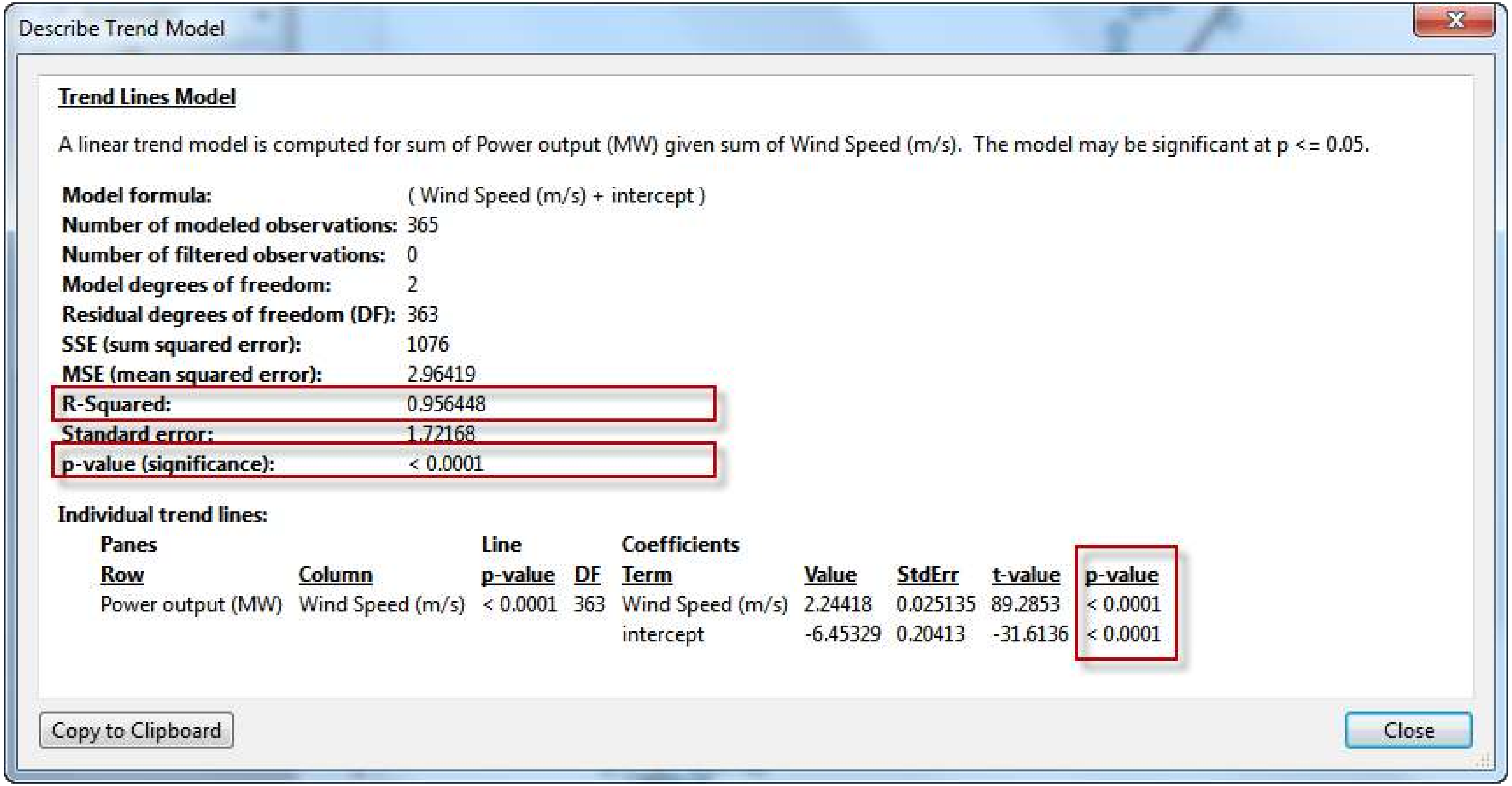
Individual trend lines:

Panes		Line		Coefficients				
Row	**Column**	**p-value**	**DF**	**Term**	**Value**	**StdErr**	**t-value**	**p-value**
Power output (MW)	Wind Speed (m/s)	< 0.0001	363	Wind Speed (m/s)	2.24418	0.025135	89.2853	< 0.0001
				intercept	-6.45329	0.20413	-31.6136	< 0.0001

Copy to Clipboard

Close

Trend Models: Evaluating Model Fit



Describe Trend Model

Trend Lines Model

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:	(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):	1076
MSE (mean squared error):	2.96419
R-Squared:	0.956448
Standard error:	1.72168
p-value (significance):	< 0.0001

Individual trend lines:

Panes		Line		Coefficients				
Row	Column	p-value	DF	Term	Value	StdErr	t-value	p-value
Power output (MW)	Wind Speed (m/s)	< 0.0001	363	Wind Speed (m/s)	2.24418	0.025135	89.2853	< 0.0001
				intercept	-6.45329	0.20413	-31.6136	< 0.0001

Copy to Clipboard

Close

Trend Lines

Describe Trend Model

Trend Lines Model

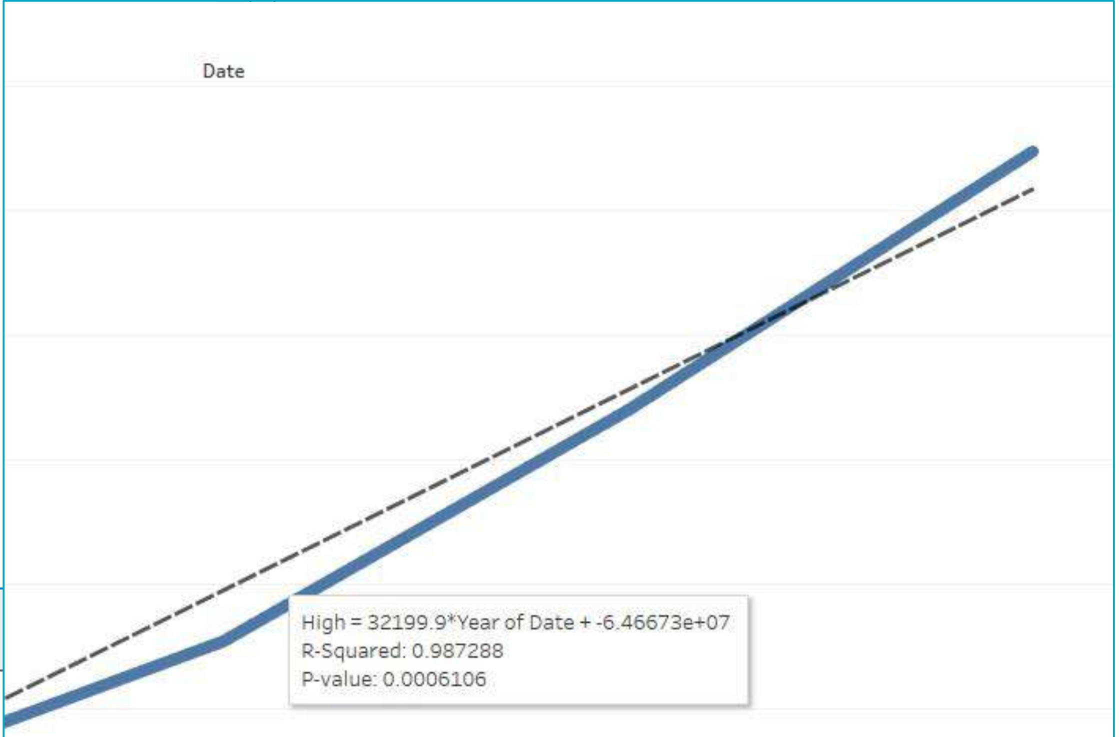
A linear trend model is computed for sum of High given Date Year. The model may be significant at p <= 0.05.

Model formula:	(Year of Date + intercept)
Number of modeled observations:	5
Number of filtered observations:	0
Model degrees of freedom:	2
Residual degrees of freedom (DF):	3
SSE (sum squared error):	1.33503e+08
MSE (mean squared error):	[illegible]
R-Squared:	0.987288
Standard error:	[illegible]
p-value (significance):	0.0006106

Individual trend lines:

Panes		Line		Coefficients				
Row	Column	p-value	DF	Term	Value	StdErr	t-value	p-value
High	Year of Date	0.0006106	3	Year of Date	32199.9	2109.53	15.264	0.0006106
				intercept	-6.46673e+07	4.24437e+06	-15.236	0.000614

Copy



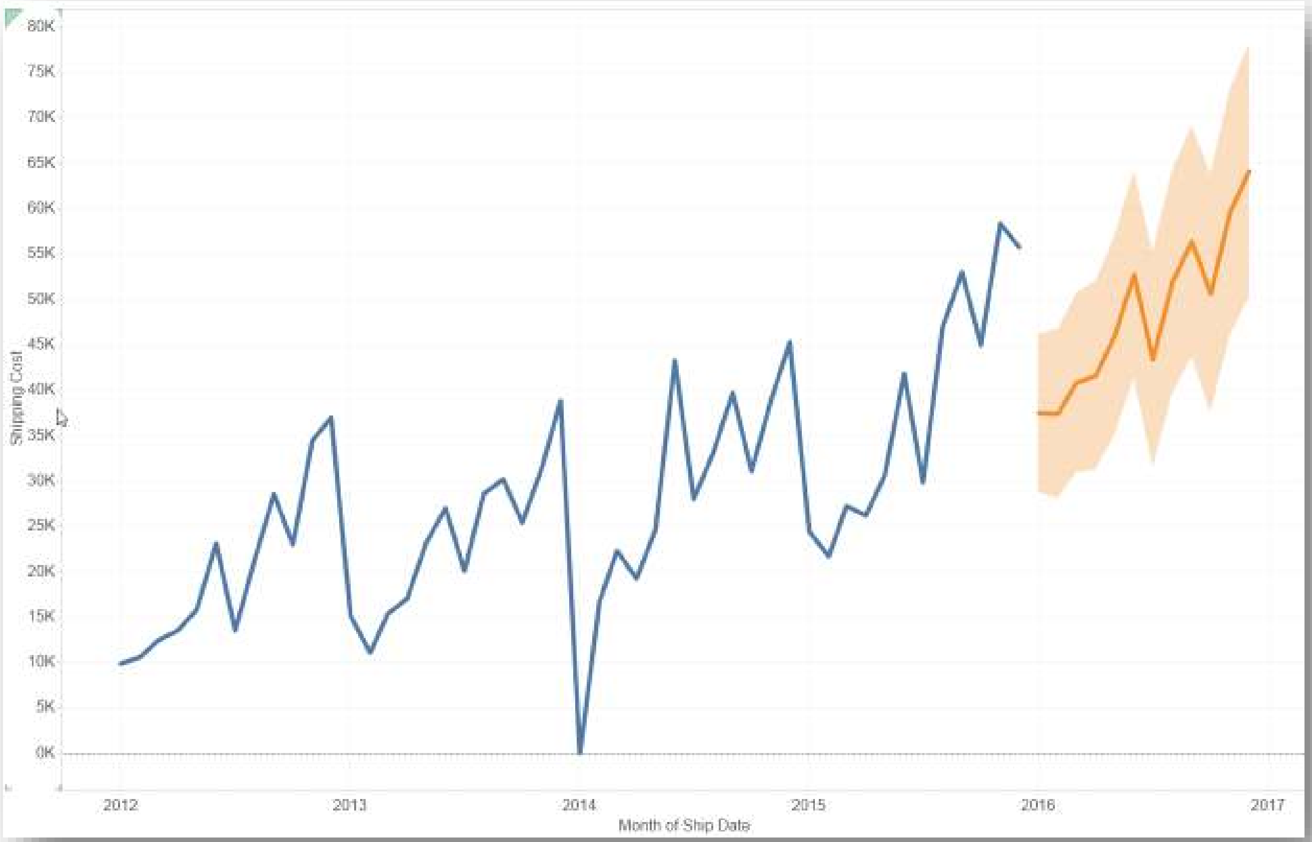
The R-Squared value shows the ratio of variance in the data, as explained by the model, to the total variance in the data. The P-value reports the probability that the equation of the line was a result of random chance. The smaller the p-value, the more significant the model is. A p-value of 0.05 or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

Forecasting

Forecast Requirements

- **At least:**
  - **One Dimension**
  - **One Measure**

- **Dimension Requirements**
  - **Date Field**
    or
  - **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

Seasonality

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

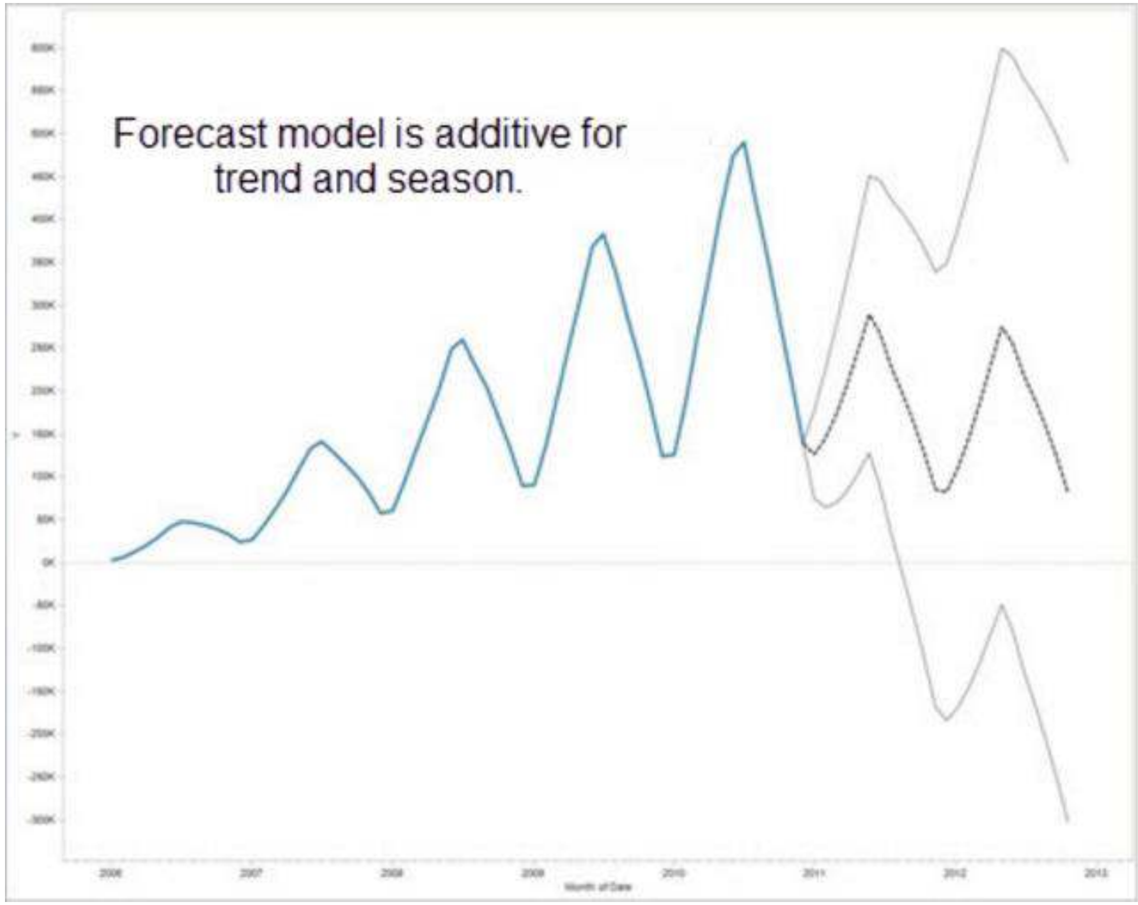
Trend
Tendency in the data to increase or decrease over time

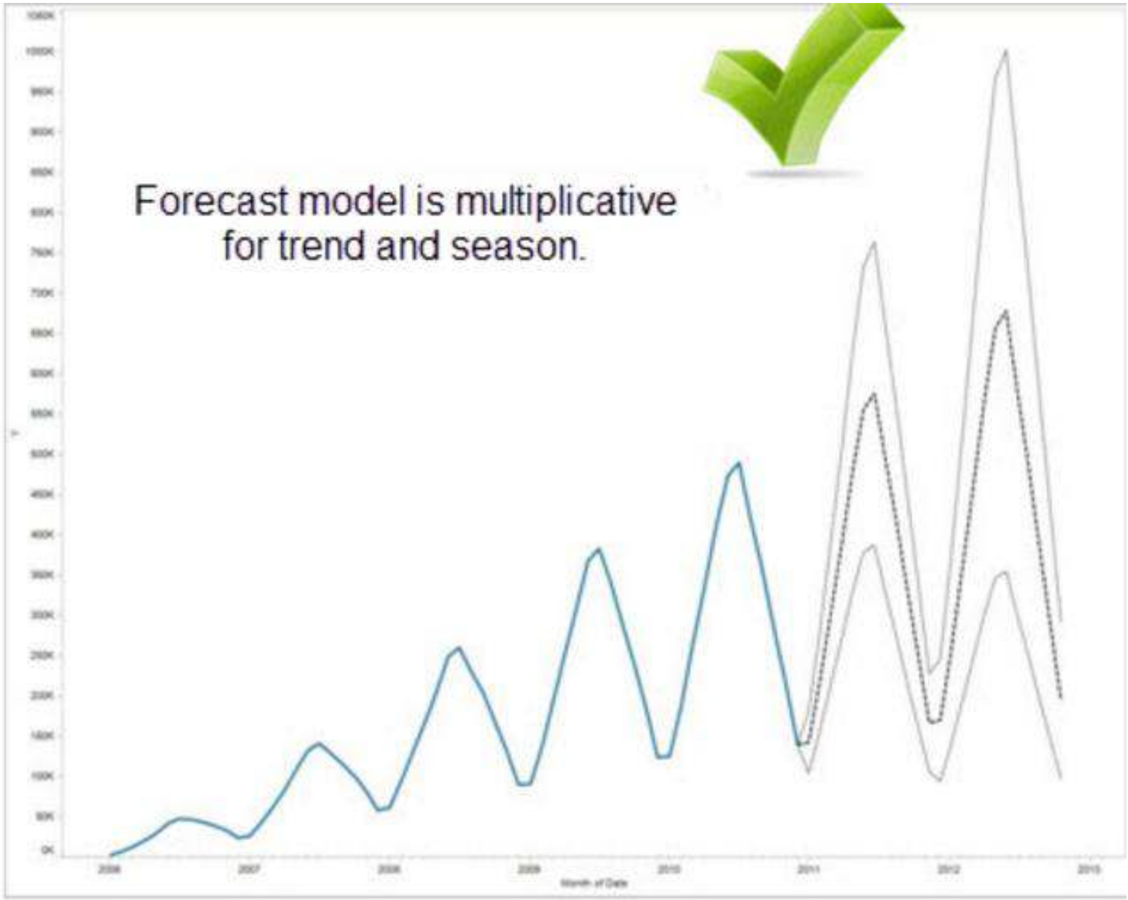
Granularity

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

Describe Forecast

Summary | Models

Options Used to Create Forecasts

Time series:	Month of Ship Date
Measures:	Sum of Shipping Cost
Forecast forward:	12 months (Jan 2016 – Dec 2016)
Forecast based on:	Jan 2012 – Dec 2015
Ignore last:	1 month (Jan 2016)
Seasonal pattern:	12 month cycle

Sum of Shipping Cost

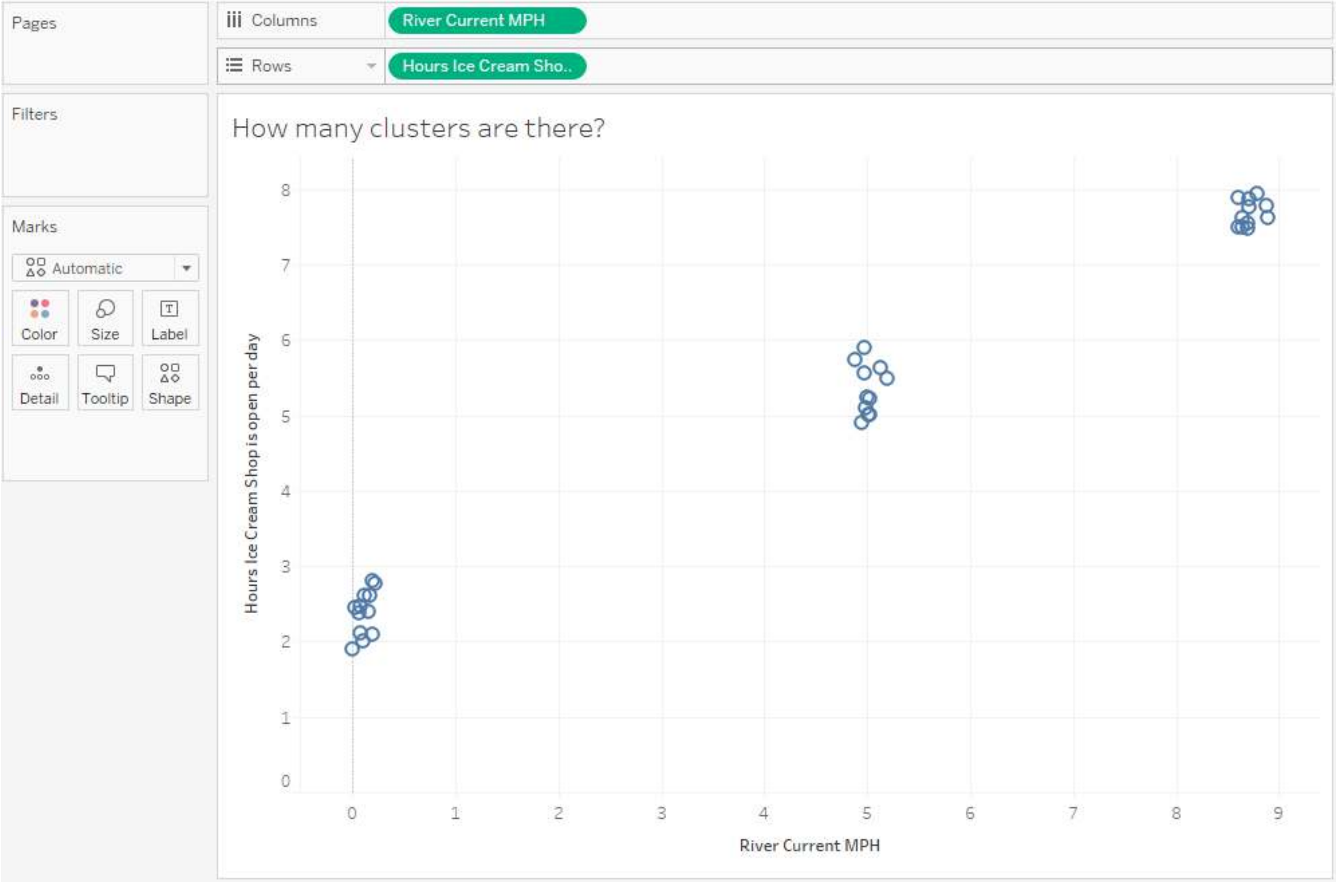
Initial Jan 2016	Change From Initial Jan 2016 – Dec 2016	Seasonal Effect High		Seasonal Effect Low		Contribution Trend	Contribution Season	Quality
37,467 ± 8,717	26,663	Dec 2016	13,163	Feb 2016	-8,527	6.1%	93.9%	Good

Analysis > Forecast > Describe Forecast

- **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?

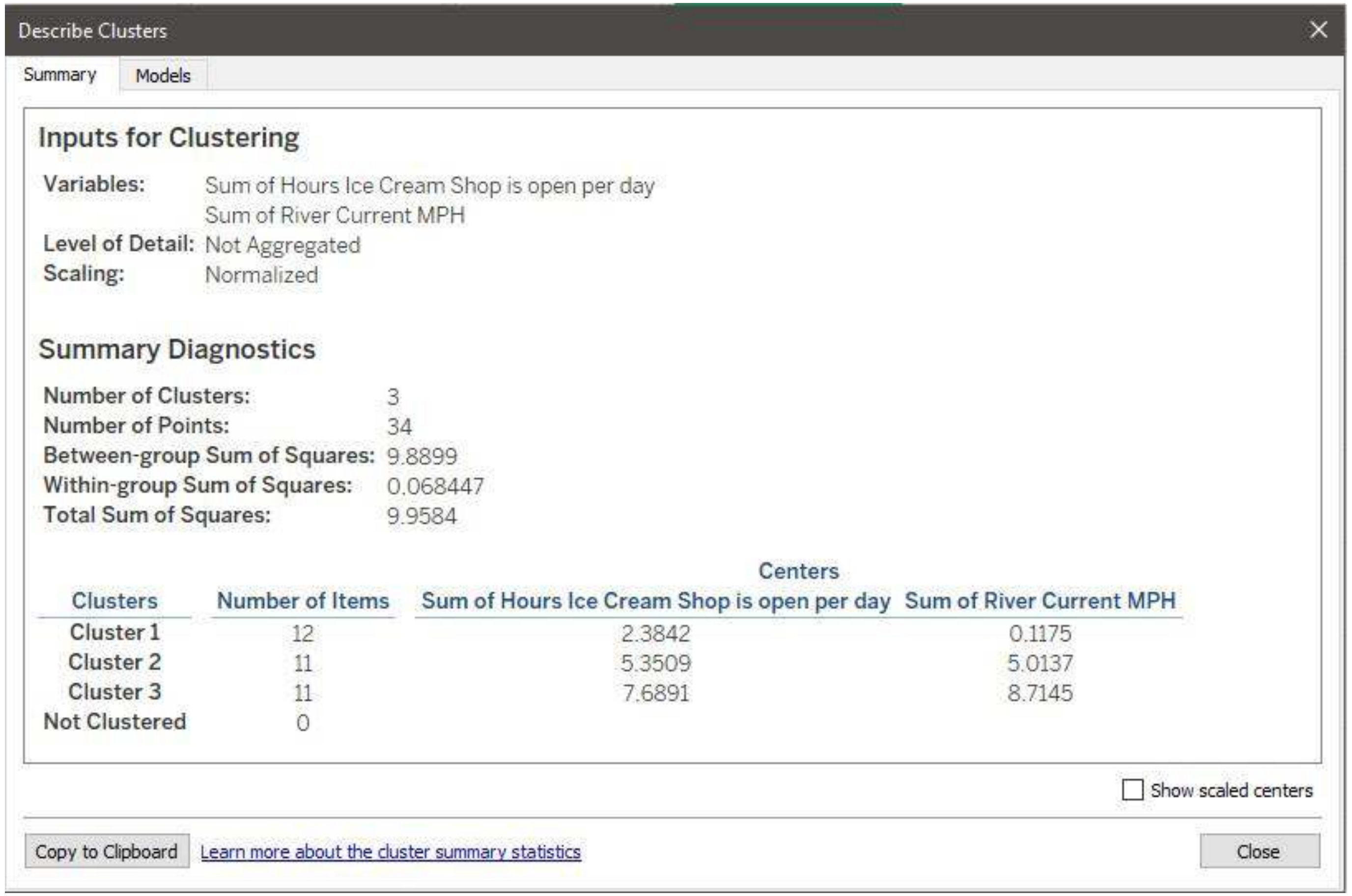
Clusters

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



Describe Clusters

Summary | Models

Inputs for Clustering

Variables:	Sum of Hours Ice Cream Shop is open per day Sum of River Current MPH
Level of Detail:	Not Aggregated
Scaling:	Normalized

Summary Diagnostics

Number of Clusters:	3
Number of Points:	34
Between-group Sum of Squares:	9.8899
Within-group Sum of Squares:	0.068447
Total Sum of Squares:	9.9584

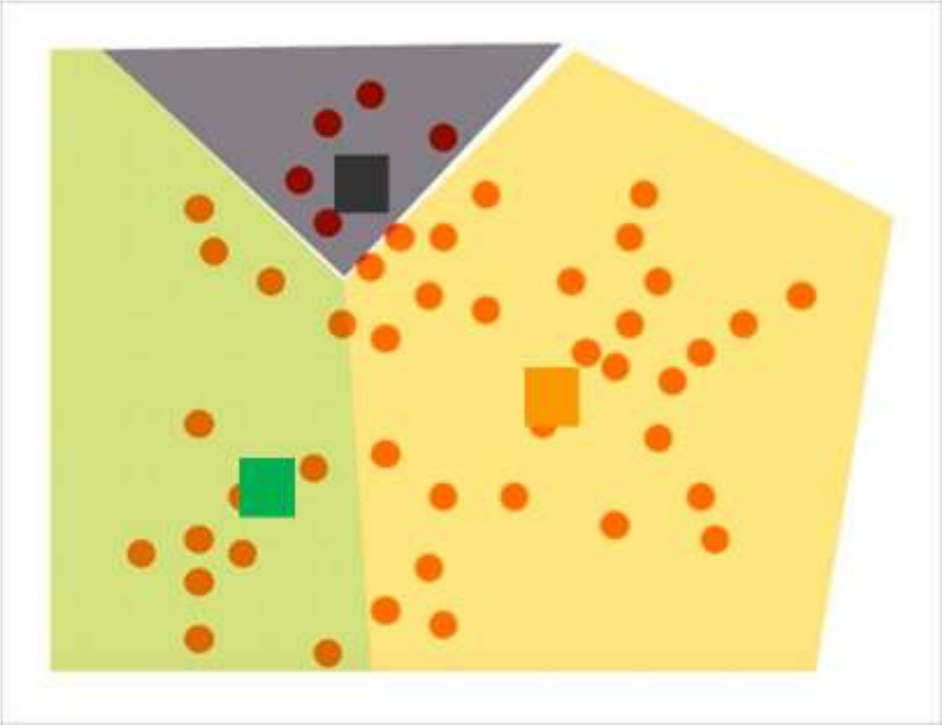
		Centers	
Clusters	Number of Items	Sum of Hours Ice Cream Shop is open per day	Sum of River Current MPH
Cluster 1	12	2.3842	0.1175
Cluster 2	11	5.3509	5.0137
Cluster 3	11	7.6891	8.7145
Not Clustered	0		

Show scaled centers

Copy to Clipboard | Learn more about the cluster summary statistics | Close

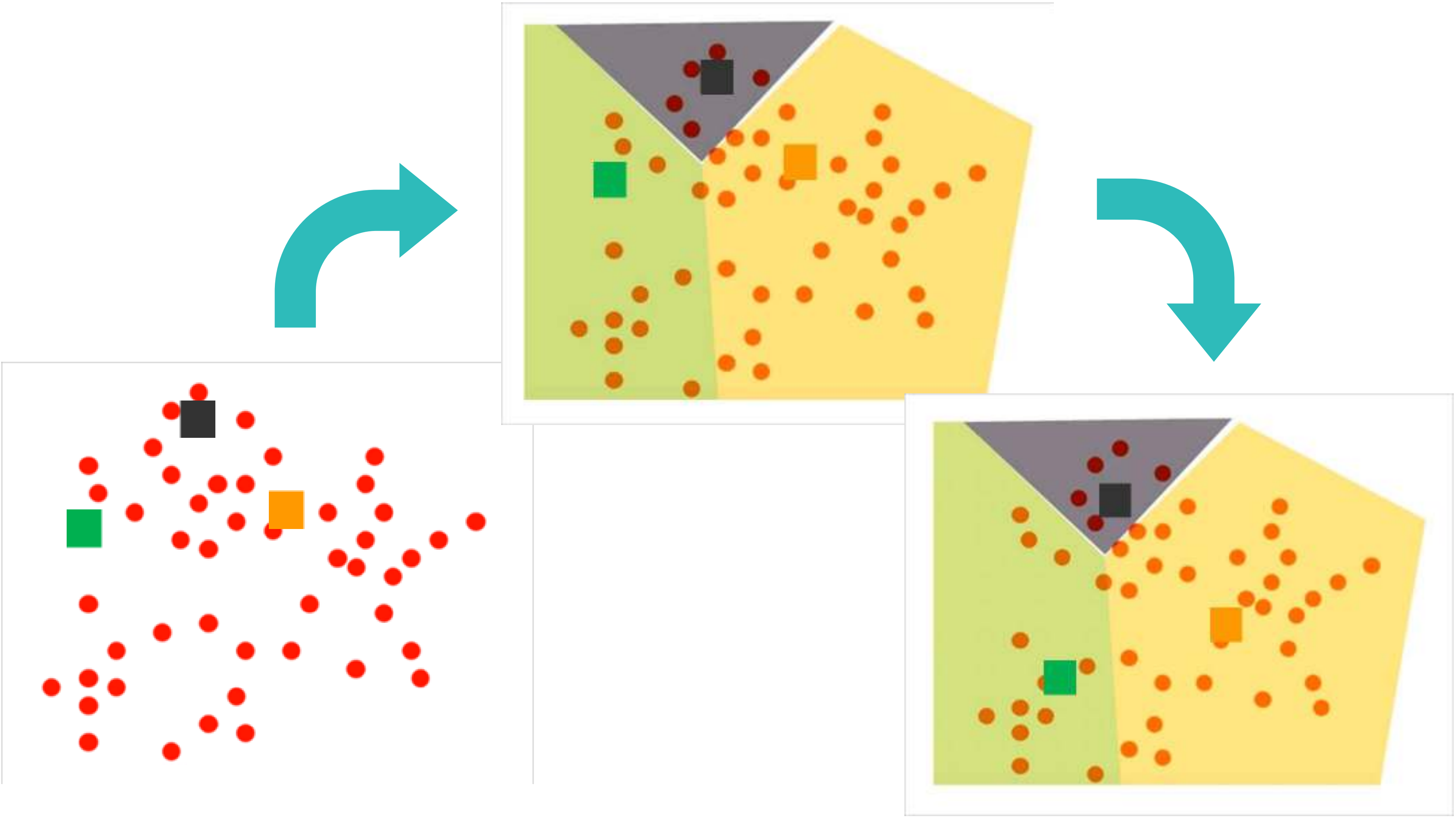
This shows the inputs to the clusters. We see our two variables, we were not aggregated and scaling was not adjusted.

Clustering

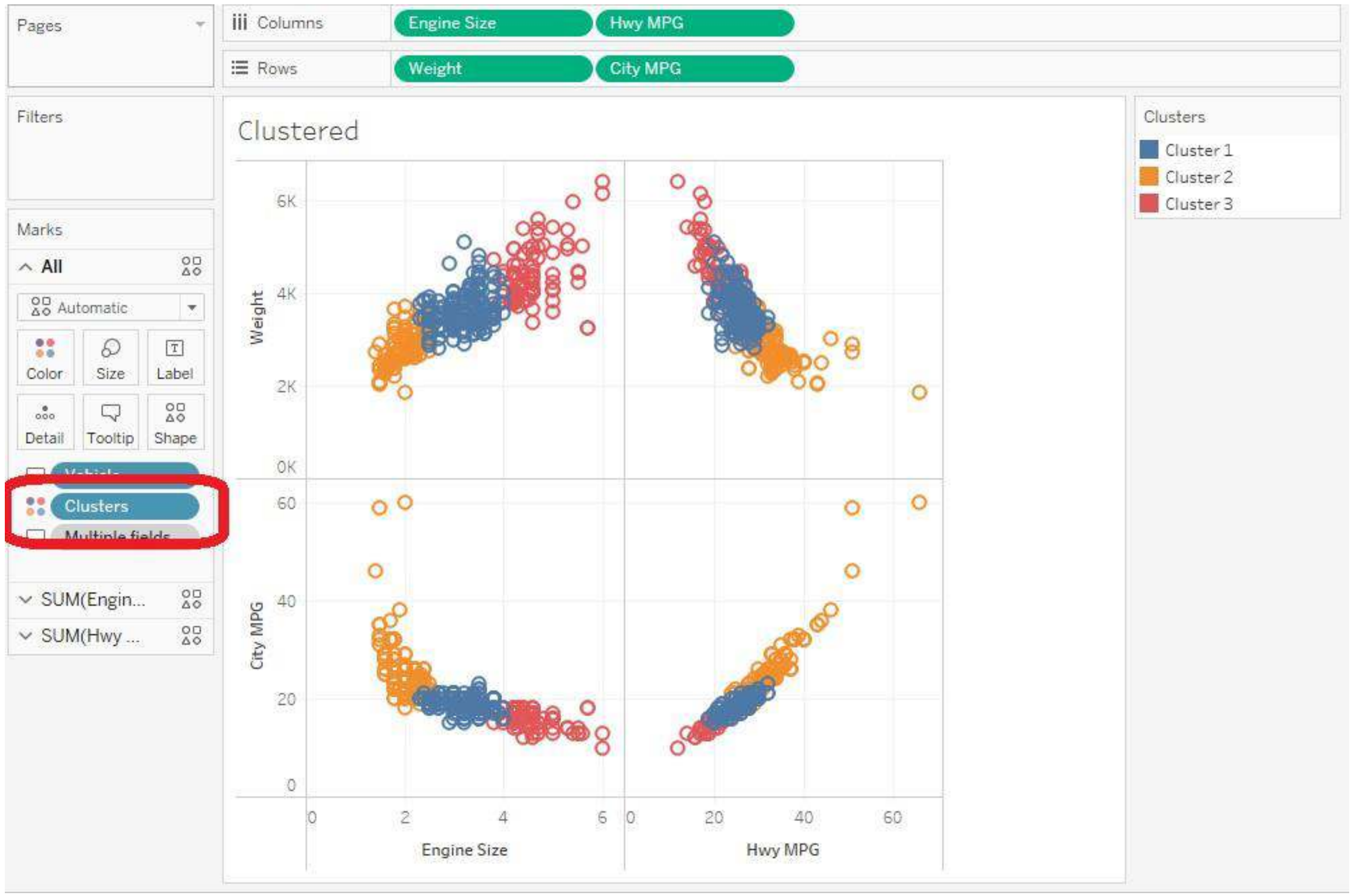


Grouping a set of objects such that marks within each cluster are more similar to one another than they are to marks in other clusters

Clustering



Clusters – Saving the results

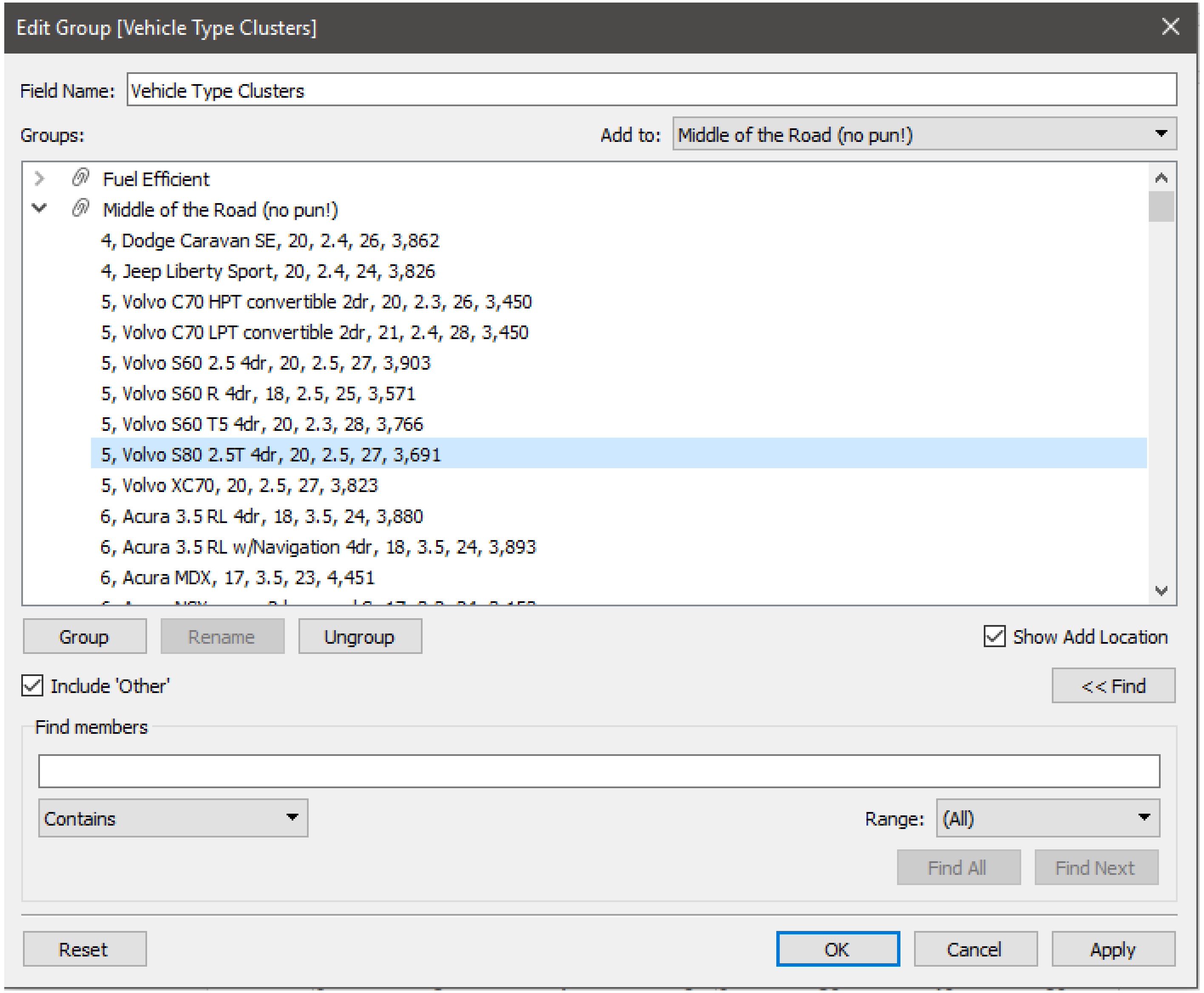


The clustering is done. Three clusters with default names and colors. Plus, if new data comes in the data gets re-clustered and results may change.

Drag/drop the clusters pill onto the Data pane.

I'm going to rename it to Vehicle Type Clusters. Notice the icon will change.

Clusters – Fine tuning the saved group

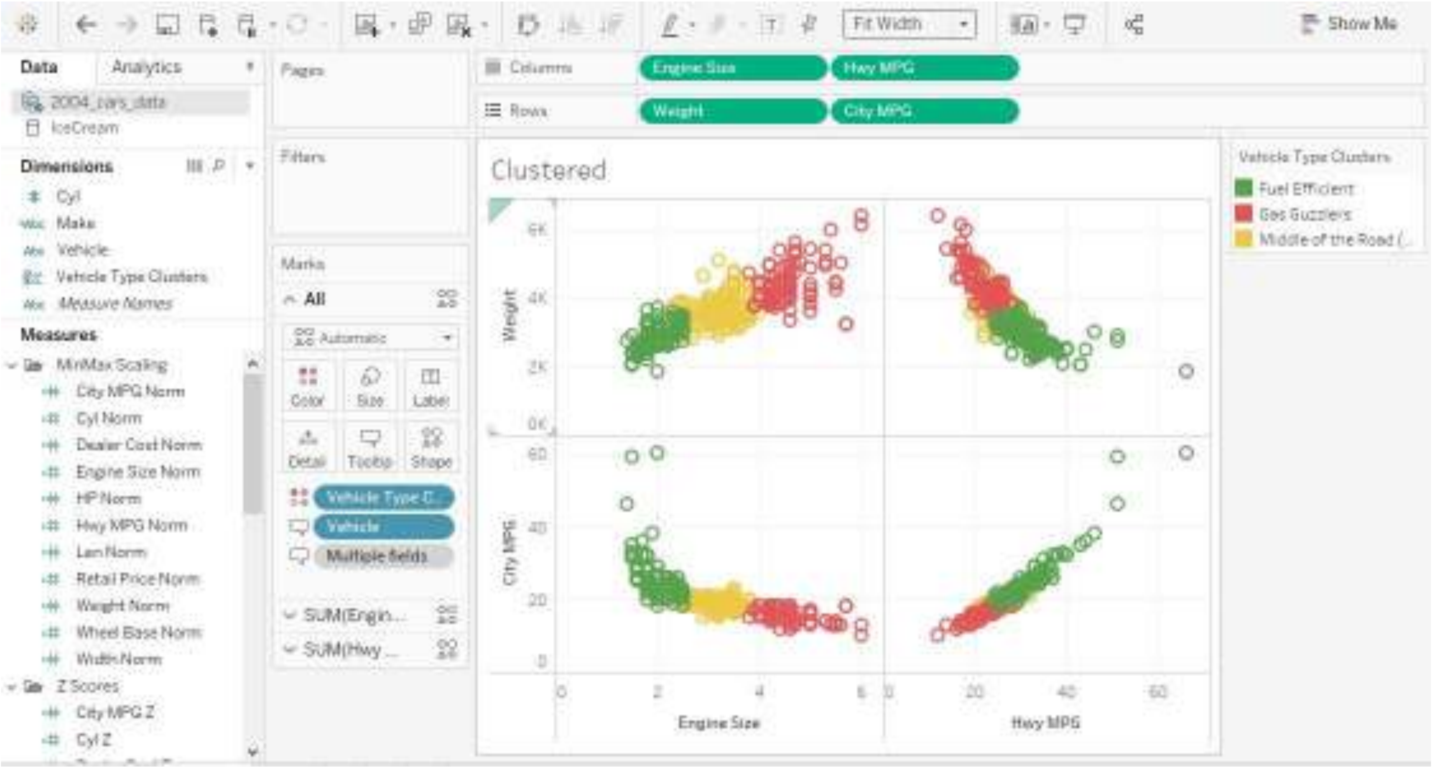


Rename the groups to something more meaningful

e.g.:

- Fuel Efficient
- Middle of the Road
- Gas Guzzlers

Clusters – Fine tuning the saved group

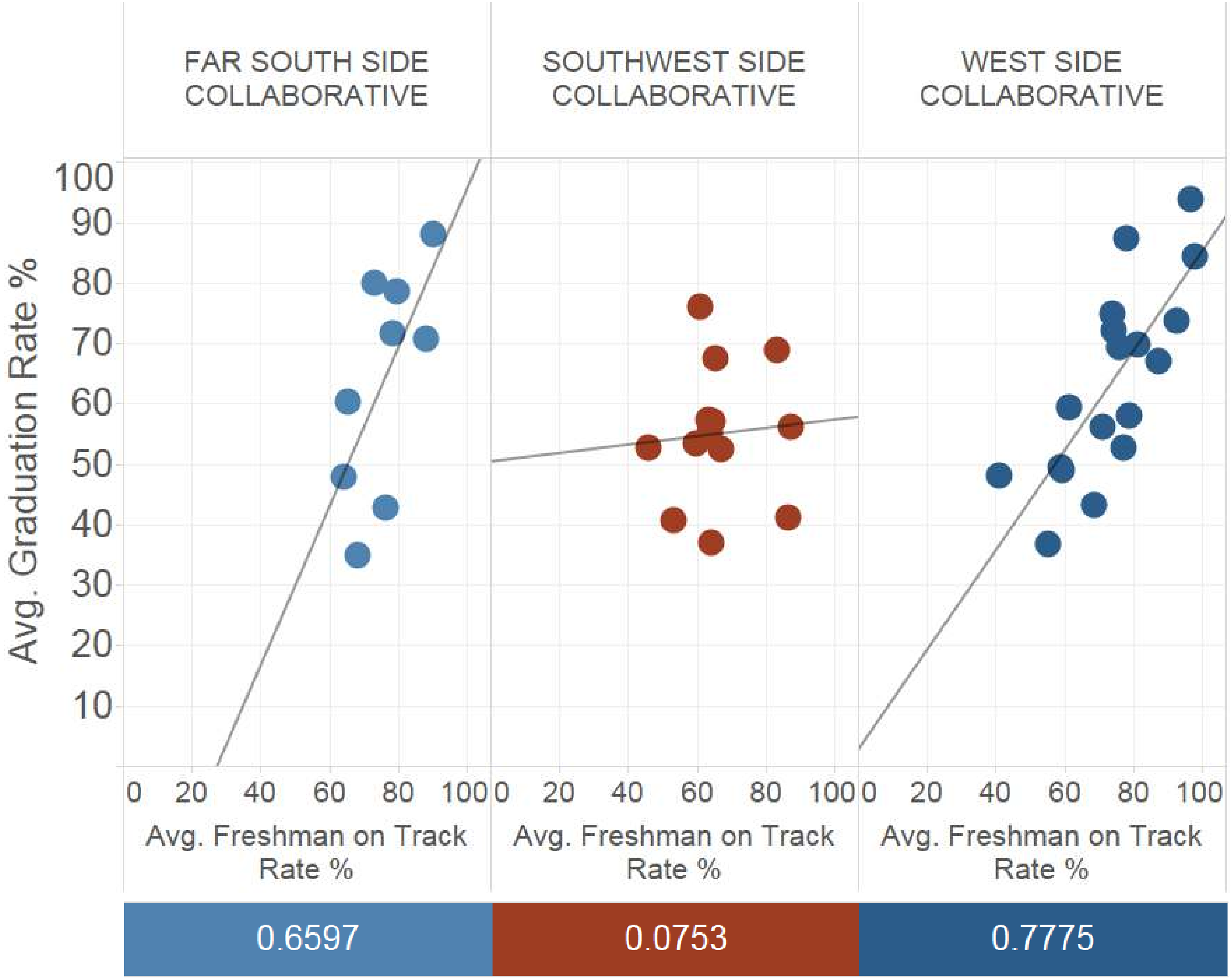


Now I can remove the ad-hoc "Clusters" group and replace it with my saved group.

Update the colors and I am done!

Correlation (is not Causation)

Correlation Coefficient

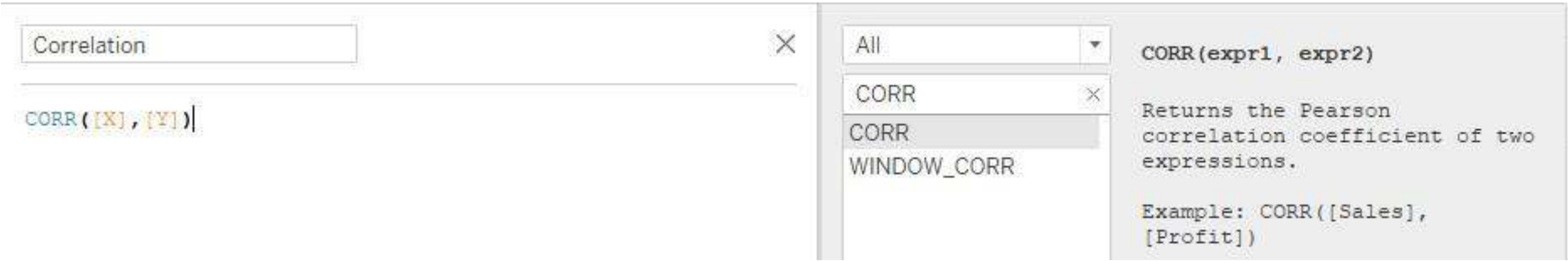


Pearson's correlation coefficient is a measure of the strength and direction of the linear relationship between two variables

Correlation computation

Correlation Function:

- Built into Tableau
- Uses Calculated Fields



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