# ANOMALY DETECTION TUTORIAL

## TABLE OF CONTENTS

Introduction: Explore Malware Database for Anomalies 1

Loading Data 1

Create a Histogram 2

Create a Boxplot 4

Create Score Anomaly Detection Graph 5

Creating a graph to show daily average Scores from the dataset 5

Updating the Graph to Show both daily max and average scores 8

Creating Reference Bands to Detect Anomalies 9

# 

# Introduction: Explore Malware Database for Anomalies

To detect anomalies in the malware entries in the dataset, explore them based on their scores.

- What is the average score?

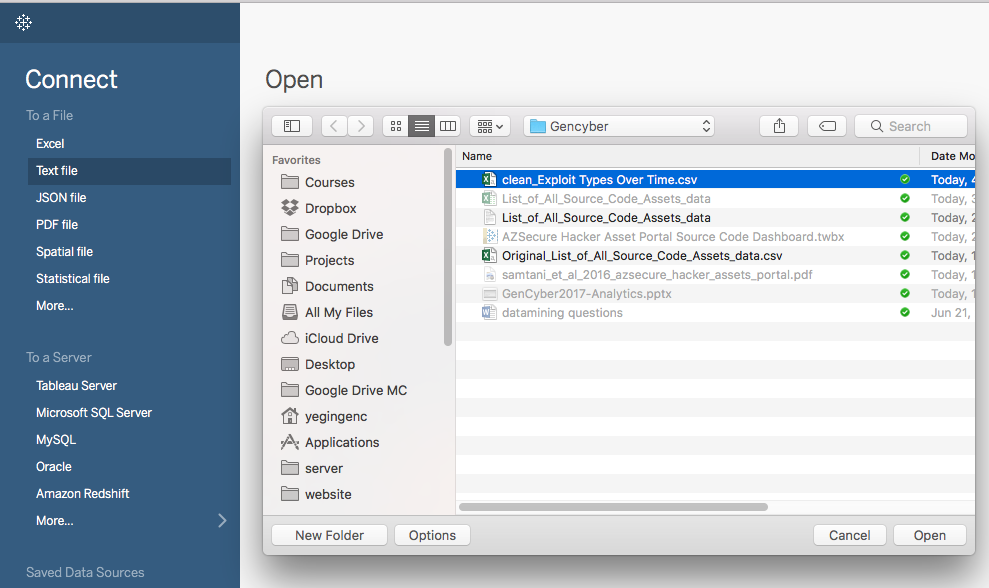
- What is the median score?

- Which one would you prefer to use to determine a typical malware in terms of its score?

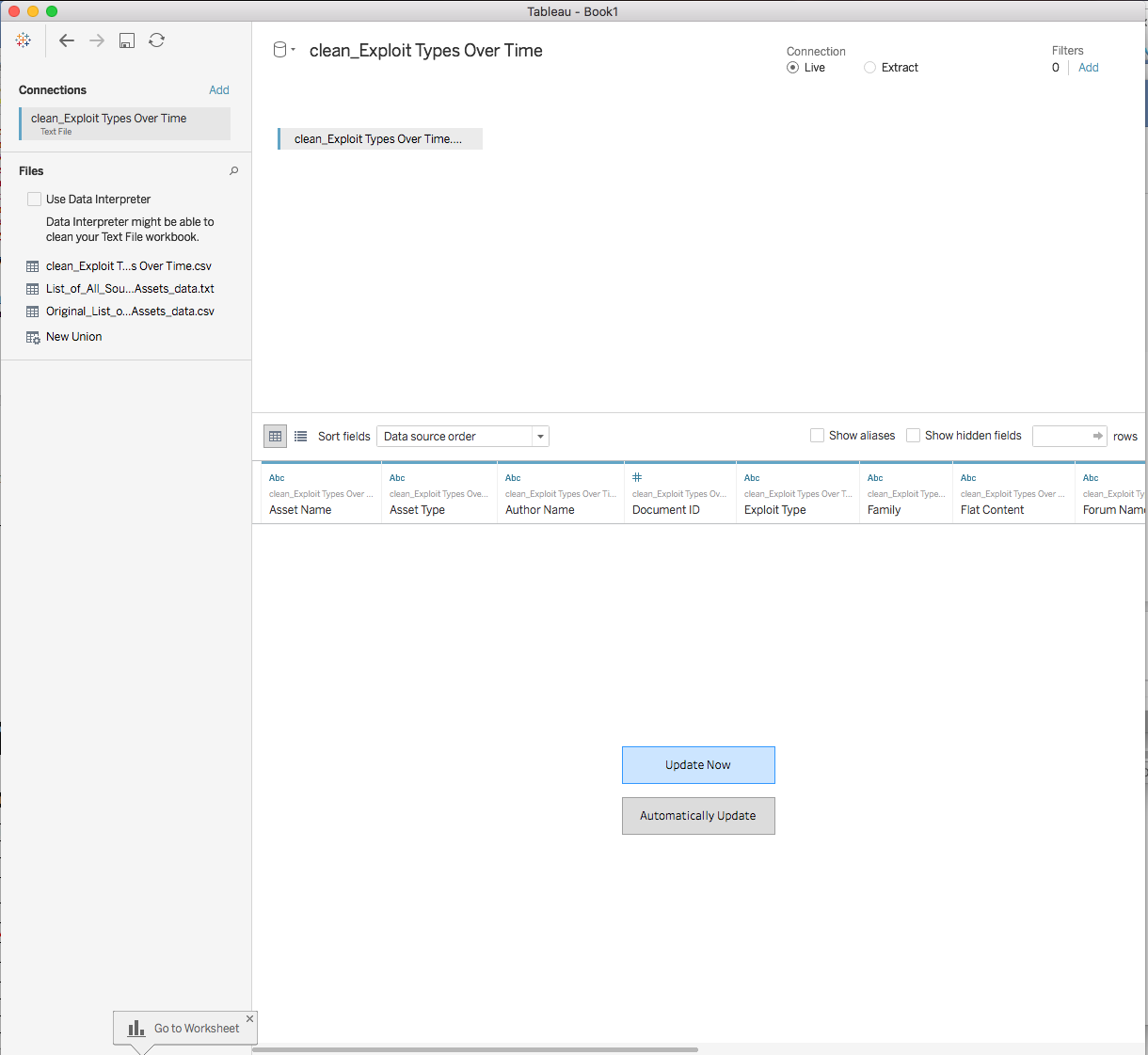
- Are there extremely high (or low) scored malware entries? (Hint: One way to detect them is creating box plots. Another is reviewing the distributions with histograms

# Loading Data

* Loading data requires ‘connecting’ to the data source. In our case, we will be loading data in file therefore we will *connect to a file*.
* Our data file is “Exploit Types Over Time.csv” which is considered as a “Text File”
* Click on *Text File* and browse to select the file in your computer

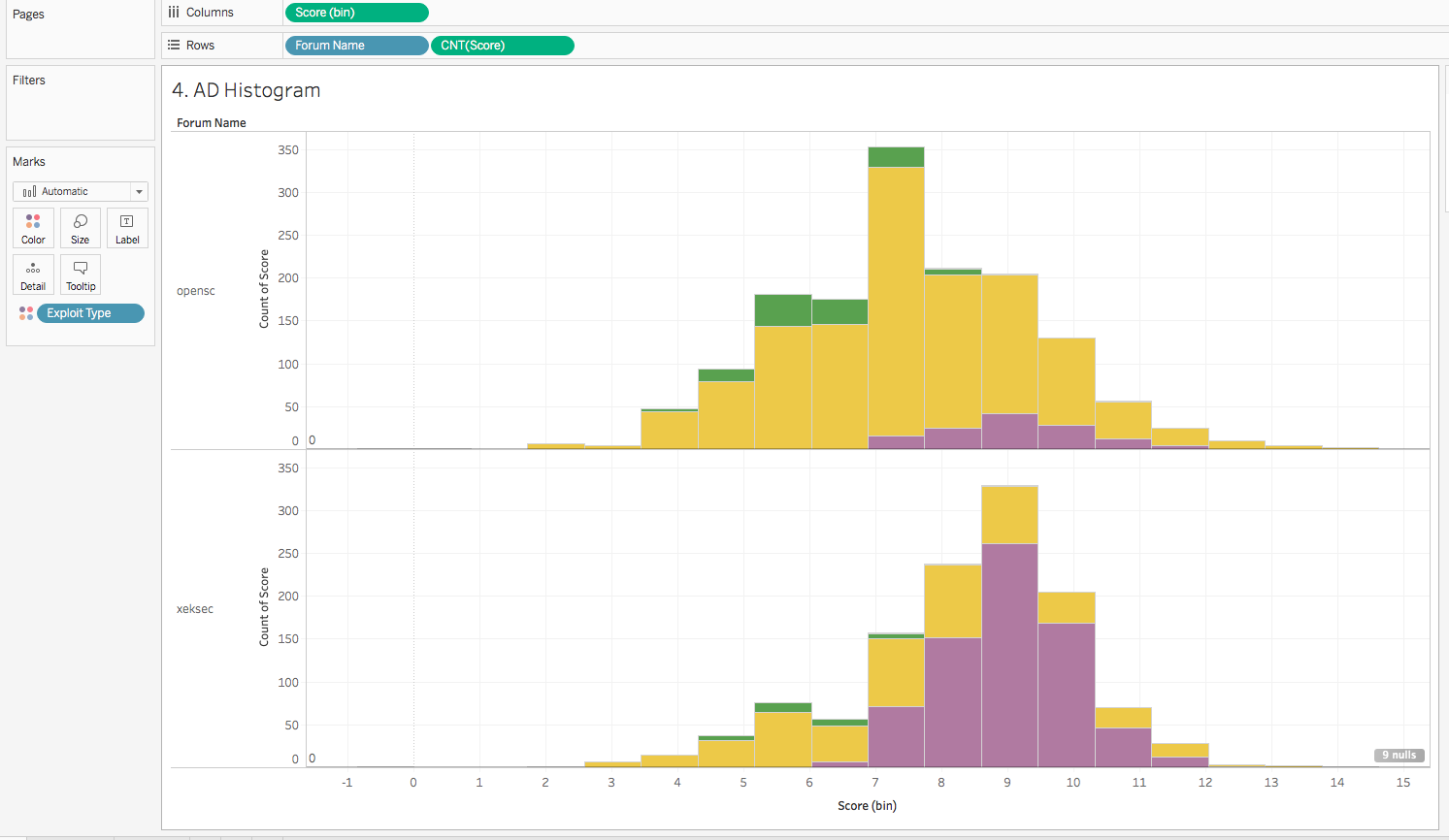


* Once connected, columns (variables) will appear on screen. Click on Update now, to import the data



# Create a Histogram

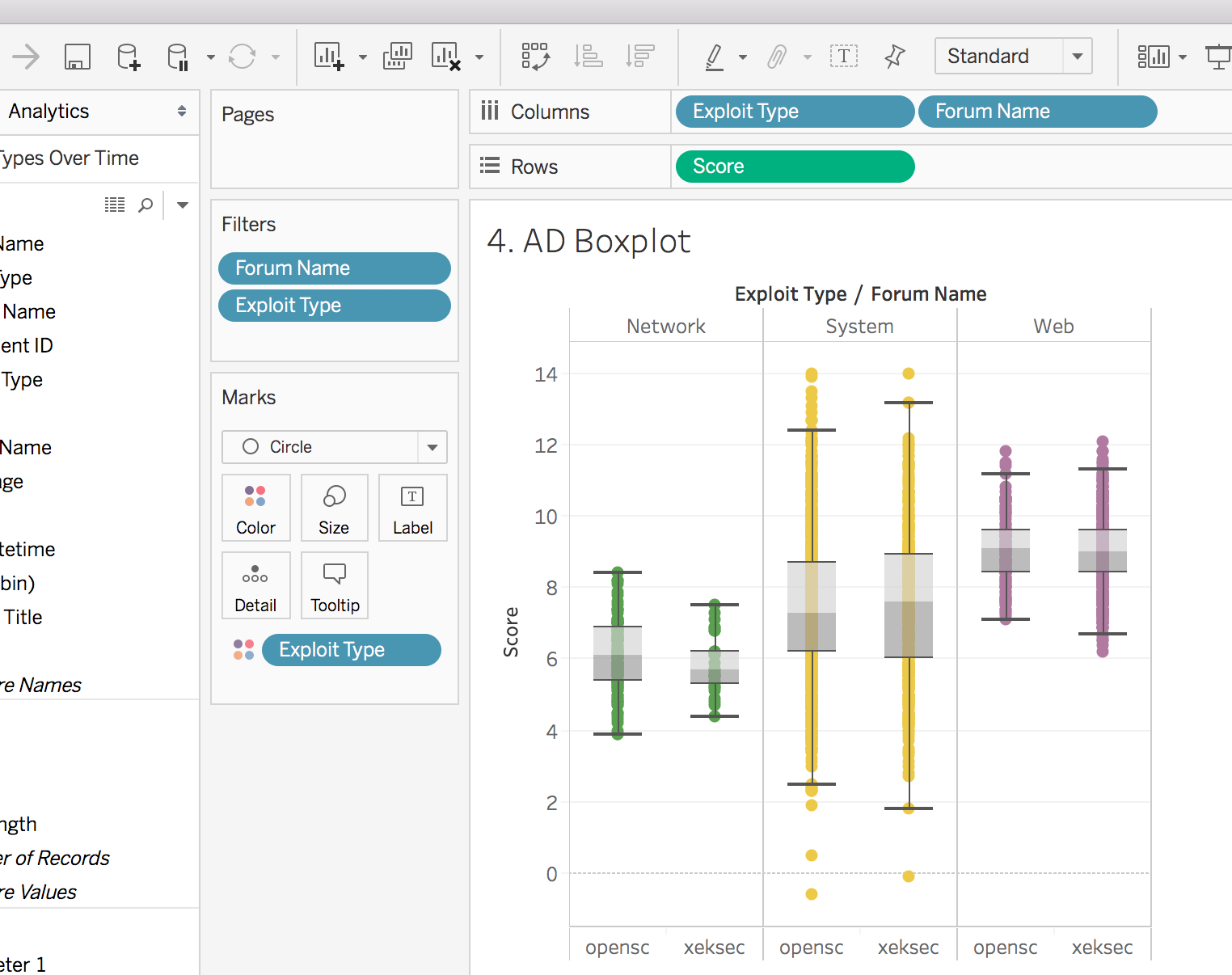
* Start a new sheet and label it as AD histogram.
* Create a histogram that will look like the figure below. (Hint You: can follow the instructions from Tutorial 2: Build a Histogram.)
* This will be a histogram for the *Score* variable, colors indicate the Exploit type,
* Add colors to show the *Exploit Type*.
* Add labels showing the percentages.
* Finally add panels such that each panel should show the *Forum* it is coming from.



* What can you tell about the distributions in your graph? Do they resemble normal distributions?
* Does this view show and differences between entries from different Forums?
* Finally try see if you can create different variations of this view by using different variables for Panels. See if you can find an interesting histogram that doesn’t look like normal.

# Create a Boxplot

* Start a new sheet and label is as AD Boxplot
* Create a boxplot that will look the figure next. (Hint You: can refer to the instructions from Tutorial 3: Build a Boxplot.)
* Add colors to show the *Exploit Type*.
* Create panels by the *Forum*
* How many outliers are there in this distribution?
* Can you tell what their Exploit Type is?
* What more can we say about these outliers? (Hint: you can select a circle – or multiple circles; right click and select view data to get more detailed information.
* Finally, add panels for Exploit type and Forum to make your graph look like the figure below.

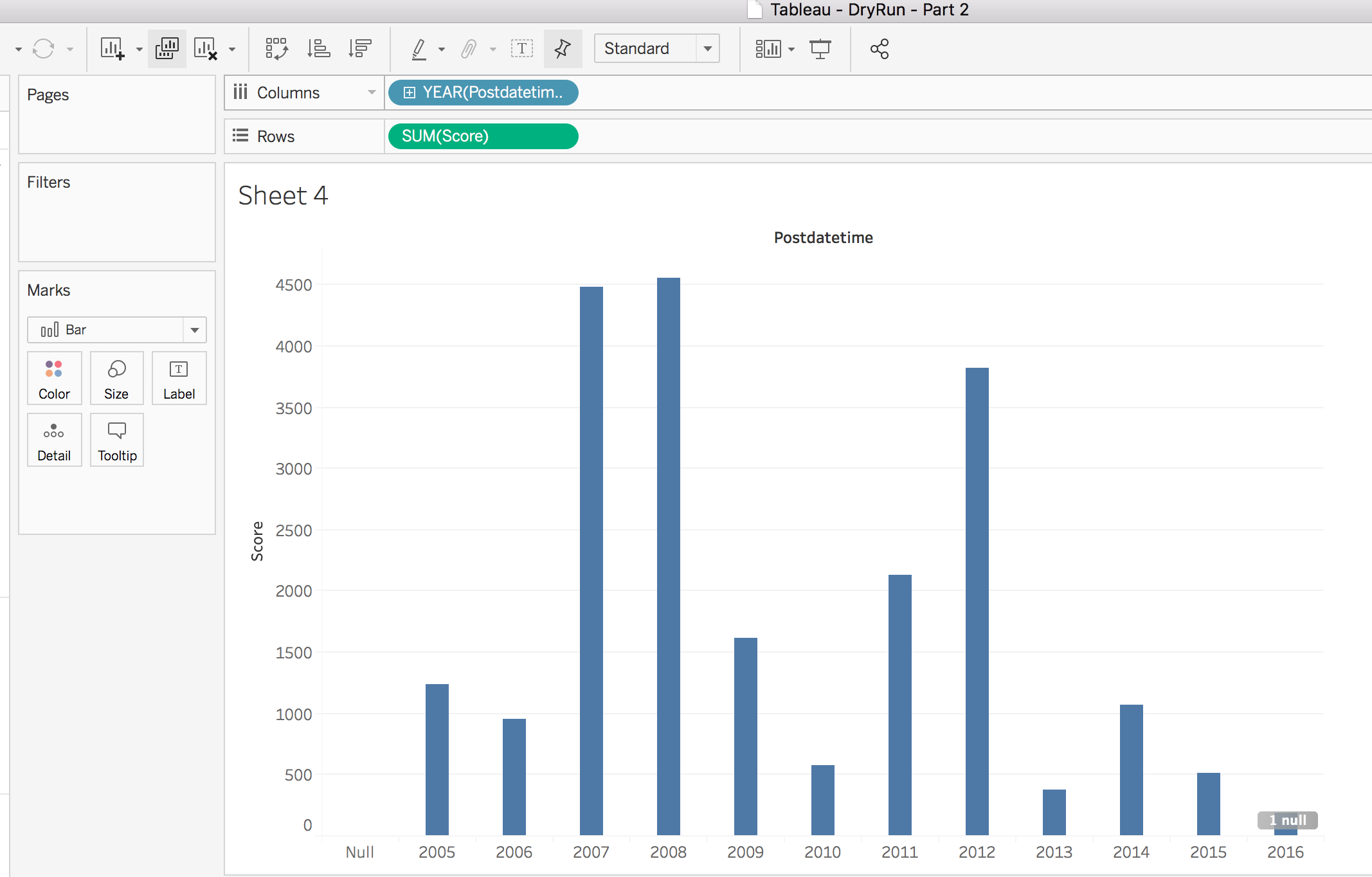


* What does this graph tell you about the distribution of scores across different types and scores?

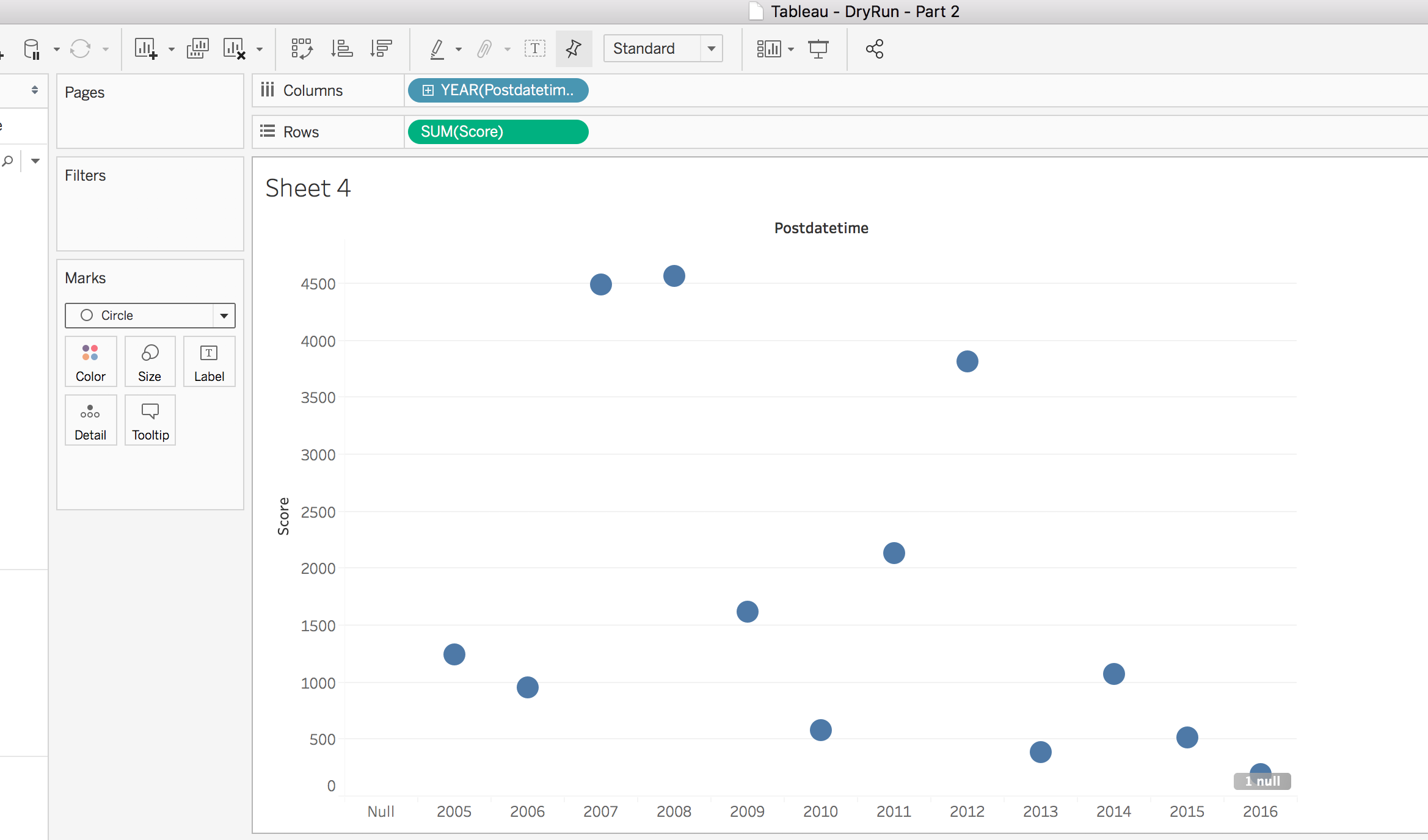
# Create Score Anomaly Detection Graph

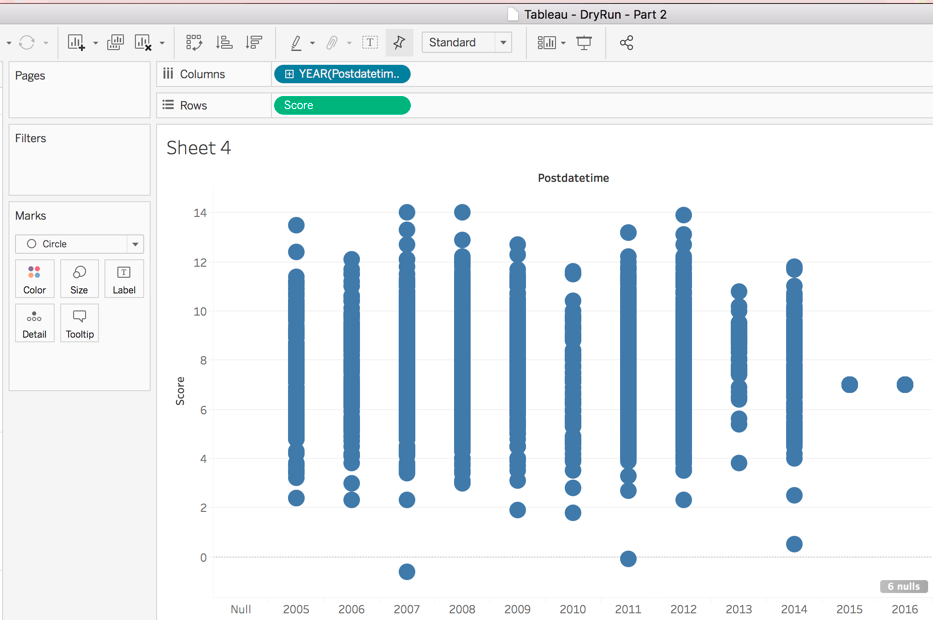
## Creating a graph to show daily average Scores from the dataset

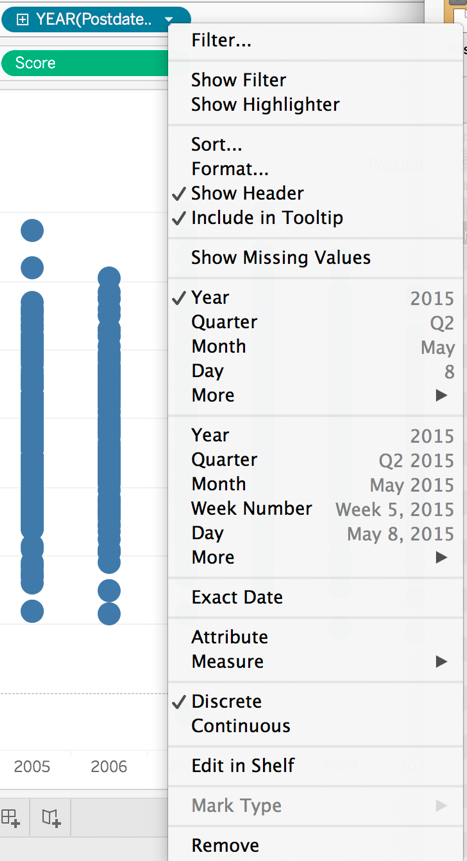
* Start a new sheet and Label it as AD Scatter plot.
* Create a Score by Postdatetime graph by dragging the Score measure to Rows and Postdatetime dimension to columns. Your graph should look like as the following:



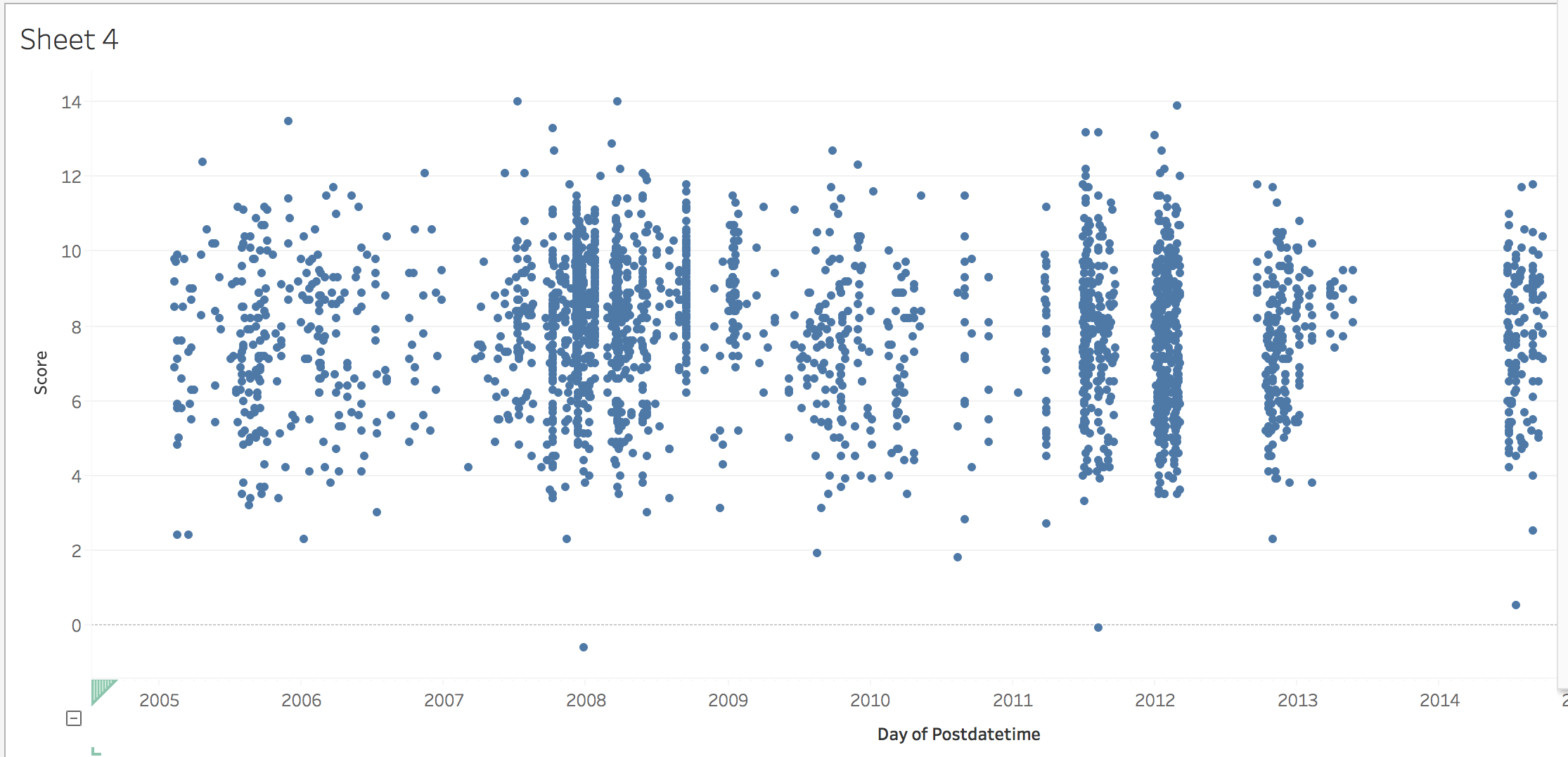
* Now let’s convert the chart type from bar chart to scatter plot. Under Marks, select *Circle* instead of *Bar*



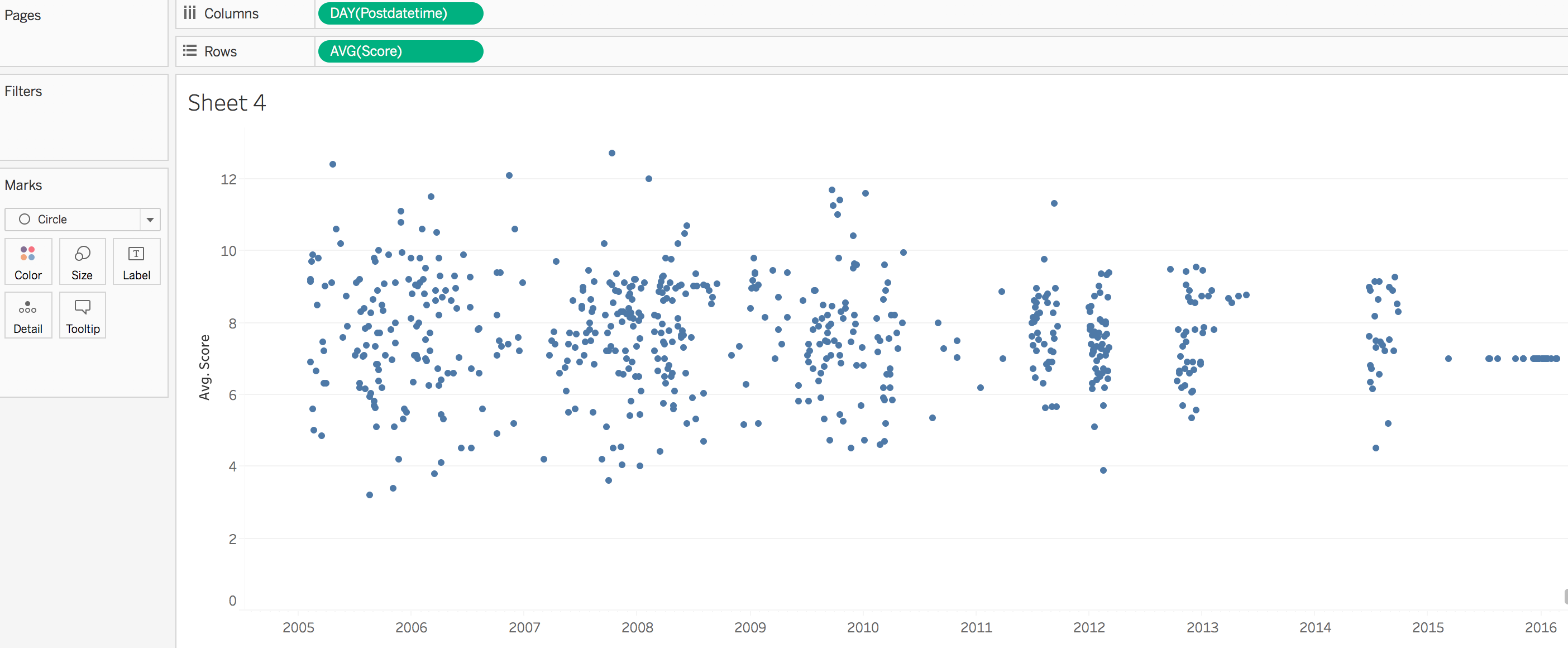
* Each circle is now representing the sum of the scores for the corresponding year. In other words, scores are aggregated to the year level. We would like to see each score individually. For that, we need to de-aggregate. Go to the top *Menu > Analysis* and deselect *Aggregate Measures.*



* Currently, the graph shows the yearly break down, we want to see the scores on a daily-basis. For that, go to *the YEAR(Postdateitem)* measure under the column section, right click on it and select *Day*.
* You will see that there are two Day options. Try both and see if you can see identify what the difference is.
* Which one should you select the scores on a daily basis between 2005 and 2016 (as in the figure below)?
* Your graph currently should like as the following:

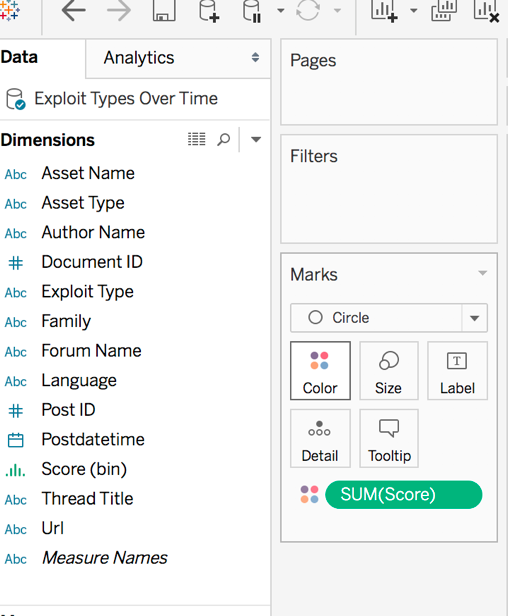


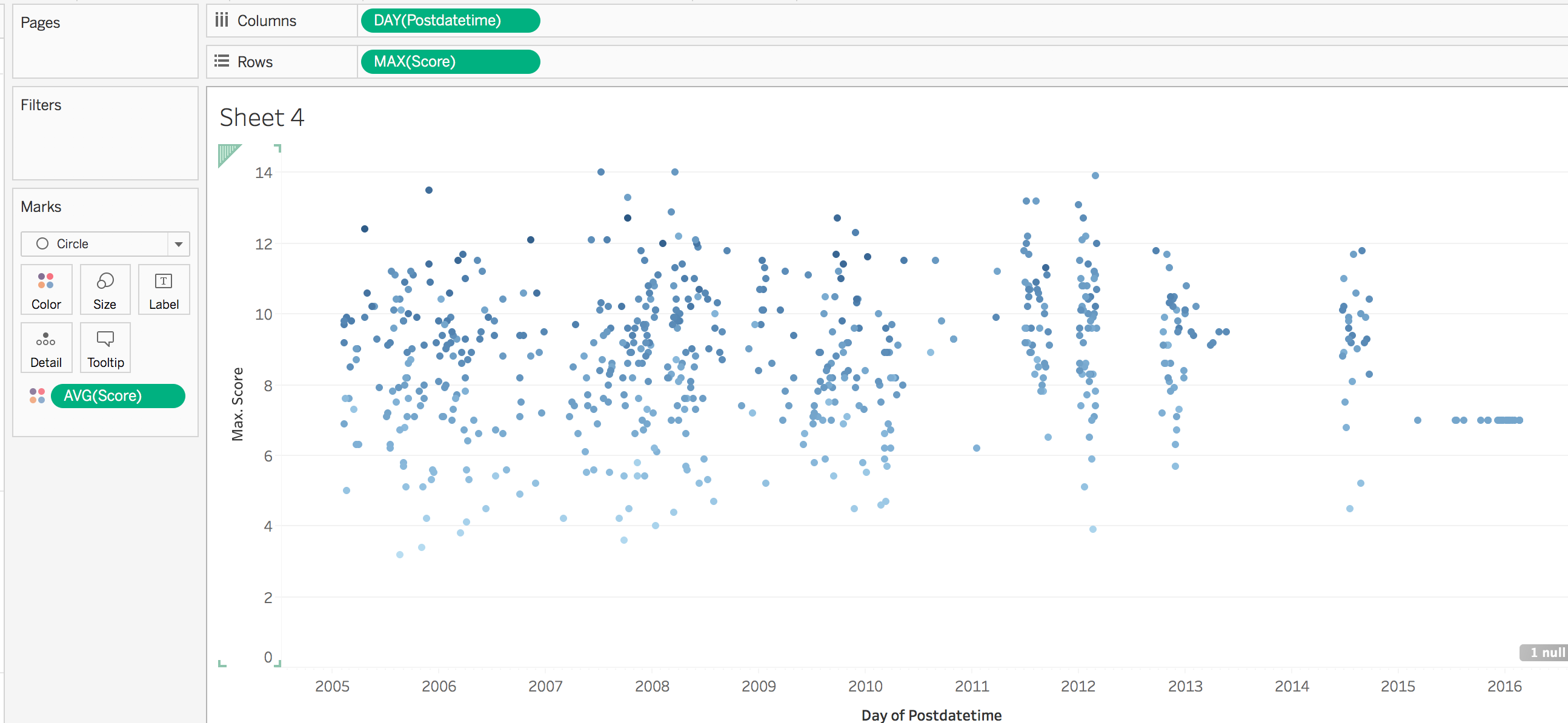
* What do you think each row represents now? Right click on one of the circles, and click on *View Data*. In the popup, select *Full Data* for more details.
* Repeat looking at the Full Data for a few more circles. How many rows are there in each case? Why?
* Next, let’s aggregate the data to the day level. That is, let’s update the graph in such a way that each circle represents the average score of all the entries for the day. For this, go to top *Menu > Analysis > Aggregate Measures*.
* The default aggregation is summation. To change it to average, click on SUM(Score) measure under Rows and Select *Measure > Average*
* Your final graph should look like as the following:



* Now right click on a few circles, click on View Data and examine the Summary and Full Data Tabs. Do you still have same number of rows under Full Data tabs for each circle? Why (not)?

## Updating the Graph to Show both daily max and average scores

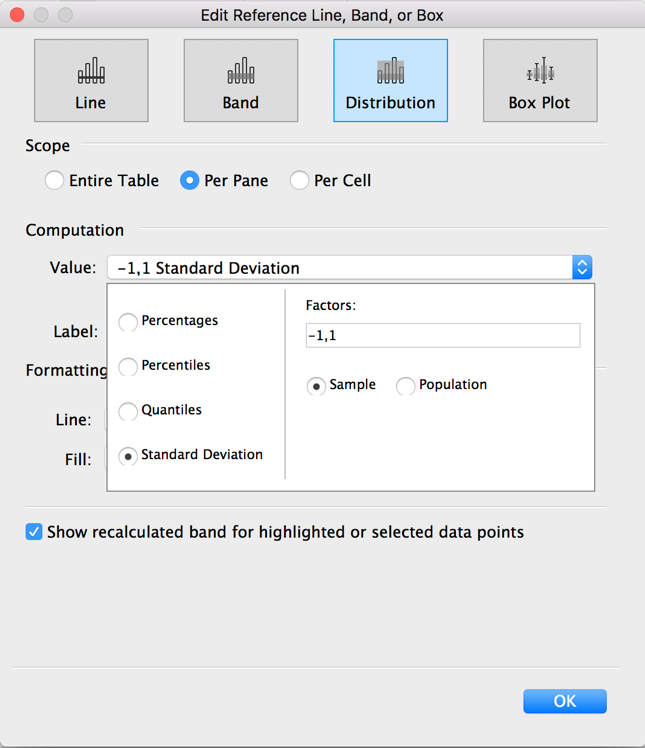
* While average scores are helpful, it also is important to be aware of high-score entities. Next we will update the graph to show daily maximum scores (instead of average) and we will use colors to indicate the average.
* To change the values from average to maximum, go to AVG(Score) under rows and select *Measure > Maximum*
* Next we will change the coloring scheme to reflect the average score. For this, Select *Score* from the Measures and drag it to Color under Marks section. Marks section should look like the figure next.
* Again, the default aggregation is summation. To change it to average, click on SUM(Score) and select *Measure > Average*
* Your graph should look like the following:



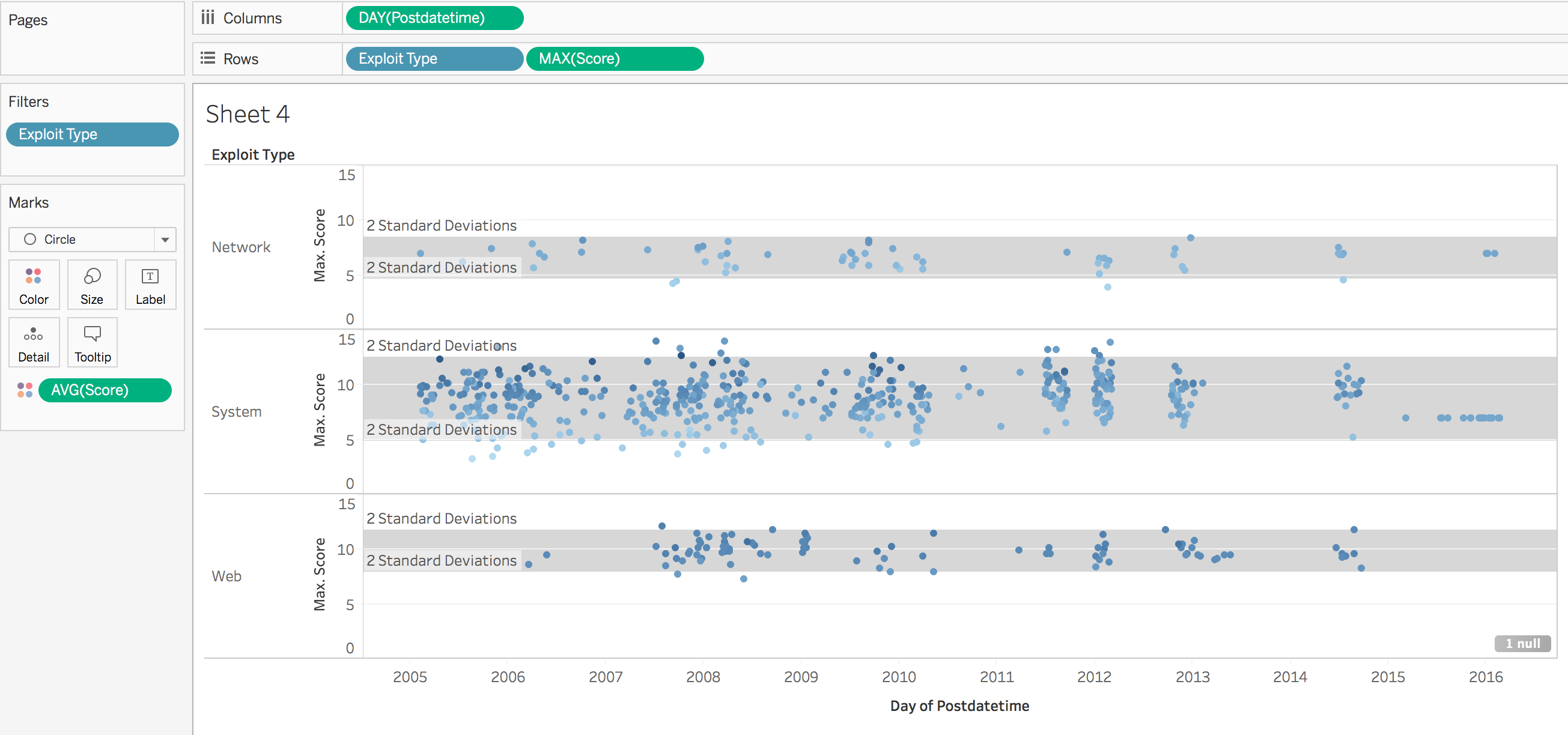
* Let’s break the graph by Exploit type. Drag *Exploit Type* to Rows, in front of MAX(Score).



## Creating Reference Bands to Detect Anomalies



* The graph so far shows, the days with exploit entries with max scores. So, if the max score is too high, then we can consider it as anomaly. But how high is too high? Usually, we decide it based on the standard deviations. If the max-score is 2 standard deviations above the average max-score than it is considered as high. (If normally distributed, what percent of the data is 2 standard deviation of the mean?)
* Distribution bands can help us detect values that are 2 sd above the mean. To create a distribution band, go to the Analytics Tab and double click on *Distribution Band.*
* In the pop-up menu, click on value and select Standard Deviation.
* Under factors, type in -2,2 and click ok.
* Your final graph should look like as follows:



* Click on a few circles that are above the band and review data. What can you say about them?