## ETHS 8460-

## Web Analytics and Business Intelligence

## Midterm

## Eunna park

## ID: 8631050

## Created: July 17, 2020

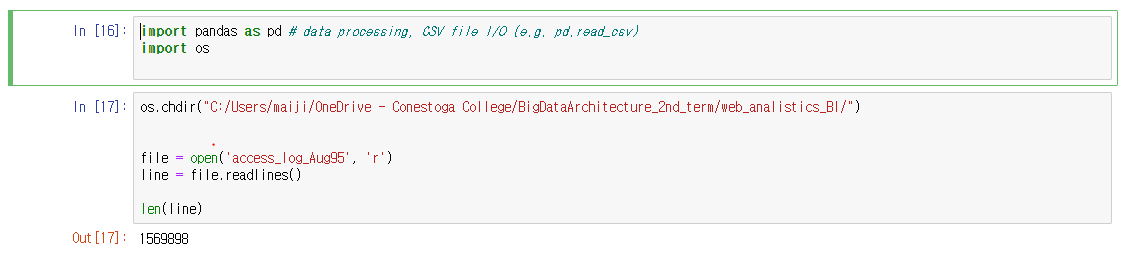
Step 01:

Data cleansing is the first processing to analysis dataset.

A data set for midterms is provided. The data came from the log files of a server, and it is judged that the data cannot be viewed and counted, such as the type of data and the number of rows as I see roughly. Firstly, after I check the data shape, type, and existing invalid data, it will filter to visually show analyzed consequences in Tableau tool.

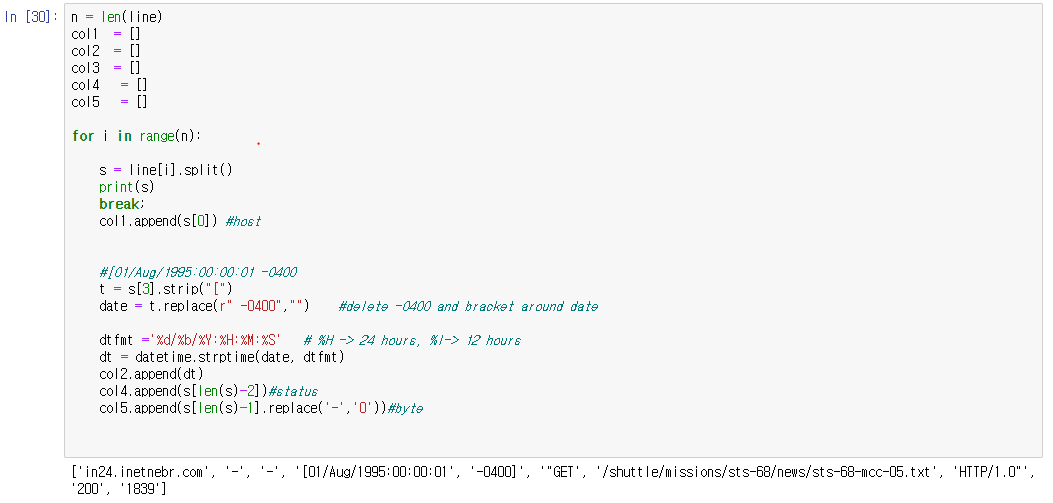
1. Loading Data

Pandas library was used to check data. The number of data is 1569898.



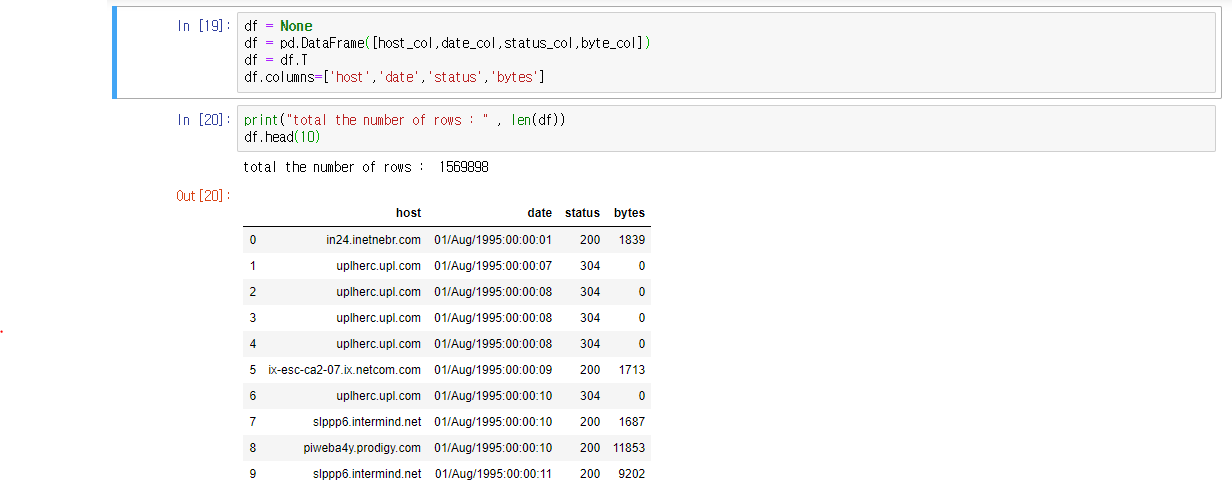
1. Split data and input data into new arrays.

As mush as the number of data rows, each row divides as each columns and input to each column. Each row has unregular blanks, spaces, and special characters, also might distribute invalid data. However, before converting Pandas dataset, looking for incorrect data is hard, so that I decided to input arrays to convert the Dataframe. When I print only 1 row as test, I can check the first data is Host, and the third index is Date, and last and before last data is each status, and bytes. (after checking, I removed the [break;]), to format the date column, I deleted special character, and meaningless string such “-400”. The Byte column located the end of line might has different data type “-“ instead of decimal type, it will change from “-“ to 0. Before last column called status is located the before end of index in a line, so it is located length of split -2. Actual coding and printed result is as below a captured picture.

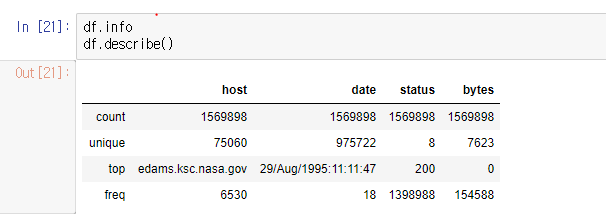


1. Convert the Dataframe in pandas library.

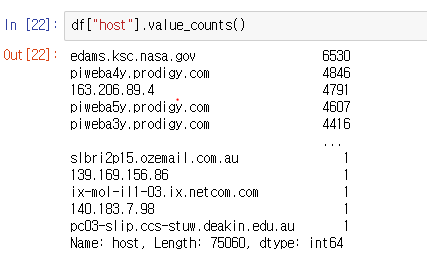
We can check the data row is still same number, and the top of data is that I expected.



1. Confirm the total data in the dataframe

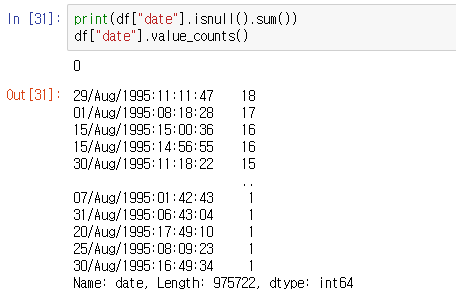


1. Conform host data

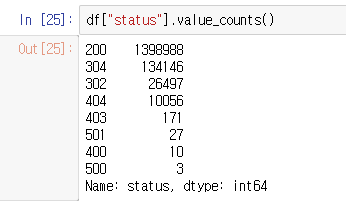


1. Confirm date

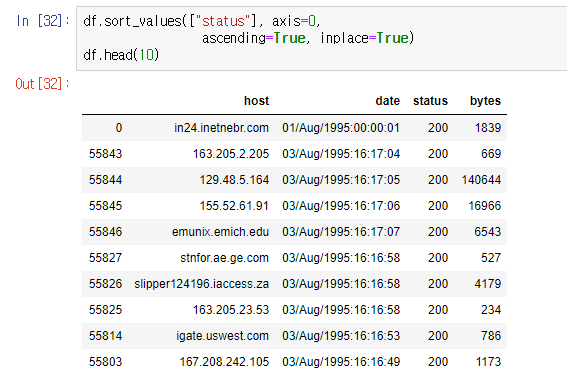
Date does not have null data and invalid data.



1. Check status



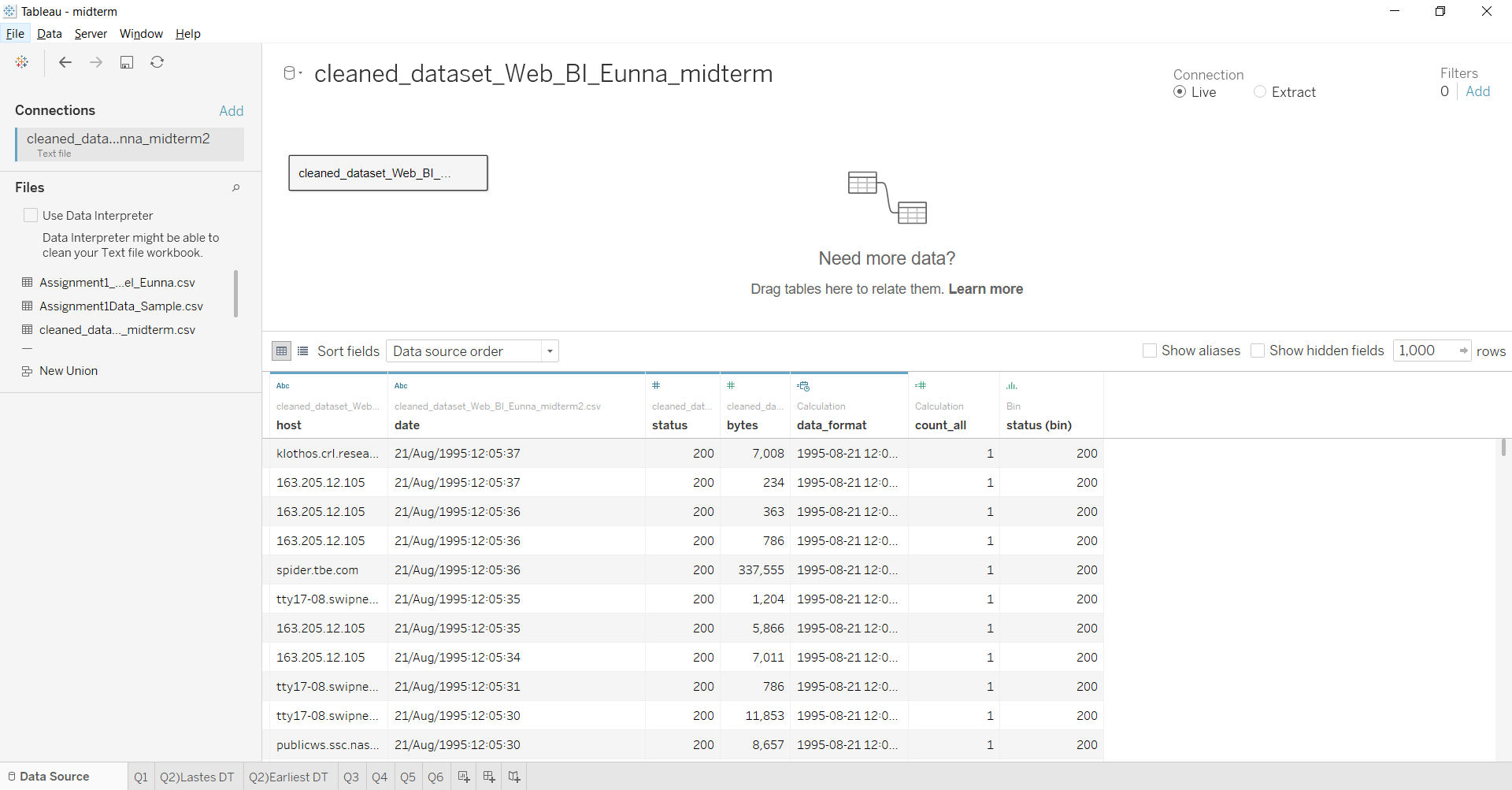
1. After sorting several times, check data for finding weird data.



1. Export to use in Tablaue

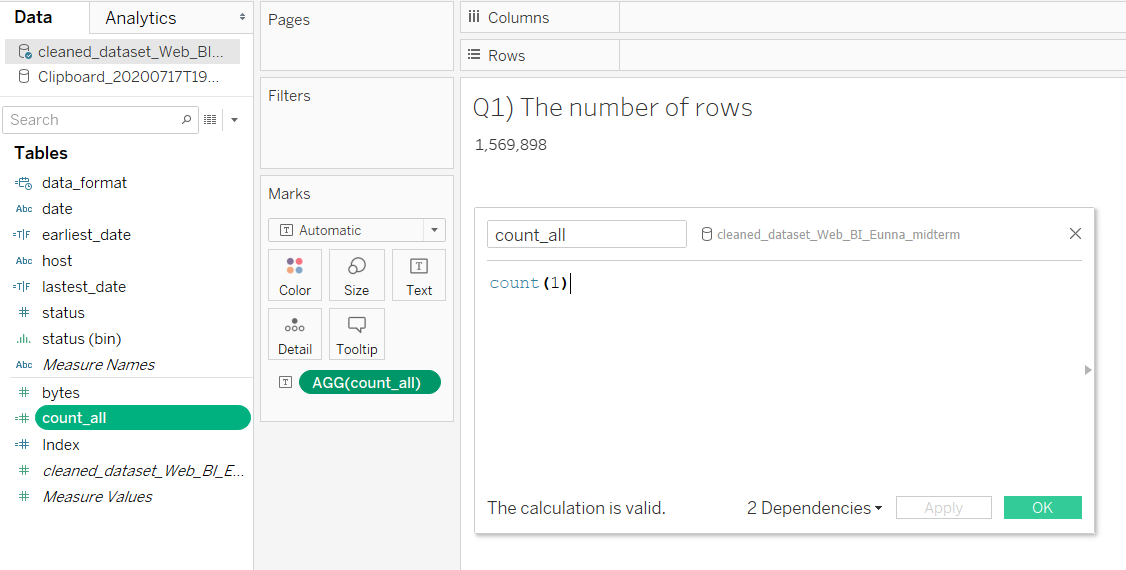
Step 02:

The second step was using Tableau and import the data:

In this step, I have used Tableau Desktop to visualize the data according to instruction.

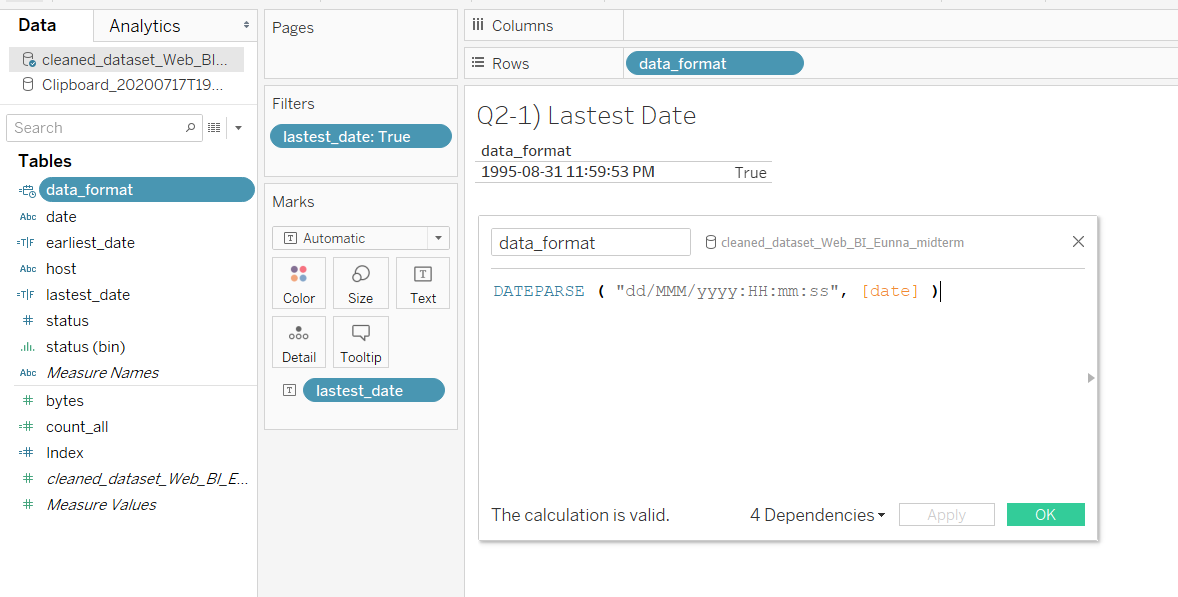
Q1 How many data records are in your final dataset?

I have created a new field as the Calculated field to count the number of rows each condition. For answering this question, this new field is used simple calculation without condition.

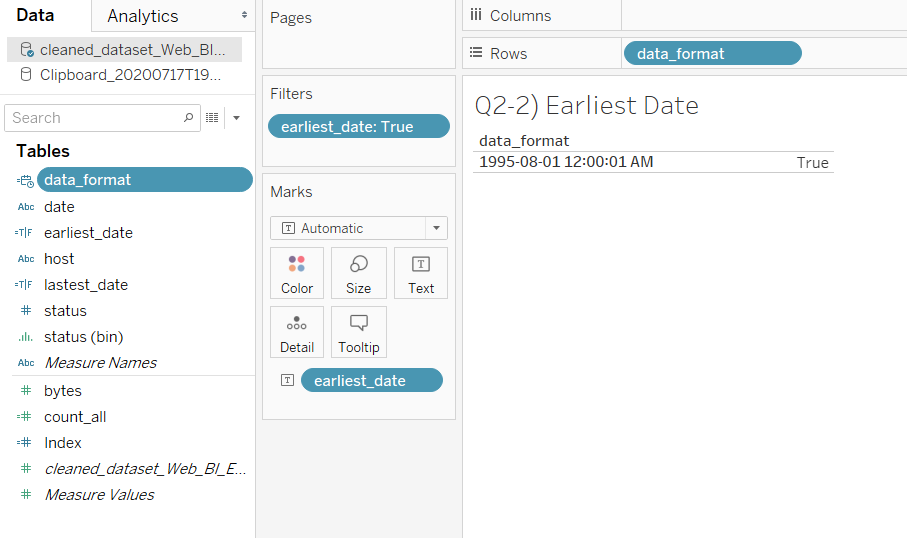


Q2. What is the timestamp for the earliest record in your dataset? What is the timestamp for the latest record in your dataset?

In this part, I have formatted date column because the original date column was string type from python, to sort, it needs to change to data type to date, so it used DATPARSE function with type “dd/MMM/yyyy:HH:mm:ss” for the latest date.

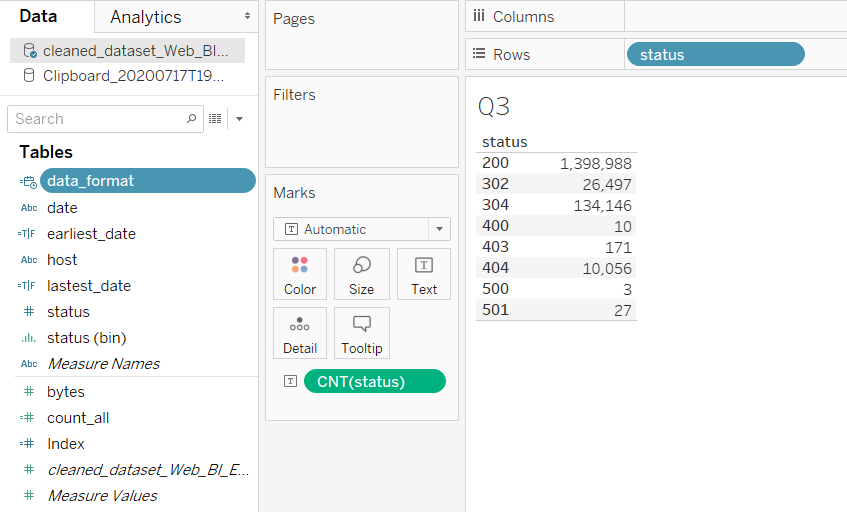


The earliest date.



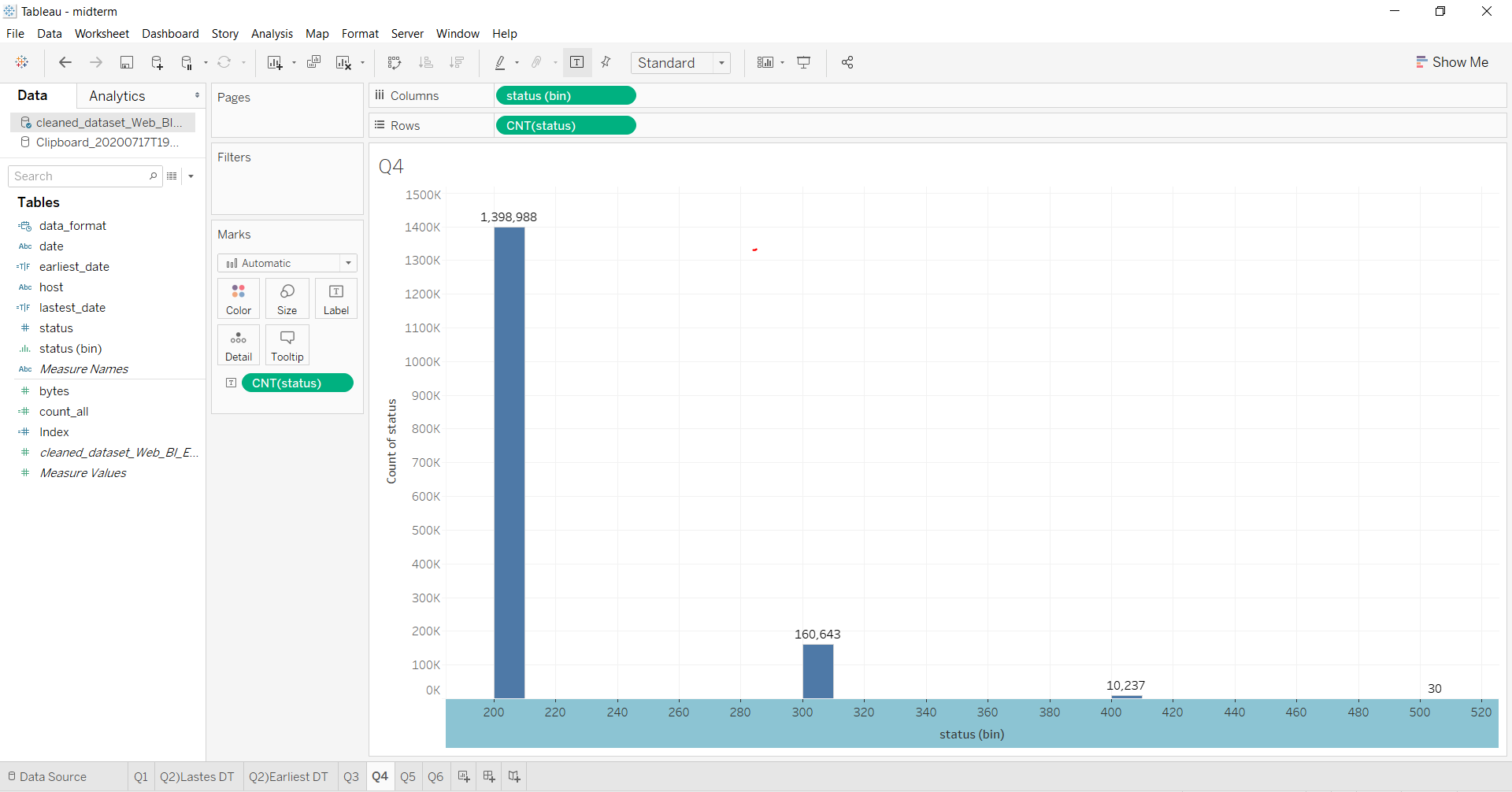
Q3. Create and display a frequency distribution table for the values contained in the “status” field.

Firstly, the status field is dragged in Row and make groups each status by using count function.

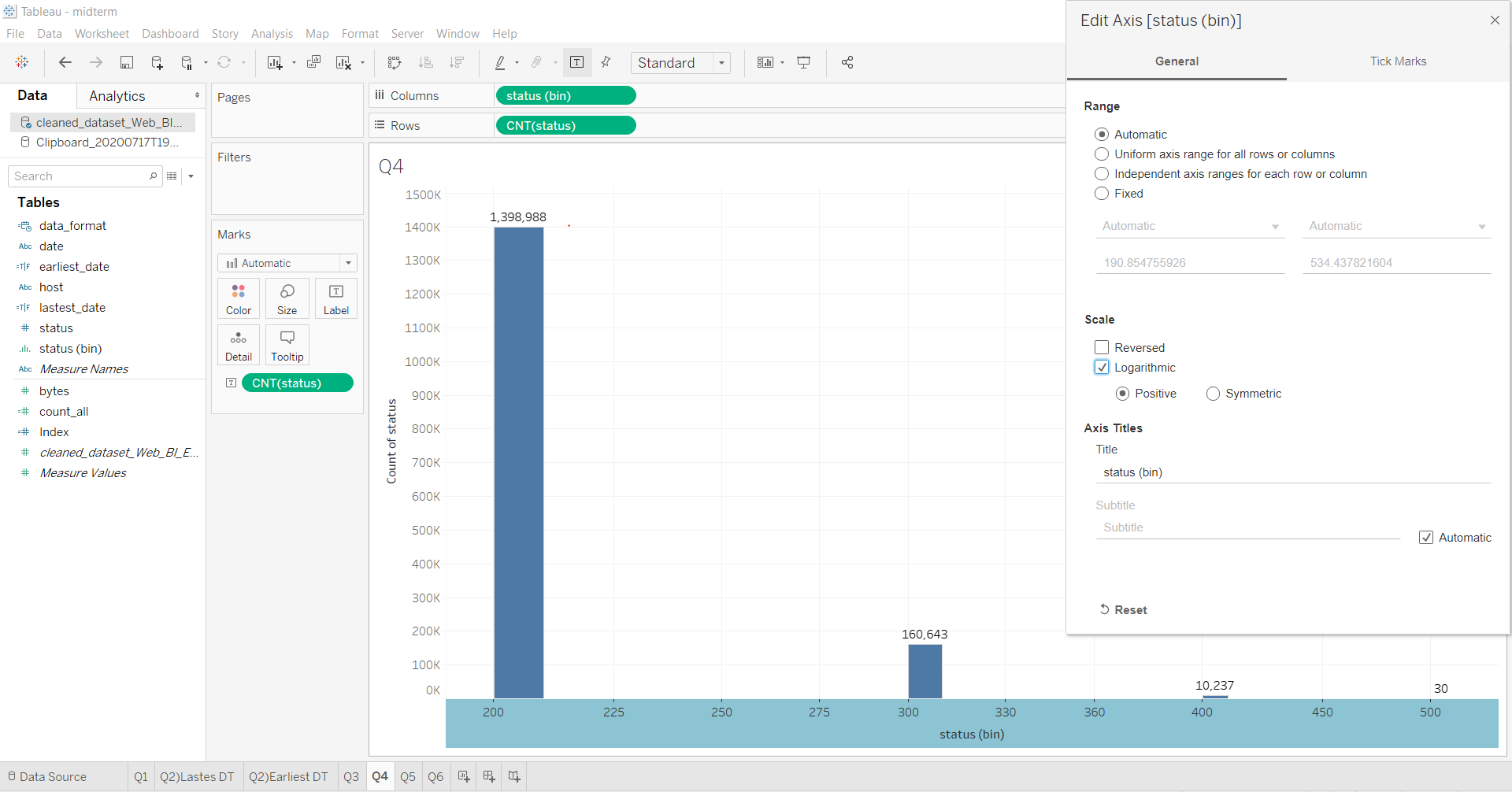


Q4. Create and display histograms for the previous frequency distribution table using both a standard scale and a log scale.

1. Standard Scale



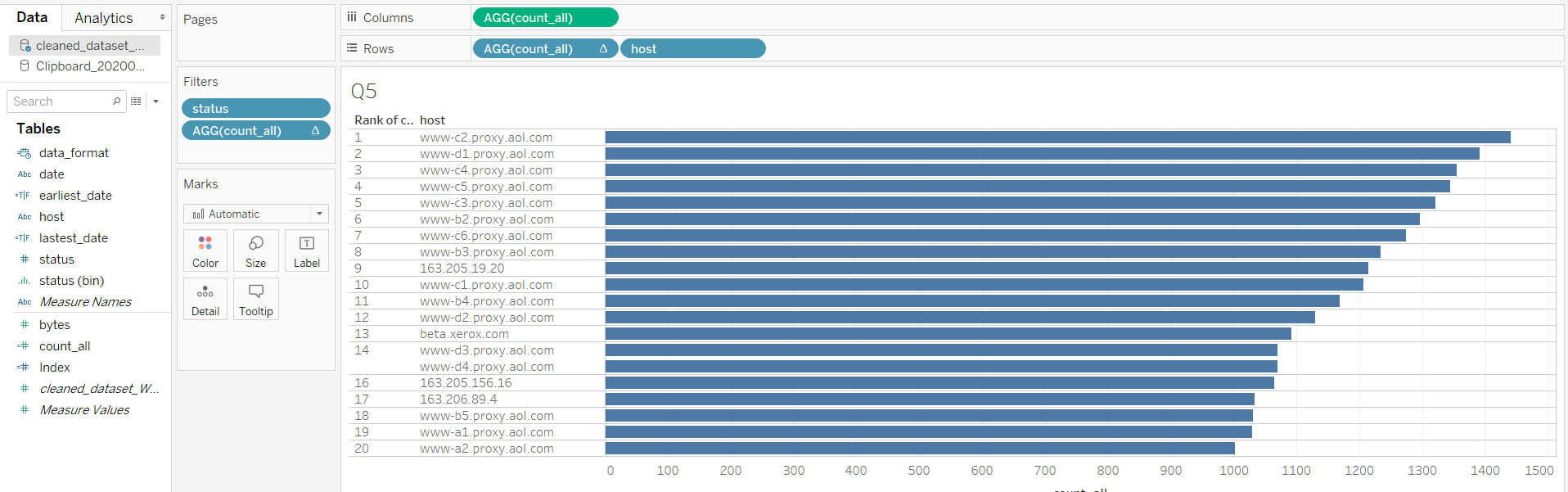
1. Log Scale



For this question, I have created bin field for status, so I could show the histogram. Setting Axis can show log scale.

Q5. Create a frequency distribution table for the endpoints that that resulted in an error. Display the results of this table using only the top 20 most frequent endpoints.

For answering this question, I have used the host and aggregation field, also the same fields in a filter. In filter, there is a condition of excepting 200 statuses. I have chosen rank function to show top 20 most frequent endpoints.



Q6. Using only the HTTP requests that resulted in a status of 200, calculate and report the following statistics for the content length.

a. minimum

b. maximum

c. mean

d. standard deviation

I have used just calculations for reporting this question. To identify the values and each requirement, Measure Names column distinguishes as name of conditions.