A Tutorial for the DAISEC Project

Fraud Detection Tutorial

Yegin Genc

KUs: Cyber Threats, Basic Data Analysis, Basic Scripting

Seidenberg School of Computer Science and Information Systems

Pace University

October 2019

Introduction: Explore Transactions Dataset for Anomalies

To detect anomalies in a collection of online transactions , we can explore them based on transaction details and certain flags provided by e-commerce providers. Some questions we can ask ourselves are:

- What kind of transactions are more likely to be fraudulent? (small vs big transactions)
- What is the amount of a typical transaction?
- What statistical measures can we use to identify fraudulent intent?
- Which one of these measures signals fraud better ?

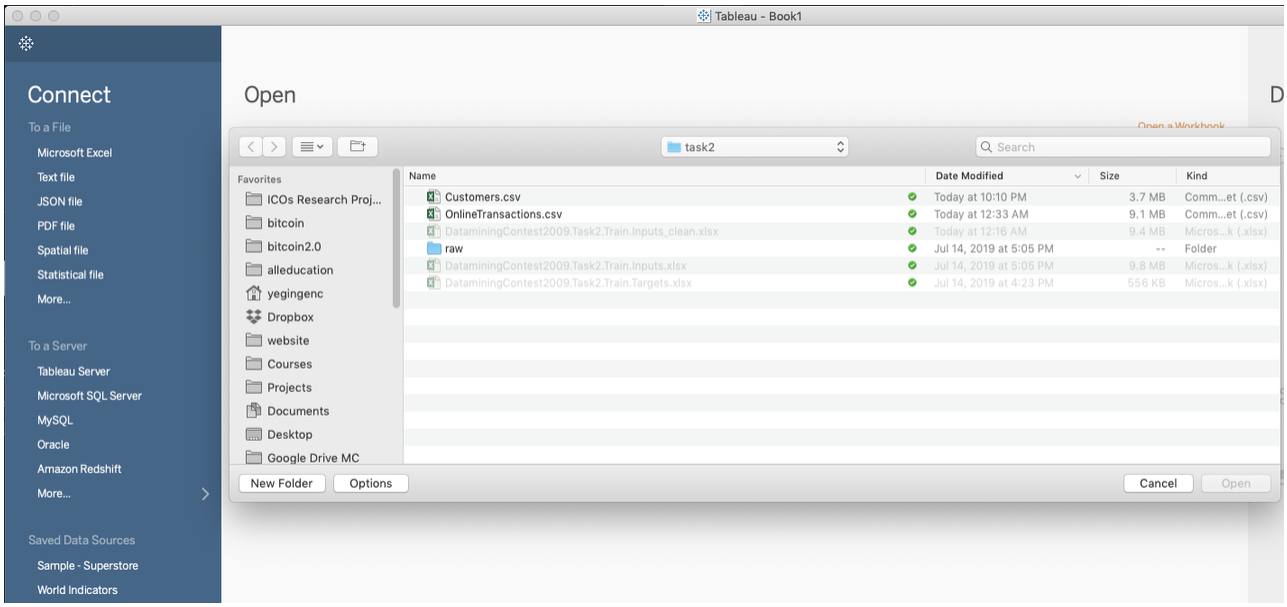
We can try to answer these questions by exploring the data with visualizations. For example, by creating a histogram of the transaction amounts, we can easily see how the amount variable is distributed. We can also see the distribution patterns by creating box plots.

This tutorial uses Tableau for exploratory analysis. You can get academic licenses following instructions at https://www.tableau.com/academic/students

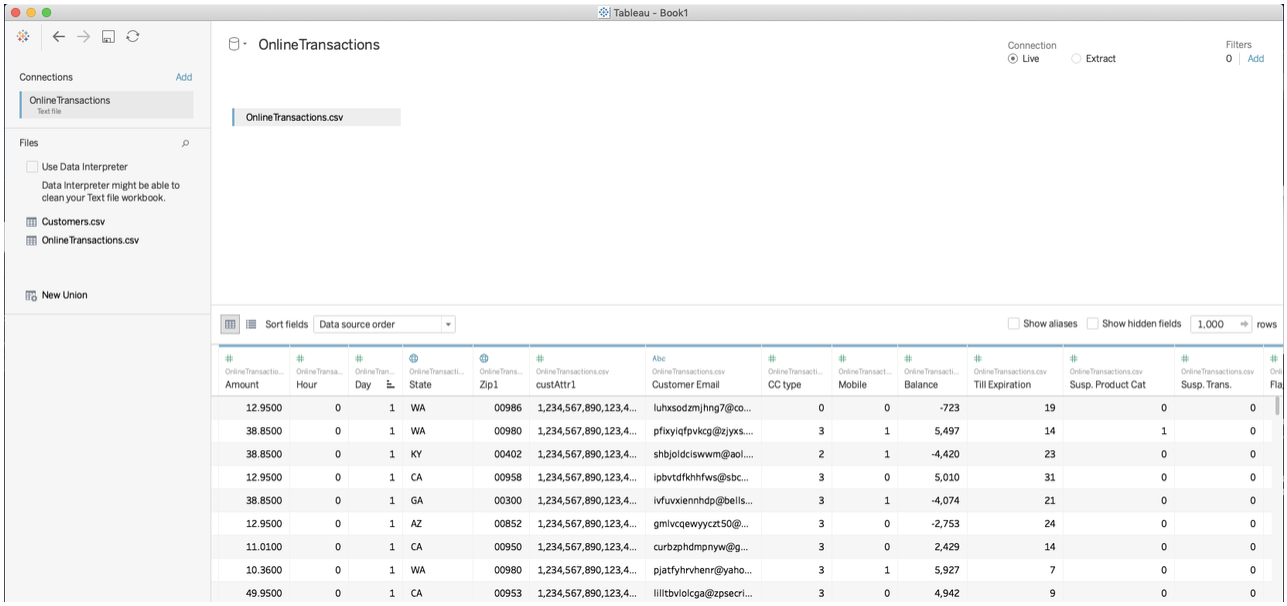
Loading Data

- 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 "**OnlineTransactions.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



Explore Transactions Amounts: Create a Histogram

In this section, we will explore if transaction amount and payment type (CC type) can help us detect fraudulent transactions.

- [ ] Start a new sheet and label it as Amount Histogram.
- [ ] Create a histogram that will look like Figure 1. (Hint You: can follow the instructions from Tutorial 2: Build a Histogram.)m

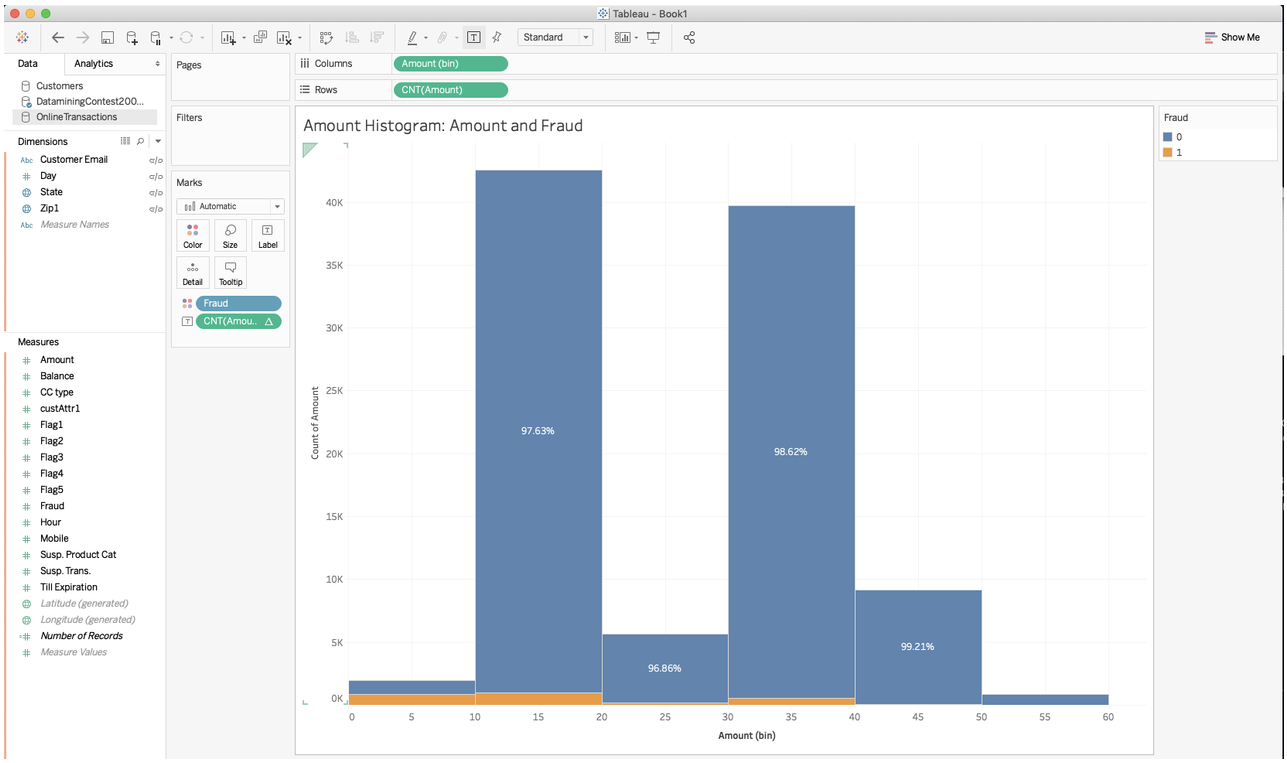


Figure 1

- [ ] This will be a histogram for the *Amount* variable, colors indicate fraud vs non-fraud transaction amounts.
- [ ] Add colors to show the *Fraud*.
- [ ] Add labels showing the percentages.
- [ ] Edit Y-axis by selecting *Logarithmic* under *Scale* (Figure 2).
- [ ] Finally, add panels such that each panel show the Credit Card Type (cc type) of the transactions. Final graph should look like Figure 3.

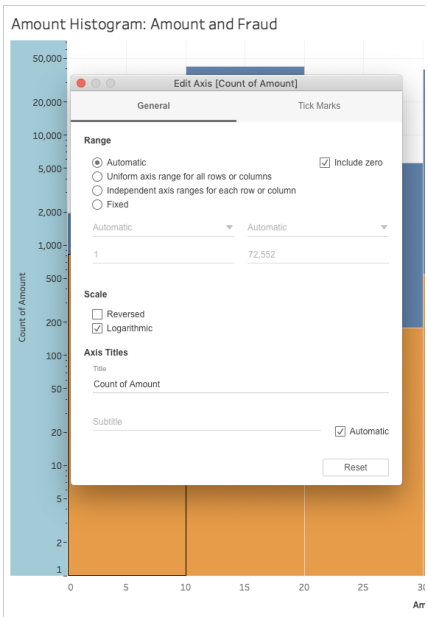


Figure 2

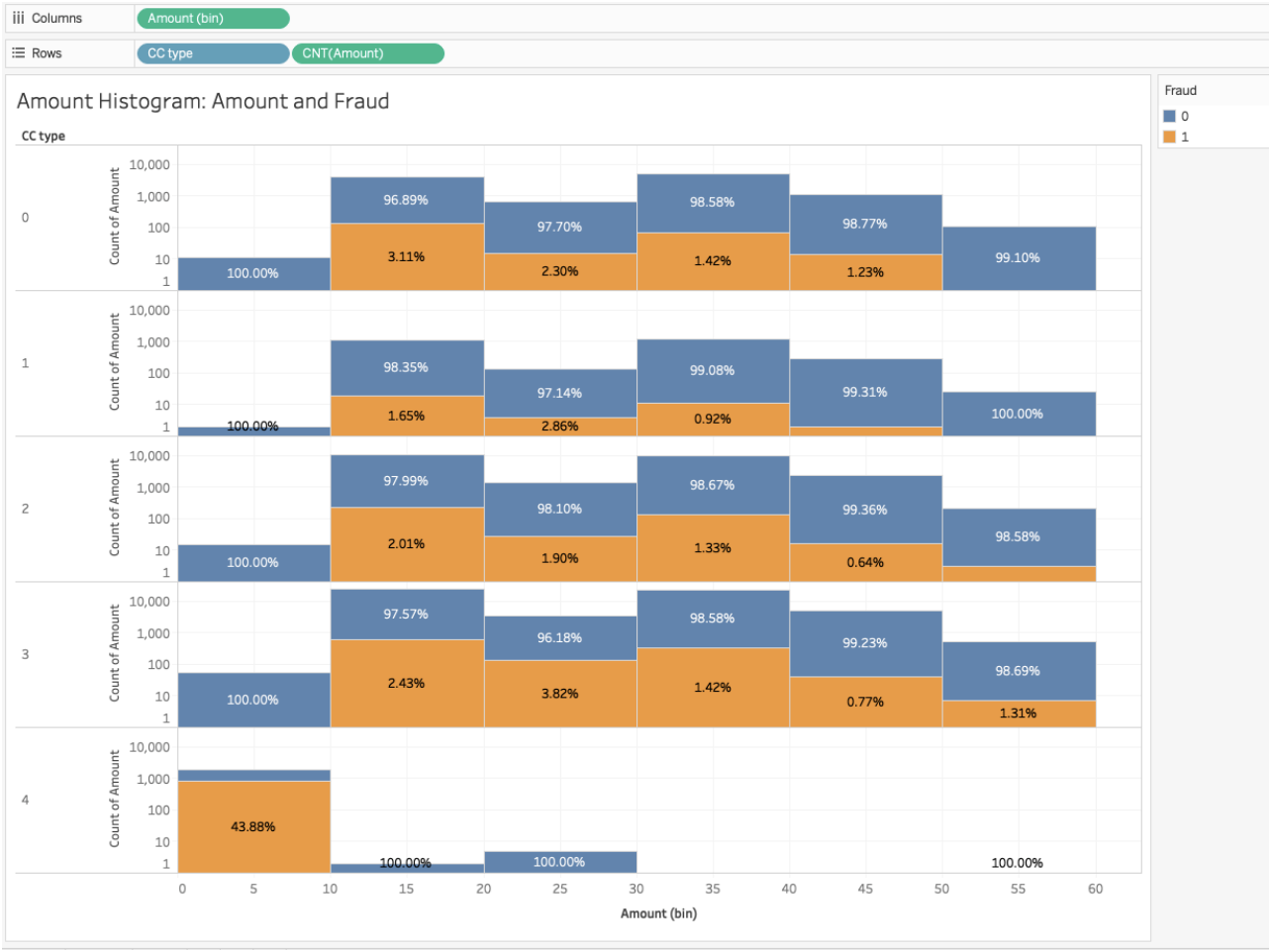


Figure 4

Histograms tell us how values vary in a dataset. We usually expect variables to be somewhat normally distributed (i.e. the bars are expected to follow a bell curve pattern). Here are some questions to consider that can help us understand nature of fraudulent transactions.

- What can you tell about the distributions in your graph? Do they resemble normal distributions?
- Does this view show any differences between transactions from different credit card types? If they are different, what can be the reason for the differences between subgraphs ?
- Finally, if we need to focus, what kind of transactions we need to pay more attention ?

Data exploration is usually an iterative process. You can continue exploring the data by checking the distributions when the data is divided into Panels by variables other than the CC Type.

Explore Account Balances: Create an Interactive Graph

In this section, we will explore if customer account balance at the time of transaction can help us detect fraudulent transactions.

- [ ] Create a new sheet and label it as Account Balances.
- [ ] Create a new histogram for *Balance* Measure (Figure 5a).
- [ ] Drag *Balance* Measure to Filter.
- [ ] Select *All values* in the popup menu, click *Next* and click *Ok* (Figure 5b).
- [ ] Right click on *Balance* and click *Show Filter* (Figure 5c). You should see a slider on the right pane.

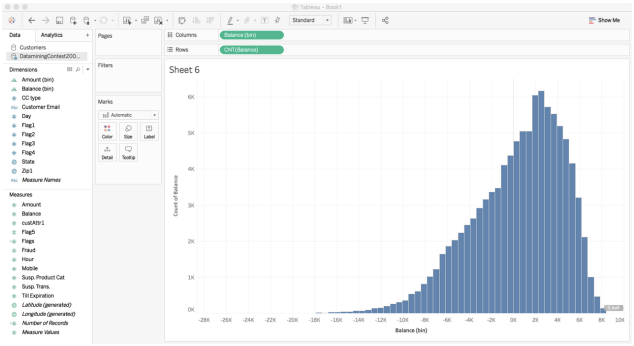


Figure 5a

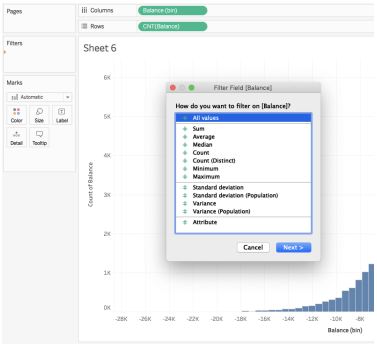


Figure 5b

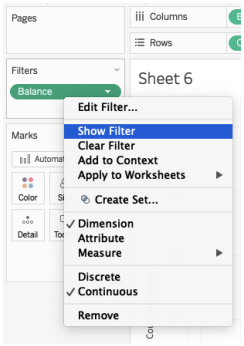


Figure 5c

- [ ] Drag *Fraud* measure to Color on Marks card.
- [ ] Right click on *Sum(Fraud)* that appeared, and change the calculation to average (Figure 6a).
- [ ] Click on the Color icon under Marks, click on Edit Colors, and change the Palette to Temperature Diverging (Figure 6b).

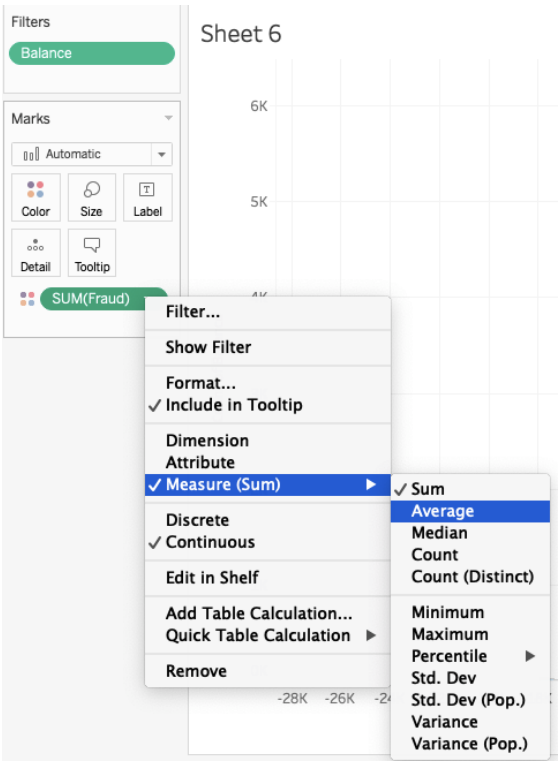


Figure 6a

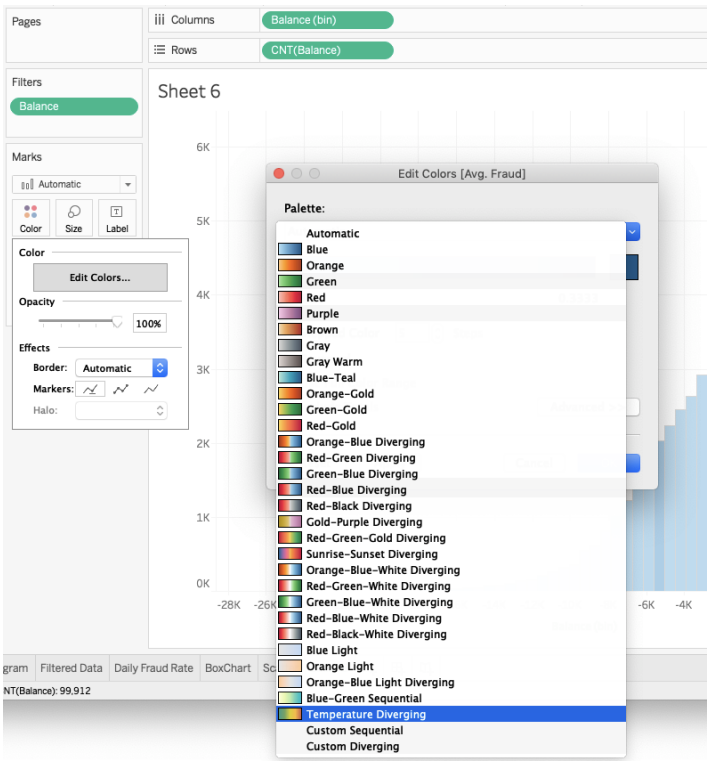


Figure 6b

- [ ] Drag Balance Measure under Filters, right click on it and select Show Filter.
- [ ] Using the balance slider, find the balance intervals that have on average more fraudulent transactions than others (see Figure 7)

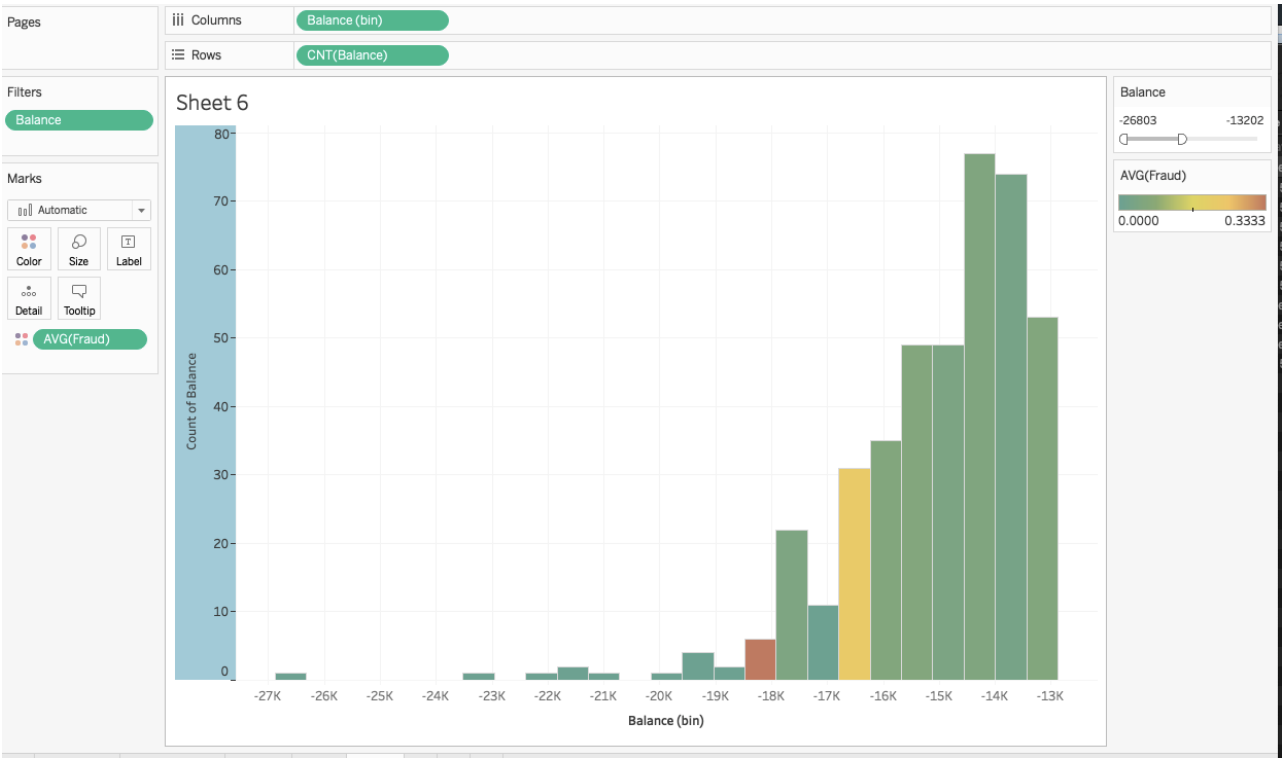


Figure 7

Here are some questions to consider that can help us further understand nature of fraudulent transactions.

- Is there a balance range that has a significant higher fraudulent transaction rate? Can you think of any reason why this range has more fraudulent transactions ?

Daily Fraud Rate: Create Fraud Anomaly Detection Graph

In this section, we will explore if there are particular days with significantly high fraudulent transaction rates.

- [ ] Start a new sheet and label it as Daily Fraud Rate.
- [ ] Drag the *Day* measure under Dimensions. This way we can aggregate *Number of Records* and number of Fraudulent transactions per day. (Another option is to drag Day from Measures to Columns, right click and check *Dimension*)
- [ ] Drag the *Fraud* measure to Rows. Results should look like Figure 8. The default function that is applied on Fraud variable is sum. Since the variable value is 1 for fraudulent transaction and 0 otherwise, the result of the sum shows the total number of fraudulent transactions. Hence, this graphs shows number of Fraudulent transactions per day.

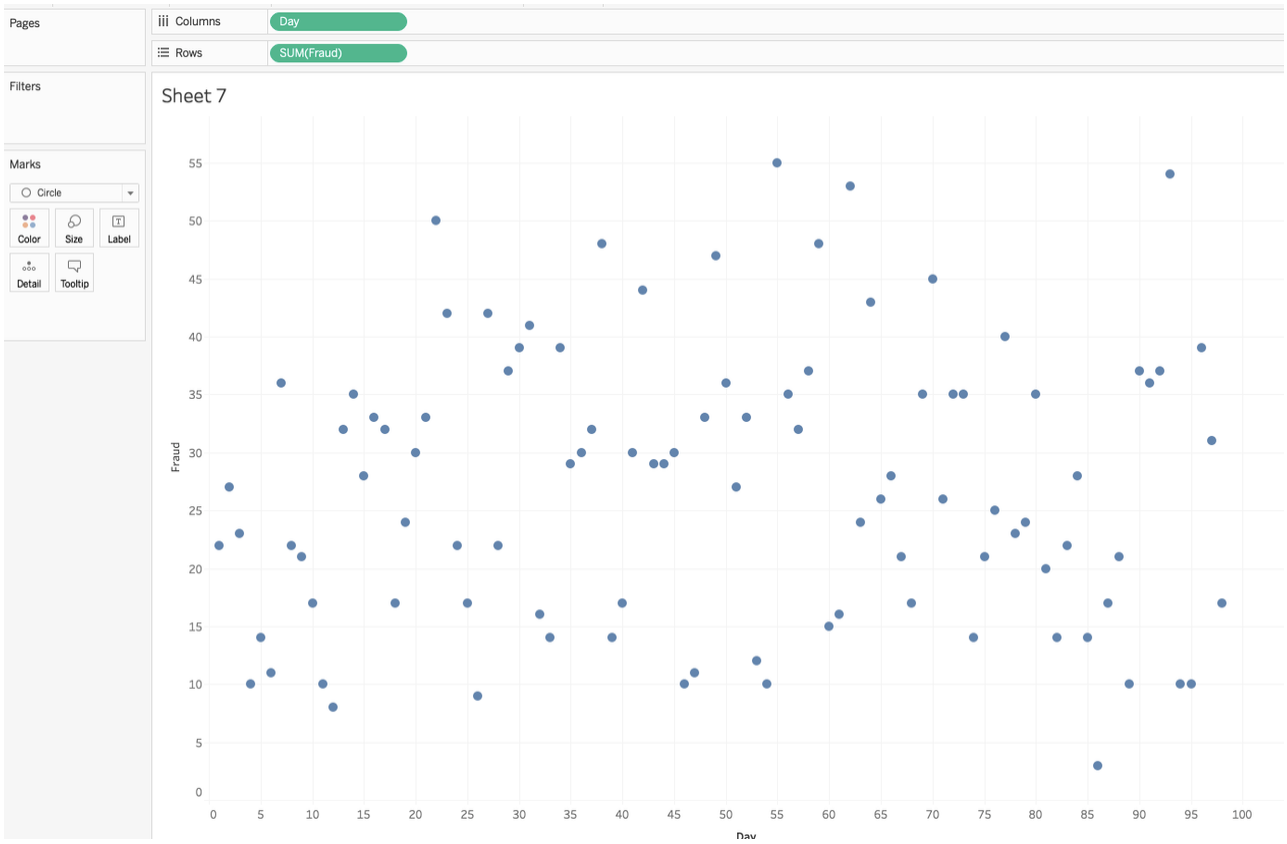


Figure 8

- [ ] Next we will update the row calculation to reflect the percentage of transactions that are fraudulent. To that end, right click on *Sum(Fraud)* under Rows and select Edit in Shelf.
- [ ] Update the formula to the following.

```
SUM([Fraud]) / SUM([Number of Records])
```

Number of records is a dummy variable that is set to 1 in each cell. By applying the sum function, we get the number of records (transactions) per day. As a result of this formula, each dot shows the percentage of fraudulent transactions per day .

- [ ] You can change the Y axis label by right clicking and selecting Edit Axis (Figure 9)
- [ ] Drag *Number of Records* measure to *Color* on *Marks* card. This way, the color of the dots will indicate the number of the transactions that day. Final result should look like (Figure 10)

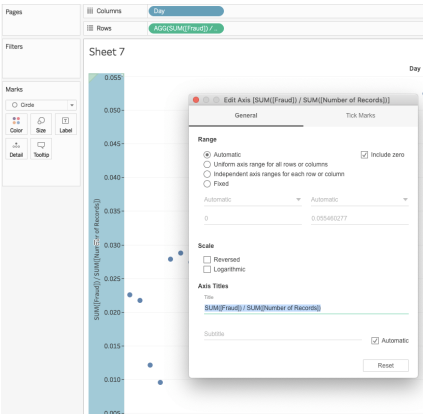


Figure 9

Figure 10

- [ ] Click on the Analytics tab and drag the Distribution Band option on the graph.
- [ ] Select *Scope > Entire Table; Computation > Value Standard Deviation;* set Factors as -2, 2 Standard Deviation (Figure 11 a); and click OK. Final graph should look like Figure 11b

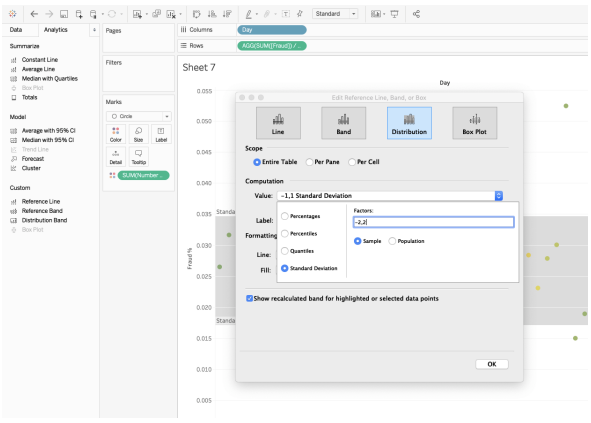


Figure 11 a

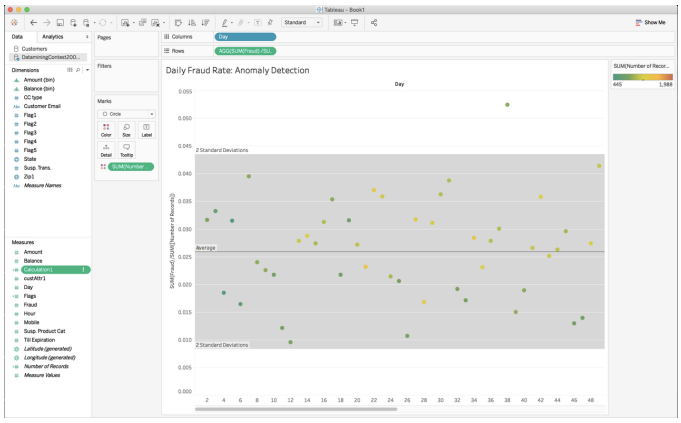


Fig 11 b

Here are some questions to consider that can help us further understand nature of fraudulent transactions.

- Transactions usually follow a recurring pattern. For example, for many businesses weekends have repeatedly lower number of transactions than weekdays; or,

summers usually have lower sales rates. These effects are often regular and predictable. We call this seasonality. Do we see similar patterns for fraud rates? (Sometimes the answer is no.)

Fraud rates by State: Build a Simple Map

https://onlinehelp.tableau.com/v2018.3/pro/desktop/en-us/maps_howto_simple.htm

You can build graphs on maps as long as your data has location indicators. In our dataset, the *state* variable can be used to create location based graphs.

- [ ] Highlight *State* under *Dimensions*, and select maps option under *Show Me*
- [ ] Drag the *Fraud* measure to *Color* on *Marks* card.
- [ ] Drag the *Number of Records measure to Size* on the *Marks* card.
- [ ] Click on Marks >Size and increase the dot size using the slider.
- [ ] Change the color palette to Temperature Diverging. Final results should look like Figure 12.

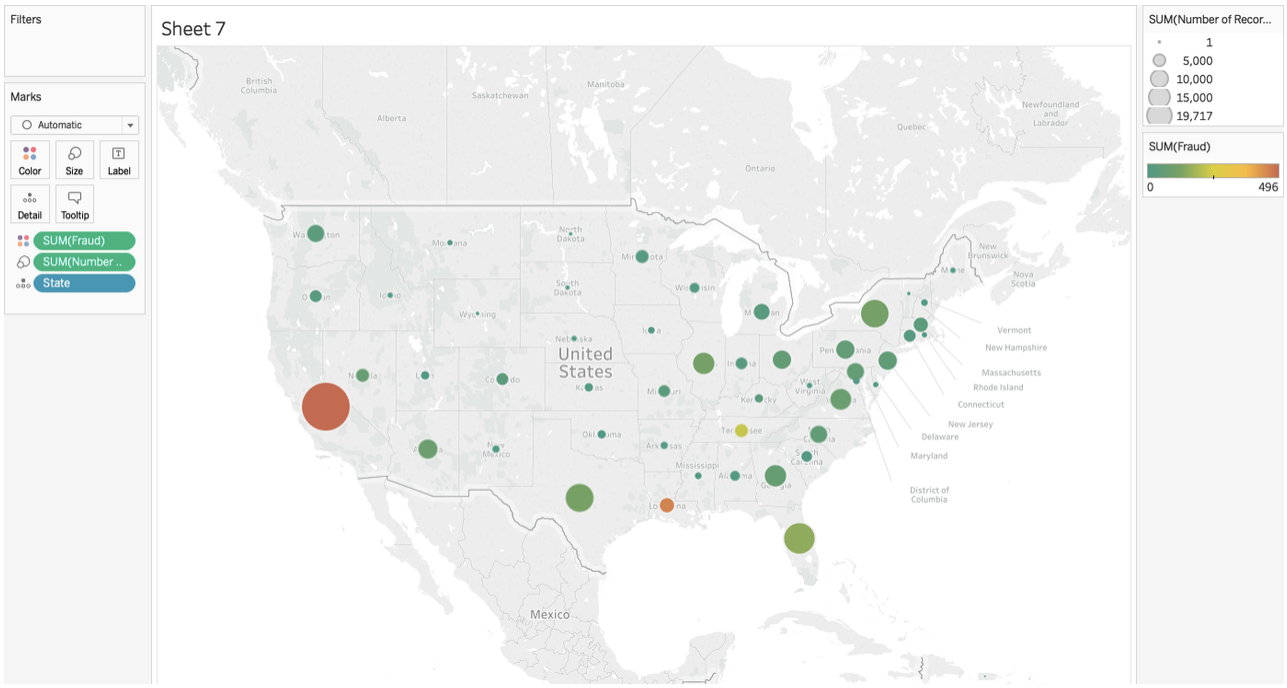


Figure 12

The graph shows the number transactions (dot size) and the number of fraudulent transactions (green: low, red: high). We expect larger size dots to be closer to red

than to green. That is, in the states with higher number of transactions, we expect to see more fraudulent transactions.

Here are some questions to consider that can help us further understand nature of fraudulent transactions:

- Which state(s) confirm this assumption ? and which states do not fit (anomalies) ?

Adding Demographic Information: Expanding Map Graph

We can update the dot size to show proportion of the fraudulent transactions.

- [ ] Right click on *Sum(Fraud)* under *Marks* and select *Edit in Shelf* (Figure 13)
- [ ] Update the equation as

```
SUM([Fraud])/SUM([Number of Records])
```

- [ ] Remove *SUM(Number ..* on Marks card. The graph should look like Figure 14a.
- [ ] Fraudulent transaction proportions are represented with color marks. Drag *AGG(SUM([Fra..* option to *Size* on *Marks card.* This way, the same property is shown with dot size (Figure 14b).

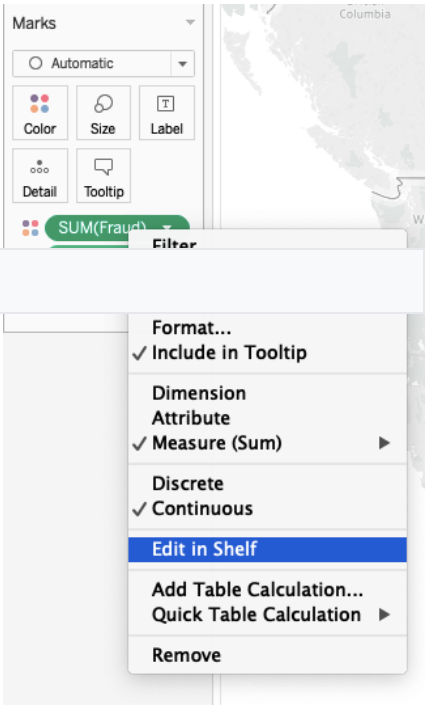


Figure 13

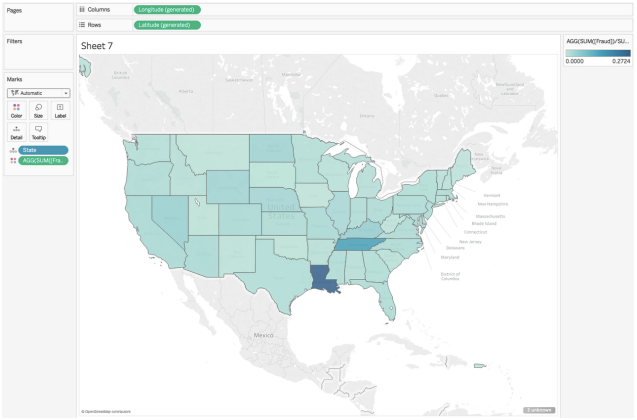


Figure 14a

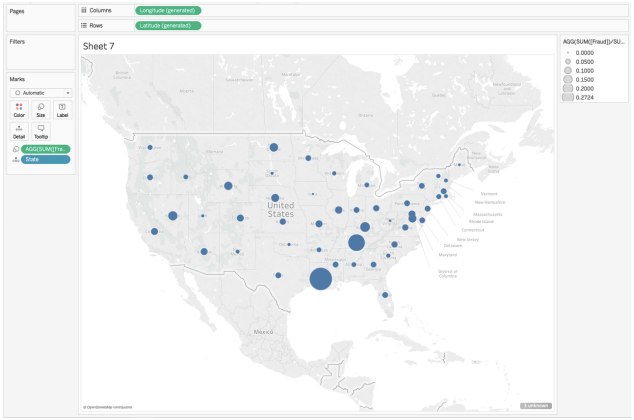


Figure 14b

- [ ] From the top menu, select Map > Map Layers.
- [ ] Under the *Map Layers* tab, expand the *Layer* dropdown and select *Per Capita Income (Figure 15).*

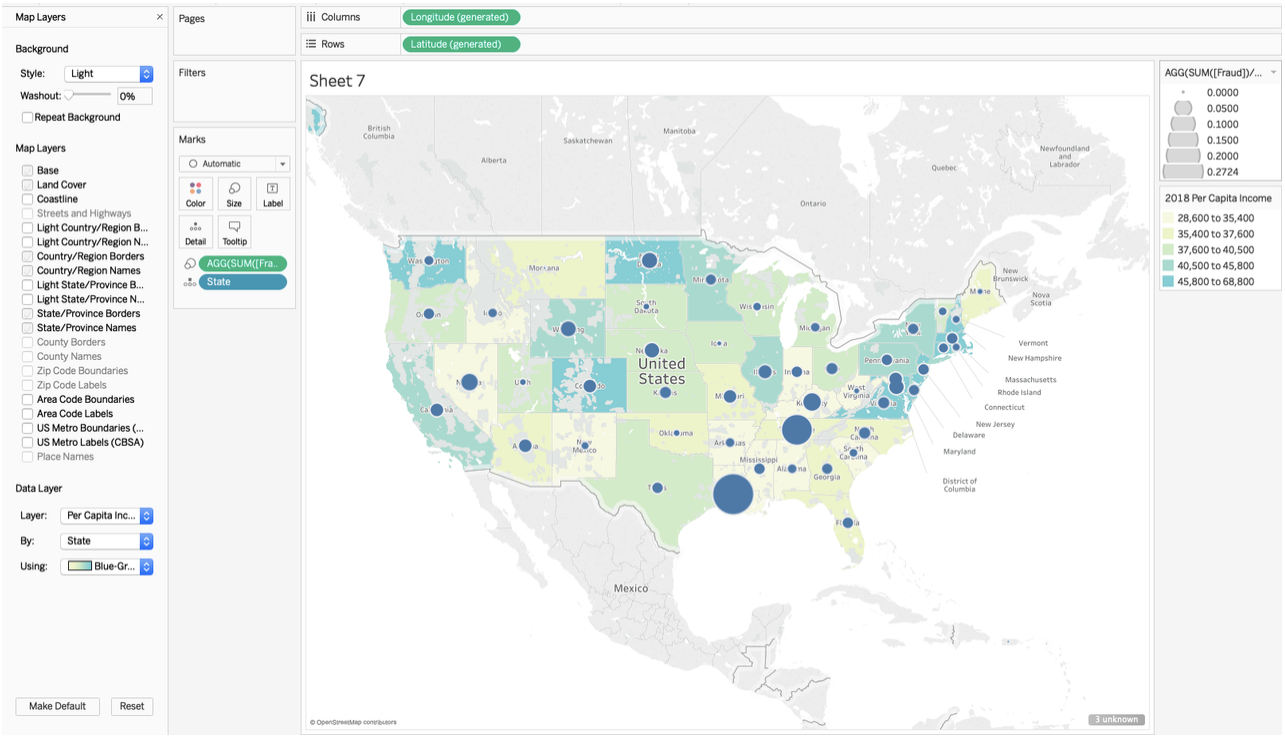


Figure 15

- Can you see any relationship between per capita income and fraud rates per states?

Create a Calculated Field

https://onlinehelp.tableau.com/current/pro/desktop/en-us/calculations_calculatedfields_formulas.htm

Although fraudulent transaction ratio (Fraud Ratio) is not a direct measure that can be recorded as a variable, we often need it in our graphs. We can store it as a calculated measure to reuse in multiple graphs.

- [ ] Right click under the *Measures* tab and select *Create Calculated Field*
- [ ] On the pop up menu, rename the field to Fraud Ratio.
- [ ] Type in the following formula and click ok.

```
Sum([Fraud])/ Sum([Number of Records])
```

Fraud Rates by the Hour of the Day: Building a Combination Chart

https://onlinehelp.tableau.com/current/pro/desktop/en-us/qs_combo_charts.htm

Combination charts are views that use multiple mark types in the same visualization. For example, you may show total number of Transaction per hour as bars and add Fraudulent Transaction Rate with a line across the bars.

- [ ] Drag the *Hour measure* to *Dimensions*
- [ ] Then, drag the *Hour dimension* to *Columns shelf.*
- [ ] Drag the *Fraud Ratio* measure to *Rows.* The graph shows Fraudulent Transaction Ratios for the hour of the day.
- [ ] Drag the *Number of Records measure* to *Rows shelf.* You should see two graphs that show the hourly break down for Fraud Ratio and Number of Records.
- [ ] On the Rows shelf, right click on *Sum(Number of Records)* and select Dual Axis (Figure 16a). The results should look like Figure 16b

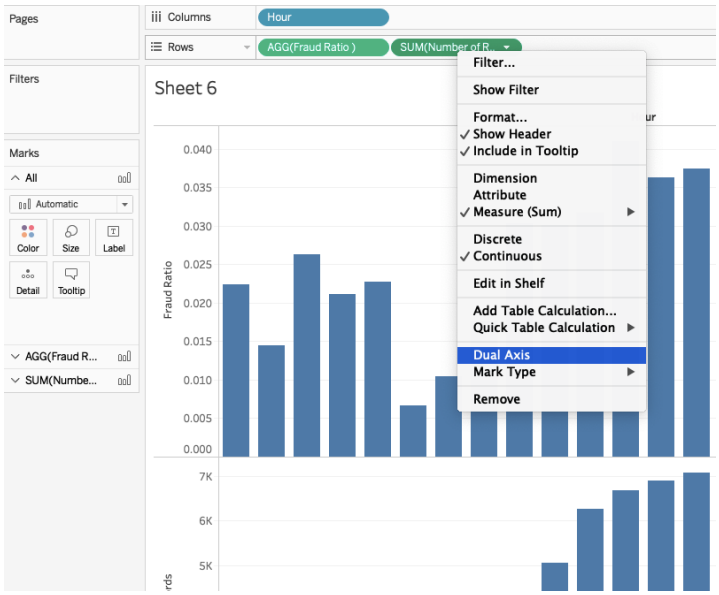


Figure 16a

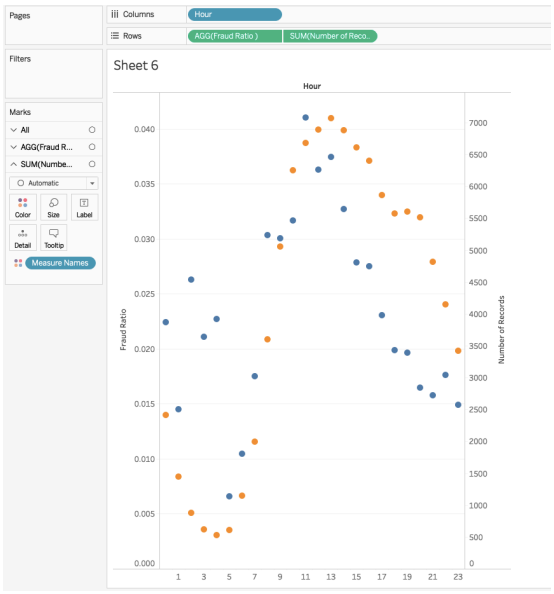


Figure 16b

- [ ] Click on *Agg(Fraud Ratio)* under Marks card, and change the graph type from Automatic to Line (Figure 17a)
- [ ] Click on Sum(Number of Records) under Marks card and change the graph type from Automatic to Bar (Figure 17b)
- [ ] On the Rows shelf, move *Sum(Number of Records)* before *Agg(Fraud Ration)* to move the bars behind the line chart. Final result should look like Figure 17c

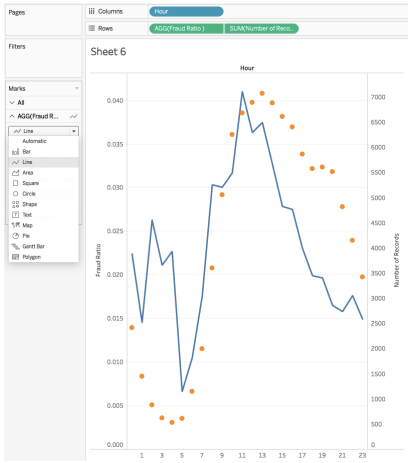


Figure 17a

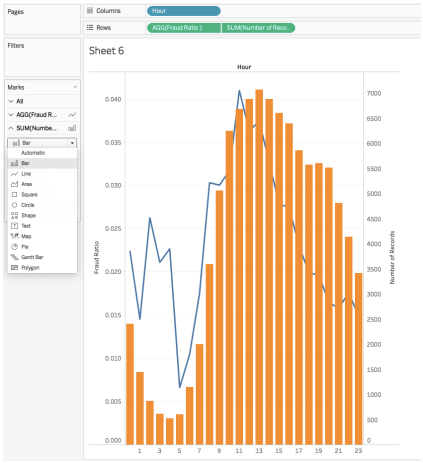


Figure 17b

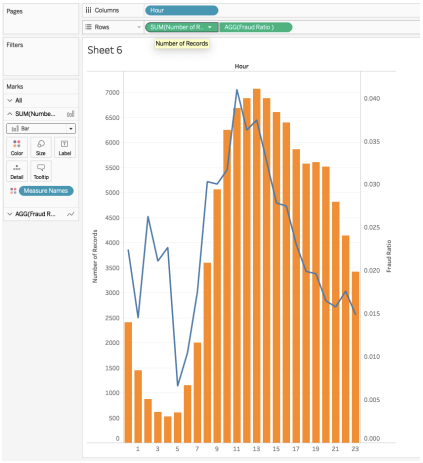


Figure 17c

- [ ] Drag the *Mobile* measure to *Dimensions section*
- [ ] Then drag Mobile from *Dimensions* to Rows shelf, in front of *Sum(Number of Records)* (Figure 18)

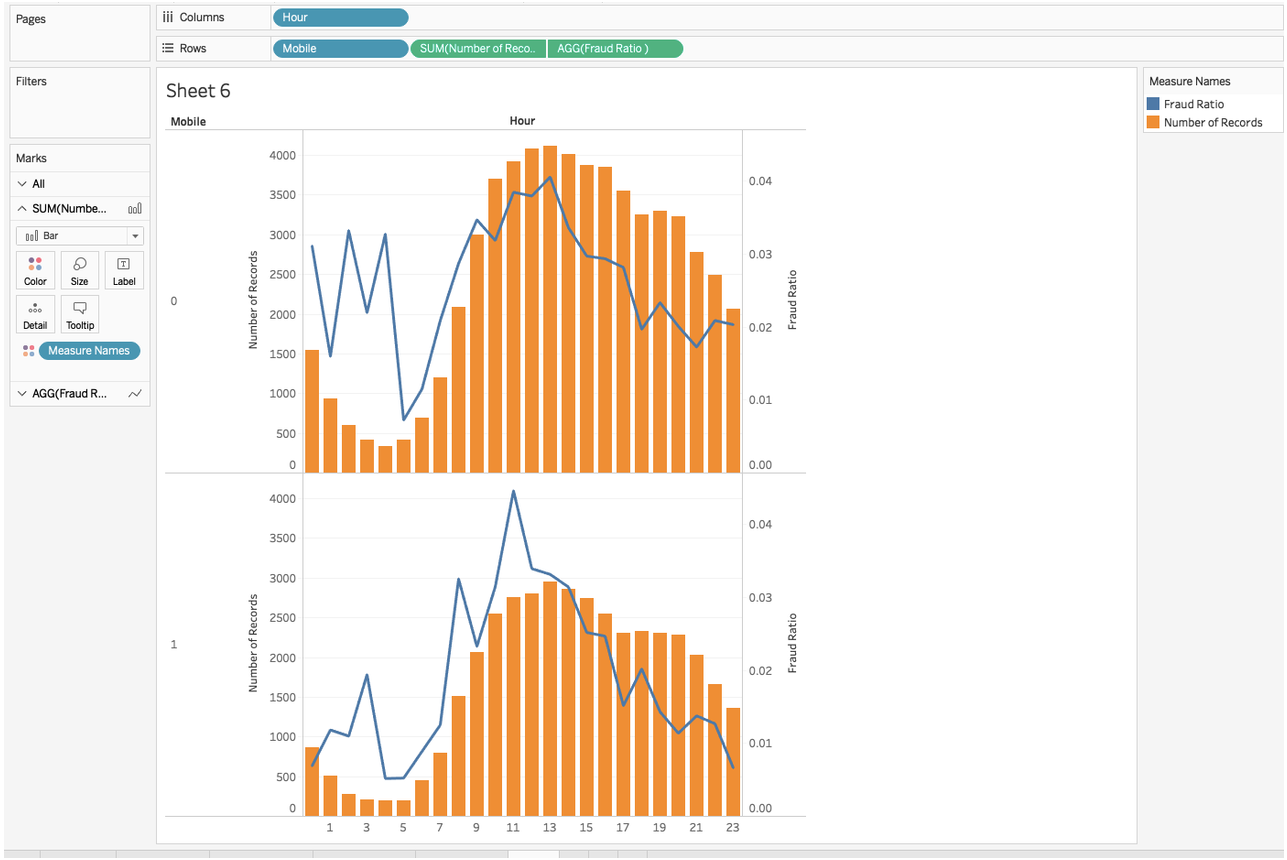


Figure 18

Here are some questions to consider that can help us further understand nature of fraudulent transactions:

- What time of the day has higher fraud rates?
- Are the ratios the similar for both mobile and desktop transactions? What can potentially explain the difference ?

Finding Malicious Domains: Building a Bubble Chart

We will use *Customers.csv*. This dataset is a report generated from all transactions by summarizing customer details. Customer email addresses are used as unique identifiers. We are going to find if we can identify email providers to which fraudulent transactions owners are subscribed. To that end, we will first have to extract the domain name from the user email variable

- [ ] Load Customers.csv file as a text file.

- [ ] Create a new sheet and label it as Domains.
- [ ] Under Data tab select Customers
- [ ] Under *Dimensions* tab, right click and select *Create Calculated Field.*
- [ ] On the pop up menu, set the label to *Domain* and below input the following formula

```
SPLIT([Customer], '@',2)
```

Next we will create another calculated variable called *Fraud Ratio*. We created the same field while working with Transactions. We need to create the same variable for Customers dataset.

- [ ] Under *Measures* tab, right click and select *Create Calculated Field.*
- [ ] On the pop up menu, set the label to *Fraud Ratio* and below input the following formula and select OK

```
sum([Fraud])/sum([Number of Transactions])
```

- [ ] Drag *Domain* dimension to *Columns* and in the popup message select *Add all members.*
- [ ] Drag *Fraud* measure to Rows shelf.
- [ ] Under Show Me select bubble chart icon (bottom right). The graph should look like Figure 19.

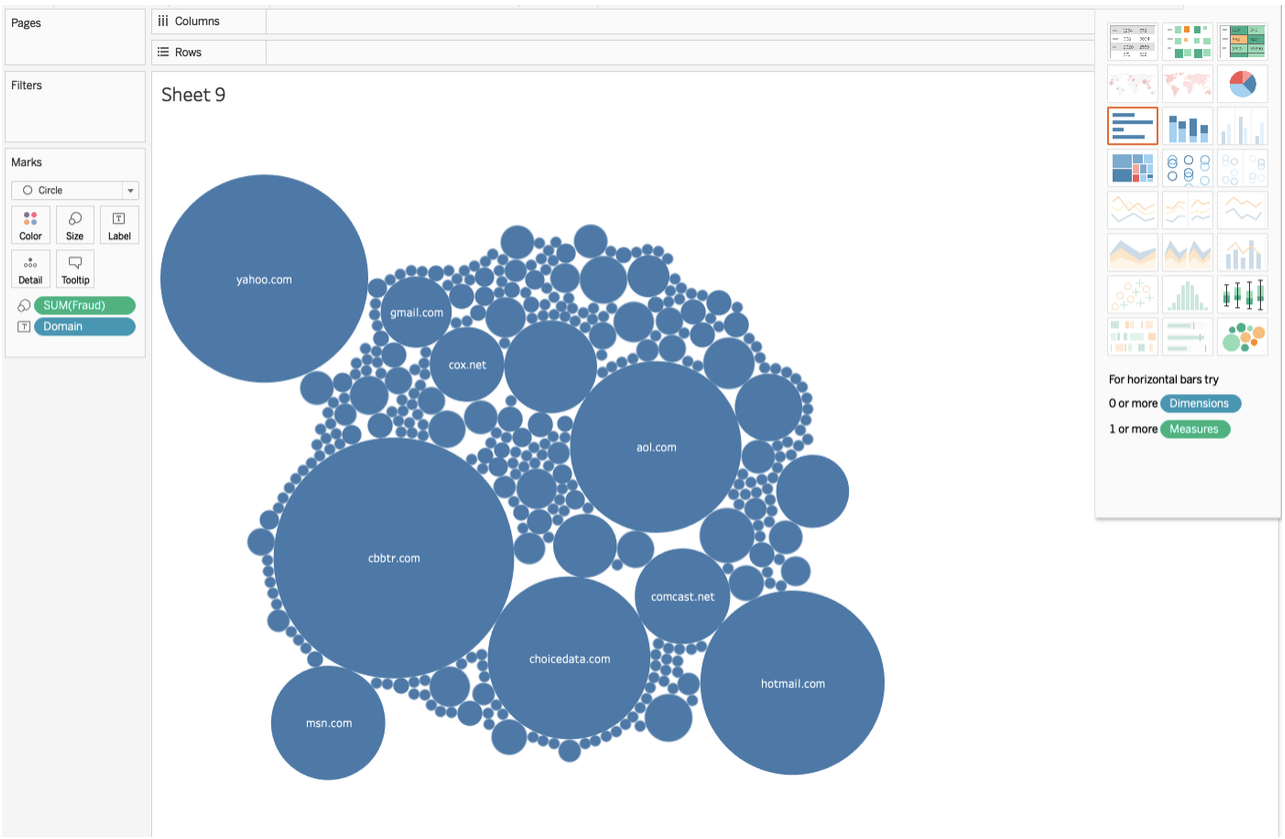


Figure 19

The circle sizes indicate the number of fraudulent transactions with email accounts originated from each domain. However it makes sense for domains like “yahoo” or “hotmail” to have more fraudulent transactions since they are larger email providers and they account for more transactions in general. We will need to add color to show the percentage of fraudulent transactions (Fraud Ratio)

- [ ] Drag *the Fraud Percent* measure to color on Mark. The graph should look like Figure 20

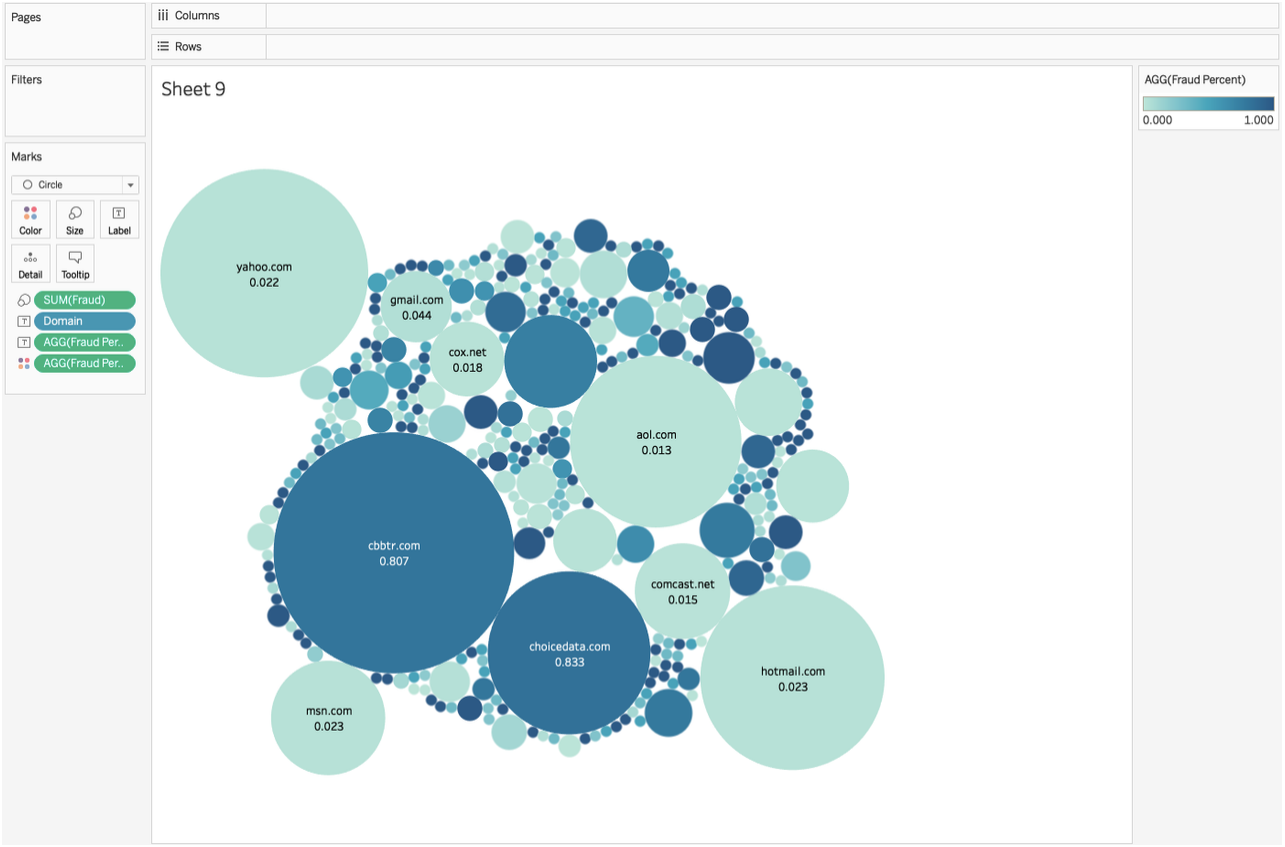
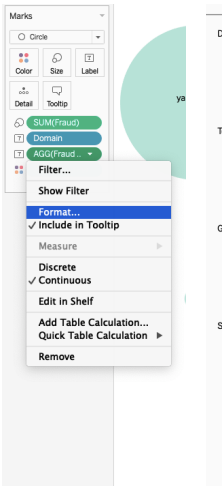
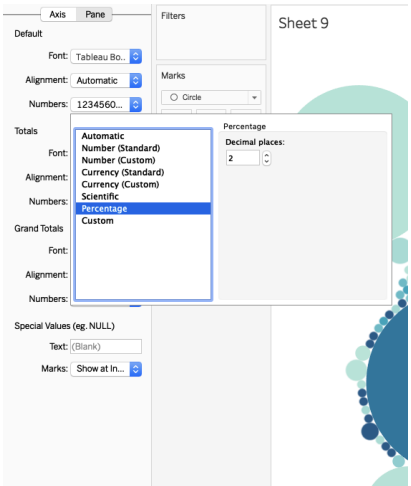


Figure 20

- [ ] Right click on *Agg(Fraud Percentage)* Text Mark (Figure 21a)and select Format.
- [ ] Change *Numbers* formatting to *Percentage* (Figure 21b) and the final graph should look like Figure 21 c.





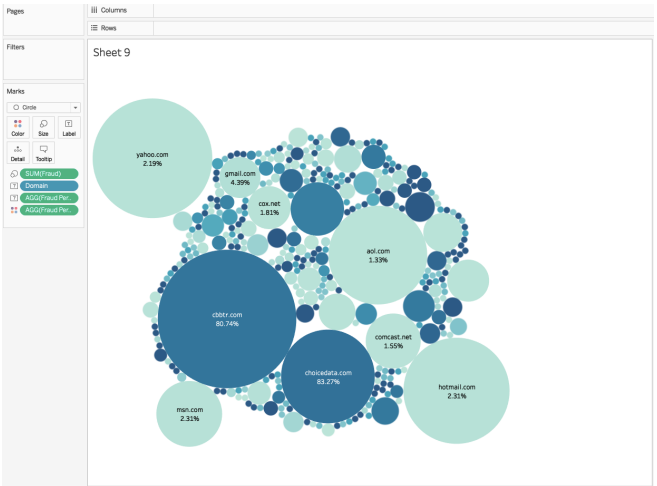


Figure 21a *Figure 21b* *Figure 21c*

Acknowledgement

The author(s) would like to acknowledge the support from the National Security Agency and the National Science Foundation under Grant No. H98230-19-1-0240. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Security Agency, National Science Foundation or the U.S. government.