

Final report

Final Project



March 17, 2022

William Jackson, Julie nguyen, matt wehnert

CSI 475

Introduction (Matt, Julie, Jackson)

# Question

What can we learn about how Europe and USA handled COVID-19? Who handled it better and who had the most vaccinations?

# Hypothesis

We believe that Europe has the most vaccination counts than the U.S. because their emergency protocol called for immediate vaccinations regardless of status.

# Intention

Seeing the comparison between both datasets and understanding the implications.

# Prerequisites

1. Must have compatible datasets for comparison.
2. Must have completed data preparations and ETL process, please refer to lab 3 document.
3. Must have basic understanding of Tableau features and functions, please refer to lab 4 document.
4. Must have somewhat adequate comprehension of LOD and table calculations, please refer to lab 5 document (not required in this lab but is a bonus).
5. Must have basic understanding of R programming language for scripting.
6. Must have completed all the above steps leading to this final procedure.

Procedure (Matt, Julie, Jackson)

1. Log into MySQL with your credentials to enable the connection to Tableau.
2. Boot up RStudio then type in and run each of the following commands:

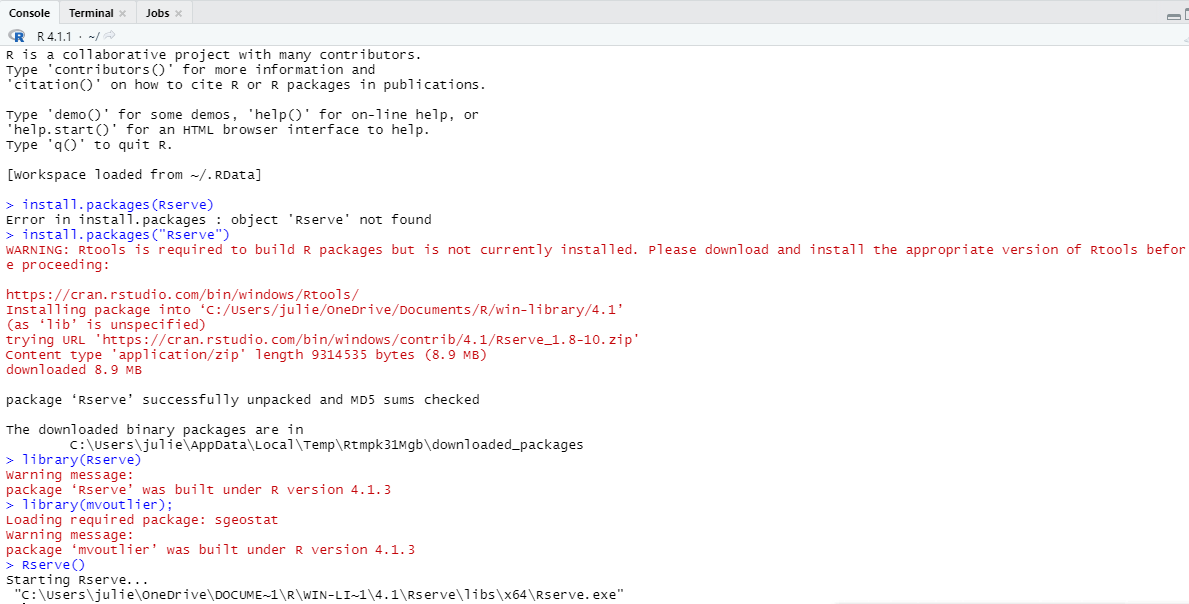
install.packages(“Rserve”)

library(Rserve)

Rserve()

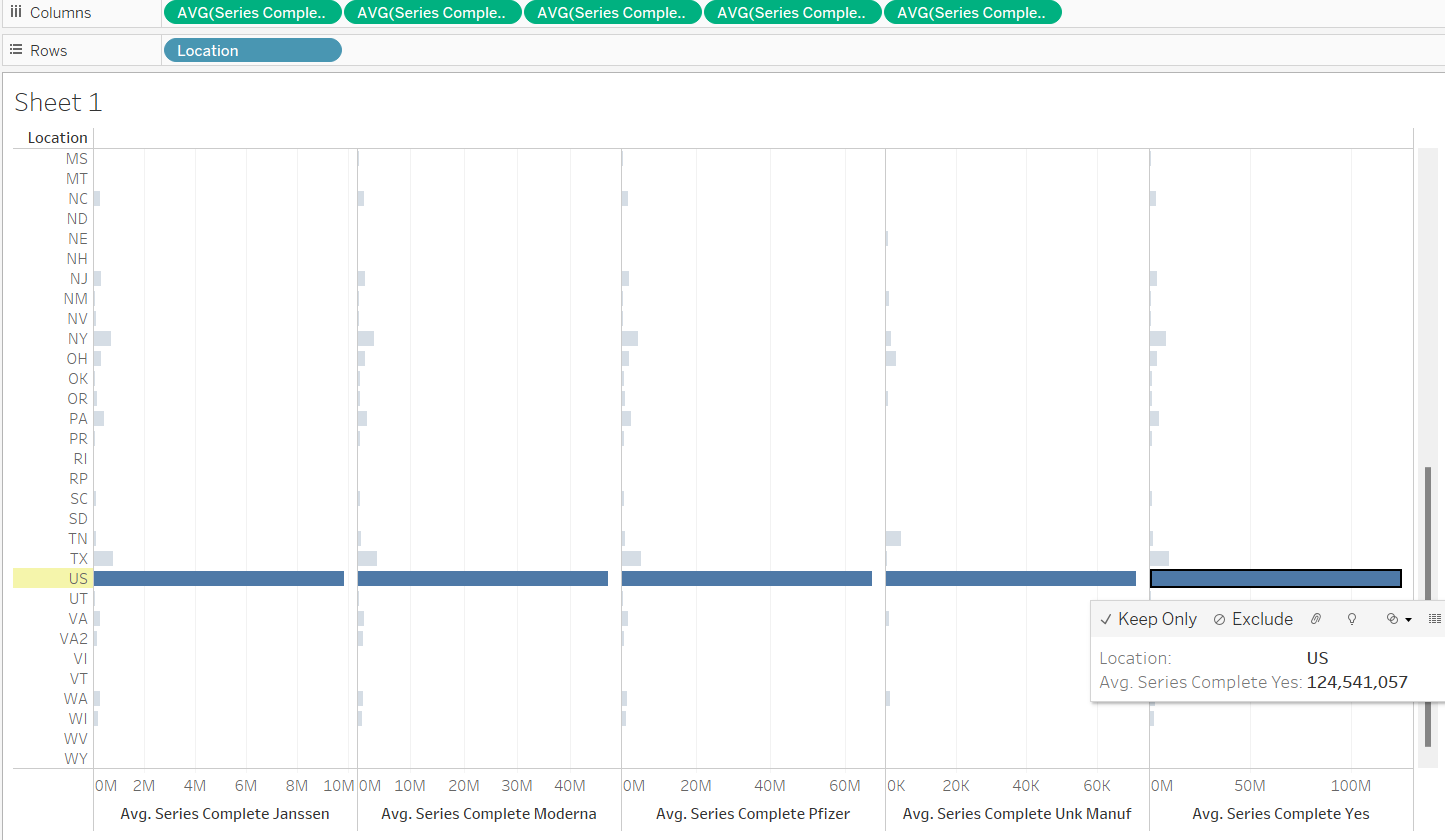
library(mvoutlier)

Your console window should look like this:

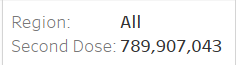


1. Boot up Tableau, once you have selected MySQL from the option panel on the left of the application proceed to login with your credentials and create a new workbook, we will call this workbook “Final”.
2. Drag and drop the US and EU vaccination datasets to connect them together to form a relationship.
3. Click on Sheet 1 located on the lower left next to Data Source. This will lead you to a blank sheet.
4. For now, in Sheet 1 we will see how many people in total got the second shot (fully vaccinated) in each vaccine category based on the US.
5. Place the Series Complete Janssen, Moderna, Pfizer, Unknown Manufacturer, and Series Complete Yes into the columns. Be sure to change the measure to average for the columns so the numbers of the US population count can roughly be more accurate.
6. Place the location in the rows. This will show the number of those vaccinated in each state and U.S. territory.

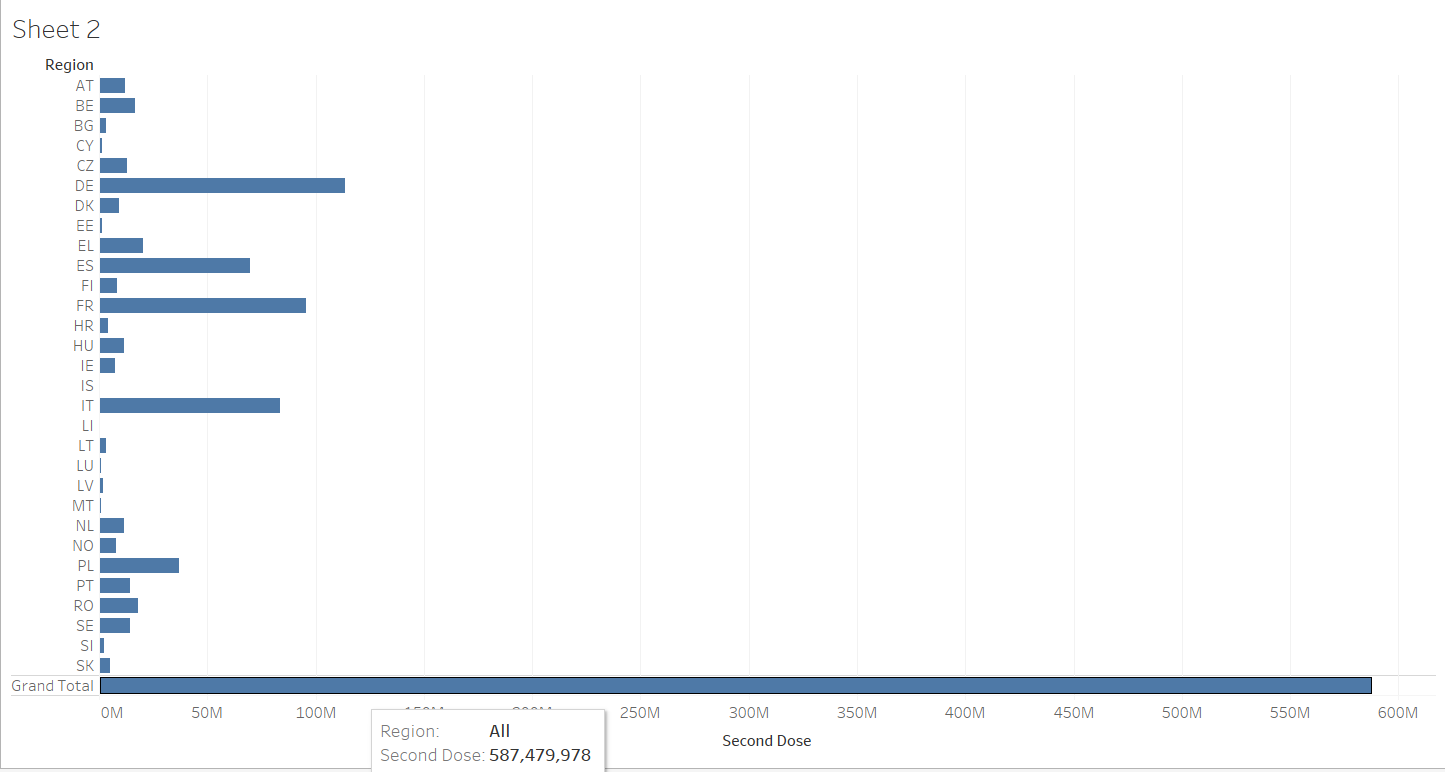
Your result should be this:



1. Next, for the Europe data we will also check for those who were fully vaccinated with the second dose. Create a new sheet, this will be Sheet 2.
2. Drag SUM(Second Dose) into columns and Region into rows.
3. Go to the Analytics tab and double-click on Totals. This will sum up all the regions who have gotten their second vaccination shot. When you click on the final pillar that is labeled Grand Total on the left, on the bottom the little box should state:

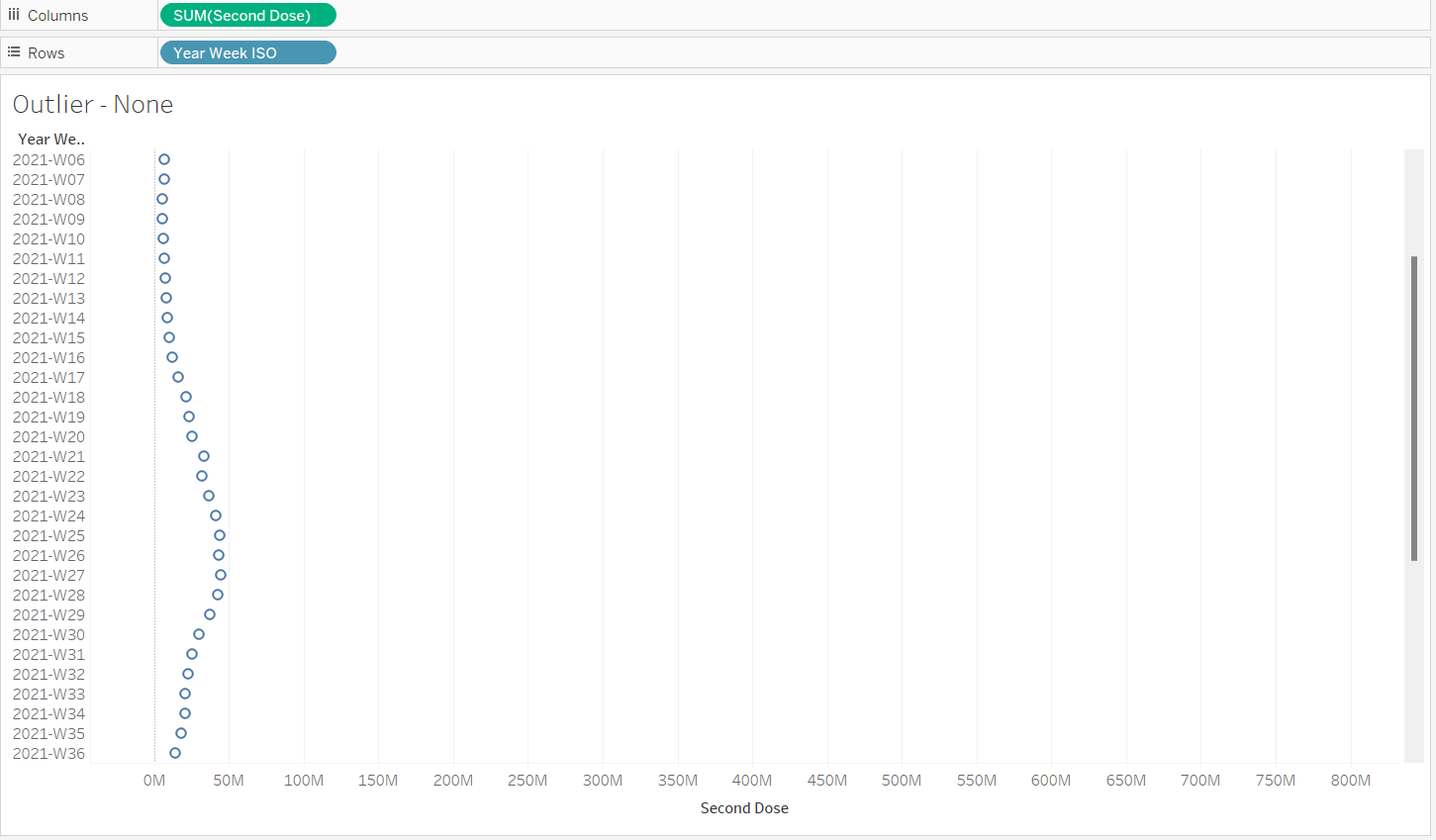


1. We notice that the population of those who got the second dose seems incorrect, we can fix the numbers by excluding the extraneous regional labels through filters.
2. Click on Region and Edit Filter. From there you will see a list of regions but the extra regions will have numbers and letters attached (this means different areas in the countries and different jurisdictions), click to exclude them but leave the distinct regions with their original value. Your result should be like this image below once you have applied the filter:

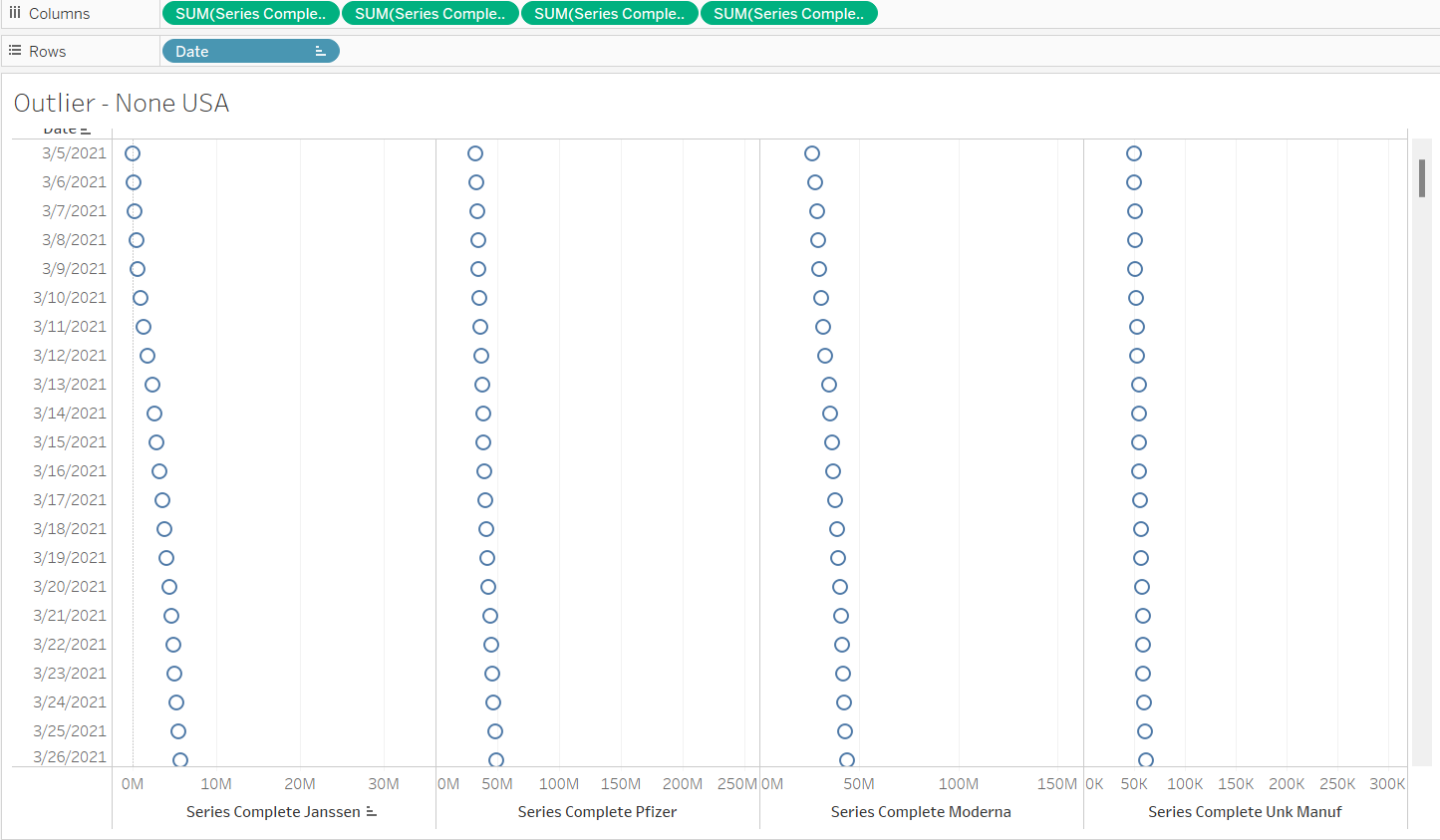


Finding Outliers (Matt, Julie, Jackson):

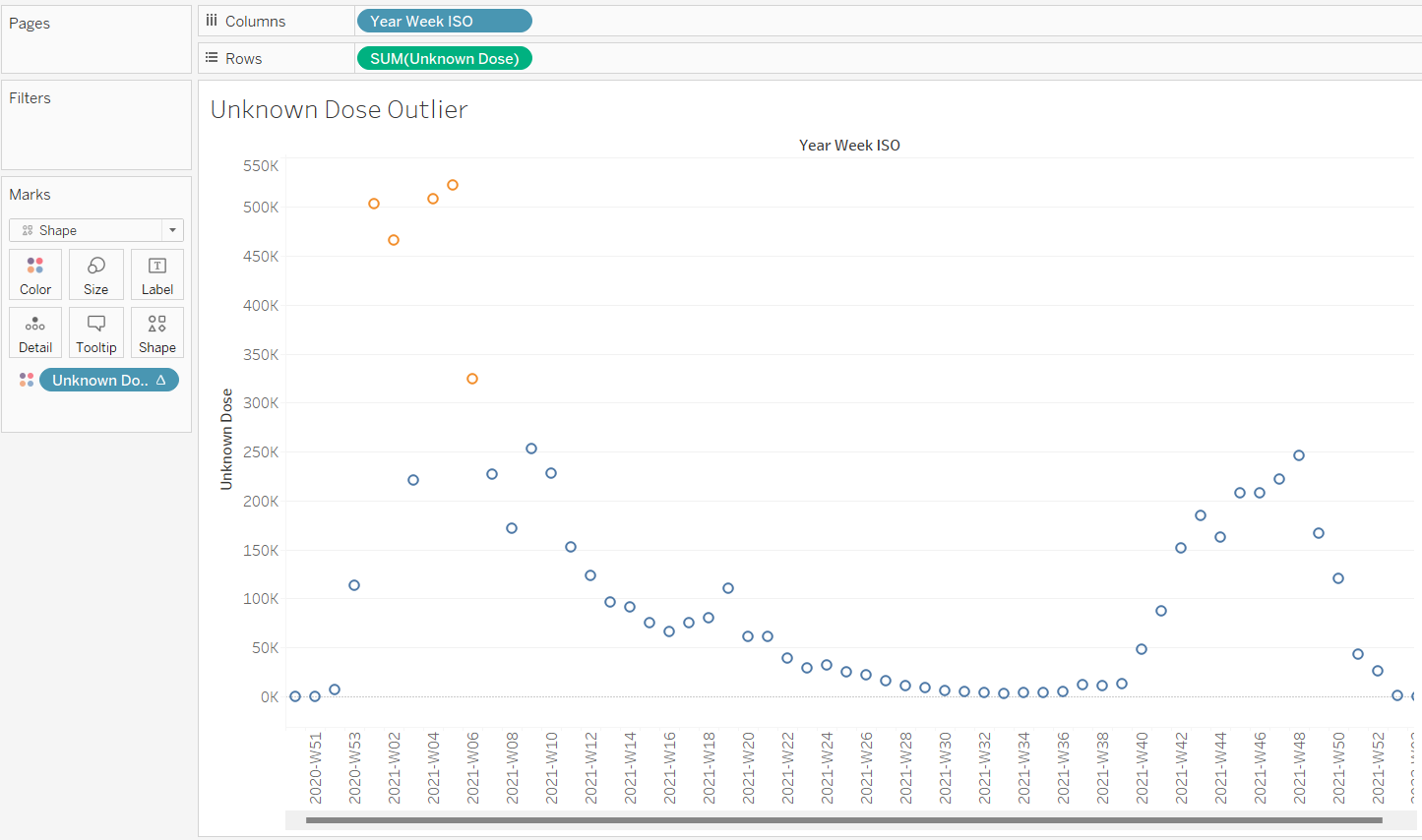
1. Create a new sheet, this will be Sheet 3. In this sheet we will try to find an outlier hidden in the EU dataset in the second dose based on the year and week. Rename the header to “Outlier – None EU” the reason why comes later on.
2. Create a calculated field dedicated to the EU set using the R Script (refer to appendix).
3. Once you have completed the calculated field, drag and drop the calculated field into the color block within the Marks section. You will notice an error saying: “More than 50% equal values in one or more variables!”. You can remove the calculated field from the slot.
4. This calculation cannot be applied because the pattern the datapoint exhibits is reminiscent to a roughly normal distribution shape. The datapoints are closely knit together and so there are no outliers to be found. Please refer to the picture below.



1. Create a new sheet, this will be sheet 4. In these next few steps, we will try to find an outlier for the following: Series Complete Janssen, Series Complete Pfizer, Series Complete Moderna, and Series Complete Unknown Manufacturer.
2. Put the Date in rows and the Series Complete Janssen, Series Complete Pfizer, Series Complete Moderna, and Series Complete Unknown Manufacturer into columns.
3. Create and apply a filter on the dates, exclude the dates from 12/13/2020 to 3/4/2021 as these hold zeros.
4. Create a calculated field dedicated to each vaccine by repeating step 2. We will create one to test out the script.
5. Repeat step 3 but you will get a different type of error saying “infinite or missing values in x” due to the dataset having too many zeros in one or more of the columns it was not able to compute. This is okay as our dataset originally had missing values but was replaced with zeroes instead of nulls in the cleaning process. Also, the remaining datapoints has a linear arrangement as seen in this figure, we will explain the implication of this trend in the conclusion later:

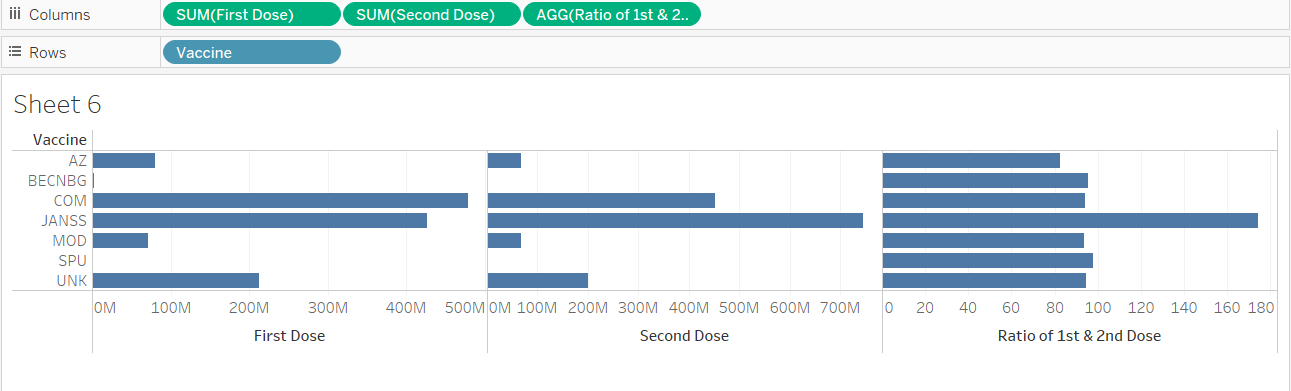


1. Add another sheet, this will be sheet 5. We will be testing out to find an outlier hidden in the EU dataset through unknown doses. You can rename this sheet to “Unknown Dose Outlier”
2. Drag and drop Year Week ISO into columns and Unknown Dose into rows.
3. Create a calculated field using R scripting to procure the outlier repeated in earlier steps call this calculated field: “Unknown Dose Outlier”. Your result should be this:



Creating and Understanding Ratios (Matt, Julie, Jackson)

1. Create a new sheet, this will be sheet 6. Here we will create ratios. For this sheet we will use the Europe dataset. Put the First and Second Dose into the columns while the Vaccine remains in the rows. The explanation and understanding of these ratios will come later in the conclusion. The ratios will be vital as it is part of our analysis.
2. Create an aggregate function called “Ratio of 1st & 2nd Dose” through calculated fields with the following command: SUM([Second Dose])/SUM([First Dose]) \* 100
3. Drag this calculated field into the columns. Your result should be this:

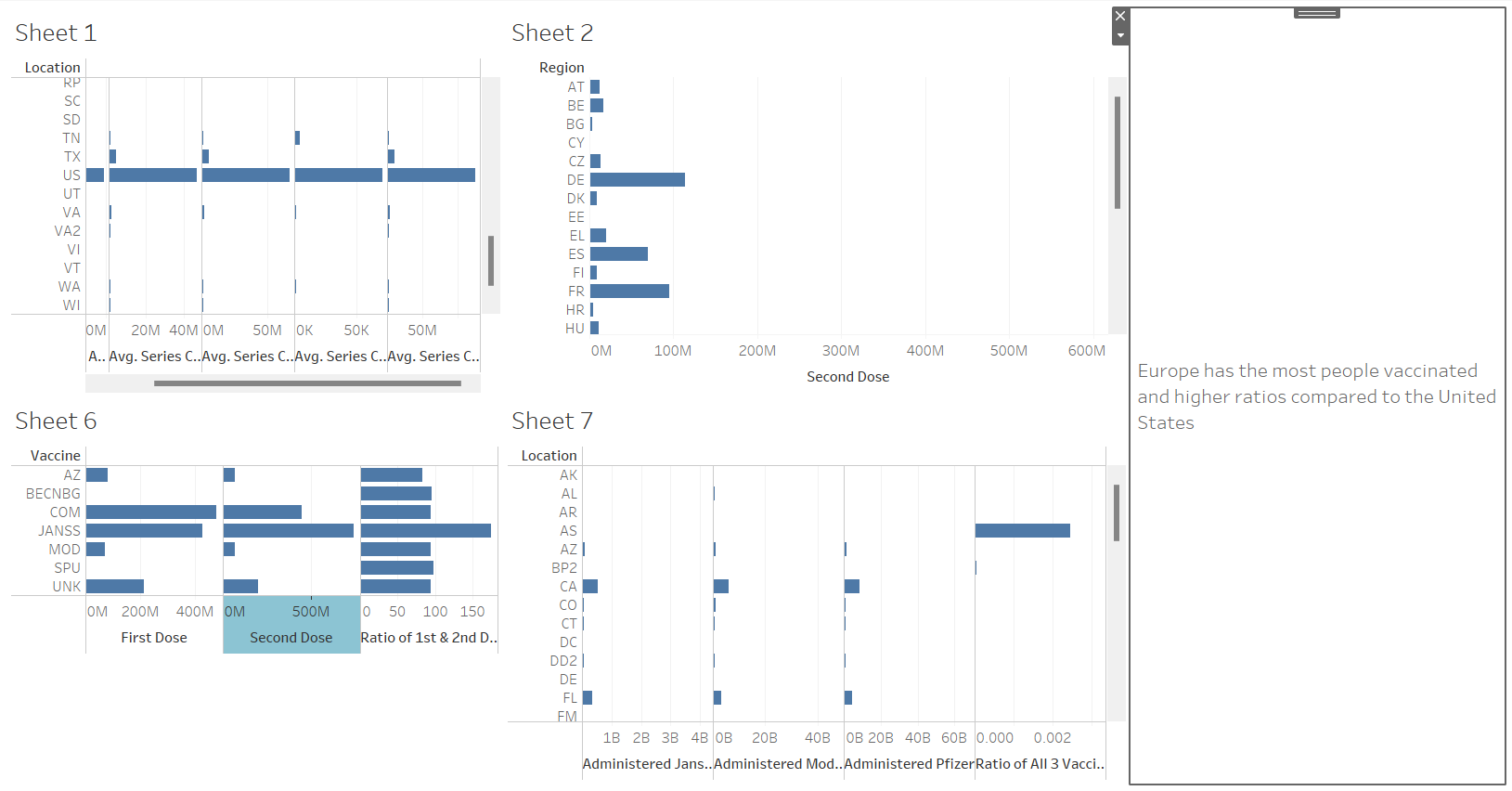


1. Create another sheet, this will be sheet 7. We will get our ratios using the same process. Create an aggregate function called “Ratio of All 3 Vaccines” through calculated fields with the following command: SUM([Administered Pfizer])/SUM([Administered Moderna])/SUM([Administered Janssen]) \* 100
2. Apply this aggregate calculation to the column, your result should be this:

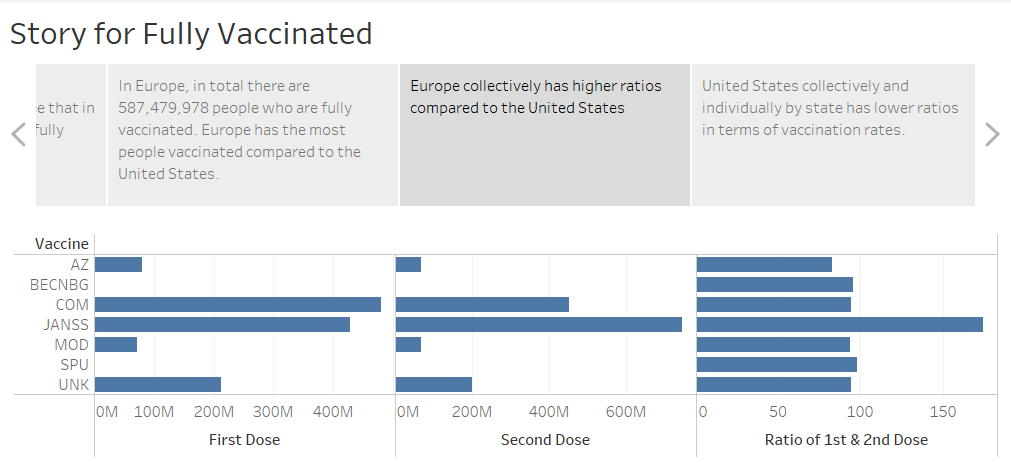


Dashboard and Story Creation (Matt, Julie, Jackson)

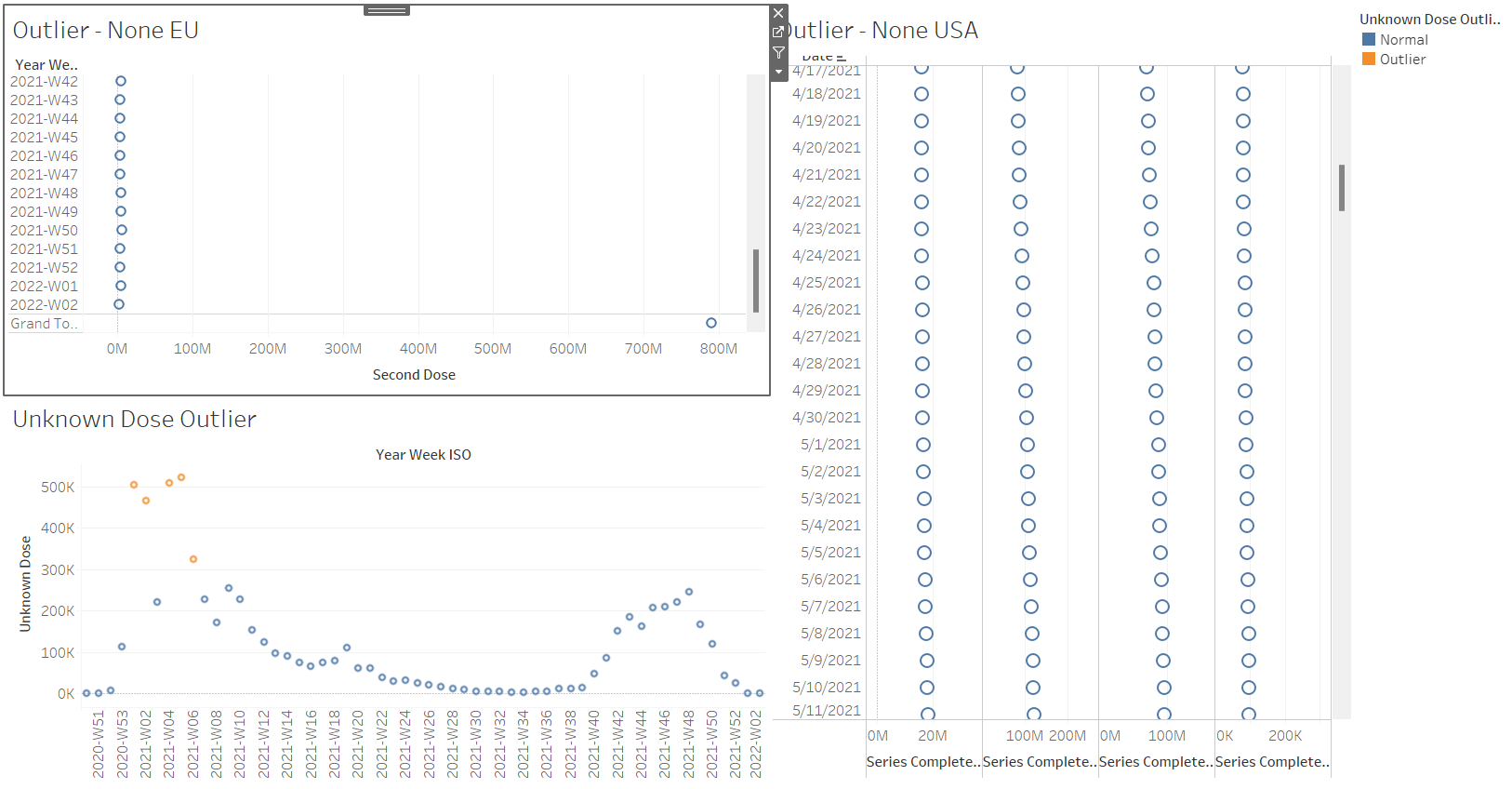
1. Create a new dashboard, this is originally called “Dashboard 1” but rename this to “Dashboard for Fully Vaccinated” this gives additional context based on our analysis.
2. Drag Sheets 1 and 2 into the empty space as halves. Readjust the size as needed for orientation.
3. Next, drag sheets 6 and 7 on the bottom half of the quadrant to split the dashboard into fourths. Readjust the size to fit as necessary.
4. Drag Text to the right of the dashboard, fill the textbox with information as needed. The end result should be this:



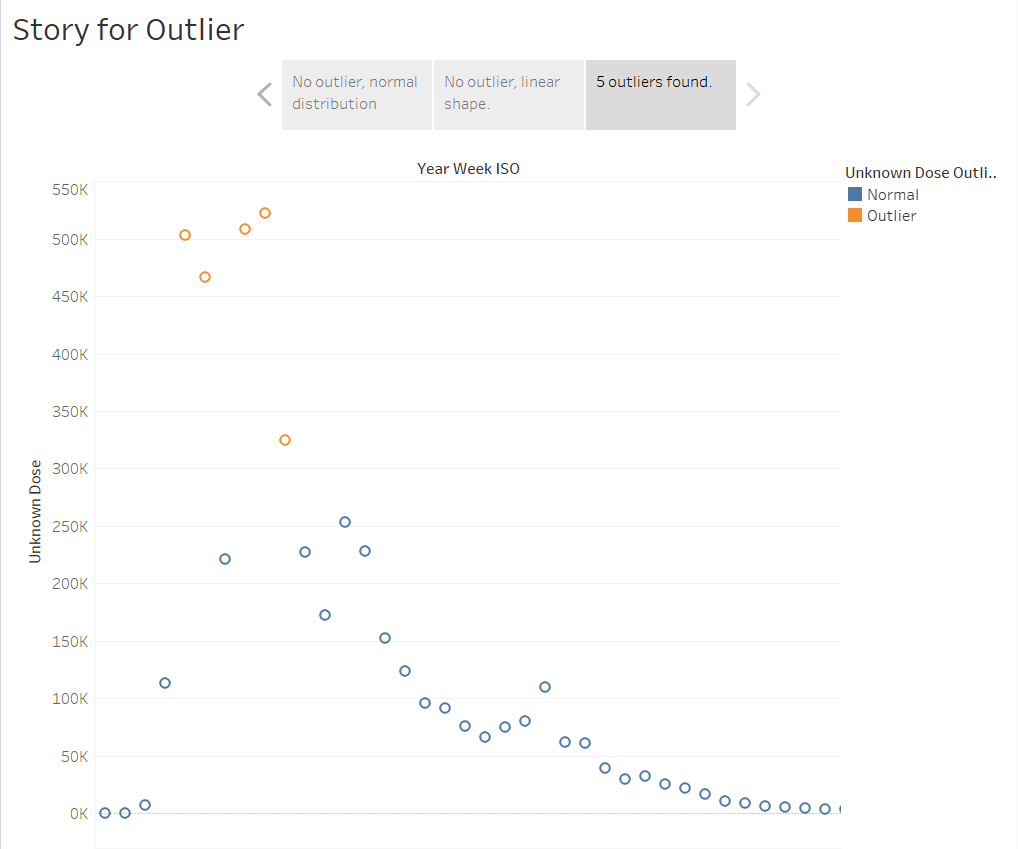
1. Create a story to link more context to the dashboard created in the previous steps, call this story “Story for Fully Vaccinated”.
2. Drag sheet 1 into the space and create a caption that provides additional context.
3. Under New Story Point, click blank. This creates a new page within the current story.
4. Drag sheet 2 into the space and create a caption that provides additional context.
5. Repeat steps 7 and 8 for the ratios. Your result should be this as each caption block leads you to the correct sheet information by navigation:



1. Repeat steps 1 and 2 to create a Dashboard for the outliers. Drag sheets 3, 4, and 5 into the dashboard space. Adjust the size of each sheet block as needed. Your end result should be:



1. Create a story for the outliers, call this story “Story for Outlier”.
2. Drag sheet 3 into the story space and create a caption to provide additional context.
3. Click blank under New Story Point to add more pages.
4. Repeat steps 12 to 13 two more times for sheets 4 and 5. Your end result should be this:



1. Save your work as Final.twb before exiting.

# Appendix (Matt Wehnert)

IF SCRIPT\_REAL("library(mvoutlier);

sign2(cbind(.arg1))$wfinal01",

sum([<field name here>]))= 0

THEN "Outlier"

ELSE "Normal"

END

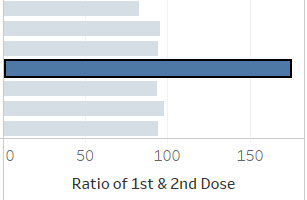
Conclusion (Jackson and Julie)

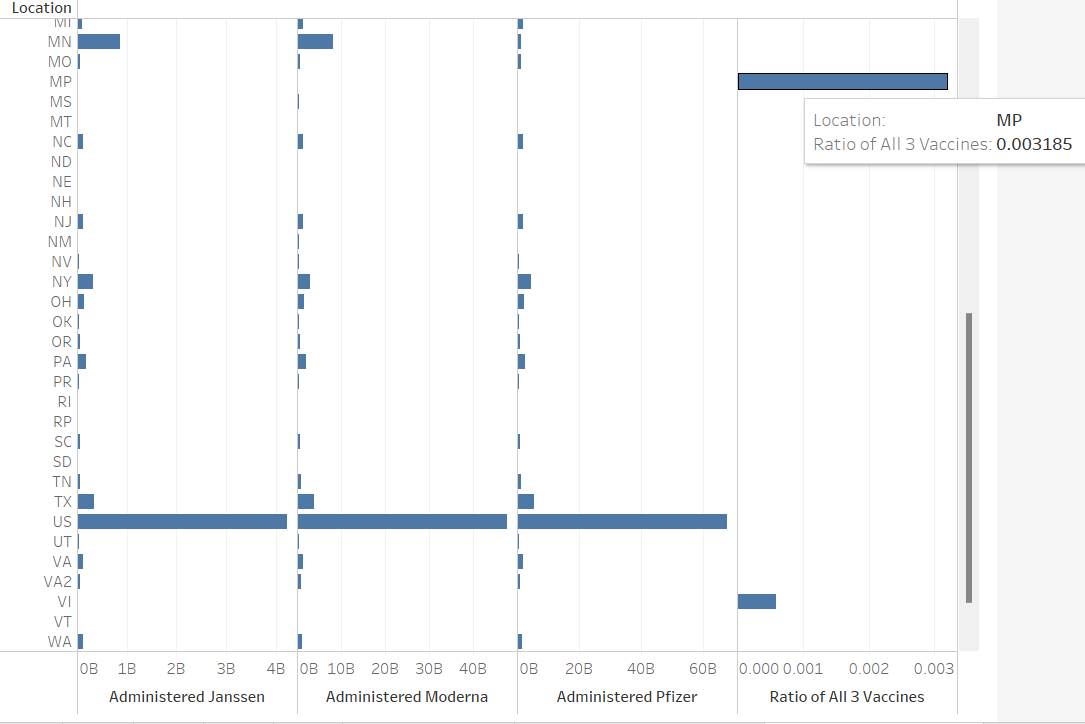
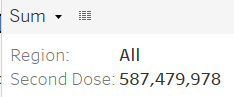
BI Inquiry (Hypothesis): What can we learn about how Europe and USA handled COVID-19? Who handled it better and who had the most vaccinations? We intend to see comparisons between the two continents.

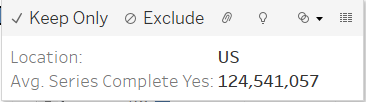
Our question was to see who handled the epidemic better and based on our analysis we can clearly see that EU handled the situation better than the United States based on the numbers and ratios. Our hypothesis was correct as we inferred that EU had more vaccinations. Not only were we able to see comparisons but we saw a trend as well.

# Analysis

Based on looking at the ratio we can see that Janssen has a ratio of 174.42 which is the highest value compared to the other available vaccines in the EU. Also, by looking through the list of states the ratio values are much smaller. Refer to images below:



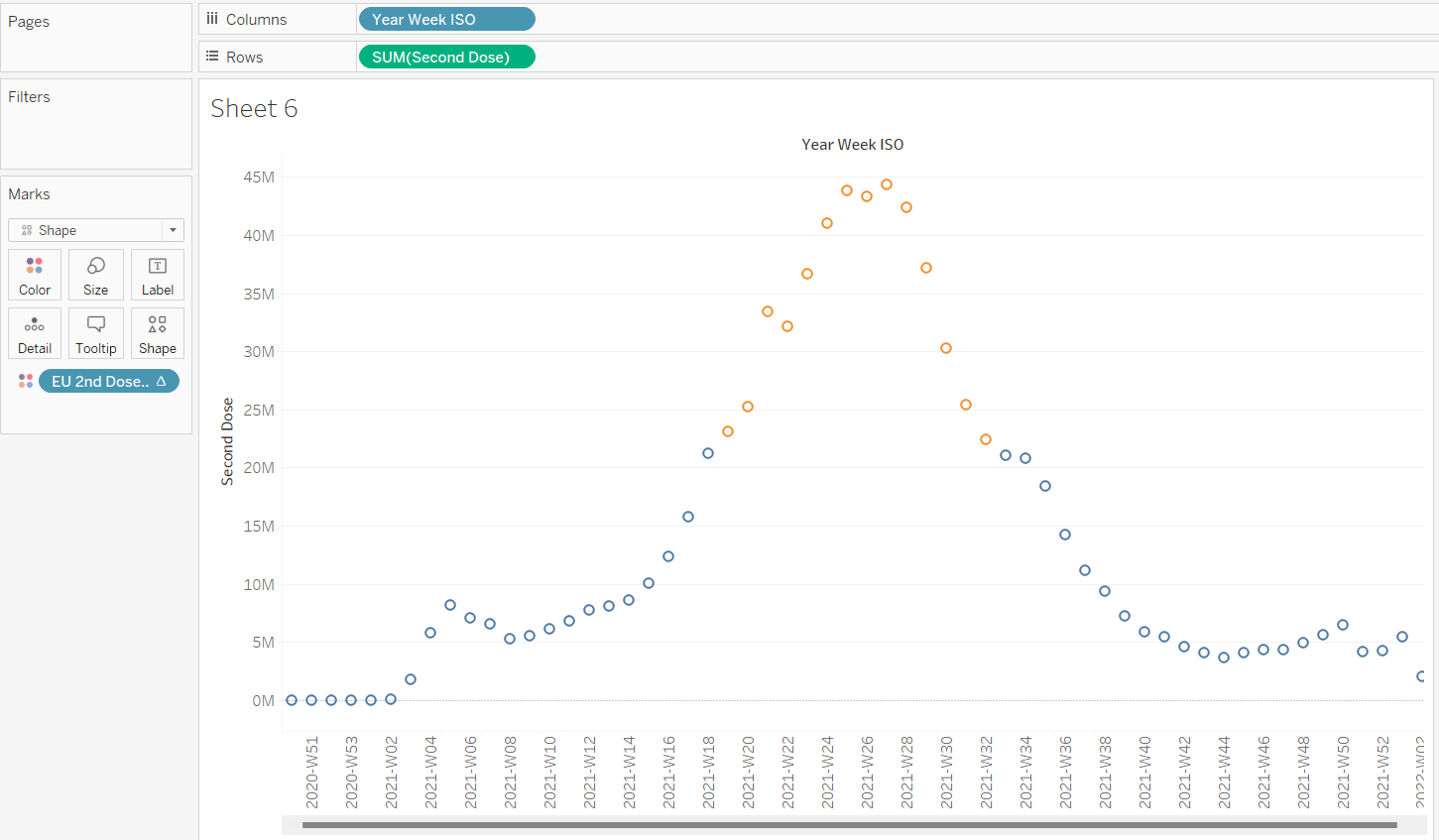




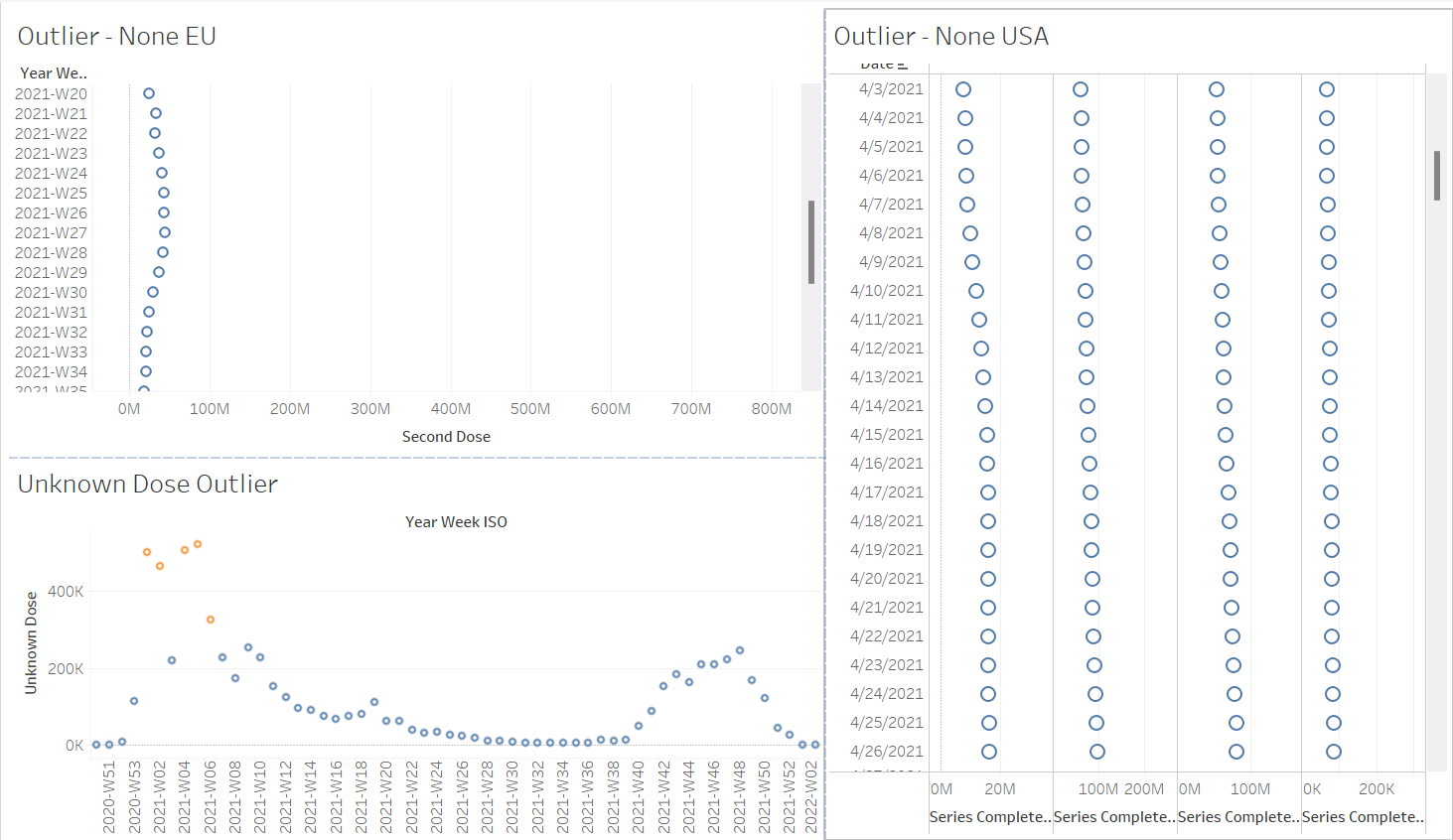
Not just the ratios depict the frequency of how many people and how often vaccines were being given, we can see from the total of people in the U.S completing the series of their vaccine is 124,541,057 whereas EU had 587,479,978 people fully vaccinated. We will talk about the factors that come into play later in the Insights and Findings section next.

# Insights and Findings

My group and I found a strange pattern and are unsure if this is an anomaly or a miscalculation/bug on Tableau’s part (refer to picture below). The shape represents an almost unimodal distribution but can be argued as bimodal or multimodal due to the multiple curves the data is exhibiting. Based on our inference, the implication of this data means that there was a surge in vaccinations peaking between week 20 of 2021 and going into a slow decline in week 32 of 2021 during the vaccine shortage that happened in Europe. Europe was facing a vaccine shortage which meant that those who already had the first dose of Moderna, Pfizer, and or some other vaccine can use Janssen as a substitute to fulfill their vaccination status. We believe this is why we see an increase over 100 percent for a second dose of a single dose vaccine in connection to the ratio.



We also notice that out of the three vaccines in the USA, Pfizer is the most taken vaccine with 67,475,657 people who completed their series.



In the picture above, based on our inference, we believe that at first, we see a slow growth in the data and this indicates the period of time where the vaccine was restricted to certain vulnerable groups and first responders. Then we see an increase as it is opened up to different age groups and then everyone. At a certain point, towards the end of the dataset we see a plateau in the data and this tells us that we had hit a point where everyone who wanted to receive the vaccine had already gotten it. Not to mention, there were political tensions during the pandemic.

**Factors at play leading to discrepancies:**

* Region size – Europe is a big continent; it is a conglomerate of independent nations of the EU.
* Population size – One has a bigger population than the other
* Undocumented person(s) – Unreported information could have influenced the data.
* Census updates – Census may have not been updated accurately and on time due to rapid infections and deaths
* Morbidity and Mortality Weekly Reports may differ from country to country.
* State of Emergencies – Some countries treat the pandemic critically and heavily encourage vaccinations despite the odds of a shortage.
* We did not know that in the story we can create multiple story points as we usually would create multiple new stories. The new story points serve as additional pages to the current story that was made. We thought this was a convenient feature.

# Challenges/Resolution

1. When trying to perform the outlier and anomaly detection using R script some fields were incompatible by first impression. Upon further investigation, we have determined that because the data was closely knit together displaying a normal distribution curve or a linear fashion this caused the script to not work. There was nothing we could have done, and even if we tried to clean the data more this would affect the integrity of the dataset harshly. Originally this dataset had countless blank cells and we filled the empty cells with zeros. We also made sure the null values were filled with zeros. This is to avoid problems with Tableau in case it would give us errors regarding null values as it cannot compute null values.
2. Had a bit trouble figuring out how to split the dashboard into pieces to fit together the sheets but we referred to the Tableau tutorials on YouTube to better guide us on how dashboards and stories worked.
3. Sometimes Tableau would give us different results despite the same steps were followed. We don’t understand the cause of this but we recorded it as something interesting if it was coherent.

# Reflection

Overall, we did encounter hardships in regards to learning certain concepts but through teamwork and guidance we were able to understand bits and pieces of the concept. We not only learned how to do outlier detection but also how to do some R scripting which was interesting. Even though we did not apply much of LOD in this lab we wanted to focus more on the totals as a whole without the extraneous numbers. We also do not have much experience in R as we still feel as if R is its own language that needs more time to comprehend its functionalities. We as a team struggle with this programming language but feel as if there is a significance behind this language as we need this language to be able to import certain packages and libraries for our work in Tableau.

# Further Future Investigations

We feel that in this project, if there were any future investigations that could have been done, we would want to look deeper into COVID-19 mortality rates between both continents of North America (USA portion) and Europe to see if there is a connection or a reason why some numbers were not recorded correctly. We also see a drop on 1/13/2022 which implies that due to some sort of political disagreements or controversy there are those who have second thoughts on the vaccine.