CIS 4560-01 Term Project Tutorial

Author: **Kevin Duran**

Instructor: Jongwook Woo

Date: 5/18/2025

Lab Tutorial

Kevin Duran(kduran26@calstatela.edu)

Analysis of US Weather Using Big Data

Objectives

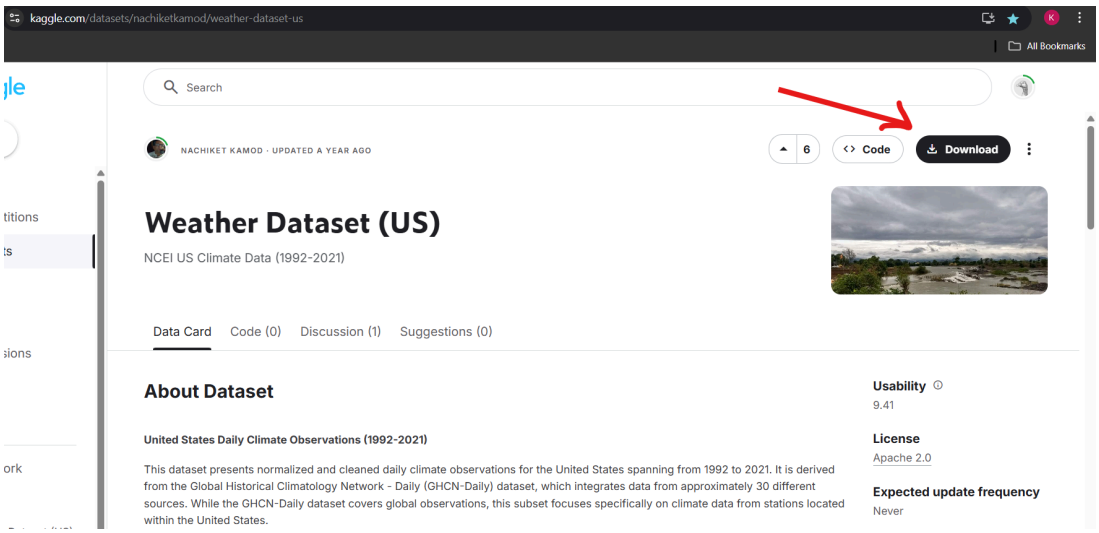
In this hands-on lab, you will learn how to:

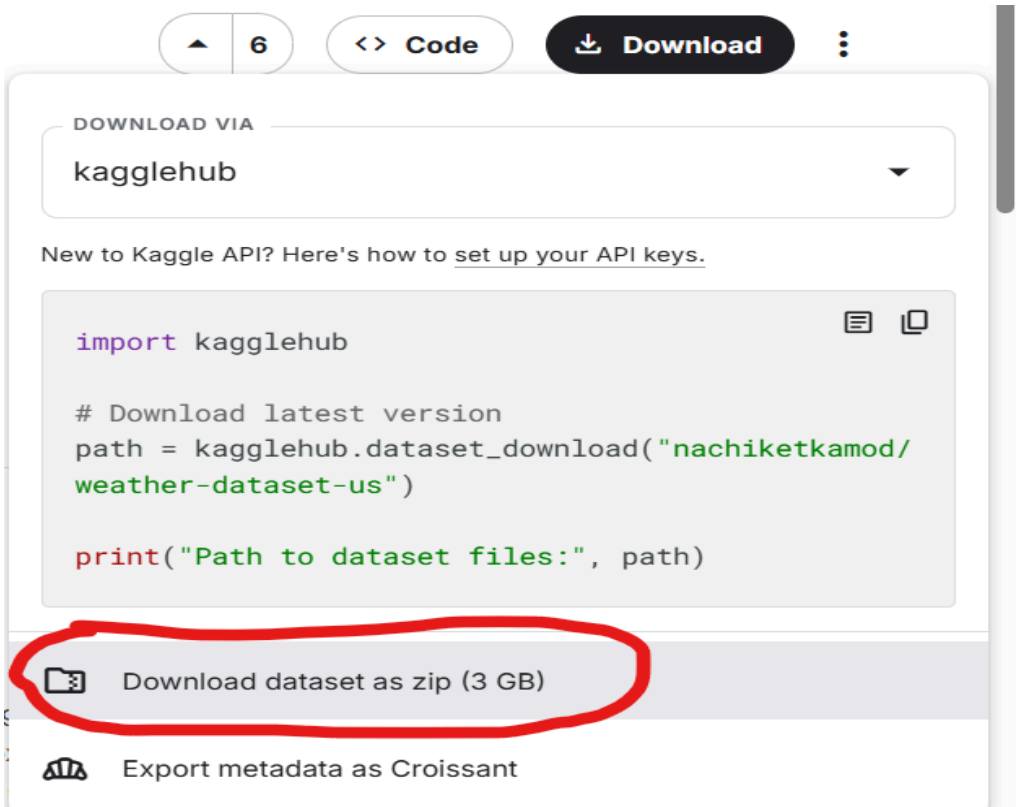
- Upload weather data into HDFS
- Develop Hive tables to clean the dataset
- Derive climate label indications (average daily temperature and volatility)
- Export Final CSV to local machine to use for analysis in Tableau and Excel
- Create 3D Power maps with Excel with clean data to visualize temperature and precipitation across the U.S.
- Create effective charts and graphs with Tableau to analyze global warming

Step 1: Download and Prepare Dataset from Kaggle

To begin this lab tutorial we must first acquire the zipped dataset from Kaggle. This will allow us to export it to Apache Hadoop and Hive.
It's essential that on steps with directories/usernames to use your own information to ensure consistency.

1. Download Zipped dataset as this is the data we'll be doing this lab with.
   https://www.kaggle.com/datasets/nachiketkamod/weather-dataset-us





2. After the download has finished, you want to now upload this zipped data to your HDFS Hadoop environment. This is done to ensure we have a backup file of the raw zip available. We'll begin with signing into the environment with SSH. **(Be sure to use your own designated Hadoop username and server's IP address/hostname.**

```
SSH kduran26@111.11.11.111
```

3. Now we can begin setting up our project folders and uploading our files to HDFS, starting with creating a directory for the zip file and then uploading it.

```
hdfs dfs -mkdir -p /user/kduran26/us_weather
hdfs dfs -put Weather_Dataset_US.zip /user/kduran26/us_weather/
```

4. Next, since we have our backup on hdfs saved, we can unzip the zipper file locally(On your own PC not remote server Hadoop)and repeat the process. (The directory listed before with HDFS path /user/kduran26/us_weather/ was only used to maintain a backup file of the zip.) This means we use HDFS dfs -mkdir -p /user/kduran26/us_weather/csv_raw_weather as our true path for the Hive directory.

```
hdfs dfs -mkdir -p /user/kduran26/us_weather/csv_raw_weather

hdfs dfs -put ~/us_weather_extracted/Weather\ Data\ \(US\).csv
/user/kduran26/us_weather/csv_raw_weather/
```

Step 2: Apache Hive Beeline for making tables.

Now that we have uploaded the unzipped data onto a directory within our Hadoop environment we can sign into Beeline from HDFS to begin our cleaning and formatting of the dataset:

1. We begin with signing into Beeline after being signed into HDFS.

   ```
   beeline
   ```

2. Now that we are in the hive beeline environment we'll use the following query to create our raw dataset into a table. During this process Hive is reading the raw CSV data as if it were a table. By using 'EXTERNAL' keyword we can make sure that the file is readable by Hive but not movable or alterable. Using this table we can then move on to the next step where we create a clean finalized version of the table. **Make sure you revise Location to the path you want to create the table in.**

```
USE kduran26;


DROP TABLE IF EXISTS weather_raw;


CREATE EXTERNAL TABLE weather_raw (
  id STRING,
  `date` STRING,
  tmax STRING,
  tmin STRING,
  evap STRING,
```

```
  prcp STRING,
  latitude STRING,
  longitude STRING,
  elevation STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/user/kduran26/us_weather/csv_raw_weather/'
TBLPROPERTIES ("skip.header.line.count"="1");
```

3. Moving on, we'll now create our clean Hive Table with our Derived attributes **avg_temp_celsius and temp_range_celsius**, along with categorical labels like **avg_temp_label**, and **temp_range_volatility.** This step is essential as this cleaned table helps filter out invalid records, like nulls and blanks, it also configures our date range along casting attributes in their appropriate data types. Analysis cannot be done later without this step.

```
    DROP TABLE IF EXISTS weather_cleaned;
    CREATE TABLE weather_cleaned AS
    SELECT
      id,
      `date`,
      CAST(tmax AS INT) AS tmax_tenths,
      CAST(tmin AS INT) AS tmin_tenths,
      CAST(prcp AS INT) AS prcp_tenths,
      CAST(latitude AS DOUBLE) AS latitude,
```

```
  CAST(longitude AS DOUBLE) AS longitude,

  CAST(elevation AS DOUBLE) AS elevation,

  (CAST(tmax AS FLOAT) + CAST(tmin AS FLOAT)) / 20.0 AS
avg_temp_celsius,

  (CAST(tmax AS FLOAT) - CAST(tmin AS FLOAT)) / 10.0 AS
temp_range_celsius,

  CASE

    WHEN (CAST(tmax AS FLOAT) + CAST(tmin AS FLOAT)) / 20.0 < 5
THEN 'COLD'

    WHEN (CAST(tmax AS FLOAT) + CAST(tmin AS FLOAT)) / 20.0
BETWEEN 5 AND 25 THEN 'MODERATE'

    WHEN (CAST(tmax AS FLOAT) + CAST(tmin AS FLOAT)) / 20.0 > 25
THEN 'HOT'

    ELSE 'UNKNOWN'

  END AS avg_temp_label,

  CASE

    WHEN (CAST(tmax AS FLOAT) - CAST(tmin AS FLOAT)) / 10.0 < 5
THEN 'STABLE'

    WHEN (CAST(tmax AS FLOAT) - CAST(tmin AS FLOAT)) / 10.0
BETWEEN 5 AND 15 THEN 'NORMAL RANGE'

    WHEN (CAST(tmax AS FLOAT) - CAST(tmin AS FLOAT)) / 10.0 > 15
THEN 'EXTREME RANGE'

    ELSE 'UNKNOWN'

  END AS temp_range_volatility

FROM weather_raw

WHERE tmax IS NOT NULL AND tmax != ''

  AND tmin IS NOT NULL AND tmin != ''

  AND prcp IS NOT NULL AND prcp != ''
```

```
AND latitude IS NOT NULL AND latitude != ''

AND longitude IS NOT NULL AND longitude != ''

AND elevation IS NOT NULL AND elevation != ''

AND `date` IS NOT NULL AND `date` != ''

AND `date` RLIKE '^[0-9]{1,2}/[0-9]{1,2}/20(1[0-9]|20)$';
```

Step 3: Prepare cleaned table as a CSV for Local machine for Excel and Tableau analysis

1. This step starts with exporting the cleaned table previously made into HDFS as a Text Output. This essentially writes the content of our weather_cleaned table into an HDFS folder as comma separated values in which we'll use for our CSV file download.

```
INSERT OVERWRITE DIRECTORY
'/user/kduran26/us_weather/weather_final_2010_2020'

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

SELECT * FROM weather_cleaned;
```

2. Within this next block of code, we'll be acquiring the exported output from HDFS to create the final CSV. The first part of the code deletes any possible existing local folder with the same name to ensure we avoid possible file conflicts. The next part '-get', downloads the cleaned dataset from HDFS to the local directory. Echo was used in order to develop a proper header row for the CSV file, and lastly 'cat' helped append the data that was downloaded underneath the header which helps create only one complete file

that is also structured. After completing this step, you will now have your final CSV for analysis in your local Linux machine.

```
rm -r weather_final_2010_2020

hdfs dfs -get /user/kduran26/us_weather/weather_final_2010_2020
./weather_final_2010_2020

echo
"id,date,tmax_tenths,tmin_tenths,prcp_tenths,latitude,longitude,elevat
ion,avg_temp_celsius,temp_range_celsius,avg_temp_label,temp_range_vola
tility" > weather_2010_2020.csv

cat weather_final_2010_2020/* >> weather_2010_2020.csv
```

3. Finally we can download our final CSV from the Linux server to our Local Machine. It is essential that you still replace the ‘kduran26’ with your own user along with the IP address to your own server as well, you also need to ensure the destination path is also to wherever you want to have the file saved locally. It is also important that you enter this code in your local terminal Bash not SSH.

```
scp kduran26@144.24.46.199:/home/kduran26/weather_2010_2020.csv
~/Downloads/
```

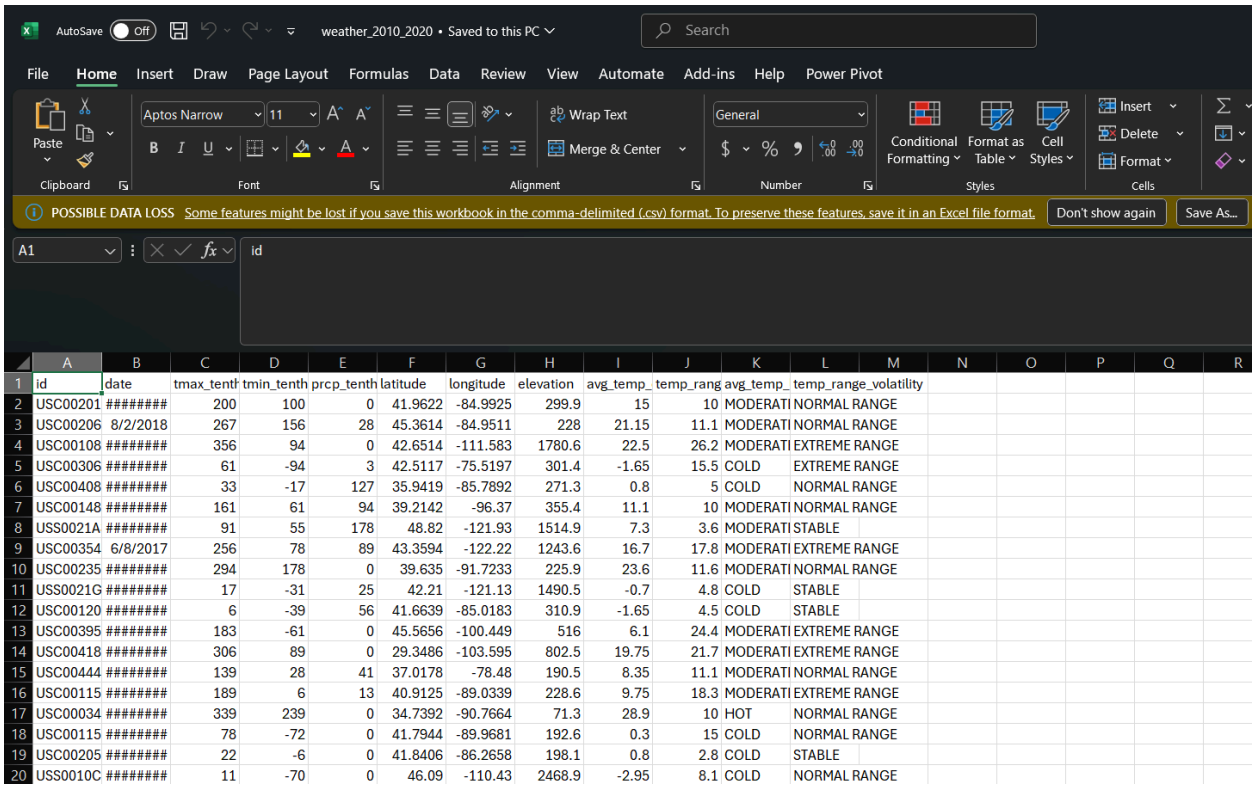
4. You now have your cleaned csv file on your local machine ready for analysis using Excel and Tableau

Step 4: Analyze the data using various Data Analysis Tools

Excel Power Maps

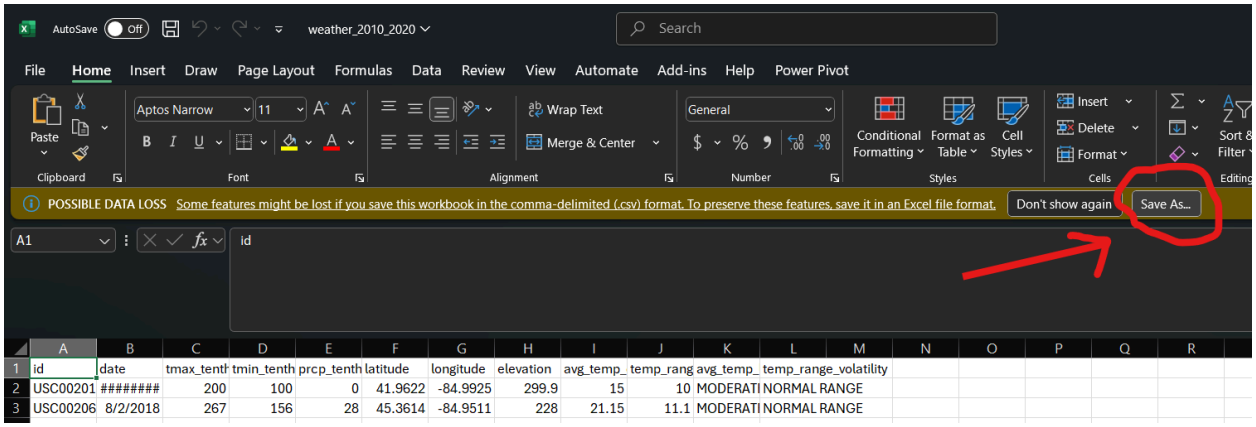
Using Excel Power Maps, we can generate and visualize this weather data over a period of time. This can help, especially during weather data analysis as we are able to understand exactly how the weather is changing over time. To begin we’ll start by converting our newly created .csv file into an xlxs file(Excel File).

1. Open the .csv file you downloaded to your local machine it should look like the figure below.



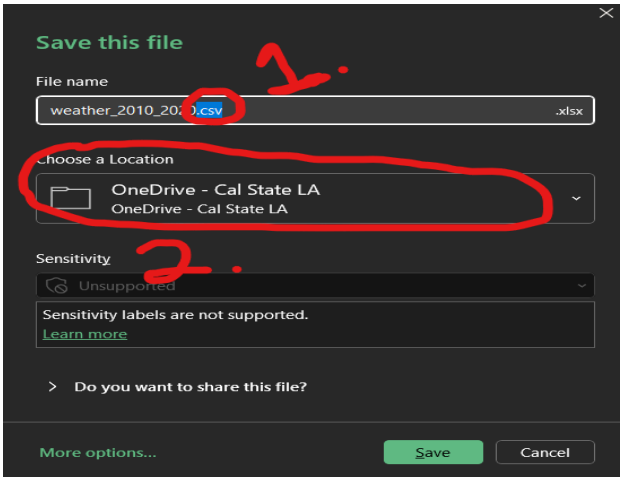
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	date	tmax_tenth	tmin_tenth	prcp_tenth	latitude	longitude	elevation	avg_temp_	temp_rang	avg_temp_	temp_range_volatility	
2	USC00201	########	200	100	0	41.9622	-84.9925	299.9	15	10	MODERATI	NORMAL RANGE	
3	USC00206	8/2/2018	267	156	28	45.3614	-84.9511	228	21.15	11.1	MODERATI	NORMAL RANGE	
4	USC00108	########	356	94	0	42.6514	-111.583	1780.6	22.5	26.2	MODERATI	EXTREME RANGE	
5	USC00306	########	61	-94	3	42.5117	-75.5197	301.4	-1.65	15.5	COLD	EXTREME RANGE	
6	USC00408	########	33	-17	127	35.9419	-85.7892	271.3	0.8	5	COLD	NORMAL RANGE	
7	USC00148	########	161	61	94	39.2142	-96.37	355.4	11.1	10	MODERATI	NORMAL RANGE	
8	USS0021A	########	91	55	178	48.82	-121.93	1514.9	7.3	3.6	MODERATI	STABLE	
9	USC00354	6/8/2017	256	78	89	43.3594	-122.22	1243.6	16.7	17.8	MODERATI	EXTREME RANGE	
10	USC00235	########	294	178	0	39.635	-91.7233	225.9	23.6	11.6	MODERATI	NORMAL RANGE	
11	USS0021G	########	17	-31	25	42.21	-121.13	1490.5	-0.7	4.8	COLD	STABLE	
12	USC00120	########	6	-39	56	41.6639	-85.0183	310.9	-1.65	4.5	COLD	STABLE	
13	USC00395	########	183	-61	0	45.5656	-100.449	516	6.1	24.4	MODERATI	EXTREME RANGE	
14	USC00418	########	306	89	0	29.3486	-103.595	802.5	19.75	21.7	MODERATI	EXTREME RANGE	
15	USC00444	########	139	28	41	37.0178	-78.48	190.5	8.35	11.1	MODERATI	NORMAL RANGE	
16	USC00115	########	189	6	13	40.9125	-89.0339	228.6	9.75	18.3	MODERATI	EXTREME RANGE	
17	USC00034	########	339	239	0	34.7392	-90.7664	71.3	28.9	10	HOT	NORMAL RANGE	
18	USC00115	########	78	-72	0	41.7944	-89.9681	192.6	0.3	15	COLD	NORMAL RANGE	
19	USC00205	########	22	-6	0	41.8406	-86.2658	198.1	0.8	2.8	COLD	STABLE	
20	USS0010C	########	11	-70	0	46.09	-110.43	2468.9	-2.95	8.1	COLD	NORMAL RANGE	

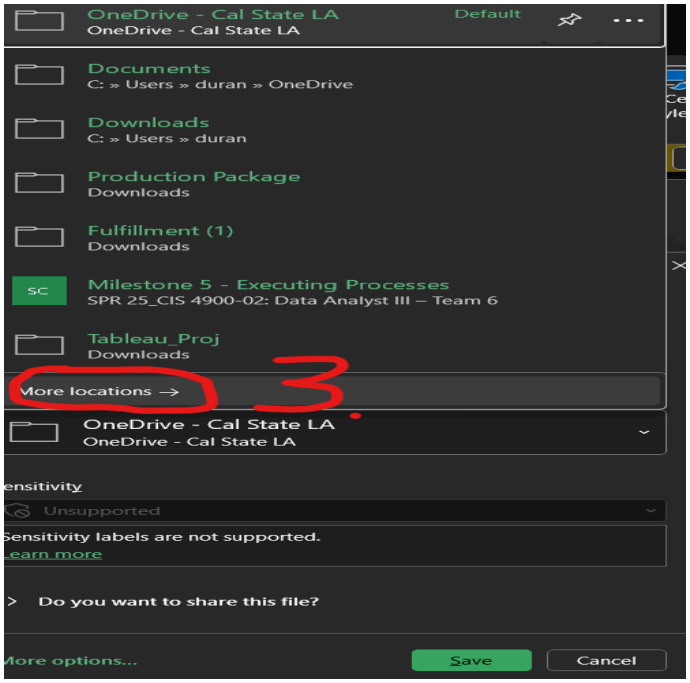
2. After the file has completely opened, we're going to now save this .csv using "Save as", Click the "Save as" button located in the "POSSIBLE DATA LOSS" warning.



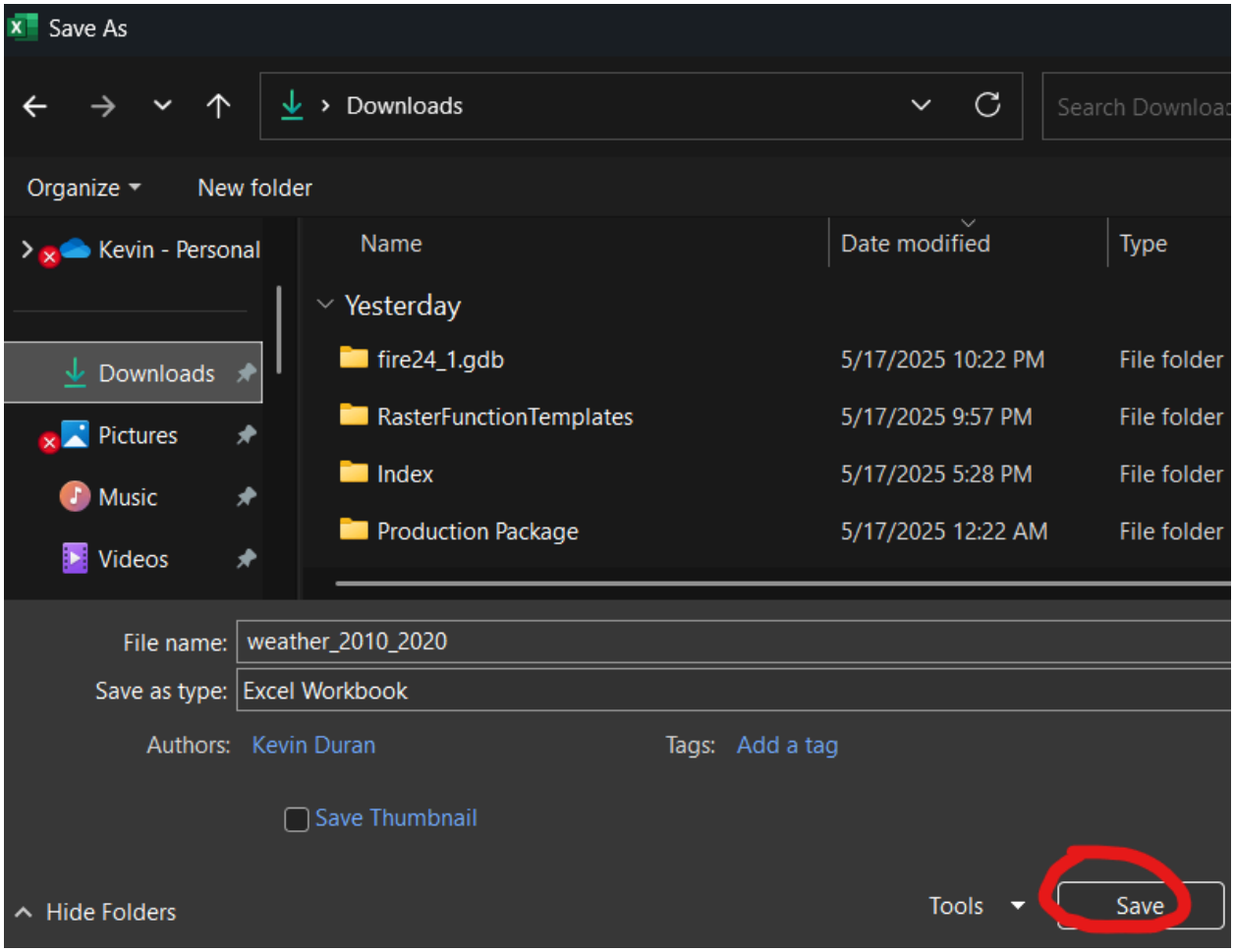
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	date	tmax_tenth	tmin_tenth	prcp_tenth	latitude	longitude	elevation	avg_temp_	temp_rang	avg_temp_	temp_range_volatility	
2	USC00201	########	200	100	0	41.9622	-84.9925	299.9	15	10	MODERATI	NORMAL RANGE	
3	USC00206	8/2/2018	267	156	28	45.3614	-84.9511	228	21.15	11.1	MODERATI	NORMAL RANGE	

3. Once in the “Save this File” window, first you want to remove the “.csv” file extension from your file.(Shown by “1.”), then you want to choose a location(Shown by “2.”). After this click on “More locations” (Shown by “3.), this will allow you to save this .xlxs file on your computer, it should be saved somewhere that's an easily accessible location if you choose not to save in the downloads as I did.





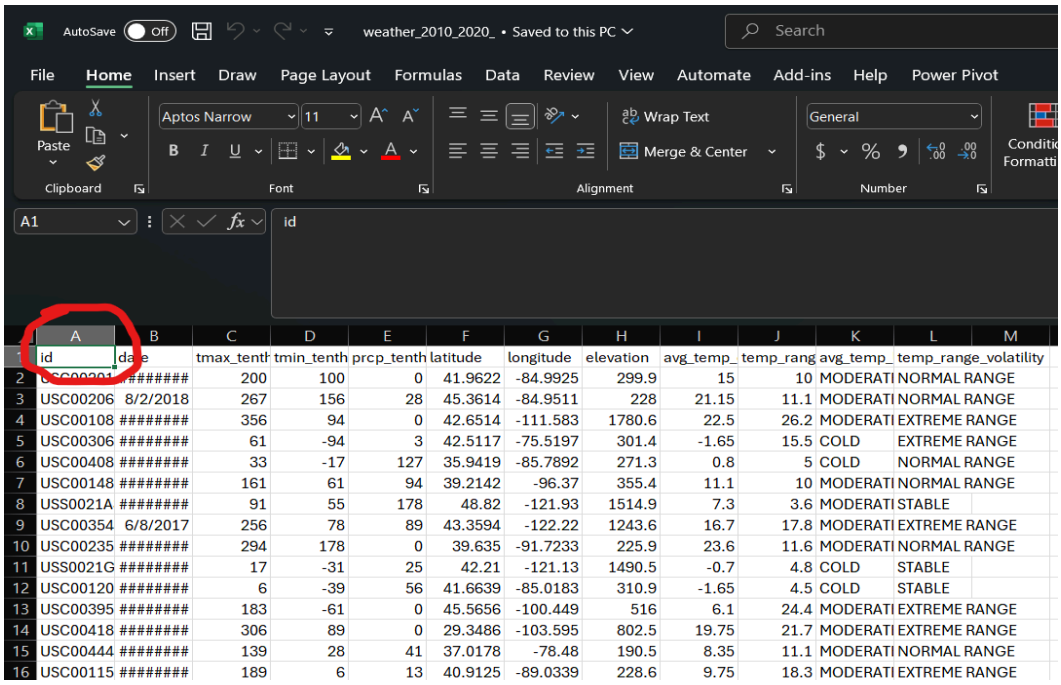
Optional location to store .xlxs file mine was saved in my downloads:



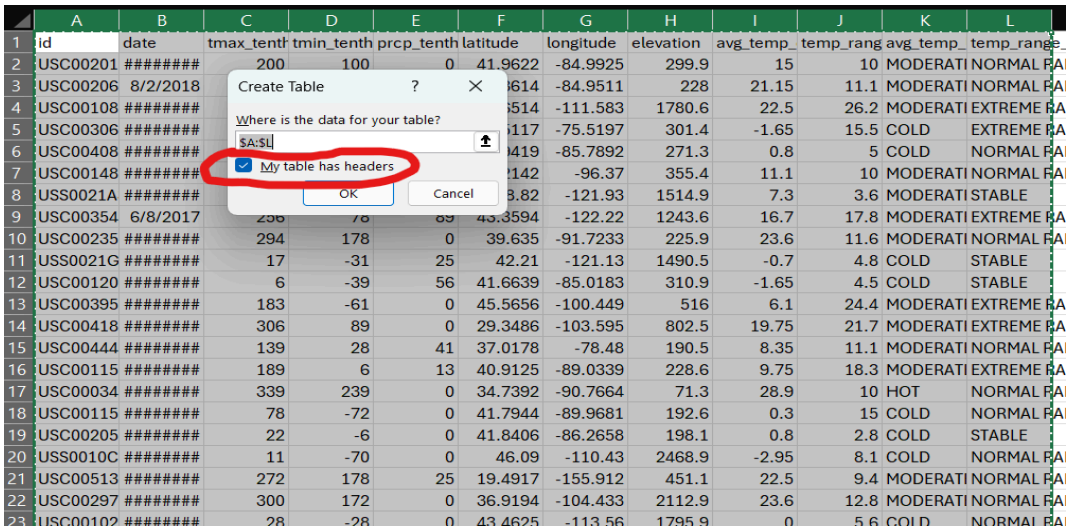
After performing this conversion, this file should automatically open the data in the .xlxs file in Excel so we can prepare for 3D Power Maps. If it didn't automatically open, you will have to go to your File Explorer and go to the location in which you saved this file.

Prepare Data in .xlxs file:

1. Now that were in the .xlxs file let's quickly set up our table. Click any cell within the dataset, then you follow the keyboard shortcuts starting with “Ctrl + A” this will select all the data in the dataset. After this, you’ll follow it up with a “Ctrl + T”, just two quick shortcuts will bring us to the “Create Table” window. MAKE SURE TO SELECT “My table has headers” and leave it checked before hitting “Okay”. This is because the first row in our data contains our headers.

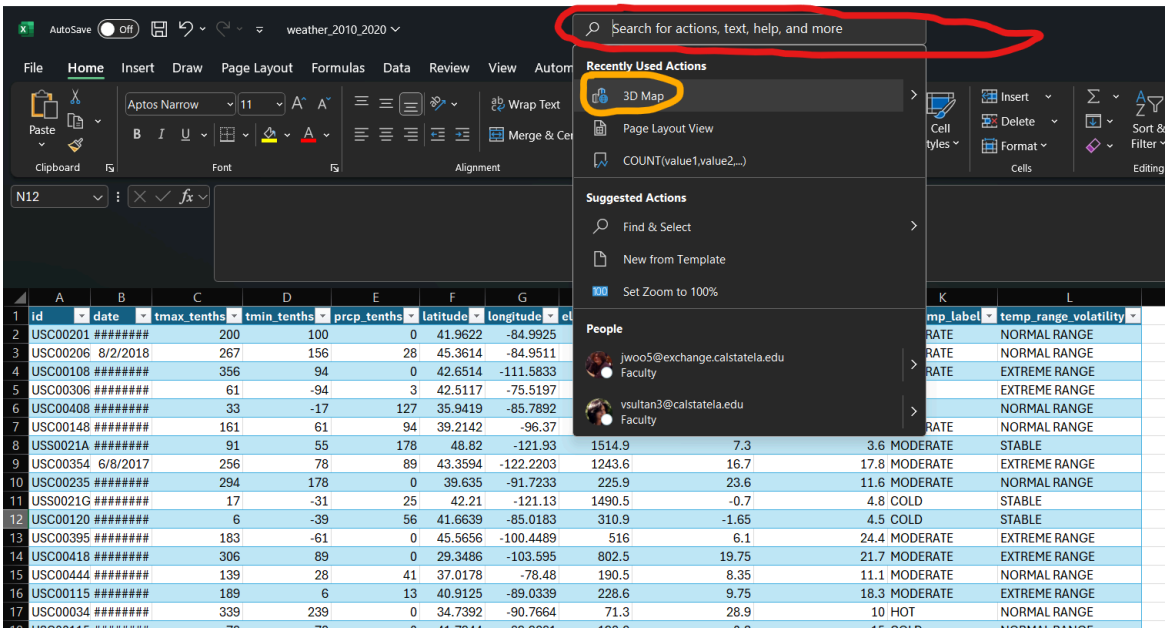


	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	date	tmax_tenth	tmin_tenth	prcp_tenth	latitude	longitude	elevation	avg_temp_	temp_rang	avg_temp_	temp_range_volatility	
2	USC00201	########	200	100	0	41.9622	-84.9925	299.9	15	10	MODERATI	NORMAL RANGE	
3	USC00206	8/2/2018	267	156	28	45.3614	-84.9511	228	21.15	11.1	MODERATI	NORMAL RANGE	
4	USC00108	########	356	94	0	42.6514	-111.583	1780.6	22.5	26.2	MODERATI	EXTREME RANGE	
5	USC00306	########	61	-94	3	42.5117	-75.5197	301.4	-1.65	15.5	COLD	EXTREME RANGE	
6	USC00408	########	33	-17	127	35.9419	-85.7892	271.3	0.8	5	COLD	NORMAL RANGE	
7	USC00148	########	161	61	94	39.2142	-96.37	355.4	11.1	10	MODERATI	NORMAL RANGE	
8	USS0021A	########	91	55	178	48.82	-121.93	1514.9	7.3	3.6	MODERATI	STABLE	
9	USC00354	6/8/2017	256	78	89	43.3594	-122.22	1243.6	16.7	17.8	MODERATI	EXTREME RANGE	
10	USC00235	########	294	178	0	39.635	-91.7233	225.9	23.6	11.6	MODERATI	NORMAL RANGE	
11	USS0021G	########	17	-31	25	42.21	-121.13	1490.5	-0.7	4.8	COLD	STABLE	
12	USC00120	########	6	-39	56	41.6639	-85.0183	310.9	-1.65	4.5	COLD	STABLE	
13	USC00395	########	183	-61	0	45.5656	-100.449	516	6.1	24.4	MODERATI	EXTREME RANGE	
14	USC00418	########	306	89	0	29.3486	-103.595	802.5	19.75	21.7	MODERATI	EXTREME RANGE	
15	USC00444	########	139	28	41	37.0178	-78.48	190.5	8.35	11.1	MODERATI	NORMAL RANGE	
16	USC00115	########	189	6	13	40.9125	-89.0339	228.6	9.75	18.3	MODERATI	EXTREME RANGE	



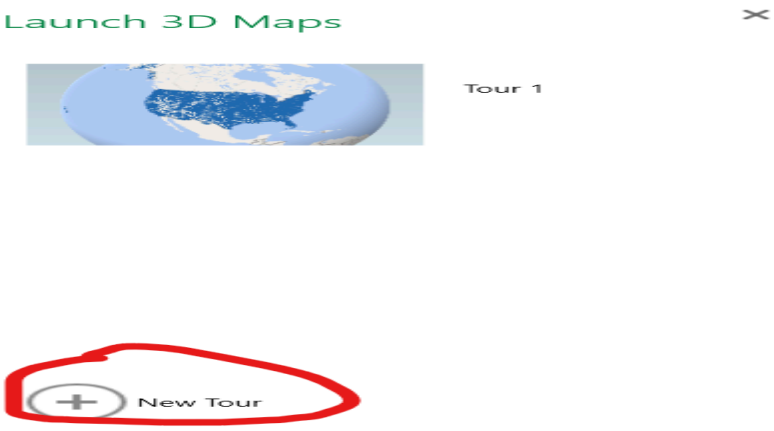
	A	B	C	D	E	F	G	H	I	J	K	L
1	id	date	tmax_tenth	tmin_tenth	prcp_tenth	latitude	longitude	elevation	avg_temp_	temp_rang	avg_temp_	temp_range_
2	USC00201	########	200	100	0	41.9622	-84.9925	299.9	15	10	MODERATI	NORMAL RA
3	USC00206	8/2/2018	267	156	28	45.3614	-84.9511	228	21.15	11.1	MODERATI	NORMAL RA
4	USC00108	########	356	94	0	42.6514	-111.583	1780.6	22.5	26.2	MODERATI	EXTREME RA
5	USC00306	########	61	-94	3	42.5117	-75.5197	301.4	-1.65	15.5	COLD	EXTREME RA
6	USC00408	########	33	-17	127	35.9419	-85.7892	271.3	0.8	5	COLD	NORMAL RA
7	USC00148	########	161	61	94	39.2142	-96.37	355.4	11.1	10	MODERATI	NORMAL RA
8	USS0021A	########	91	55	178	48.82	-121.93	1514.9	7.3	3.6	MODERATI	STABLE
9	USC00354	6/8/2017	256	78	89	43.3594	-122.22	1243.6	16.7	17.8	MODERATI	EXTREME RA
10	USC00235	########	294	178	0	39.635	-91.7233	225.9	23.6	11.6	MODERATI	NORMAL RA
11	USS0021G	########	17	-31	25	42.21	-121.13	1490.5	-0.7	4.8	COLD	STABLE
12	USC00120	########	6	-39	56	41.6639	-85.0183	310.9	-1.65	4.5	COLD	STABLE
13	USC00395	########	183	-61	0	45.5656	-100.449	516	6.1	24.4	MODERATI	EXTREME RA
14	USC00418	########	306	89	0	29.3486	-103.595	802.5	19.75	21.7	MODERATI	EXTREME RA
15	USC00444	########	139	28	41	37.0178	-78.48	190.5	8.35	11.1	MODERATI	NORMAL RA
16	USC00115	########	189	6	13	40.9125	-89.0339	228.6	9.75	18.3	MODERATI	EXTREME RA
17	USC00034	########	339	239	0	34.7392	-90.7664	71.3	28.9	10	HOT	NORMAL RA
18	USC00115	########	78	-72	0	41.7944	-89.9681	192.6	0.3	15	COLD	NORMAL RA
19	USC00205	########	22	-6	0	41.8406	-86.2658	198.1	0.8	2.8	COLD	STABLE
20	USS0010C	########	11	-70	0	46.09	-110.43	2468.9	-2.95	8.1	COLD	NORMAL RA
21	USC00513	########	272	178	25	19.4917	-155.912	451.1	22.5	9.4	MODERATI	NORMAL RA
22	USC00297	########	300	172	0	36.9194	-104.433	2112.9	23.6	12.8	MODERATI	NORMAL RA
23	USC00102	########	28	-28	0	43.4625	-113.56	1795.9	0	5.6	COLD	NORMAL RA

2. Now that our data is within a table allowing us to move forward with 3D Power Map.
    - First we have to open the 3D Power Map, we'll do this by first clicking the "Search for actions, text, help, and more" which below is circled in RED, once searching you'll search for the action by typing "3D Map" into the search bar, which if you look at the below image, is circled in YELLOW.

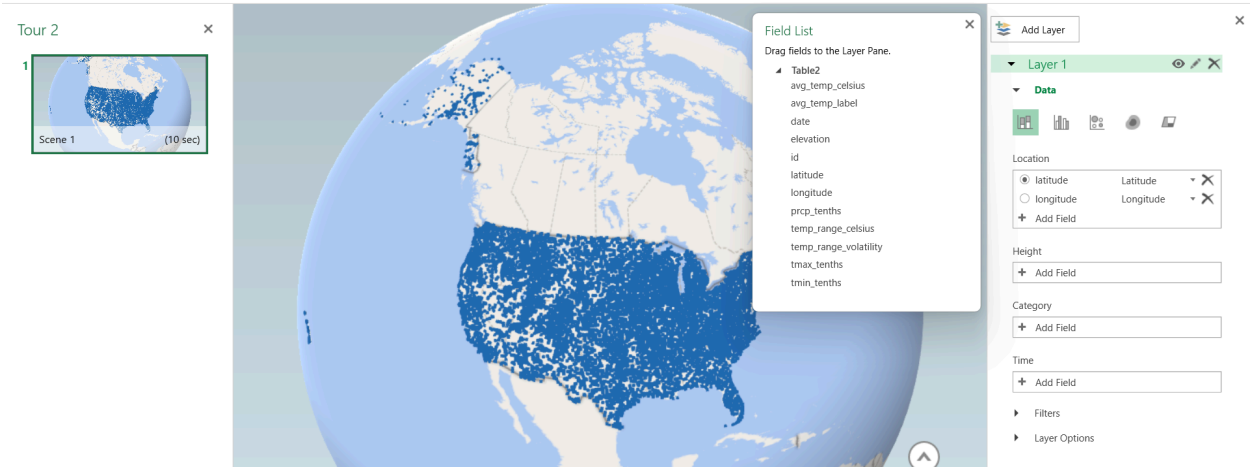


	A	B	C	D	E	F	G	K	L
1	id	date	tmax_tenths	tmin_tenths	prcp_tenths	latitude	longitude	mp_label	temp_range_volatility
2	USC00201	########	200	100	0	41.9622	-84.9925	RATE	NORMAL RANGE
3	USC00206	8/2/2018	267	156	28	45.3614	-84.9511	RATE	NORMAL RANGE
4	USC00108	########	356	94	0	42.6514	-111.5833	RATE	EXTREME RANGE
5	USC00306	########	61	-94	3	42.5117	-75.5197		EXTREME RANGE
6	USC00408	########	33	-17	127	35.9419	-85.7892		NORMAL RANGE
7	USC00148	########	161	61	94	39.2142	-96.37	RATE	NORMAL RANGE
8	USS0021A	########	91	55	178	48.82	-121.93	1514.9	7.3	3.6	MODERATE	STABLE
9	USC00354	6/8/2017	256	78	89	43.3594	-122.2203	1243.6	16.7	17.8	MODERATE	EXTREME RANGE
10	USC00235	########	294	178	0	39.635	-91.7233	225.9	23.6	11.6	MODERATE	NORMAL RANGE
11	USS0021G	########	17	-31	25	42.21	-121.13	1490.5	-0.7	4.8	COLD	STABLE
12	USC00120	########	6	-39	56	41.6639	-85.0183	310.9	-1.65	4.5	COLD	STABLE
13	USC00395	########	183	-61	0	45.5656	-100.4489	516	6.1	24.4	MODERATE	EXTREME RANGE
14	USC00418	########	306	89	0	29.3486	-103.595	802.5	19.75	21.7	MODERATE	EXTREME RANGE
15	USC00444	########	139	28	41	37.0178	-78.48	190.5	8.35	11.1	MODERATE	NORMAL RANGE
16	USC00115	########	189	6	13	40.9125	-89.0339	228.6	9.75	18.3	MODERATE	EXTREME RANGE
17	USC00034	########	339	239	0	34.7392	-90.7664	71.3	28.9	10	HOT	NORMAL RANGE

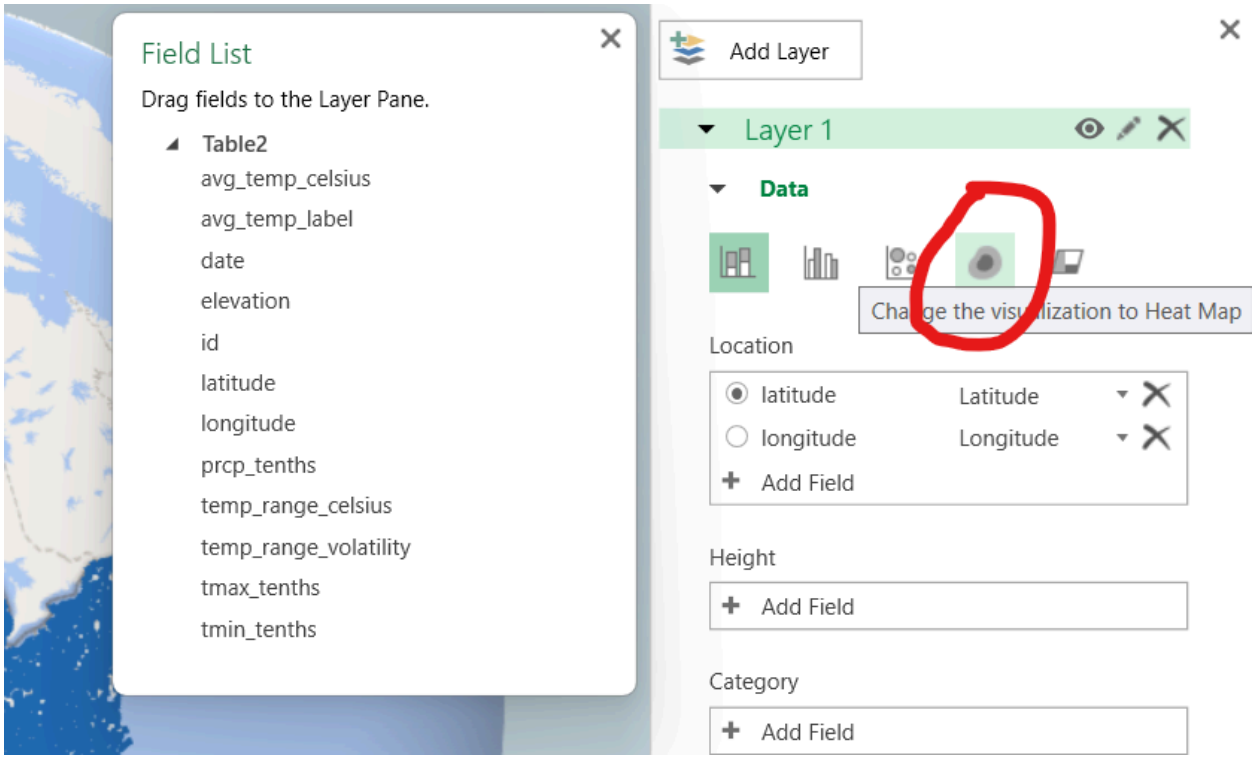
3. Once you click on the 3D power map after searching, this window will open just create a new tour circled in RED shown below.



4. This is how 3D power map should look after clicking “New Tour”, (See Figure Below)

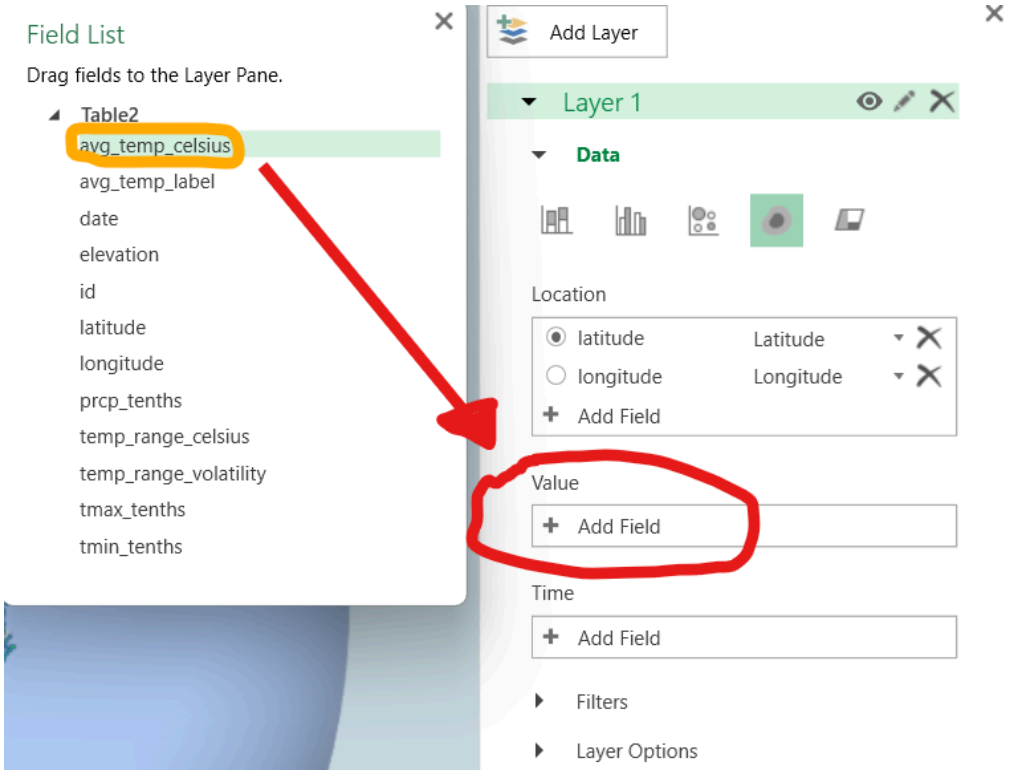


- Change to Heat Map (Shown Below).

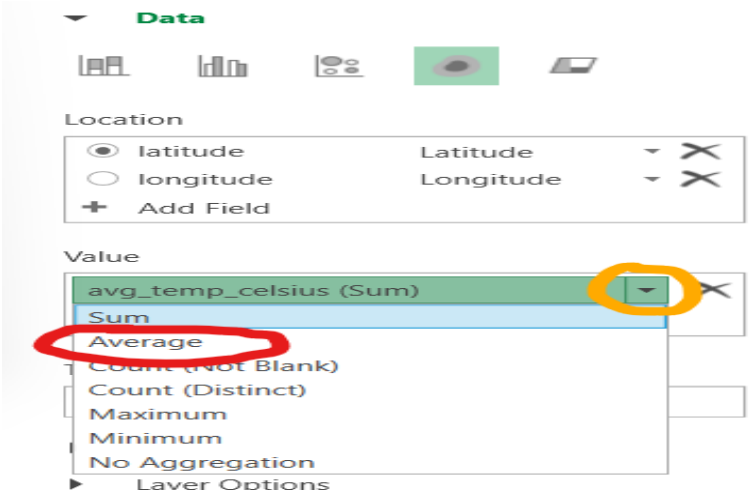


5. Now let’s get to creating the heat map,

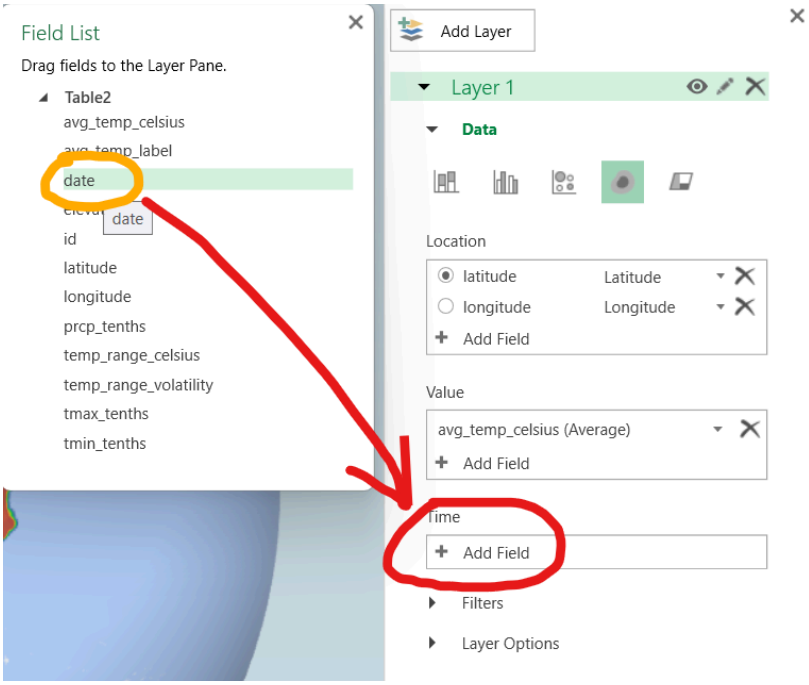
- The first step will be to drag from the “Field List” box the field “average_temp_celsius” shown circled YELLOW, into the “Value” field circled in RED.



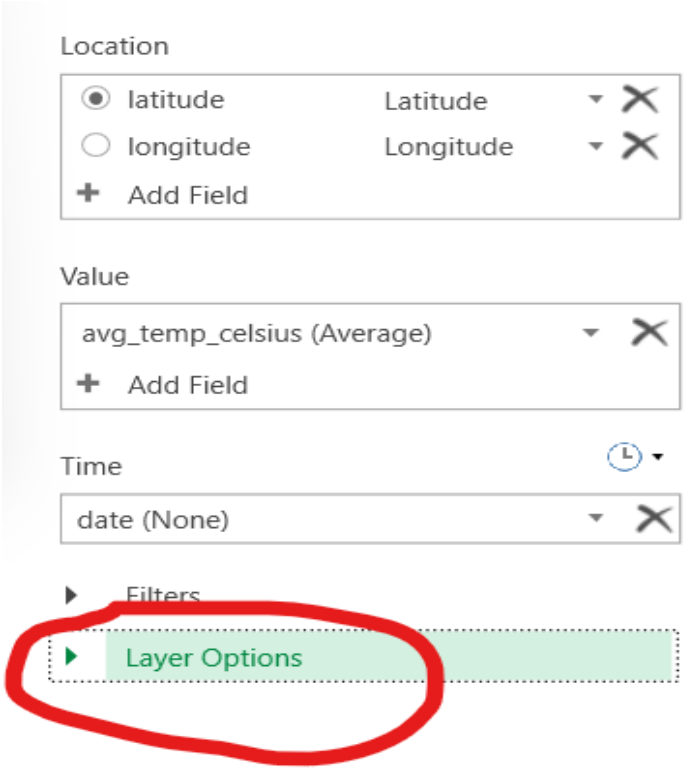
- After dragging the field you'll notice that it automatically got calculated as "SUM", however we need to change this to "Average". Below in the image, you'll first click the triangle to open the drop down which is circled in YELLOW. Then once the drop box opens select "Average" shown circled in RED.



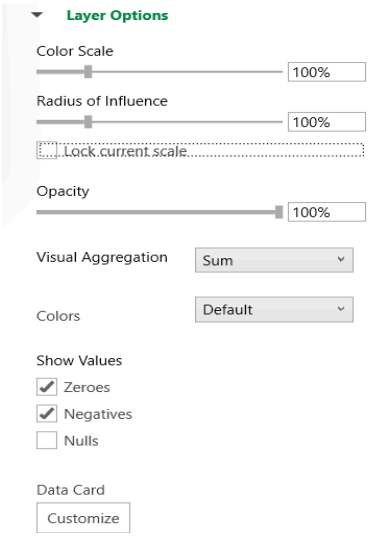
-Next, we have to go back to the field list box, this time we'll be draggin the "date"(Shown circled YELLOW) field to within "Time" which is circled in RED.



-Now we have to select “Layer options” circled in RED. This will open up the layer options

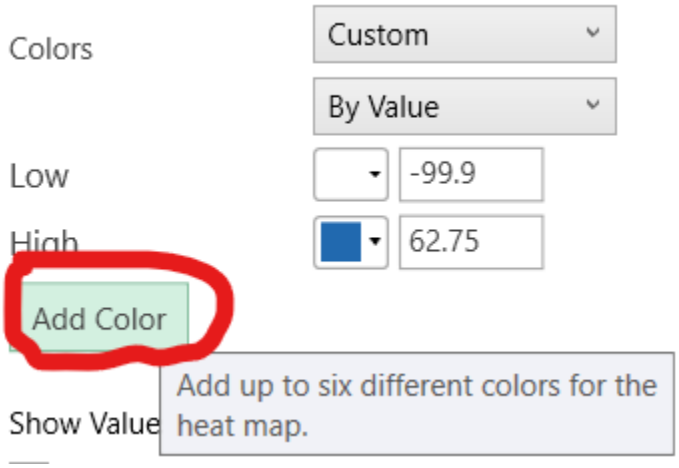


- After opening Layer Options it should show the image below which are all the default values.



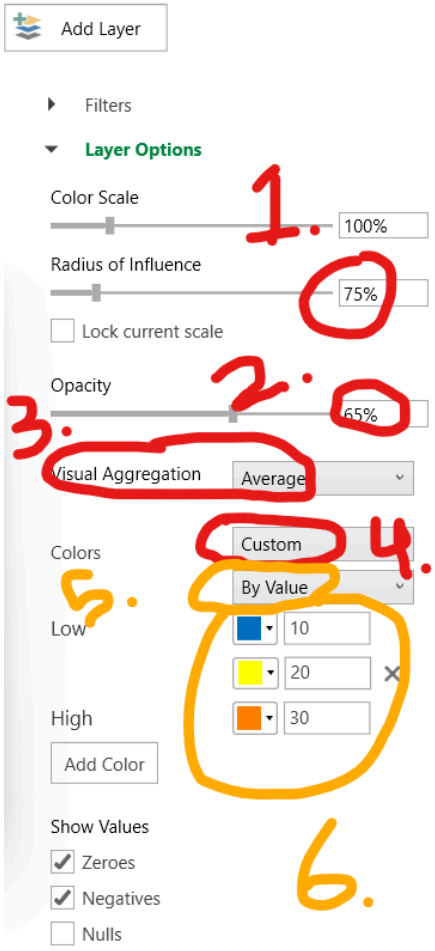
- You'll want to revise and edit these layer options following the image below. The image below demonstrates six circles that occur in an order going from 1-6. First begin with (1), then follow in order up to (6).

1. Drop "Radius of Influence" to 75% so we can visualize the data better on the map without clusters taking too much capacity.
2. Drop "Opacity" to 65% this helps us on the map by allowing us to better see the geographic names and info.
3. For "Visual Aggregation" change this to average since our "avg_temp_celsius" is averaged.
4. For "Colors" we're going to do custom colors to better represent a Heat Map classification. After selecting "custom" a new box appears under it with a default value of "Auto", we'll change this to "By Value". Since the default value of the number of colors is only two we must select "Add Color" shown below circled in RED. This will give us the three color options shown in the screenshot with the correct Layer options.

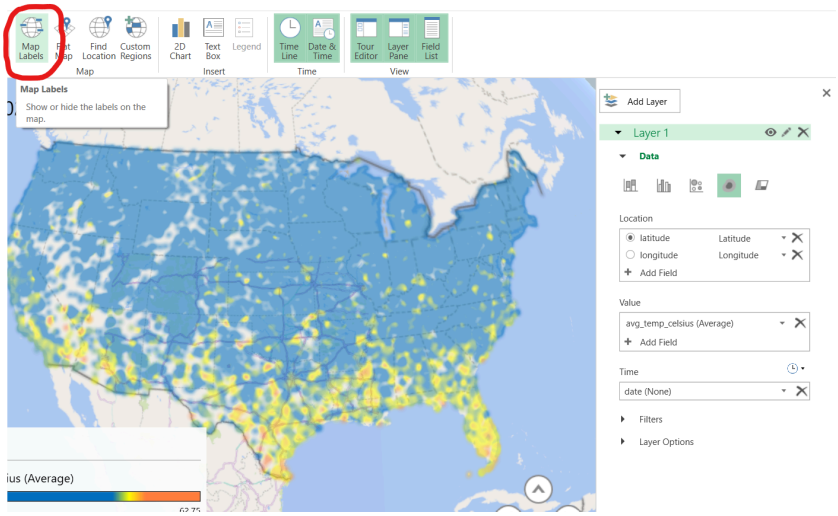


5. If not done so already, mark the “Auto” box to “By Value”
6. Now fill in the color ranges as well as fill in the appropriate colors.

   (Blue=10(COLD)Yellow=20(MODERATE)Orange=30(HOT)



- Now that we have our ready 3D power map make sure you turn on map labels so you can gather more geographic knowledge while looking at the map. Circled in RED.



- The 3D power map is now ready for viewing. Circled in RED in the image below, is the “Time Line” by selecting any place on this line you can view that point in time’s Heat Map. You can also initiate and select the Play button within it, which will show the Heat map over time going through every recorded date in the data. Remember to also save the file when done to save changes made.

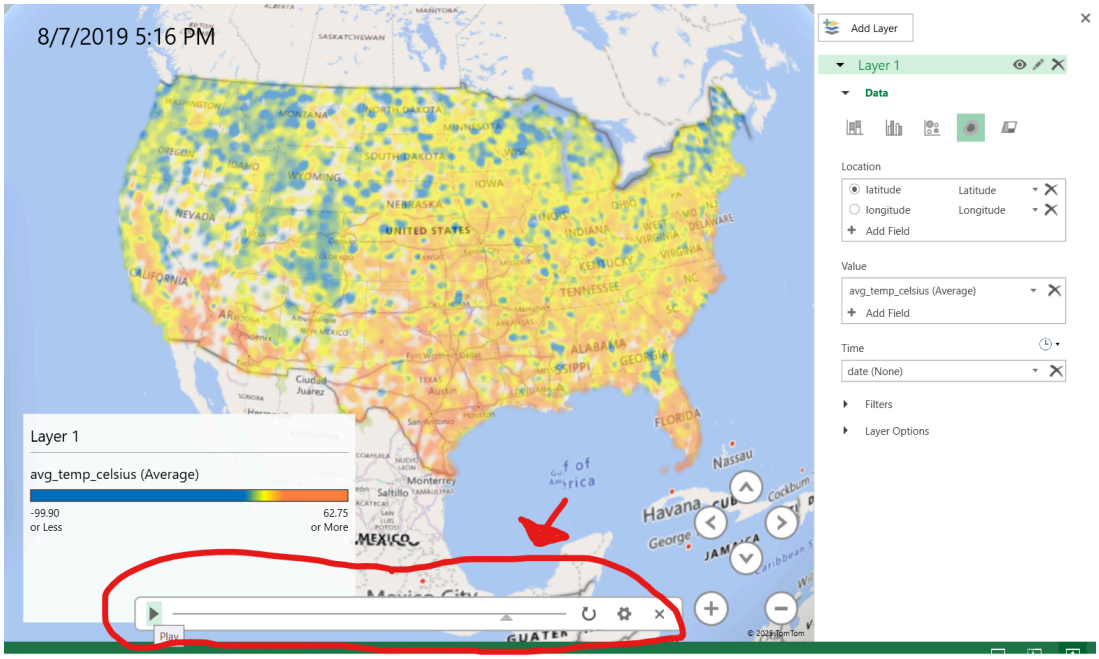
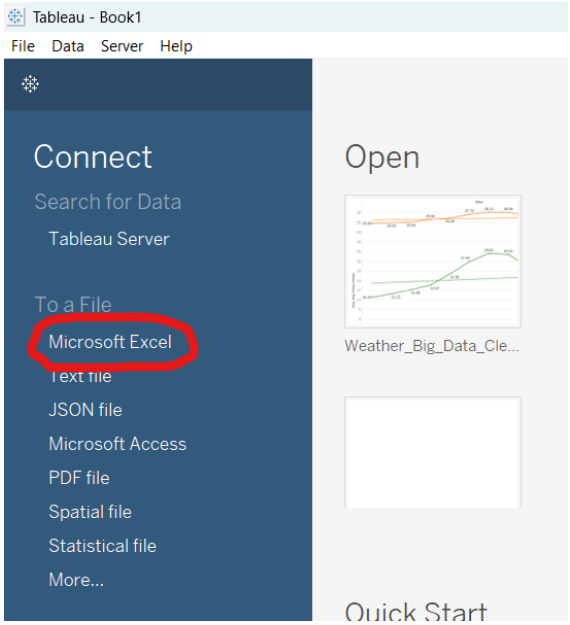


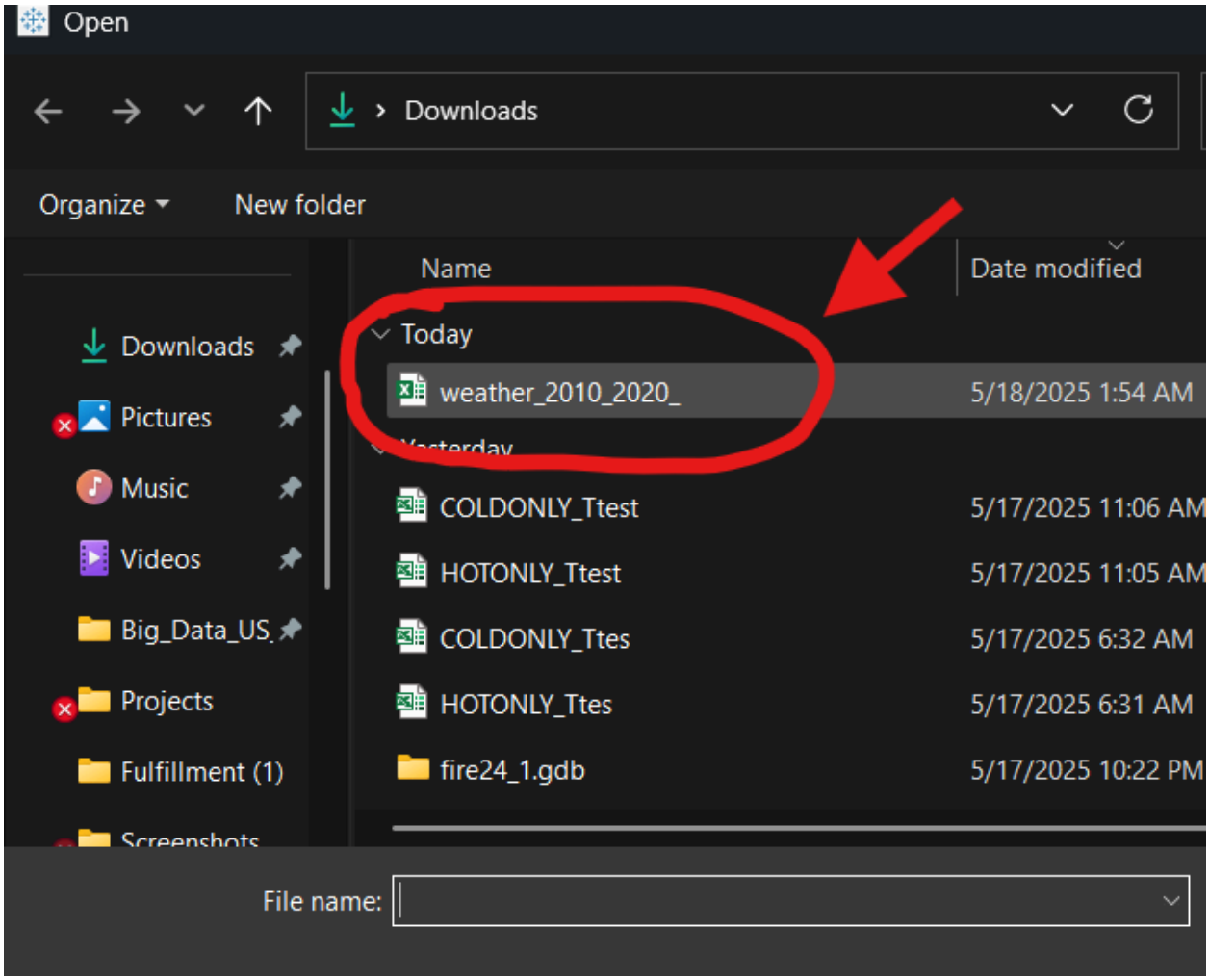
Tableau Data Analysis:

With 3D power maps now finished, we can move on to data analysis within Tableau. We'll begin with connecting to our .xlxs file on the Tableau Homescreen. Then we'll go over the process of bar chart creation to analyze precipitation in the early decade(2011-2015) along with the late decade(2016-2020) to look at how the precipitation has changed. Finally we'll create our Diurnal temperature range over time which is a great metric to measure Global Warming as it itself is greenhouse-induced warming. Diurnal Temperature range(DTR) is our temp_range_celsius attribute, the DTR is configured through the difference between daily max and min temperatures(tmax, tmin) which is already configured in temp_range_celsius. Let's get into the Tableau process.

1. We begin on the Tableau Homescreen where we can connect Tableau to a file, specifically the Excel file that we already have worked with that you saved after creating the Heat map on 3D Maps. To do this refer to the image below, you can connect directly to the file by selecting "Microsoft Excel" on the left of the homescreen circled in RED.

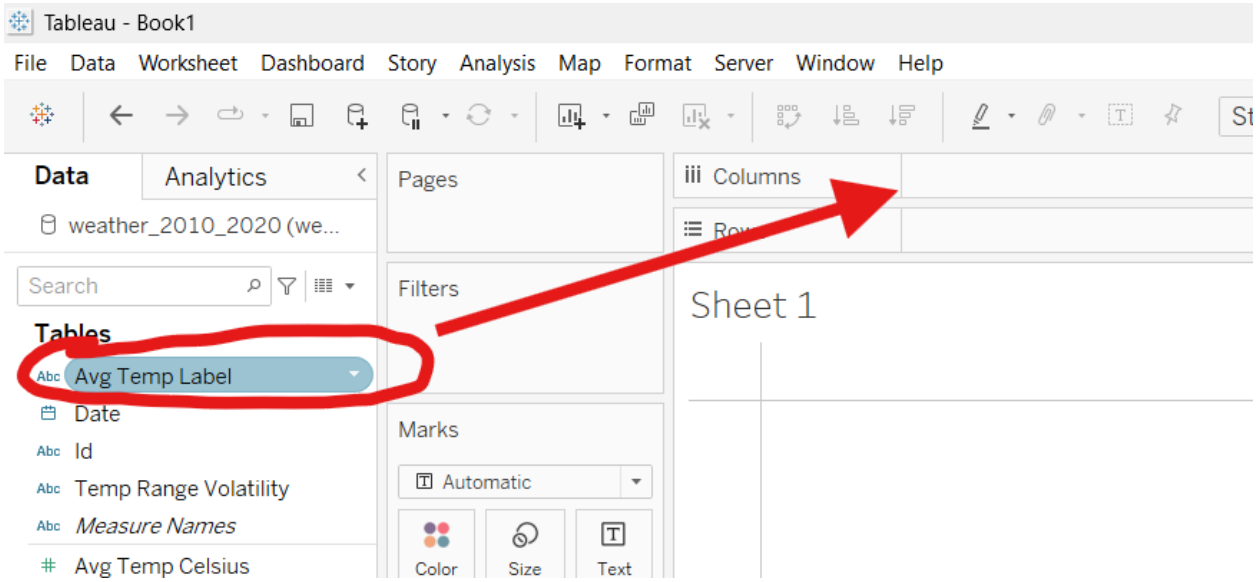


2. Once you selected “Microsoft Excel”, your file explorer window will open, now go to the location where you saved your file,(mine was in my downloads) and select the file to connect with Tableau using a double-click when hovering your cursor over the file selection.

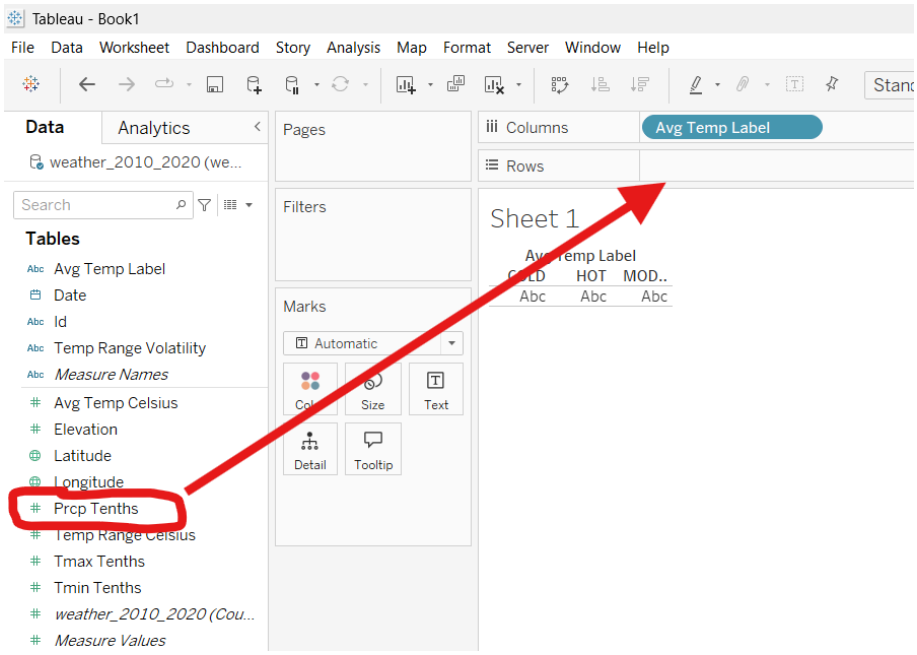


3. Once you have selected your file, Tableau will establish a connection with it and you should be taken to the Tableau data source connections screen. Looking towards the bottom left of the connections screen, shown circled in RED is the “Sheet 1”, select this as this is where we’ll create our first visualization.

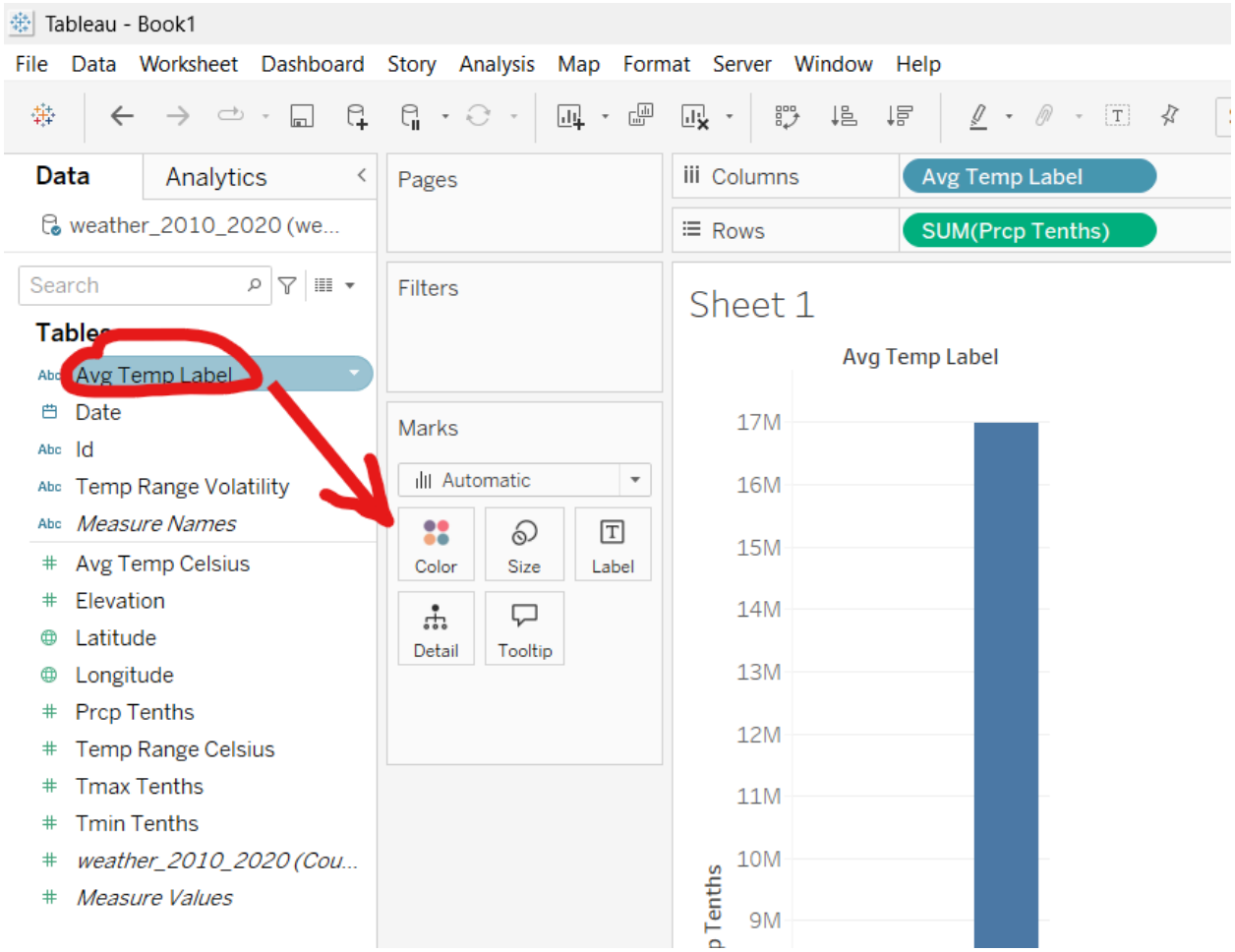
4. Now that we're within the sheet, select "Avg Temp Label" and drag it (shown circled in RED) into your "Columns" box(Shown with RED arrow).



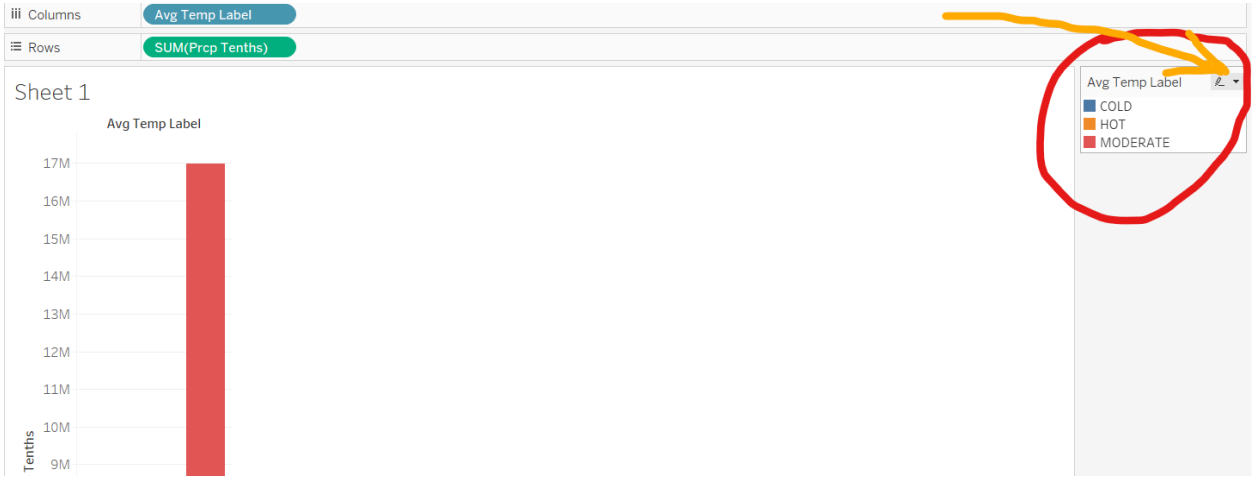
5. Now that "Avg Temp Label" is in our columns box, we can now select "Prcp Tenths" and drag it (shown circled in RED) into your "Rows" box(Shown with RED arrow).



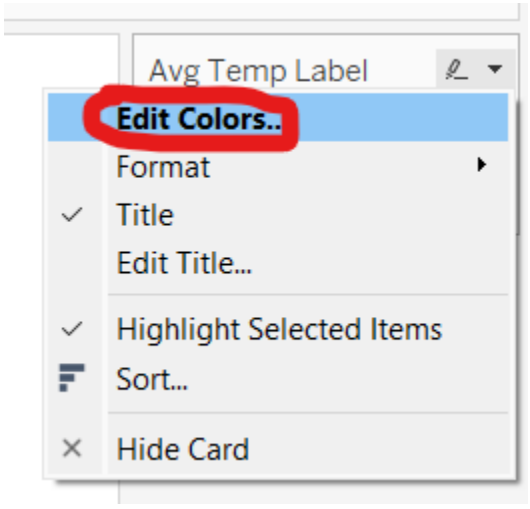
6. Now our bar chart is made and split into three columns, one for “COLD”, “HOT”, and “MODERATE”. Let’s begin adding some more depth to it to enhance the visualization and analysis, starting by selecting “Avg Temp Label” from the “Tables” section again on the left side, and drag it (shown circled in RED) into your “Color” located in the “Marks” box(Shown with RED arrow).



7. Now the three columns of our bar chart are colored differently, in order to avoid any confusion when looking at the visualization, we are going to edit the color for the “MODERATE” bar so that it’s not being associated with hotter temperatures since red and warmer colors tends to be associated with “hot”. Look to the right side of your screen and select the “triangle”(Shown with YELLOW arrow) drop down in the top right corner of the “Avg Temp Label”(Shown circled in RED). (NOTE THAT THE TRIANGLE DOESN”T APPEAR UNTIL YOU ARE HOVERING YOUR CURSOR OVER IT)



8. Now that the drop down has opened select “Edit Colors” shown circled in RED.

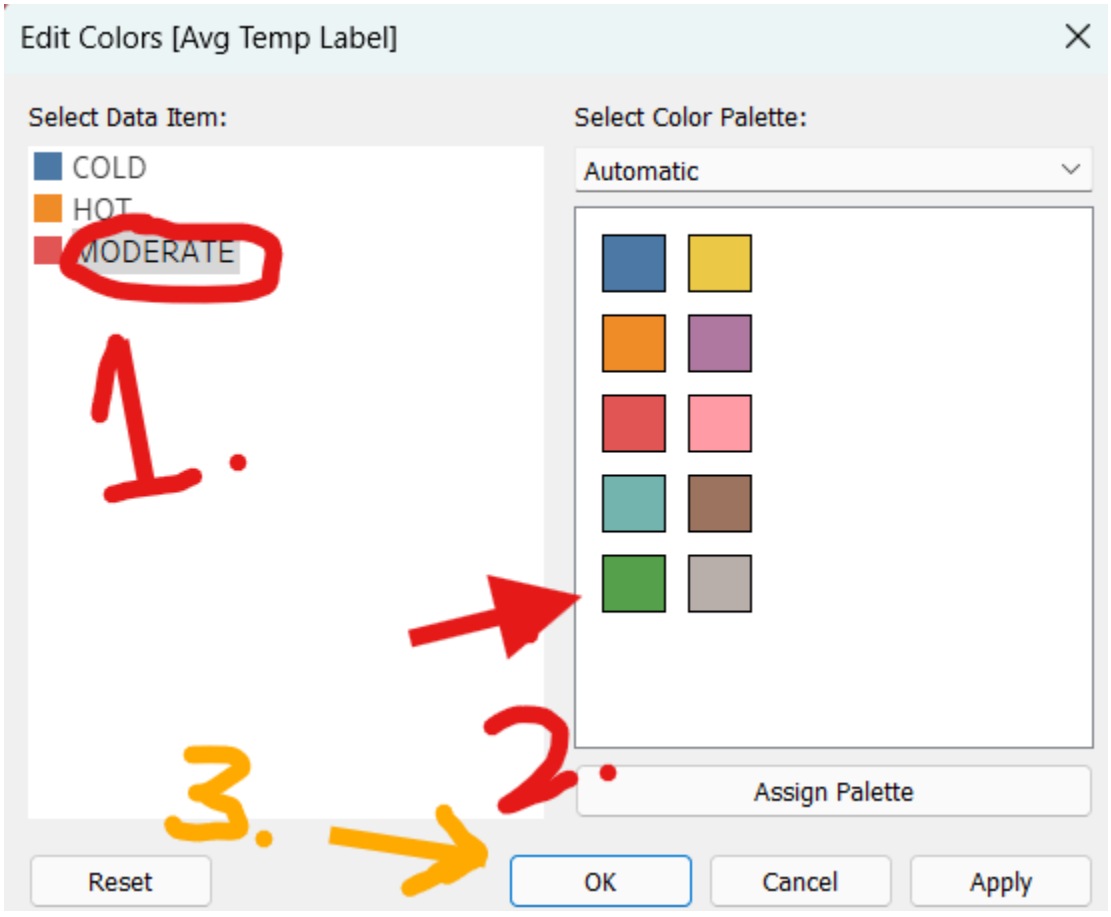


9. Now that the “Edit Colors” window we can edit the color for “MODERATE”

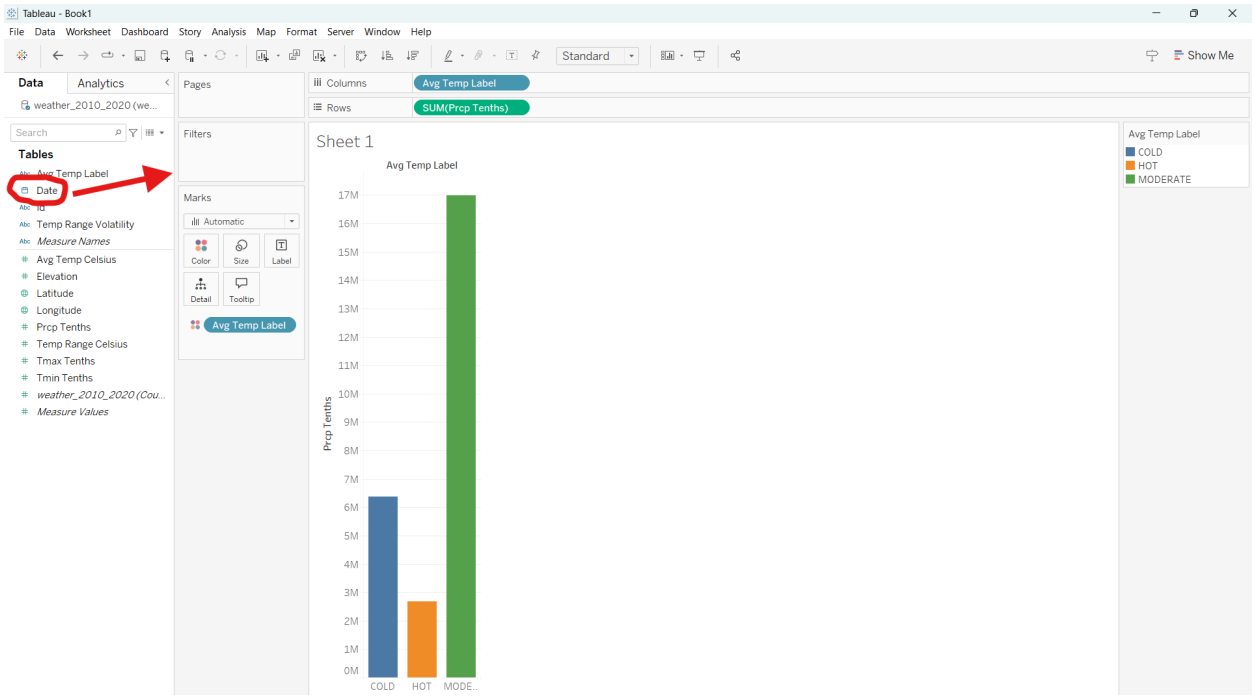
   - (1)First make sure “MODERATE” is selected within the “Select Data Item”(Shown circled in RED) section on the left side where it shows all of our temperatures and the current colors being associated with them.

   - (2)Now that “MODERATE” has been selected we can change the color to green, by selecting the green square (Shown with RED arrow) on the right of the window where all the colors are located. (You’ll know if done correctly if within the “Select Data Item” section “MODERATE” should now show a green square instead of the red one originally.)

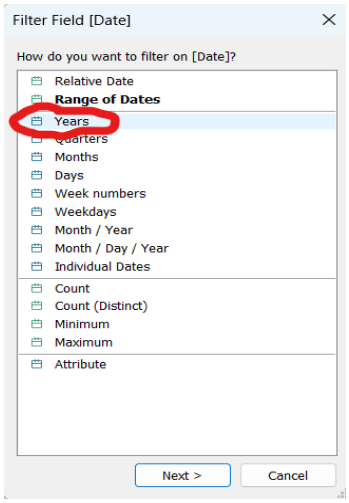
   - (3)To apply the changes to the chart, select “OK” shown with the YELLOW arrow.



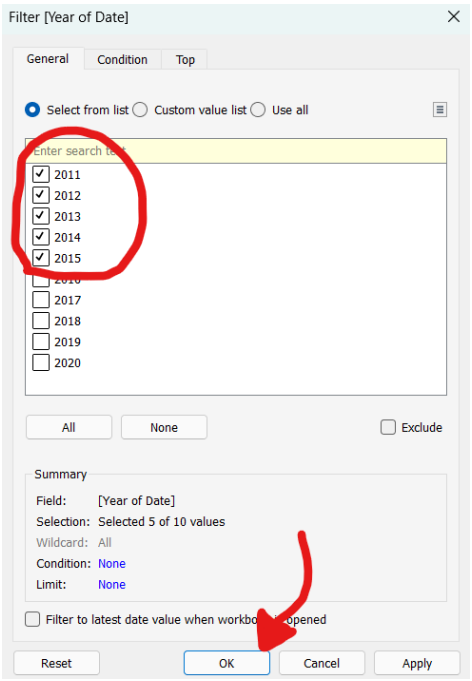
10. Our colors are now applied to our chart, lastly let's filter the date so that this visualization only shows early decade years. To do this go back to the “Tables” section on the left side of the screen and select and drag the “Date” dimension into your “Filters” box.



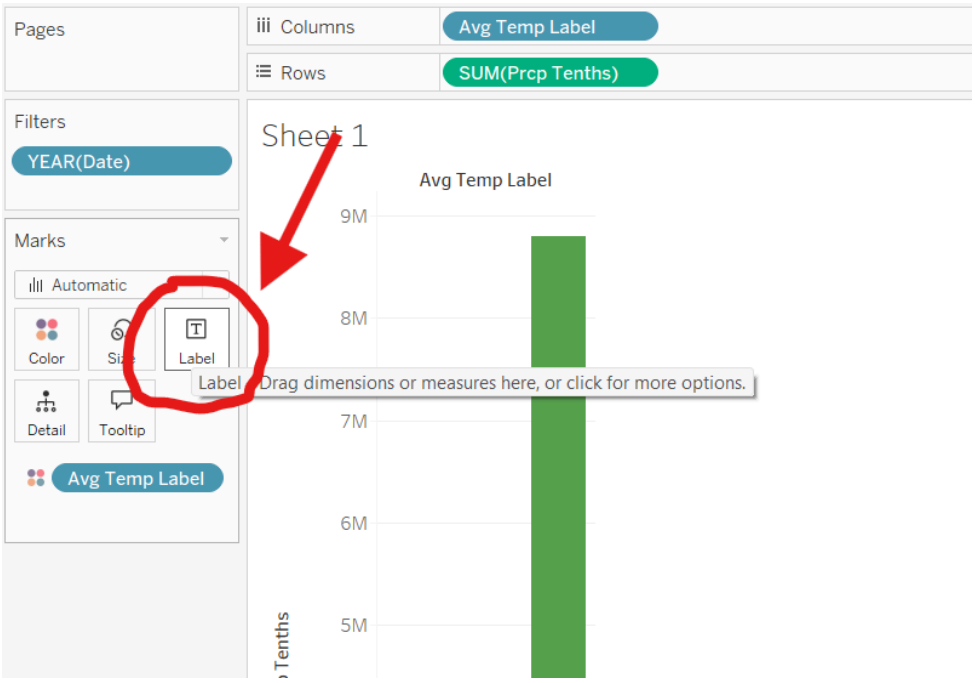
11. Once done, a new “Filter Field [Date]” window opens asking how do we want the date filtered. You must select “Years”(shown circled in RED) which is the third option from the list. To select simply double click, or select “Years” and select “Next” at the bottom.



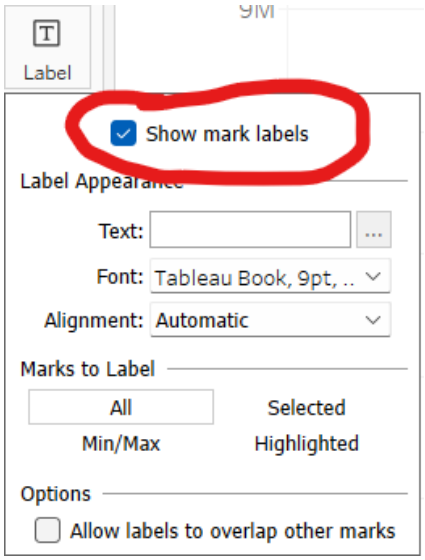
12. Now copy the selections of the years with the image below(circled in RED) 2011-2015, this will be our early decade chart. After selecting all five values select “OK” to apply the filter.(Shown with RED arrow)



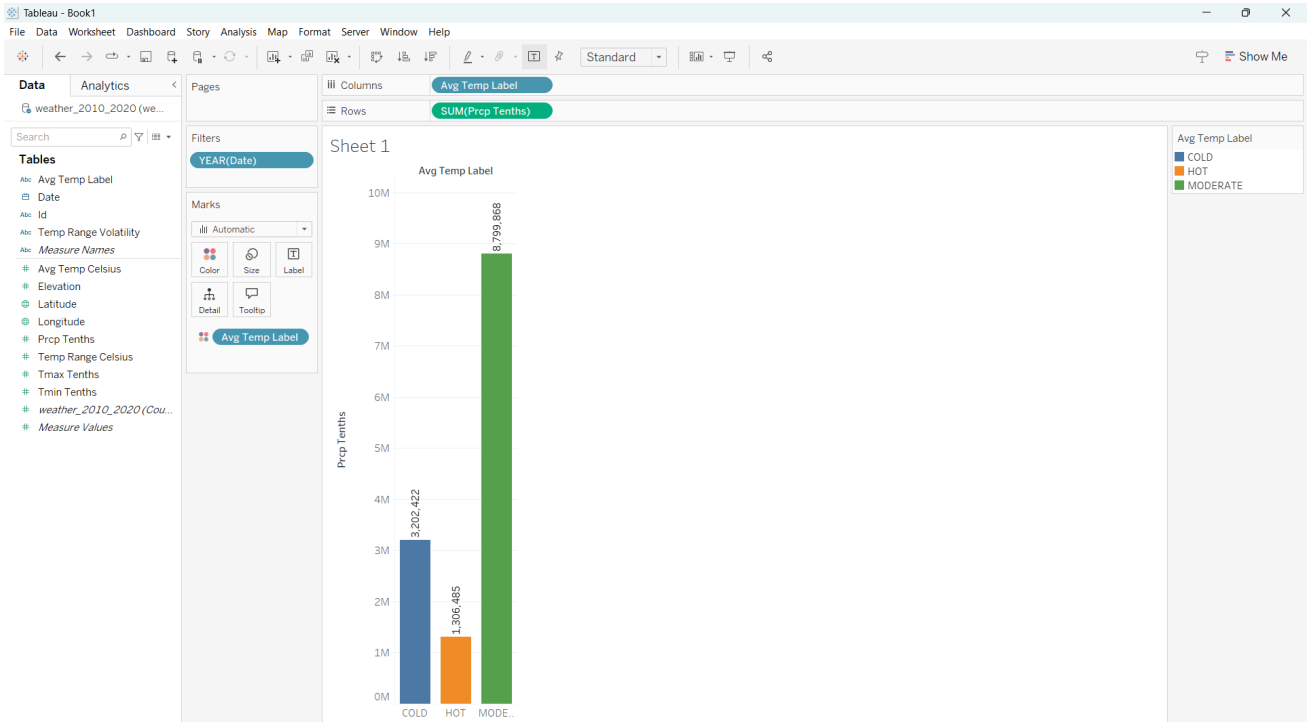
13. Select the “Labels” icon within the Marks box



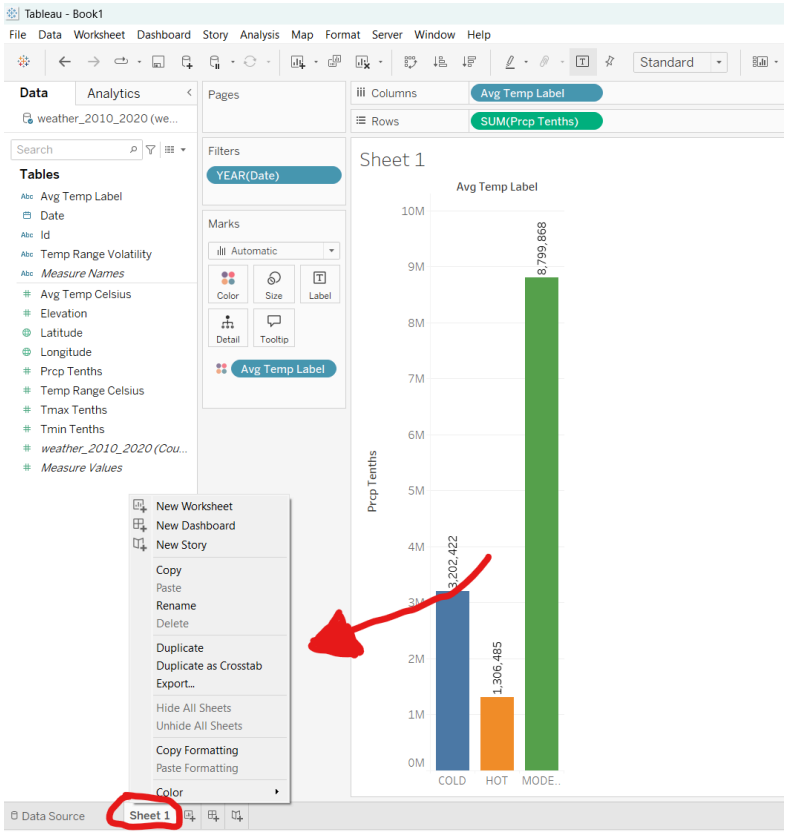
14. After selecting the “Label” icon a small window will open for it. Make sure “Show mark labels” is checked.



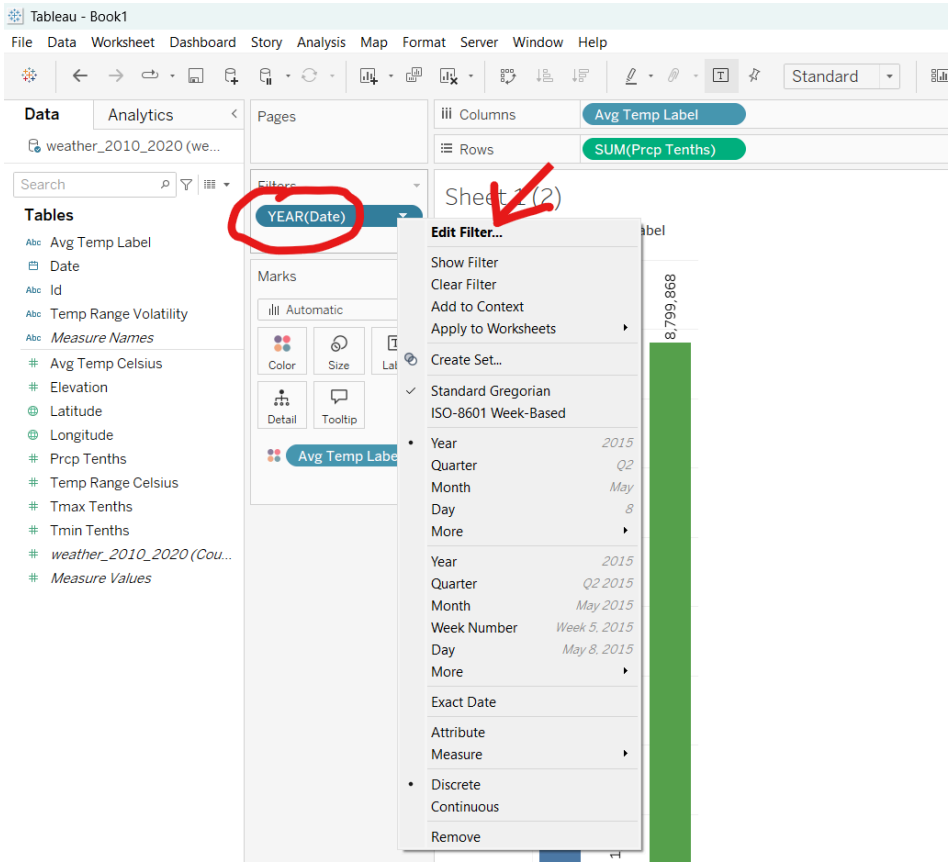
15. Now we can see how precipitation is spread through 2011-2015 in different temperature labels.



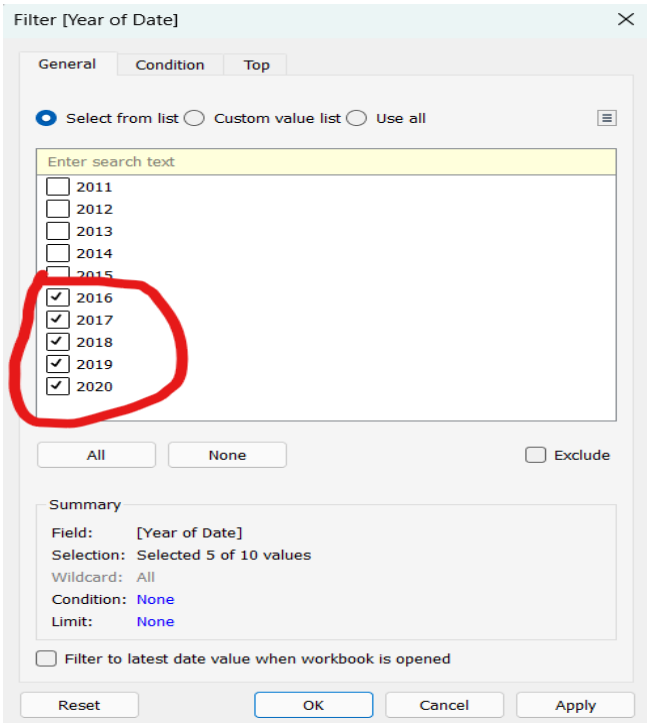
16. To compare early decade with late decade we'll be duplicating this sheet and changing our year filter from 2011-2015, into 2016-2020. First Right-click on the bottom of the screen within "Sheet 1"(Shown circled in RED), it should open up a section above where you right-clicked, this is where you select "Duplicate"(Shown with RED arrow) so all changes from this chart are replicated to the next sheet.



17. Now that you've duplicated the chart, we only have to change the filter now since all other chart elements have been duplicated already. First Right-click the "YEAR(Date)"(Shown circled RED) within your filters, this will open up the section where you can select "Edit Filter"(Shown with RED arrow)



18. Now that we're within the "Filter [Year of Date]" window, we can change the previous selection of 2011-2015 into 2016-2020 like the image shown below circled in RED. Then select "OK" to apply changes and close the filter window.



19. Now we have both charts one for early decade precipitation(First Sheet), and one for late decade precipitation(Second Sheet) so that we can compare the two to see how the precipitation is distributed throughout this decade.

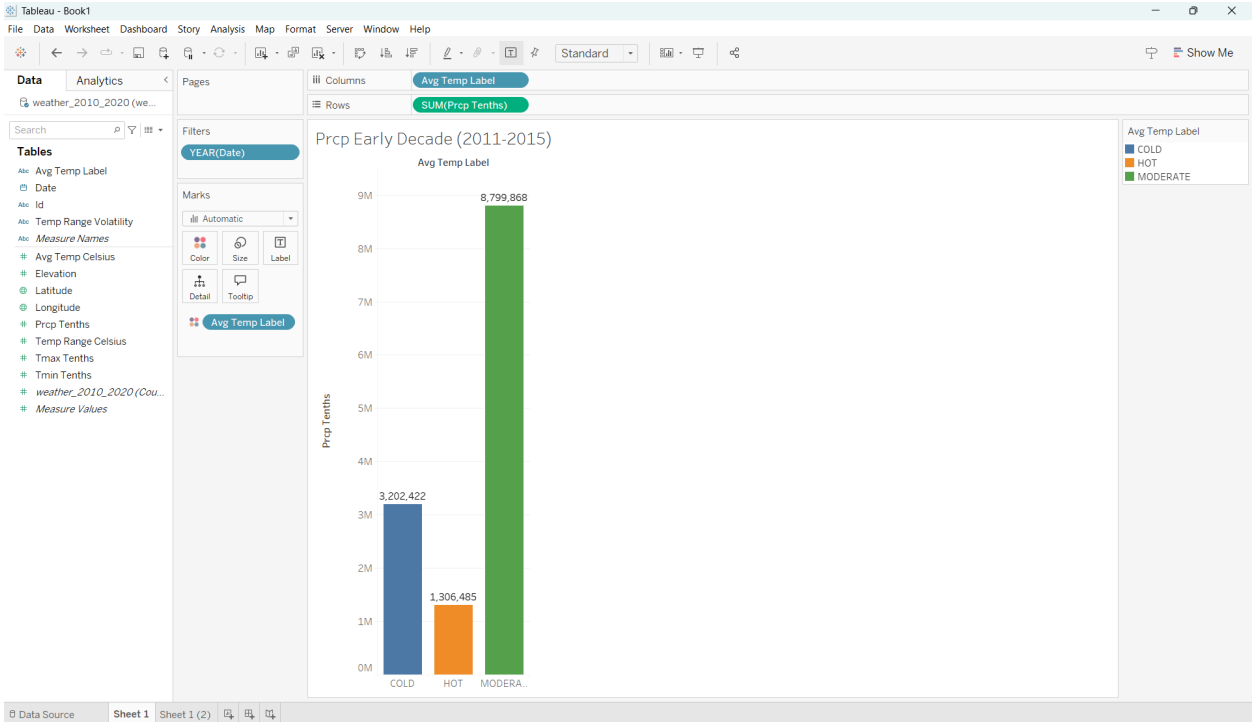


Tableau - Book1
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Data
Analytics
weather_2010_2020 (we...
Search
Tables
Abc Avg Temp Label
Date
Abc Id
Abc Temp Range Volatility
Abc Measure Names
Avg Temp Celsius
Elevation
Latitude
Longitude
Prcp Tenths
Temp Range Celsius
Tmax Tenths
Tmin Tenths
weather_2010_2020 (Cou...
Measure Values
Pages
Filters
YEAR(Date)
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Avg Temp Label
Columns
Avg Temp Label
Rows
SUM(Prcp Tenths)
Prcp Early Decade (2011-2015)
Avg Temp Label
Prcp Tenths
9M
8M
7M
6M
5M
4M
3M
2M
1M
0M
3,202,422
1,306,485
8,799,868
COLD
HOT
MODERA..
Avg Temp Label
COLD
HOT
MODERATE
Data Source
Sheet 1
Sheet 1 (2)

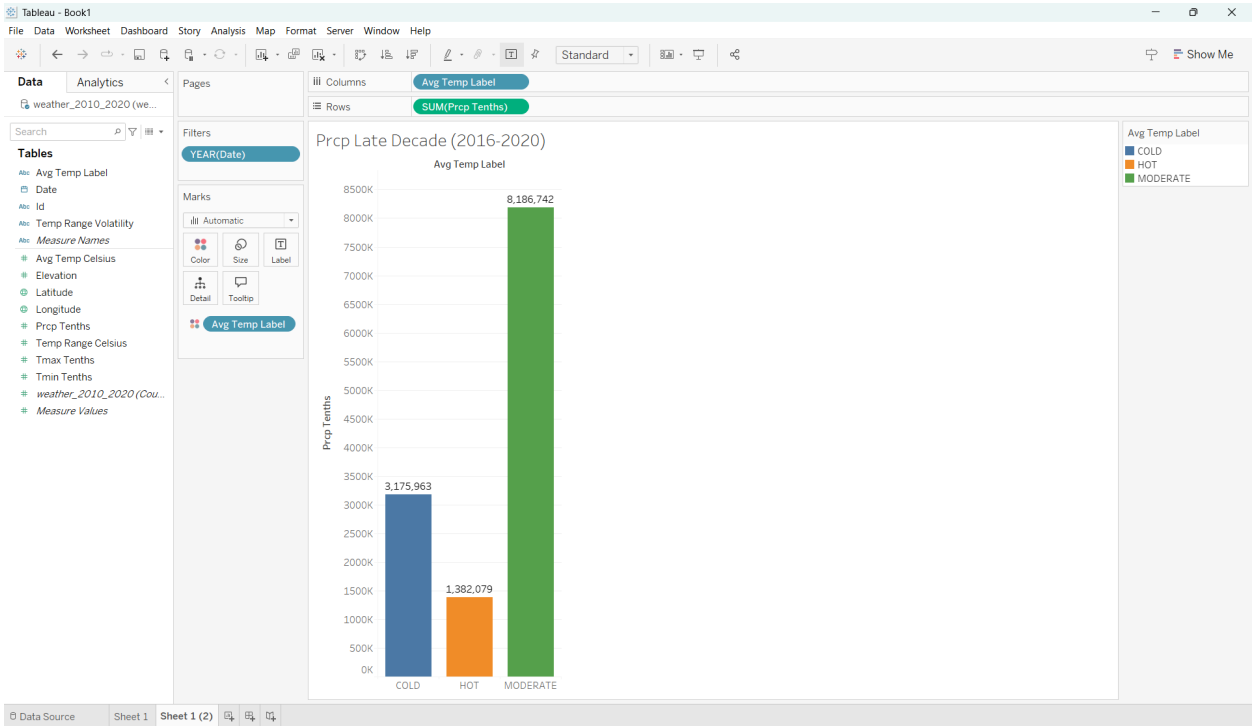
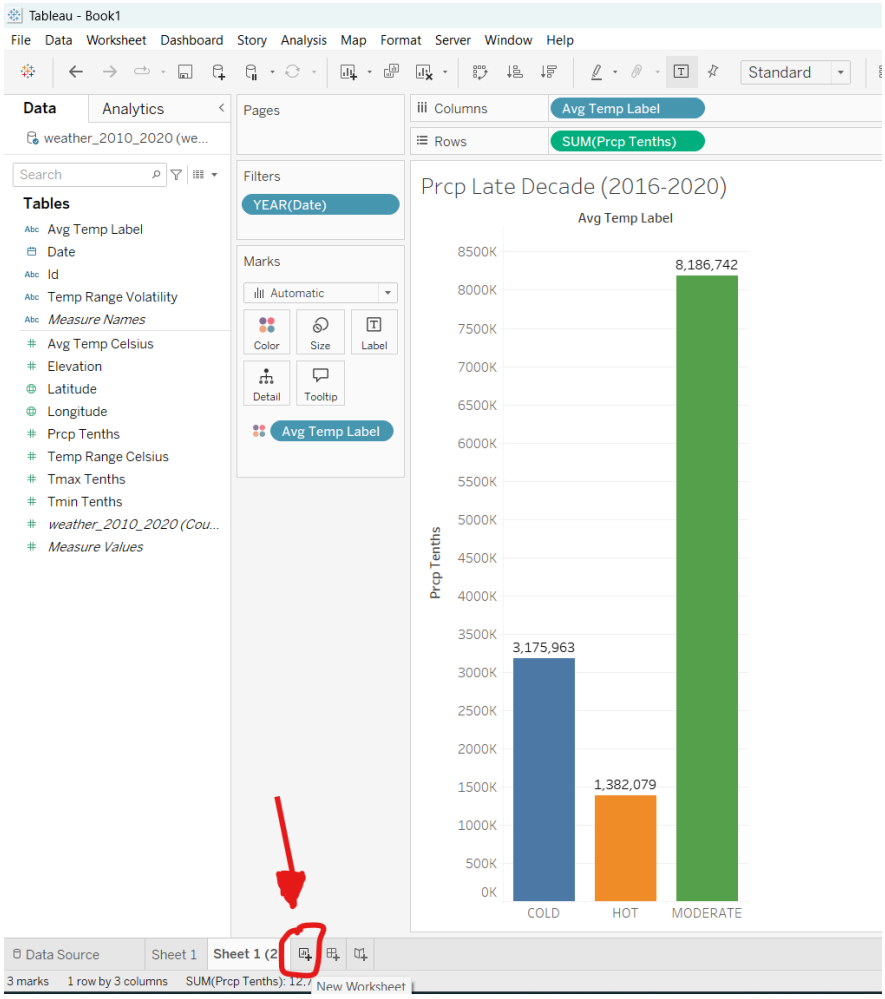


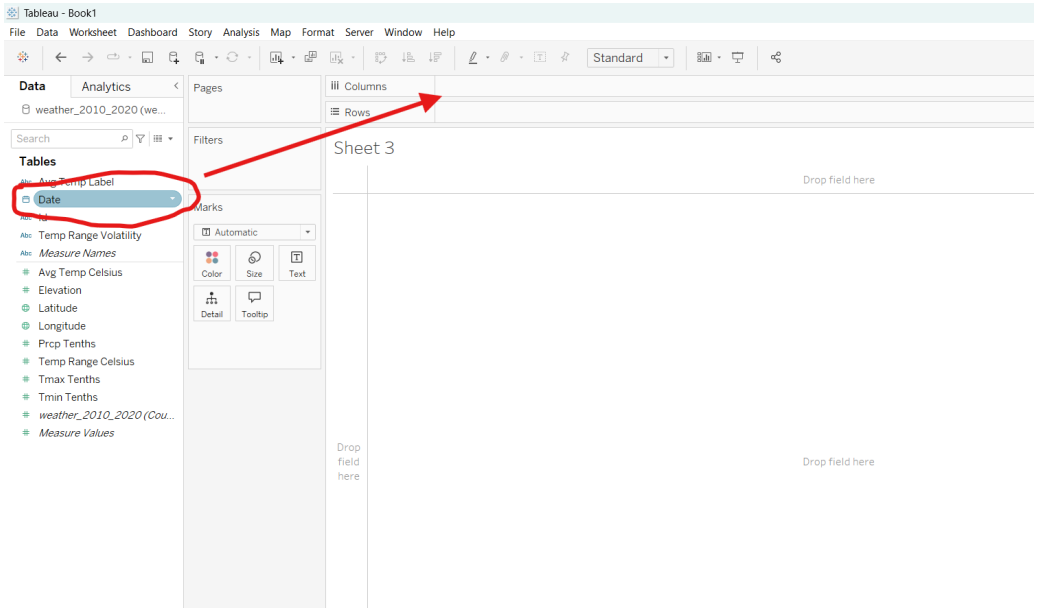
Tableau - Book1
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Data
Analytics
weather_2010_2020 (we...
Search
Tables
Abc Avg Temp Label
Date
Abc Id
Abc Temp Range Volatility
Abc Measure Names
Avg Temp Celsius
Elevation
Latitude
Longitude
Prcp Tenths
Temp Range Celsius
Tmax Tenths
Tmin Tenths
weather_2010_2020 (Cou...
Measure Values
Pages
Filters
YEAR(Date)
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Avg Temp Label
Columns
Avg Temp Label
Rows
SUM(Prcp Tenths)
Prcp Late Decade (2016-2020)
Avg Temp Label
Prcp Tenths
8500K
8000K
7500K
7000K
6500K
6000K
5500K
5000K
4500K
4000K
3500K
3000K
2500K
2000K
1500K
1000K
500K
0K
3,175,963
1,382,079
8,186,742
COLD
HOT
MODERATE
Avg Temp Label
COLD
HOT
MODERATE
Data Source
Sheet 1
Sheet 1 (2)

20. The final visualization we'll be creating is the diurnal temperature range (DTR) over time.

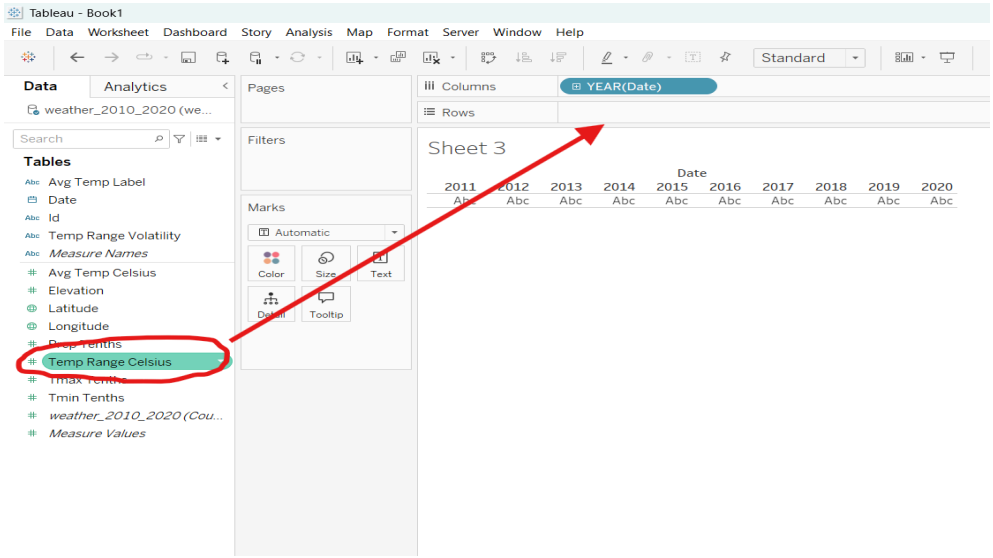
    - To begin select the new sheet Icon(shown circled in RED with RED arrow)



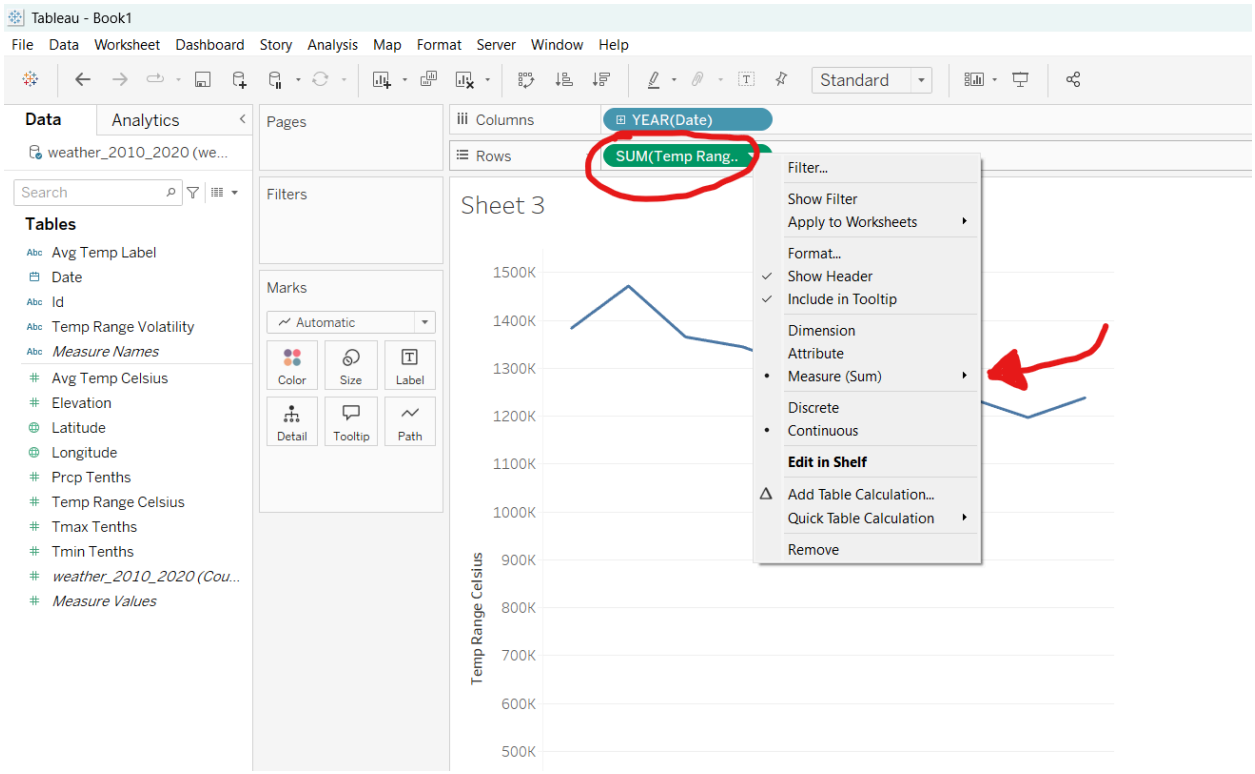
21. Look to the Tables on the left and you want to select and drag the “Date” dimension(Shown circled RED) and drop it into your columns.(Shown with RED arrow).



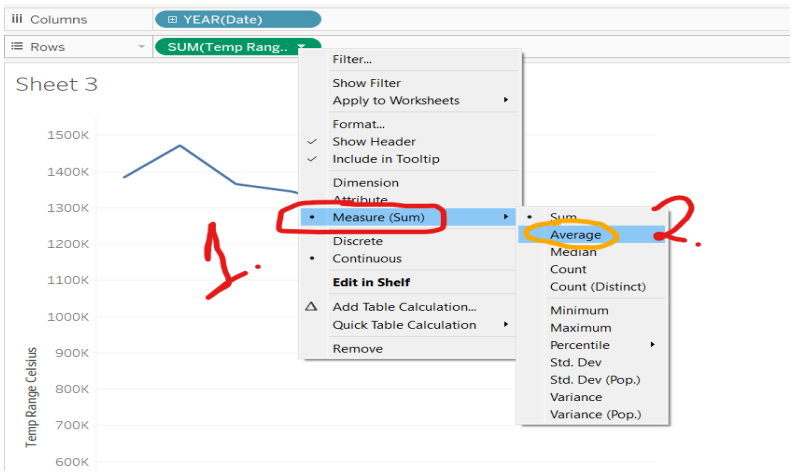
22. Now we’ll select and drag “Temp range Celsius” measure(Shown circled RED) from the Tables as well but, drop this into your “Rows”(Shown with RED arrow)



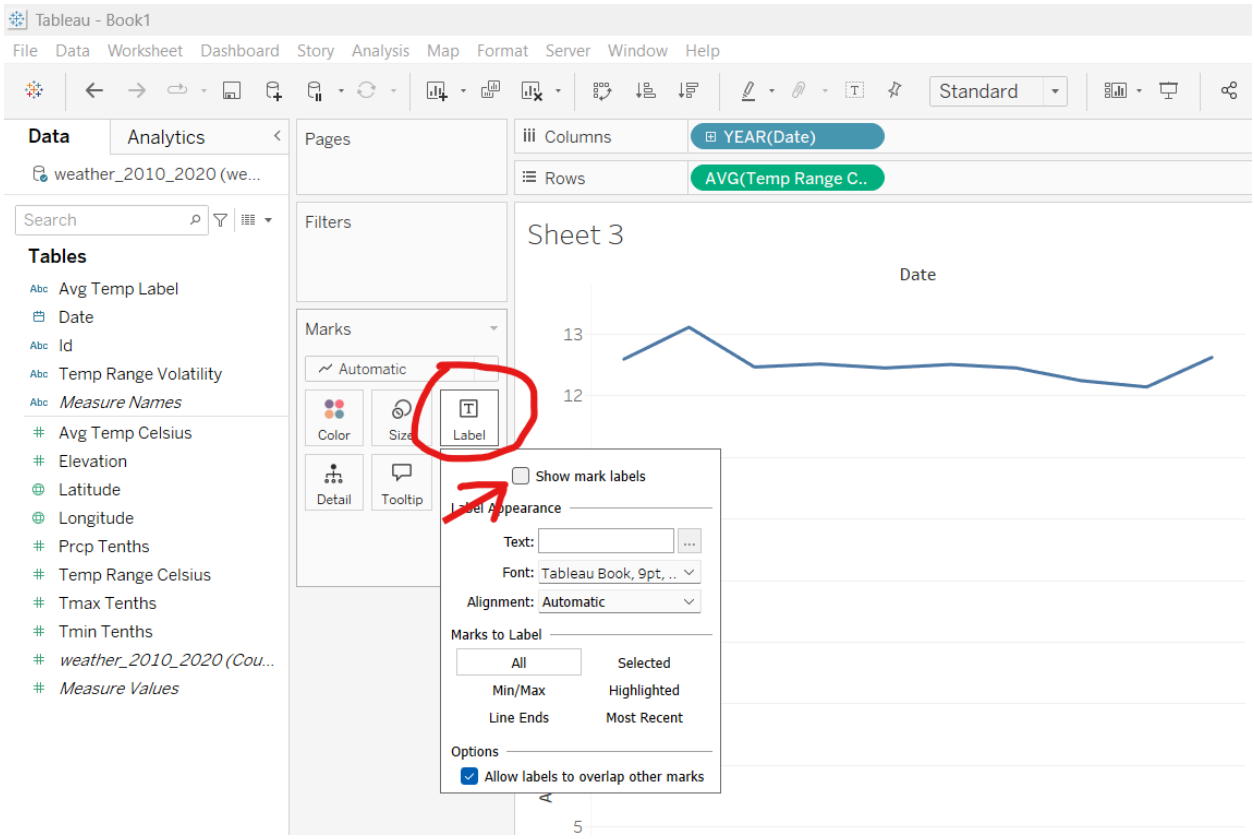
23. This visualization isn't done yet, now we'll right-click the "SUM(Temp Range." within our "Rows" this will open up the drop down so we can hover over "Measure (Sum)(shown by RED arrow)



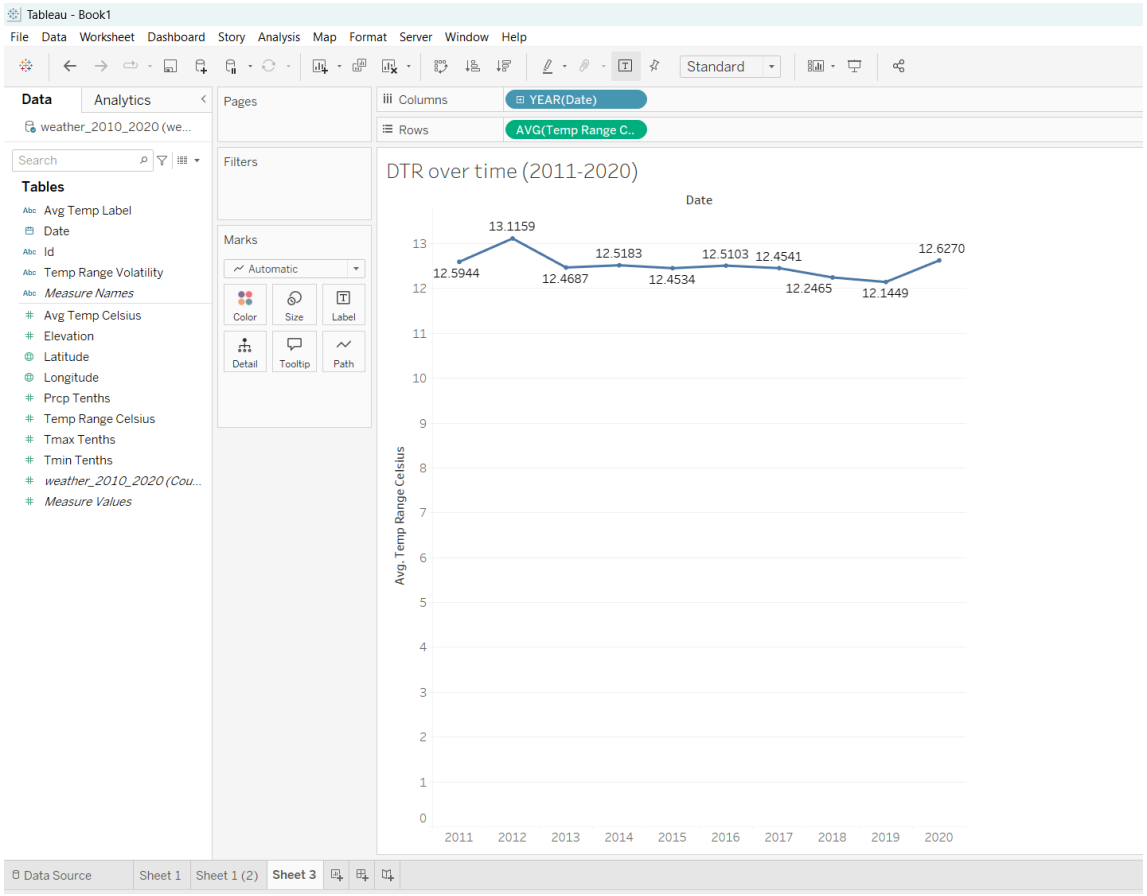
24. After hovering over "Measure (Sum) another side section will open up where we will select "Average" to change the measure from "Sum".



25. Now we have our DTR over time Line chart, but to help with the visualization, Select your “Label” icon in the marks box and make sure to check and select “Show mark Labels”



26. This successfully completes our DTR over time visualization in which we can see 2018 and 2019 showing a significant decrease in DTR over time compared to earlier years which is a demonstration of the climate change taking effect as these drops are much larger than normal. This shows how the fluctuation of temperature is actually decreasing which can signify warmer nights are happening within these years thus contributing to the DTR.



Conclusion

To conclude, within this tutorial we successfully walked through a full data pipeline for analyzing Weather data from 2010-2020. We uploaded the data to HDFS, using this data we created Hive tables to prepare and clean the data, and even derived new climate indicators such as average temperature label or even volatility label and even derived DTR as well. Finally we exported our clean dataset into the local machine so we can create a Heat Map in Excel Power Map, along with comparing precipitation early decade with late decade with Bar charts, and lastly establish a line chart showing the DTR over time within this time period.

References

1. URL of Data Source,

   https://www.kaggle.com/datasets/nachiketkamod/weather-dataset-us/data

2. URL of your Github

   https://github.com/KevMD-23/CIS-4560-Weather

THIS IS THE END OF THE TUTORIAL