

# WILDFIRES AND ITS IMPACTS



**Parikh Malav Arunkumar  
30708974**

**Lab 04 Monday 8 AM  
Fahimeh Sadat Saleh and Pratik Bhumkar**

## Introduction

Wildfires have been a dominant force of causing destructions across the world since long and wildfires in United States of America (USA) are a common phenomenon. Though wildfires are a natural disaster, there are multiple reasons due to which wildfires occur and global warming is one of them. Global warming has been one big reason due to which wildfires are occurring frequently as it has resulted in extreme weather conditions which causes wildfires. A statistic about wildfire suggest that, California is one of the states which has the highest number of wildfires across USA and in recent years it has increase five folds and the intensity of these fires has been severe. Wildfires cause damage to vegetation as it starts in an area of combustible vegetation, but vegetation is not the only thing which is affected due to it. It causes damage on multiple fronts and thus, this report is aimed to draw some comparisons and show trends of wildfires and its impacts and its relation to various other things.

The main questions answered here are:

1. Is there a relation between rise in temperatures and subsequent rise in number of wildfires?
2. Do wildfires result in higher CO2 emission which in turn cause deaths due to respiratory diseases and is there a significant trend to showcase the relationship between all the three?
3. Do wildfires impact the economy (GDP) and population? And whether it increase the number of business closures?

The motivation behind exploring the wildfires is to see how wildfires have a domino effect and the need to act against it. The motivation is to see how in the recent years, the phenomenon of wildfires has impacted on multiple fronts.

## Data Wrangling

To answer the above-mentioned questions, multiple datasets are chosen. The datasets chosen are mentioned below:

1. **Population dataset** contains information about population by years and zip code  
<https://data.world/lukewhyte/us-population-by-zip-code-2010-2016>
2. **Zip code dataset** contains data about every zip code. Its location, state and every related data. <https://simplemaps.com/data/us-zips>
3. **Number of Wildfires** contains month, year and state wise data of number of wildfires and acres burned [https://www.ncdc.noaa.gov/societal-impacts/wildfires/month/12?params\[\]=acres&params\[\]=fires](https://www.ncdc.noaa.gov/societal-impacts/wildfires/month/12?params[]=acres&params[]=fires)
4. **Wildfires by state** contains data about number of wildfires and acres burned in every state every year <https://www.iii.org/table-archive/23284>
5. **GDP** contains data about GDP of US, GDP per capita, inflation and unemployment rate  
[https://www.imf.org/external/pubs/ft/weo/2019/02/weodata/weorept.aspx?sy=1980&ey=2019&scsm=1&ssd=1&sort=country&ds=.&br=1&pr1.x=21&pr1.y=11&c=111&s=NGDP\\_RPCH%2CNGDP%2CNGDPPC%2CPCIPCH%2CLUR%2CBCA\\_NGDPD&grp=0&a=#download](https://www.imf.org/external/pubs/ft/weo/2019/02/weodata/weorept.aspx?sy=1980&ey=2019&scsm=1&ssd=1&sort=country&ds=.&br=1&pr1.x=21&pr1.y=11&c=111&s=NGDP_RPCH%2CNGDP%2CNGDPPC%2CPCIPCH%2CLUR%2CBCA_NGDPD&grp=0&a=#download)
6. **CO2 emissions by state** contains year and state wise data of CO2 emissions  
<https://www.eia.gov/environment/emissions/state/>
7. **Respiratory Disease** contains data about the number of deaths caused by respiratory diseases across states in different years

[https://www.cdc.gov/nchs/pressroom/sosmap/lung\\_disease\\_mortality/lung\\_disease.htm](https://www.cdc.gov/nchs/pressroom/sosmap/lung_disease_mortality/lung_disease.htm)

8. **Temperature by State and Year** contains data about state wise and year wise temperatures

<https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data>

9. **Business Data (Table Name: State)** contains data about every state wise and year wise new business establishments, closures, job creations, job closings and all related data

<https://www.census.gov/programs-surveys/bds/data/data-tables/legacy-establishment-characteristics-tables-1977-2014.html>

The datasets are a mix of textual and tabular data alongside having spatial information in the form of comma separated values and xlsx format. The datasets collected are raw and require significant data wrangling to answer the questions discussed earlier. The data wrangling procedure is carried on question to question basis as each question requires multiple datasets in a different form that the other questions.

The datasets used to answer first question are 3, 4 and 8; dataset 3,4,6,7 is used for the second question and dataset 1,3,4,5,9 is used to answer question 3. The data related to number of wildfires was divided into 12 csv files of 1 per month from 2000 to 2020 and the data was read using readLines and the data was as shown below (left):

```
[1] "January U.S. Wildfires (2000-2020)"
[2] "Date,Number of Fires,Acres Burned,Acres Burned per Fire"
[3] "2000,\"2,796\", \"40,757\", \"14.6\""
[4] "2001,\"1,231\", \"44,334\", \"36.0\""
[5] "2002,\"1,383\", \"10,079\", \"7.3\""
[6] "2003,\"1,964\", \"18,818\", \"9.6\""
[7] "2004,\"1,740\", \"15,386\", \"8.8\""
[8] "2005,\"1,674\", \"9,735\", \"5.8\""
[9] "2006,\"3,507\", \"330,447\", \"94.2\""
[10] "2007,\"387\", \"4,597\", \"11.9\""
```




Date	Number.of.Fires	Acres.Burned	Acres.Burned.per.Fire	Month
2000	2,796	40,757	14.6	1
2001	1,231	44,334	36.0	1
2002	1,383	10,079	7.3	1
2003	1,964	18,818	9.6	1
2004	1,740	15,386	8.8	1
2005	1,674	9,735	5.8	1
2006	3,507	330,447	94.2	1
2007	387	4,597	11.9	1
2008	1,380	40,804	29.6	1

Figure 1 January data of wildfires (2000-2020)

The data was transformed into the above shown in figure above (right) format by removing the extra rows and by adding appropriate header to columns and a month column was added for further use. The data shown here is only of the month of January, but similar transformation was done for the rest of months data. The dataset related to temperatures was in a format as shown below (left) which had extra columns and a combined column of date which was separated into different columns upon wrangling and the data was grouped by month and year as shown in the figure below (right). Further all the data of month which were stored in different files were merged into one file by using rbind() function. A combined dataset of wildfires and temperatures was developed by performing inner join operation to give the final dataset to provide appropriate visualization for question 1. Following a similar process all the datasets were read and transformed into appropriate formats.

Most of the datasets had different ways of storing the location as some had names in the form of state name, some stored it with state id, some with only numbers ranging from 1 to 56 for the states, these datasets were joined appropriately by using join function alongside the dataset 2 which contained location information in all forms: zip code, state name, state id and latitude and longitude. As the dataset (number 2) of zip codes was containing information about every city of one state, it had the latitude and longitude of every city so to find the state's location the dataset was grouped using group by function and summarized using summarise function to find the mean latitude and longitude of every location as shown in figure 3.

dt	AverageTemperature	AverageTemperatureUncertainty	State	Country
1/01/2010	4.617	0.172	Alabama	United States
1/02/2010	5.014	0.156	Alabama	United States
1/03/2010	11.073	0.233	Alabama	United States
1/04/2010	18.244	0.149	Alabama	United States
1/05/2010	23.488	0.137	Alabama	United States
1/06/2010	27.717	0.164	Alabama	United States
1/07/2010	28.656	0.220	Alabama	United States
1/08/2010	28.844	0.248	Alabama	United States




Country	Month	Year	avg. temperature
United States	1	2010	-1.3087647
United States	1	2011	-1.9585882
United States	1	2012	1.5889412
United States	1	2013	0.2029020
United States	2	2010	-0.4912549
United States	2	2011	0.5156471
United States	2	2012	2.9857451
United States	2	2013	0.8715882

Figure 2 Temperature dataset for every month every year (2010-2013)

Column names were changed using `colnames()` to enable datasets to be joined by names using the join function. All the dataset were not sorted in an alphabetical orders of state names which were formatted to an ascending order using `order()`. Datasets like population and GDP were transposed in order to get the year column as row to easily enable the join. Alongside that these column names had prefixes of X before the year value, so they were formatted using `substring()`.

zip	lat	lng	city	state_id	state_name	zcta	parent_zcta	population
601	18.18004	-66.75218	Adjuntas	PR	Puerto Rico	TRUE	NA	1724
602	18.36073	-67.17517	Aguada	PR	Puerto Rico	TRUE	NA	3844
603	18.45439	-67.12202	Aguadilla	PR	Puerto Rico	TRUE	NA	4884
606	18.16724	-66.93828	Maricao	PR	Puerto Rico	TRUE	NA	6434
610	18.29032	-67.12243	Anasco	PR	Puerto Rico	TRUE	NA	2704
612	18.40699	-66.70803	Arecibo	PR	Puerto Rico	TRUE	NA	6034
616	18.41753	-66.66814	Bajadero	PR	Puerto Rico	TRUE	NA	1074
617	18.44124	-66.55017	Barceloneta	PR	Puerto Rico	TRUE	NA	2304



state_id	state_name	population	latitude	longitude
AK	Alaska	737979	61.65628	-153.33722
AL	Alabama	4864630	32.88404	-86.81980
AR	Arkansas	2990472	35.13441	-92.37733
AZ	Arizona	6949259	33.70583	-111.57576
CA	California	39140219	36.42590	-119.92227
CO	Colorado	5531233	39.23323	-105.34536
CT	Connecticut	3581504	41.57642	-72.76644
DC	District of Columbia	684390	38.90071	-77.02629

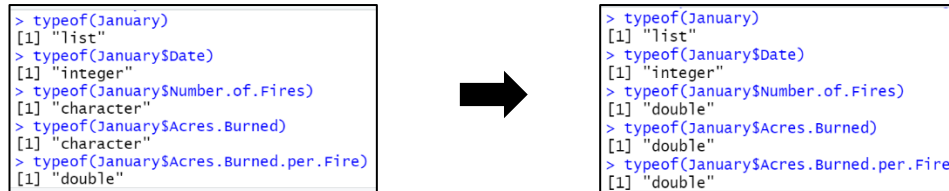
Figure 3 Zip code dataset for all cities of USA

In some cases, initial datasets were broken down into individual year wise dataset using `filter()` to provide year wise data as used for the question 2 which is explained in detail below. These broken-down individual datasets were merged again for use using the `rbind()` function and checked for identical column names before using the `rbind()`. To display the visualization in Tableau, formatted datasets were again stored as csv files using the `write` function. Thorough wrangling of datasets was required to format it from its raw form to the form appropriate to answer the question framed. And each question required different format of data so multiple transformations were done to the same dataset.

**Tools Used:** All the data was wrangled and formatted in R using the following libraries: `dplyr`, `lubridate` and `tidyverse`.

## Data Checking and cleaning

Every dataset was checked for data types and as shown below some columns had data type of character for numeric values like number of fires and number of acres burned per fire so by using `as.numeric()` they were transformed into numeric (double) data type.



```

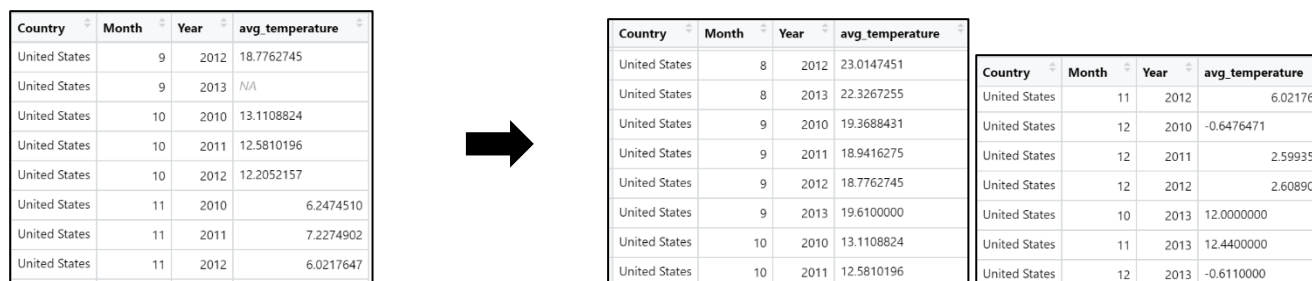
> typeof(January)
[1] "list"
> typeof(January$Date)
[1] "integer"
> typeof(January$Number.of.Fires)
[1] "character"
> typeof(January$Acres.Burned)
[1] "character"
> typeof(January$Acres.Burned.per.Fire)
[1] "double"

> typeof(January)
[1] "list"
> typeof(January$Date)
[1] "integer"
> typeof(January$Number.of.Fires)
[1] "double"
> typeof(January$Acres.Burned)
[1] "double"
> typeof(January$Acres.Burned.per.Fire)
[1] "double"

```

Figure 4 typeof used to check data types

Some datasets had NA values and some missing values like no data for month 10 and year 2013 as shown in the figure below (left) were cleaned, and appropriate data was inserted from the national data available online. The NA value was replaced by specifically narrowing down to month 9 and year 2013 and appropriate value was given to the average temperature column. While for the missing values new rows were added containing to make the dataset a complete one based on the national data information found online.



Country	Month	Year	avg_temperature
United States	9	2012	18.7762745
United States	9	2013	NA
United States	10	2010	13.1108824
United States	10	2011	12.5810196
United States	10	2012	12.2052157
United States	11	2010	6.2474510
United States	11	2011	7.2274902
United States	11	2012	6.0217647

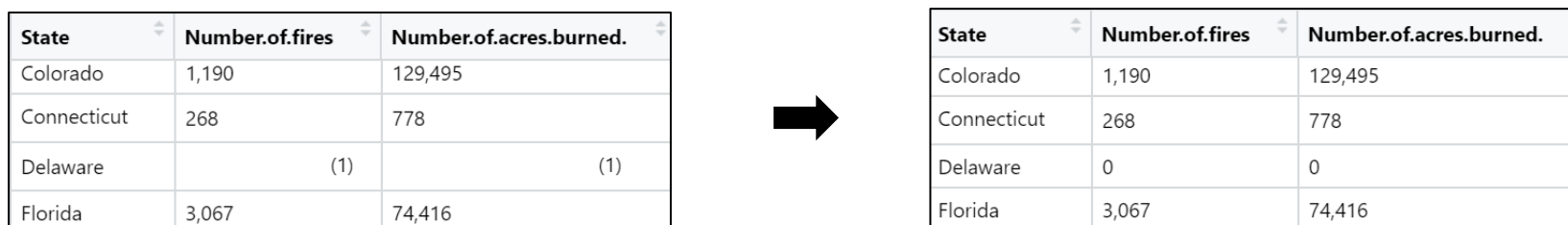
Country	Month	Year	avg_temperature
United States	8	2012	23.0147451
United States	8	2013	22.3267255
United States	9	2010	19.3688431
United States	9	2011	18.9416275
United States	9	2012	18.7762745
United States	9	2013	19.6100000
United States	10	2010	13.1108824
United States	10	2011	12.5810196

Country	Month	Year	avg_temperature
United States	11	2012	6.0217647
United States	12	2010	-0.6476471
United States	12	2011	2.5993529
United States	12	2012	2.6089020
United States	10	2013	12.0000000
United States	11	2013	12.4400000
United States	12	2013	-0.6110000

Figure 5 NA values transformation to actual values

Some datasets about wildfire information had indicated that no wildfires had been recorded in that state using -1 or (1) for the number of fires and acres burned so they were replaced by 0 as shown in the figure below (right).



State	Number.of.fires	Number.of.acres.burned.
Colorado	1,190	129,495
Connecticut	268	778
Delaware	(1)	(1)
Florida	3,067	74,416

State	Number.of.fires	Number.of.acres.burned.
Colorado	1,190	129,495
Connecticut	268	778
Delaware	0	0
Florida	3,067	74,416

Figure 6 Replacing wrong values with appropriate values

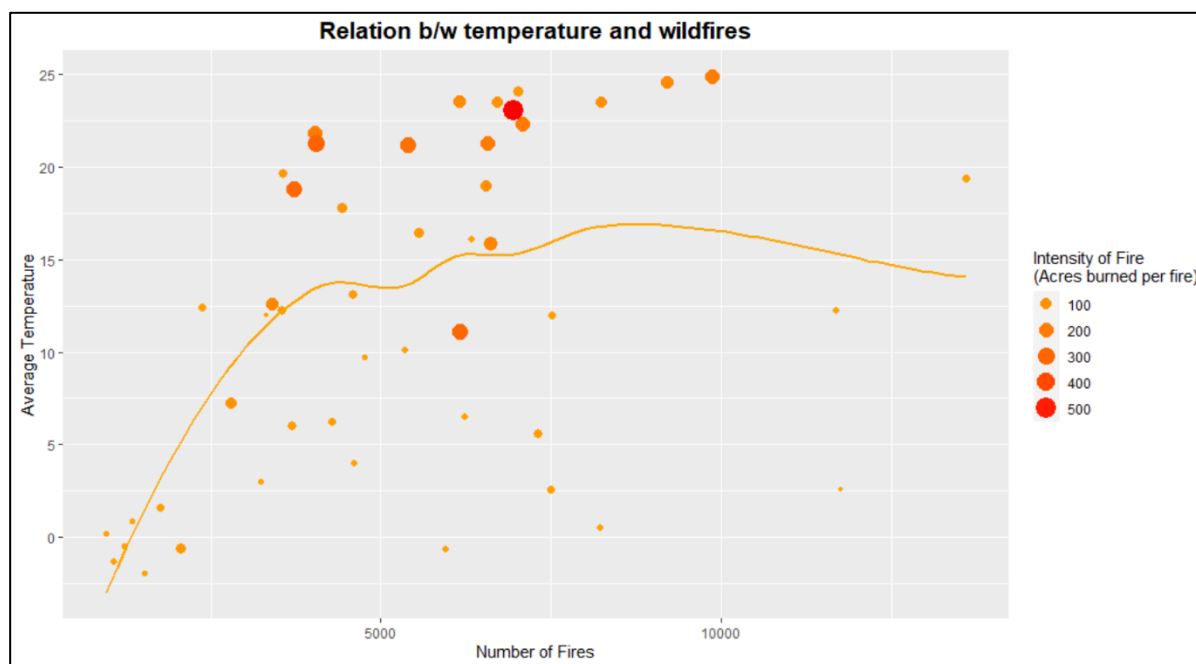
## Data Exploration

### I. Relation between temperature and number of wildfires

As we can clearly see from the below show visualization that at lower temperatures, around the 0 – 5 degree mark, the number of wildfires is quite low which is the result of a cooler atmosphere around that period while on the other hand if we look at the higher temperatures, around the 20 – 25

degree mark the number of fires is quite high as compared to the lower temperatures. So, a general trend of rise in temperature resulting in rise in number of wildfires can be justified to quite some extent as there are outliers which makes the trend line dip a bit at the extreme end. The main aim was to find a relation between temperature and wildfires but what the visualization shows is that not only when the temperature rises there is a jump in number of wildfires but the intensity of the fire which is recorded as the acres burned fire also goes up significantly. As the temperature rises, the land dries up and allows the fire to spread more rapidly and in process increasing the intensity of it.

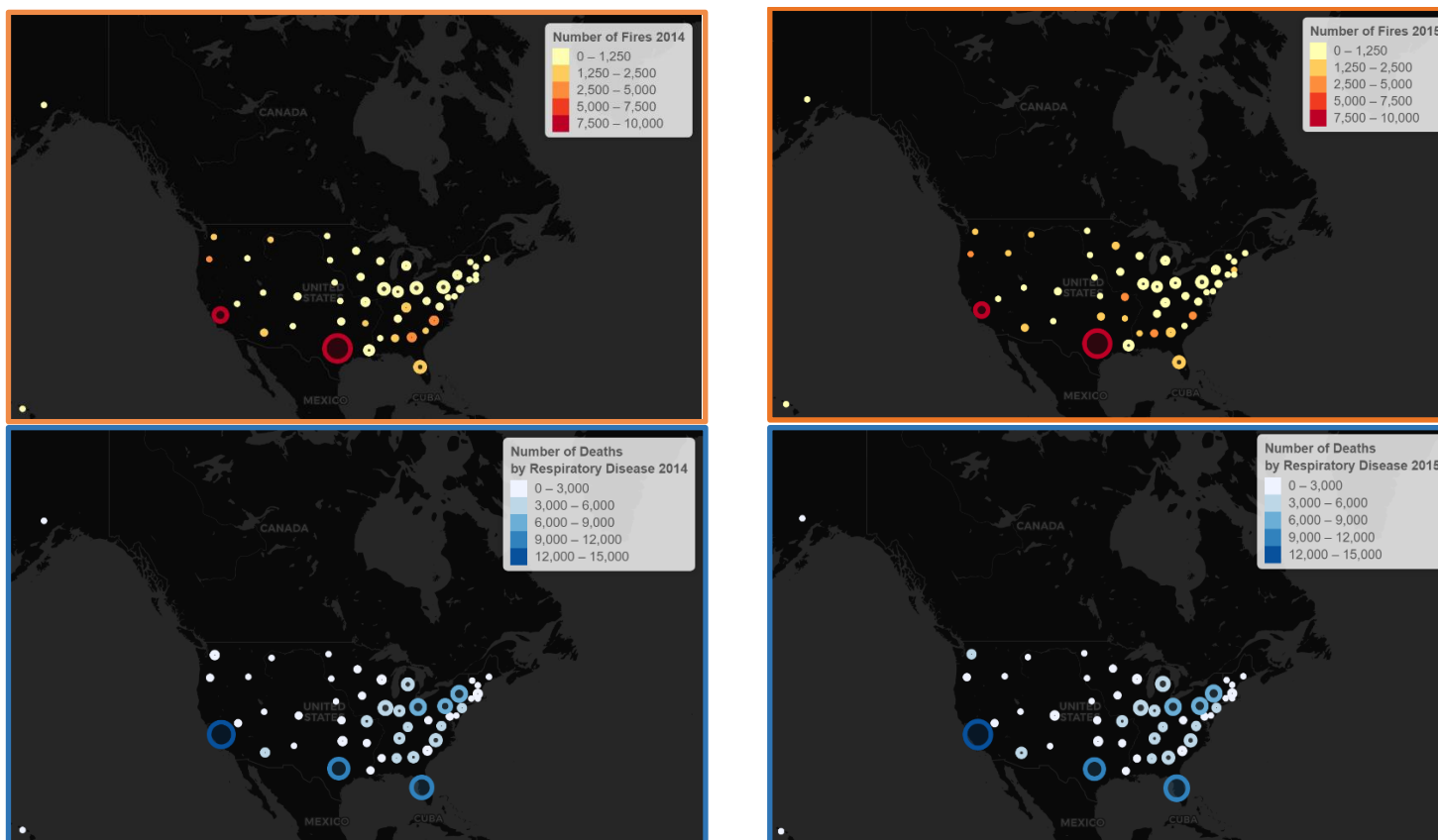
**Tools Used:** R is used to provide the below shown visualization with the help of ggplot2 library functions like ggplot(), geom\_point() and geom\_smooth().



## II. *Relation between wildfires, CO2 emissions and respiratory diseases*

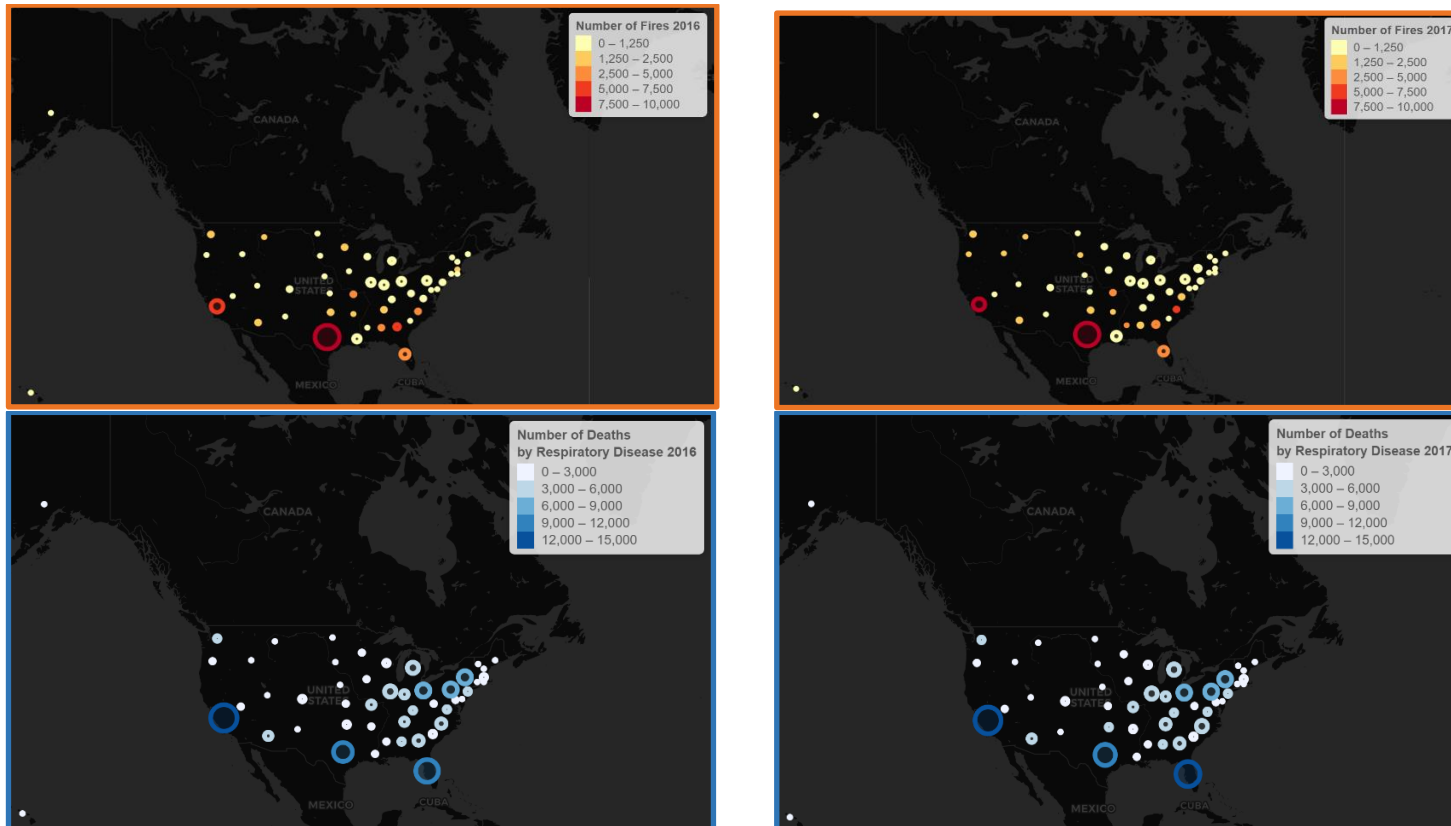
The visualizations shown below are to be seen in two parts the one with orange border showing the CO2 emissions and wildfires information and the blue bordered ones indicate the number of deaths by respiratory diseases in the same regions. The data here shown is for the year range of 2014-2017. As shown in the legend in the orange bordered ones, the number of wildfires are shown by the different color shades while the size of the circle depicts the CO2 emission levels. This visualization clearly depicts that the regions where there are a greater number of wildfires, the levels of CO2 emissions are significantly higher as compared to the regions where the number of fires is relatively low. This visualization clearly justifies the trend where rise in number of wildfires results in higher CO2 levels.





The visualization with blue borders show the number of deaths that have occurred in every state due to respiratory diseases for the year range of 2014-2017. As shown in the legend, the shade of blue depicts the number of deaths caused by respiratory diseases. The darker the shade the more number of deaths have occurred and the size of the circle too depicts the same. The visualization when seen together with the corresponding orange bordered ones, it can be clearly said that the states having higher CO<sub>2</sub> emission levels are the one which have highest number of deaths due to respiratory diseases. This trend continues for every year for the given range of 2014 to 2017 though this might be a small range to look at but the sheer number of wildfires and deaths are easy to analyze that there is a certain between the wildfire numbers and respiratory disease death numbers. The domino effect of wildfires causing the CO<sub>2</sub> levels to go up and subsequently resulting in respiratory diseases which lead to deaths can be clearly seen. Also, the global warming effect of rising temperatures resulting in dried up lands makes the case worse as the fires in those regions release significantly more amount of CO<sub>2</sub> and rapidly spreads the wildfires. This threatening trend can be seen in this visualization, the areas like Texas,

Florida and California are places having higher temperatures than other states of USA and clearly the CO2 levels and number of deaths due to respiratory disease is significantly higher in these regions.

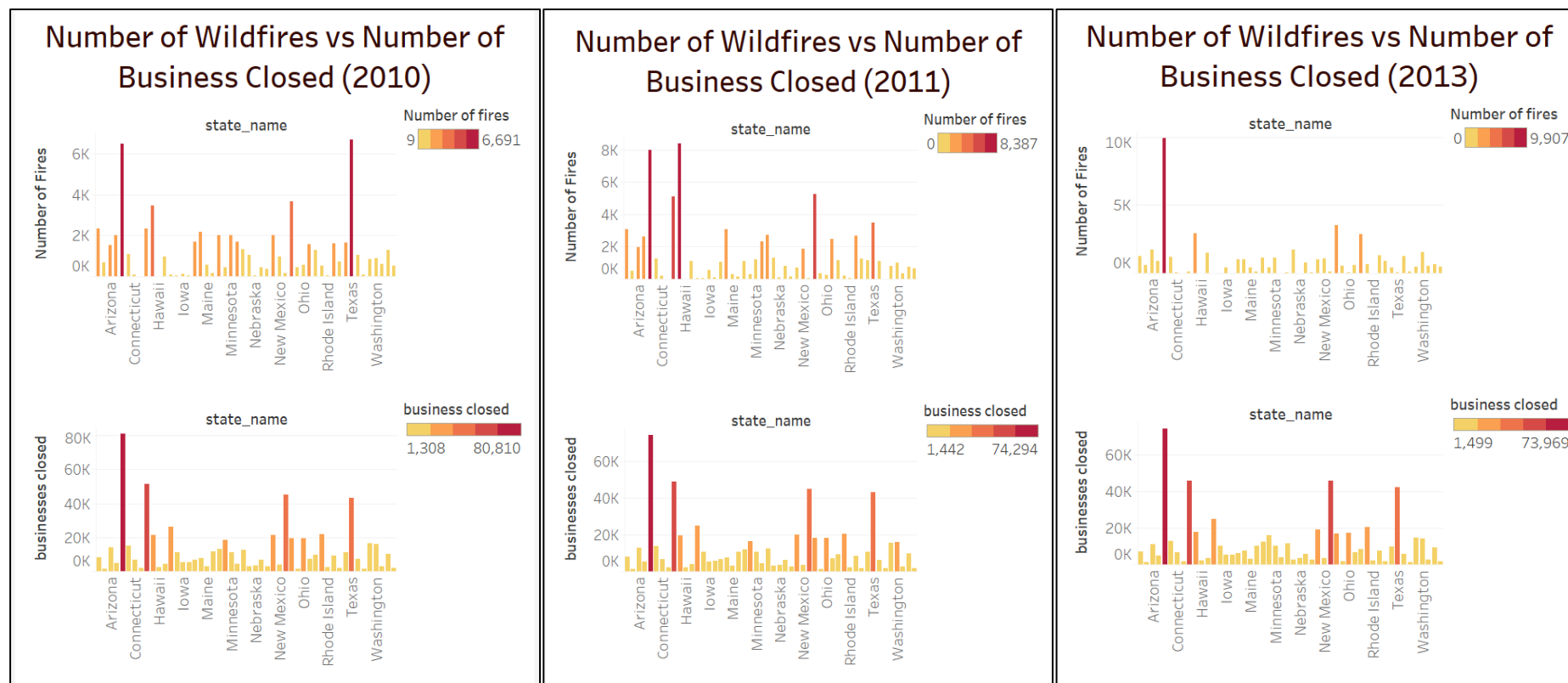


**Tools Used:** R is used to produce these visualizations with the help of libraries like leaflet, maps and htmltools.

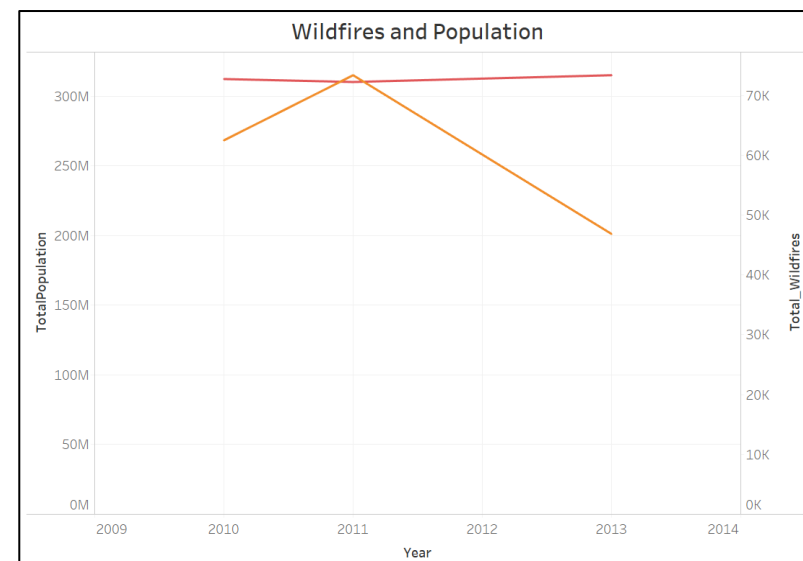
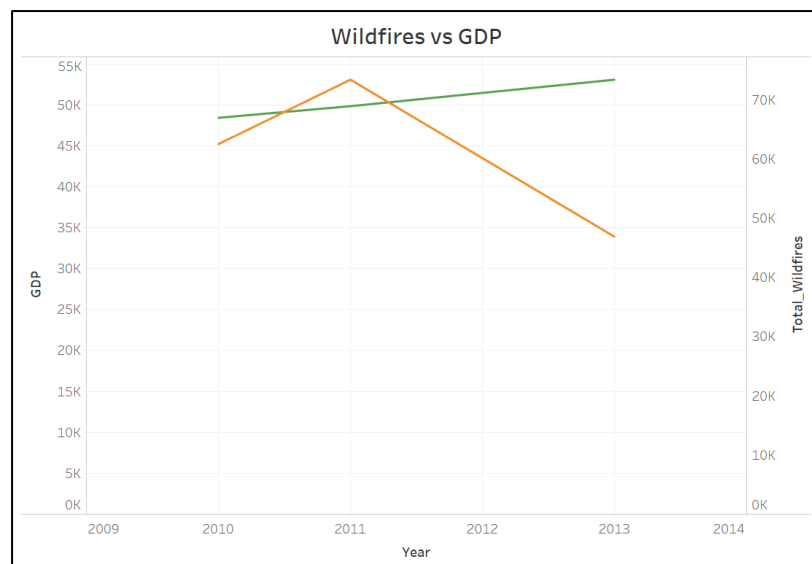
### III. *Relation between wildfires, business closures, economy and population*

The visualization shown below are indicating the number of wildfires and the number of businesses closed for the same state. In states which have a higher number of wildfires tend to have higher number of businesses closed as compared to the states where the number of wildfires are less. But it cannot be clearly distinguished that the wildfires can be a reason of business closures as some regions having higher number of wildfires have lesser number of businesses closed. So, it cannot be clearly justified from the findings of the gathered data that wildfires have a direct connection to business closures. And considering the year range is quite small as only data of three years can be compared. Also, it is noteworthy to say that these numbers are in the years followed by a great recession which was faced by the whole world.





Further, the visualizations shown below indicated the number of wildfires with respect to GDP and population. It can be clearly seen from the visualization that there is no specific trend of wildfires impacting the population as the population has remained steady through the years 2010 to 2013 even when the wildfire numbers have increased or decreased. While in case of GDP, even though the wildfires affect human habitats, animal habitats, businesses, atmosphere and farming, there is a lesser impact on the GDP of the wildfires as USA's GDP is always on the rise as shown in the visualization. Loss of life, house, business accounts to a huge amount of money but that negligibly affects the GDP of the nation as the number of businesses shut give rise to more new businesses to open up and the age old wildfire seasons continue to cause destruction of many folds and keeps on increasing every year with a minimal effect on GDP but a humongous effect on atmosphere, human lives and animal lives.



**Tools Used:** Tableau is used to depict the visualization of data wrangled in R and exported to use as csv file in Tableau. In Tableau, different sheets were created for each visualization and then combined using the dashboard feature of Tableau.

## Conclusion

The exploration and visualization performed on the dataset clearly depicts domino effect of global warming on multiple fronts. The findings from the first question justified that there is a direct relation between temperature and wildfires. As the temperatures rise, the number of wildfires too rise due to the dried up land and burnt vegetation. It not only answered our question but also a significant result was displayed in the form of the intensity of fires i.e. acres burned per fire. The intensity of fire was significantly higher when the temperatures were high and significantly lower when the temperatures were low. The findings of the second question took on from the findings of question one and justified that wildfires release large amounts of CO2 into the atmosphere which gives rise to more and more respiratory diseases and subsequently resulting in more deaths due to the same. The trends showed that regions having higher number of wildfires faced more deaths due to respiratory diseases which clearly justify our question of a direct relation between the two. The third and final question's findings tried to see a pattern where wildfires caused more business closures but a clear trend could not be identified between the two as the reason of business closures will not only depend on the wildfires but due to more than one clear factor and also our findings are from the five years after the great recession of 2008 which turned the economies upside down. So, our third and final question could not be clearly justified.

## Reflection

The learnings from this project suggest that every data has some story behind it and upon visualizing the data some clear patterns can be found out which help understand the story telling behind the data clearly. The project also reflects on the need of clean and correctly transformed data which can bring out the visualizations more clearly and in a better way than its raw form. Data wrangling and cleaning is an integral part of the visualization process. Also the project gave a better understanding of how closely the destruction of our nature by us humans affects none other than ourselves bringing unavoidable circumstances for the mankind as whole. It is utterly important to preserve our nature and rightly justify that protect nature and nature will protect you.

The certain limitations of these visualizations are due to lack of all datasets giving data about the same year range which makes our findings year range small and hence cannot clearly identify and justify the trends like in population and gdp which more data availability could have provided.

## Bibliography

- [1] RDocumentation. "base v3.6.2." RDocumentation. <https://www.rdocumentation.org/packages/base/versions/3.6.2/> (Accessed Sep. 18, 2020).
- [2] RStudio. "RStudio Cheatsheets." RStudio. <https://rstudio.com/resources/cheatsheets/> (Accessed Sep. 18, 2020).
- [3] DATA.GOV. "The home of U.S. Government's open data." DATA.GOV. <https://www.data.gov/> (Accessed Sep. 18, 2020).