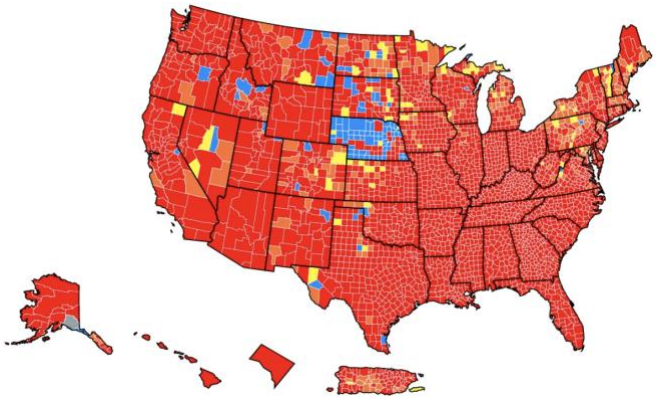
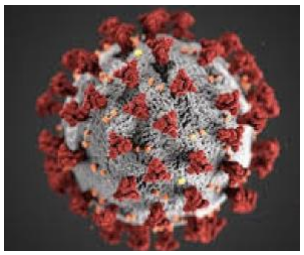
Covid-19 spread overtime in 2020 - USA

Introduction

The world has been staggered by the new virus covid-19 and its impacts. We all have been affected by the pandemic and our life hasn't been same ever since after. It affected each country's health, wealth and what not! It changed our lives in every aspect and that is why it is worth mentioning and its spread is worth analyzing. I am motivated the most to utilize my SAS knowledge on something that has left its big impact on the world, and I feel fortunate enough to be able to represent the spread in graphs and numbers with my SAS knowledge.





Data Source

Kaggle hosts massive open-source public data across various domains, and I found required dataset from the same www.kaggle.com. Data imported for this project is published at https://www.kaggle.com/sudalairajkumar/covid19-in-usa a year ago and it was originally obtained from COVID-19 Tracking project and NYTimes. It was collected under federal covid tracking project by government since number of new cases were increasing day by day around the world. I downloaded and modified the dataset from the source and uploaded under 'data' section in EPG194 in SAS. I used PROC IMPORT step to import it into pg1 library to proceed further. This dataset has information from 55 US states and 1925 county at daily level with 8,00,437 observations.

Data Exploration and preparation

I have used SAS steps to explore this dataset. It contains one csv file, “covid-19_usa_cases” that has 8,00,437 observations with number of cases and deaths for a specific state, county and date. It also has FIPS code in this data for that county. Please see figure 1 to sneak pick into raw data.

covid-19_usa_cases

date	county	state	fips	cases	deaths
2020-01-21	Snohomish	Washington	53061.0	1	0.0
2020-01-22	Snohomish	Washington	53061.0	1	0.0
2020-01-23	Snohomish	Washington	53061.0	1	0.0
2020-01-24	Cook	Illinois	17031.0	1	0.0
2020-01-24	Snohomish	Washington	53061.0	1	0.0
2020-01-25	Orange	California	6059.0	1	0.0
2020-01-25	Cook	Illinois	17031.0	1	0.0
2020-01-25	Snohomish	Washington	53061.0	1	0.0
2020-01-26	Maricopa	Arizona	4013.0	1	0.0
2020-01-26	Los Angeles	California	6037.0	1	0.0
2020-01-26	Orange	California	6059.0	1	0.0
2020-01-26	Cook	Illinois	17031.0	1	0.0
2020-01-26	Snohomish	Washington	53061.0	1	0.0
2020-01-27	Maricopa	Arizona	4013.0	1	0.0
2020-01-27	Los Angeles	California	6037.0	1	0.0

Fig.1

The CONTENTS Procedure

Data Set Name	PG1.COVID_USA_CASES	Observations	800437
Member Type	DATA	Variables	6
Engine	V9	Indexes	0
Created	12/06/2021 10:52:42	Observation Length	56
Last Modified	12/06/2021 10:52:42	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64, LINUX_POWER_64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information

Data Set Page Size	65536
Number of Data Set Pages	687
First Data Page	1
Max Obs per Page	1166
Obs in First Data Page	1121
Number of Data Set Repairs	0
Filename	/home/vmodi/EPG194/data/covid_usa_cases.sas7bdat
Release Created	V.0305M0
Host Created	Linux
Inode Number	57234
Access Permission	rw-r--r--
Owner Name	vmodi
File Size	43MB
File Size (bytes)	45088768

Variables in Creation Order

#	Variable	Type	Len	Format	Informat
1	date	Num	8	YYMMDD10.	YYMMDD10.
2	county	Char	11	$11.	$11.
3	state	Char	10	$10.	$10.
4	fips	Num	8	BEST12.	BEST32.
5	cases	Num	8	BEST12.	BEST32.
6	deaths	Num	8	BEST12.	BEST32.

Fig. 2

We can tell from the figure 2 where I performed CONTENTS procedure, that covid_usa_cases has number of observations with 6 variables named ‘date’,’county’,’state’,’fips’,’cases’,’deaths’.
Each variable has significance as shown below:
date - date of observation
county – USA state’s county name
state - USA state name
fips - federal information processing system code which uniquely identifies geographic areas.
cases – number of cases
death - number of deaths
Cases and death comes with ‘Num’ datatype whereas county and state as ‘Char’. ‘Variable in Creation Order table’ helped me explore on length and format information as well.

Creating frequency tables for the variables ‘state’ where ‘frequency’ column gave me an idea about how many observations fell into the given category. I kept ‘order=freq’ to sort it down from the higher to lower frequency. We can interpret from below table by saying that there are 60398 observations for Texas state with 7.55% of observations for Texas out of all nonmissing observations. The Cumulative Frequency column is the number of observations in the sample that have been accounted for up to and including the current row. It can be computed by adding all of the numbers in the Frequency column above and including the current row. The Cumulative Percent column is the proportion of the sample that has been accounted for up to and including that row. It can be computed by adding all of the numbers in the Percent column up to the current row. If there would be any missing data , it would have been printed right below the output of PROC FREQ results. That can also be printed by ‘/MISSING’ option right after tables statement. Here, we do not have any missing data except for deaths. So from the table below we can say

that Texas has the highest number of observations recorded that is 60398 and Guam has the lowest, being 266, among all the states. I used 'NLEVELS' option in frequency procedure that gave the number of variable levels output. Hence we can see that there are total 55 states data recorded.

The FREQ Procedure

state	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Texas	60398	7.55	60398	7.55
Georgia	41049	5.13	101447	12.67
Virginia	33428	4.18	134875	16.85
Kentucky	29692	3.71	164567	20.56
Missouri	28291	3.53	192858	24.09
Illinois	25584	3.20	218442	27.29
North Caro	25408	3.17	243850	30.46
Iowa	24670	3.08	268520	33.55
Tennessee	24415	3.05	292935	36.60
Kansas	24105	3.01	317040	39.61
Indiana	23678	2.96	340718	42.57
Ohio	22554	2.82	363272	45.38
Minnesota	21771	2.72	385043	48.10
Michigan	21186	2.65	406229	50.75
Mississipp	21122	2.64	427351	53.39
Nebraska	20189	2.52	447540	55.91
Arkansas	19208	2.40	466748	58.31
Oklahoma	19052	2.38	485800	60.69
Wisconsin	18263	2.28	504063	62.97
Florida	17730	2.22	521793	65.19
Pennsylvan	17333	2.17	539126	67.35
Alabama	17204	2.15	556330	69.50
Puerto Ric	17001	2.12	573331	71.63
Louisiana	16843	2.10	590174	73.73
Colorado	15991	2.00	606165	75.73
New York	15300	1.91	621465	77.64
California	15293	1.91	636758	79.55
South Dako	14968	1.87	651726	81.42
West Virgi	13416	1.68	665142	83.10
North Dako	12204	1.52	677346	84.62
South Caro	11987	1.50	689333	86.12
Montana	11312	1.41	700645	87.53
Washington	10436	1.30	711081	88.84
Idaho	10126	1.27	721207	90.10
Oregon	8914	1.11	730121	91.22
New Mexico	8024	1.00	738145	92.22

The FREQ Procedure

state	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Utah	7199	0.90	745344	93.12
Maryland	6545	0.82	751889	93.93
Alaska	5953	0.74	757842	94.68
New Jersey	5889	0.74	763731	95.41
Wyoming	5763	0.72	769494	96.13
Maine	4324	0.54	773818	96.67
Massachuse	4054	0.51	777872	97.18
Arizona	4009	0.50	781881	97.68
Nevada	3899	0.49	785780	98.17
Vermont	3898	0.49	789678	98.66
New Hampsh	2848	0.36	792526	99.01
Connecticu	2375	0.30	794901	99.31
Rhode Isla	1560	0.19	796461	99.50
Hawaii	1102	0.14	797563	99.64
Delaware	1035	0.13	798598	99.77
Virgin Isl	863	0.11	799461	99.88
Northern M	436	0.05	799897	99.93
District o	274	0.03	800171	99.97
Guam	266	0.03	800437	100.00

The FREQ Procedure

Number of Variable Levels	
Variable	Levels
state	55

Fig. 3

FREQ procedure also helped me explore the distribution of state with number of observations with PLOTS=FREQPLOT option where order=data is set so distrubution is presented alphabaticaly.

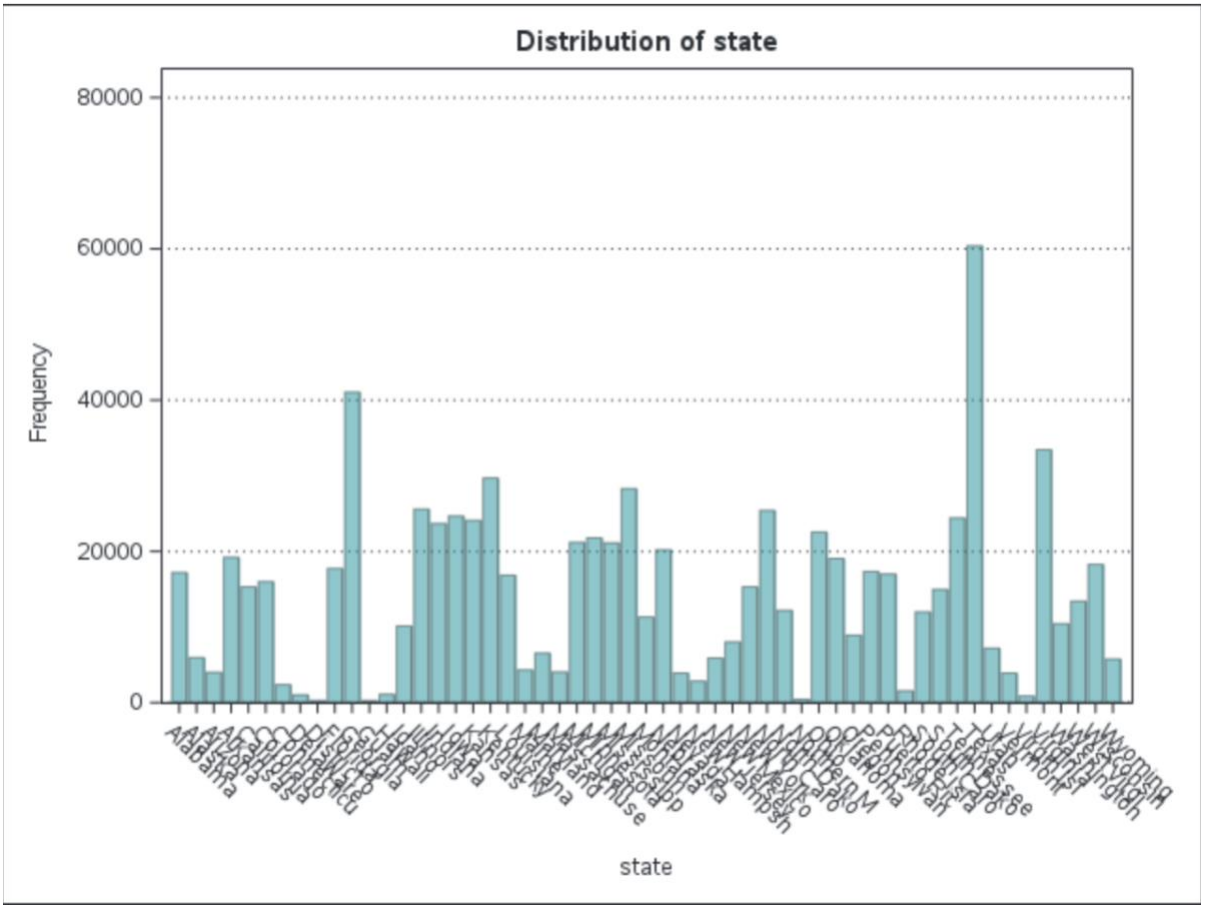


Fig. 4

After exploring county data, I found there are cases reported for 1925 counties under 55 states, which is huge and detailed data I would say! Below is the sneak pick into that.

The FREQ Procedure

Number of Variable Levels	
Variable	Levels
county	1925

county	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Washington	7971	1.00	7971	1.00
Unknown	6886	0.86	14857	1.86
Jefferson	6667	0.83	21524	2.69
Franklin	6357	0.79	27881	3.48
Jackson	5991	0.75	33872	4.23
Lincoln	5953	0.74	39825	4.98
Madison	5090	0.64	44915	5.61
Montgomery	4649	0.58	49564	6.19
Union	4540	0.57	54104	6.76
Clay	4476	0.56	58580	7.32
Marion	4304	0.54	62884	7.86
Monroe	4269	0.53	67153	8.39

Fig. 5

I also found that there are 16733 missing values of deaths in 'Puerto Rico' state which we are keeping as it is since those data is significant while checking number of cases every month. As we can see below, there are no deaths in 30% of our observations. Below is the frequency table for deaths column.

The FREQ Procedure

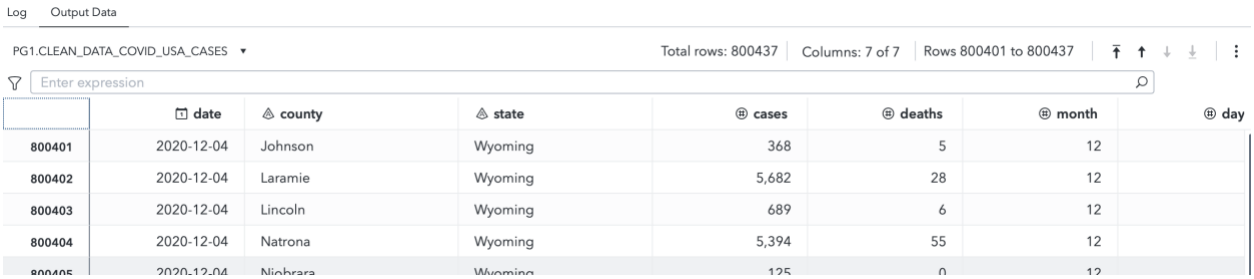
deaths	Frequency	Percent
0	235150	30.00
1	87523	11.17
2	52427	6.69

24,287	1	0.00
24,297	1	0.00
24,305	1	0.00
24,323	1	0.00
24,346	1	0.00
Frequency Missing = 16733		

Data Preparation

After exploring data and the quality of data, figured out that there are no missing values of a state or county so no need to remove nulls, as a first step, to clean data.

Also, there were no missing/Null values other than mentioned 16733 death counts for Puerto Rico hence no observation deletion requires. I, then realized that there is a column named 'date' and this whole dataset belongs to year 2020 so my further analysis requires date to be divided into month and day columns so that I can make detailed analysis and graph to see if there any trend exist in number of cases/deaths month wise. Hence, I created two new columns 'month' and 'day' and formatted 'cases' and 'deaths' to see numbers in readable format during data preparation step. I dropped 'fips' column since it does not require in further analysis that I want to perform. Now, my clean data is ready, and it looks like below.


Log Output Data

PG1.CLEAN_DATA_COVID_USA_CASES ▾ Total rows: 800437 Columns: 7 of 7 Rows 800401 to 800437

Enter expression

	date	county	state	cases	deaths	month	day
800401	2020-12-04	Johnson	Wyoming	368	5	12	
800402	2020-12-04	Laramie	Wyoming	5,682	28	12	
800403	2020-12-04	Lincoln	Wyoming	689	6	12	
800404	2020-12-04	Natrona	Wyoming	5,394	55	12	
800405	2020-12-04	Niobrara	Wyoming	125	0	12	

Fig. 6

PROC CONTENT helped me to give summary on variables and descriptive statistics on base level which says that now there are 7 variables with given Type and length and number of observation remains same 8,00,437 as before.

#	Variable	Type	Len
1	date	Num	8
2	county	Char	11
3	state	Char	10
4	cases	Num	8
5	deaths	Num	8
6	month	Num	8
7	day	Num	8

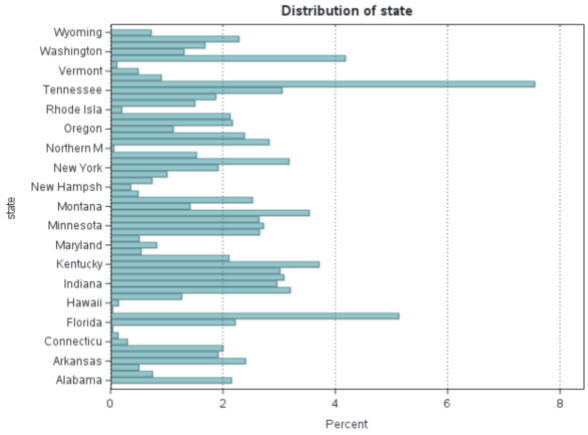
Data Analysis

From the figure below, we can say that number of cases starts with zero which means some of the states/counties had no cases and maximum cases were 4,39,538. This dataset has data from 1st month of 2020 which is January to 12th which is December and day 1 to day 31 in a month.

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
date	800437	22129.51	73.42	21935.00	22254.00
cases	800437	1589.67	8385.20	0.00	439538.00
deaths	783704	48.82	440.48	0.00	24346.00
month	800437	7.56	2.42	1.00	12.00
day	800437	15.83	8.94	1.00	31.00

Talking about observation count, distribution of state with State and its percent value is shown below. It says that maximum observation covers nearly 7.5% of the total observations in data and it seems to be Texas which is 60398 out of 800437 observations. Graph looks like below.



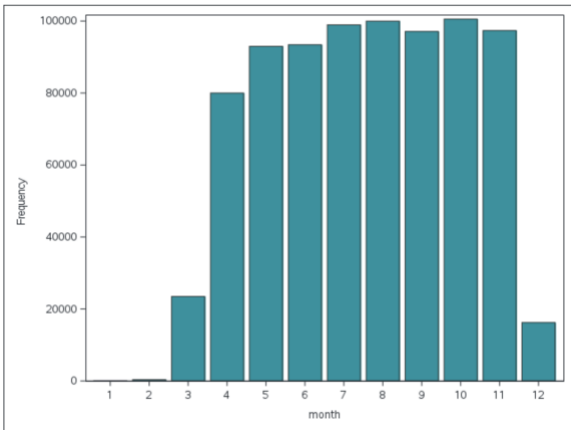
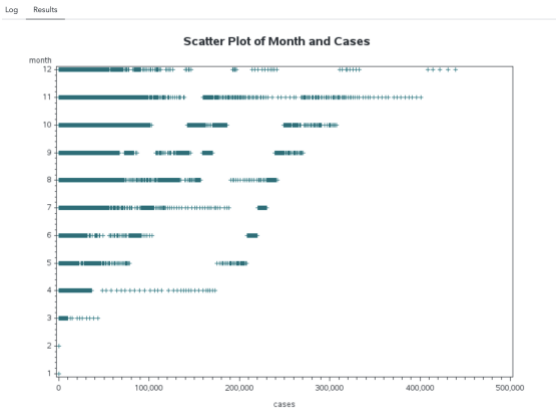


Fig. 7 Fig. 8

Figure 8 shows how the observations are distributed among months. There were very few observations recorded in first 2 months and it spiked like a rocket and highest observations seem to be recorded in August and October while December gave us little relief.

Analyzing through scatterplot makes it easy to visualize data and conclude facts that can mirror reality that happened. I was curious to know about what the scenario in January 2020 was and how gradually/non-gradually it went till December 2020. I was also curious to know graph of cases vs death, state wise cases , state wise deaths too. My all questions were simply answered by applying gplot on the dataset and getting scatterplot as result.



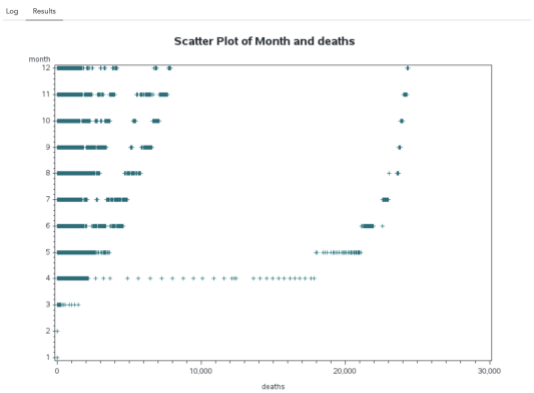


Fig. 9

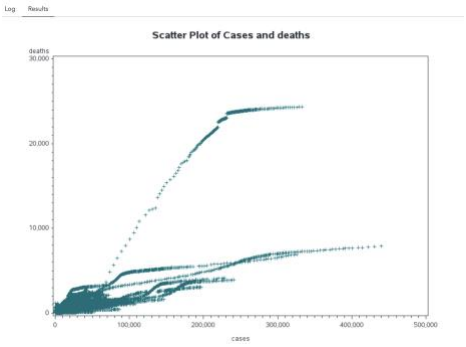
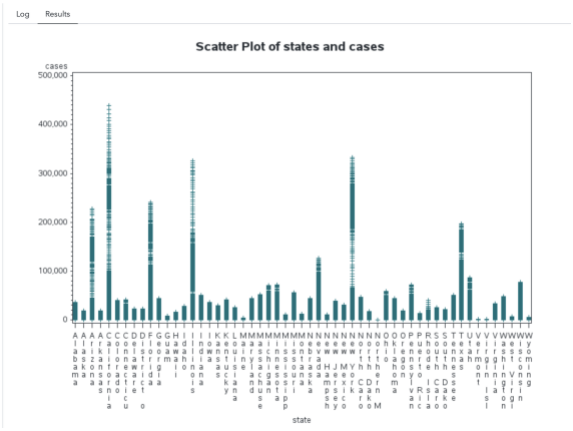


Fig. 10



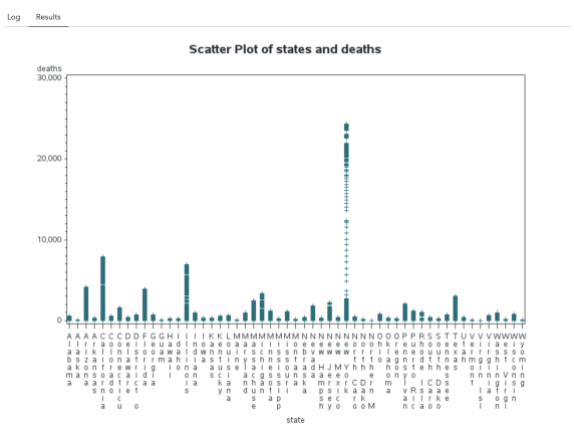


Fig. 11

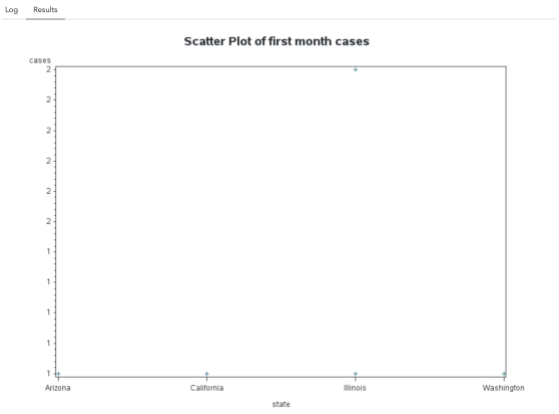
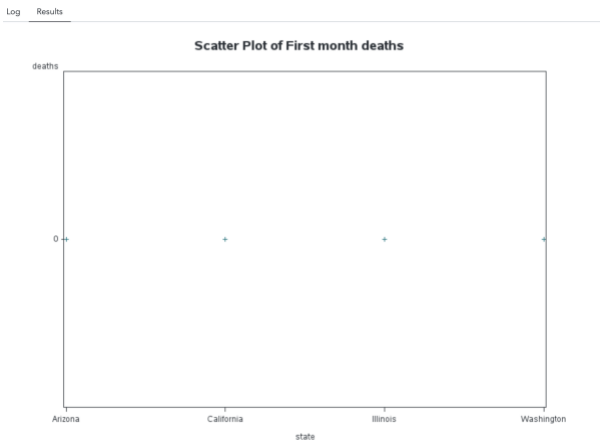


Fig. 12



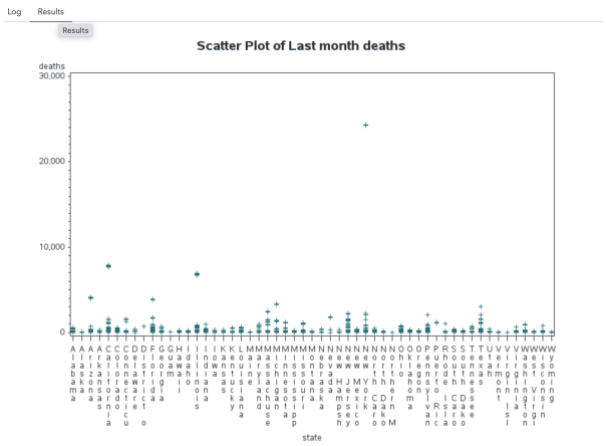


Fig. 13

From all the figured above, I could analyze those cases increased gradually every month and so the death but luckily deaths were not as high as cases produced even when it was deadly virus. I can tell from figure 10 that as cases and deaths were linearly increased. Figure 11 shows that California had the highest number of cases 4,39,538 in Los Angeles County and New York had highest number of deaths 24,346 all over the USA. There are many states that had zero cases and Maryland was one of them which has zero cases even in the last month of 2020! Texas, Illinois, Florida, Arizona joined California in massive number of cases, I found. Washington, Arizona, Illinois, and California started with 2 cases in January,2020 and no other state seen covid-19 virus even then. However, December showed all the states on the graph with huge number of cases. There were zero deaths in first month of covid-19 virus but throughout the year, ending in December, many states witnessed deaths. There are 2,35,150 observations out of all that showed zero deaths and only 665 observations showed zero cases. I exported my results and data using PROC EXPORT on EPG194 itself.

Summary

I started this project, and it took no longer than 2 minutes for me to know what kind of data I want to work on. As I got to know that we are going to choose our own goal, dataset, tasks and so on; I, right away got a thought about how exciting it would be to apply my SAS knowledge ona deadly coronavirus! I have been curious about COVID-19 data but never thought to work on it just like this. From results, I got to know more about how covid-19 affected the country, and I am glad I could do it using my SAS experience. I always had such questions like did the cases increased/decreased or what was the death graph all over the year and exactly which county/state got affected the most. Of course, news channels did show us lot of count and it is still available to see online but having the real dataset gathered by authorized organization and getting to work on it by coding on it is really a pleasure. That is why this project is even more special for me that I could convert my interest into a project work which I will cherish forever.

SAS Code

Starting the process by downloading dataset from the data source. I uploaded it into data folder of EPG194 into SAS. I used below steps in SAS code.

- PROC IMPORT- to import and then access data from local system into SAS.

- PROC CONTENTS - to see summary of dataset such as total number of observations, variable level information such as type, names, length, format.
- PROC PRINT - to print out the observations and look at raw dataset.
- PROC FREQ - to see descriptive statistics with keeping one variable as a center point and to see if that variable contains missing values or not.
- PROC MEAN - to answer question like what is minimum, maximum, or total number of cases/deaths, average number of cases/deaths.
- PROC SORT - to sort data and see highest or lowest value with WHERE statement.
- PROC GPLOT - to draw scatter plot with 'PLOT' to define x and y axis.
- DATA step - declare additional variable with specific format, change the format of existing variable and save clean data into pg1 library.
- Title – to give description of a graph or table or any result.
- ODS – to turn on graphics.
- PROC EXPORT – to export data once analysis is done.

The full SAS code is as below.

```
*Import data;
proc import datafile = '/home/vmodi/EPG194/data/covid-19_usa_cases.csv'
 out = pg1.covid_usa_cases
 dbms = CSV;
run;
*;

*Data exploration;
PROC CONTENTS DATA=pg1.covid_usa_cases order=varnum;
RUN;


*sorted data;
proc sort data=pg1.covid_usa_cases;
    by state date;
run;


*view raw data;
PROC PRINT DATA=pg1.covid_usa_cases noobs;
    by state;
RUN;


*data preparation;
data pg1.clean_data_covid_usa_cases;
    set pg1.covid_usa_cases;
    drop fips;
    month=MONTH(date);
    day=DAY(date);
    format cases comma9.;
    format deaths comma9.;
run;

*Means;
proc means data=pg1.clean_data_covid_usa_cases maxdec=2;
run;
```

```
ods graphics on;
*FREQ Table;
PROC FREQ DATA=pg1.clean_data_covid_usa_cases order=freq NLEVELS;
    TABLE county/MISSING;
RUN;

*FREQ Table;
PROC FREQ DATA=pg1.clean_data_covid_usa_cases order=data;
    TABLE state/nocum plots=freqplot(ORIENT=horizontal scale=percent);
RUN;

PROC FREQ DATA=pg1.clean_data_covid_usa_cases order=data;
    TABLE deaths/nocum plots=freqplot(ORIENT=horizontal scale=percent);
RUN;

PROC FREQ DATA=pg1.clean_data_covid_usa_cases order=data;
    TABLE cases/nocum plots=freqplot(ORIENT=horizontal scale=percent);
RUN;

proc sort data=pg1.clean_data_covid_usa_cases;
    by cases;
run;

proc sort data=pg1.clean_data_covid_usa_cases;
    by deaths;
run;

proc sort data=pg1.clean_data_covid_usa_cases;
    by descending cases;
run;

proc sort data=pg1.clean_data_covid_usa_cases;
    by descending deaths;
run;
```

```
*bar chart for cases;
proc sgplot data=pg1.clean_data_covid_usa_cases;
    vbar month/group=cases;
    title 'Cases count each month';
run;

*scatterplot;
proc gplot data=pg1.clean_data_covid_usa_cases;
    title Scatter Plot of Month and Cases;
    plot month* cases;
    run;
quit;

proc gplot data=pg1.clean_data_covid_usa_cases;
    title Scatter Plot of Month and deaths;
    plot month* deaths;
    run;
quit;

proc gplot data=pg1.clean_data_covid_usa_cases;
    title Scatter Plot of Cases and deaths;
    plot deaths* cases;
    run;
quit;

proc gplot data=pg1.clean_data_covid_usa_cases;
    title Scatter Plot of states and cases;
    plot cases* state;
    run;
quit;

proc gplot data=pg1.clean_data_covid_usa_cases;
    title Scatter Plot of states and deaths;
    plot deaths* state;
    run;
quit;
```

```
proc gplot data=pg1.clean_data_covid_usa_cases;
    title Scatter Plot of first month cases;
    plot cases* state;
    where month= 1;
run;
quit;

proc gplot data=pg1.clean_data_covid_usa_cases;
    title Scatter Plot of Last month cases;
    plot cases* state;
    where month= 12;
run;
quit;

proc gplot data=pg1.clean_data_covid_usa_cases;
    title Scatter Plot of First month deaths;
    plot deaths* state;
    where month= 1;
run;
quit;

proc gplot data=pg1.clean_data_covid_usa_cases;
    title Scatter Plot of Last month deaths;
    plot deaths* state;
    where month= 12;
run;
quit;

proc gplot data=pg1.clean_data_covid_usa_cases;
    title Scatter Plot of county and cases;
    plot cases* county;
run;
quit;

*export data;
proc export data=pg1.clean_data_covid_usa_cases
    outfile="/home/vmodi/EPG194/data/covid_data_export.csv"
    dbms=tab replace; run;
```

References

https://libguides.library.kent.edu/SAS/exploring
https://www.kaggle.com/sudalairajkumar/covid19-in-usa
https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/grstatproc/n0yjdd910dh59zn1toodgupaj4v9.htm
https://support.sas.com/rnd/datavisualization/papers/GraphBasics.pdf
https://www.tutorialspoint.com/sas/sas_bar_charts.html