The Foundations of Data Science Assignment Title Page:

Taslima Akter

Artificial Intelligence Diploma, PACE, University of Winnipeg

Course Name: Foundations of Data Science

Assignment 3: Data understanding and preparation

Student ID: 3040384

Professor's Name: Yelena Kropivnitskaya

March 21, 2022

Business/Research Objectives:

Already a major portion of the US citizens has been vaccinated by different brands.

- The main focus of the project is to measure the effectiveness of the Covid-19 vaccine in the USA.

Business Success Criteria:

- Effectiveness will be measured by comparing the death ratio of USA citizens before and after vaccination.

Previous Analysis:

- As part of the Research, Business objective and Data Collection parts have submitted with assignment-1 and assignment-2.

Current Analysis:

- This part is the continuation of previous submission. In this part of analysis, the data understanding and preparation part will be done for the ongoing research. As advance activity, the research objective will also be justified with the generated data frame.

Overview of the Tasks Covered with the Report:

Merging Dataset	Dealing with Missing Values	Data Preparation	Code	Remarks
ERD diagram and data merging explanation are provided.	In the merged table, missing values have been replaced with 'zero' and logic behind that is provided.	Five new attributes have been derived from the datasets in the 'Construct Data', 'Integrate Data' and 'Format Data' sections.	Required code is attached with the report.	For advance visualization, pandas_profiling also added with the code.

Overview of Data Preparation Steps:

Sl.No.	Steps	Action Taken	Positive Highlights	Limitations	Remarks
1	Select Data	-Initially two dataset has been selected about Covid-19. -Samples and variables have been selected from these two datasets based on the need of hypothesis. -Finally, a dataset has been generated from these two datasets. -Correlation between variables is performed.	-Data is clean. -Required Variables are available to perform analysis. -No missing data. - Datasets collected from online source and these are updated. - Numbers of data that are required to justify the hypothesis are available in the datasets. - Selected variables indicates negative correlation.	-Data collected only from online source. -No in-person survey to collect data. -Only some variables are considered for analysis because of the shortage of time.	-Details of every sections have been provided sequentially in the report. -Methods and techniques that are taken to analyze are clearly explained with visualization. -In the code, comments of every tasks are provided.
2	Clean Data	-From both the datasets some columns have been dropped, which are unnecessary.	-Pandas library is used dropped the column.	-N/A	
3	Construct Data	-In a dataset, data of USA from worldwide data following a time frame has extracted. -Changed the data type of some columns for the convenience of merging. - Derived new attribute by calculating percentage of death by comparing number of covid-19 infected with number of deaths due to Covid-19. - Derived new attributes by Summarizing data by adding them month wise from day wise data.	-Makes mathematical calculation, sorting and statistical calculation easier.	-N/A	
4	Integrate Data	-Merged two datasets and generated a new data table by these. - Applied techniques to handle new columns which have missing values.	-By generating new datasets, hypothesis is analyzed and assumed a conclusion.	-N/A	
5	Format Data	-Data is formatted converting numeric to texts and graphs.	-These techniques have made the better understanding of data.	-N/A	

1. **Select Data:**

Background of the Selected Datasets: Two datasets are selected from online source about covid-19. One dataset contains the information of COVID-19 vaccinations (Dataset-1) of the United States of America. Another dataset contains the information of Covid-19 of many countries around the world (Dataset-2).

In Dataset-1, initially there were total eight columns. These are 'location', 'date', 'vaccine', 'source_url', 'total_vaccinations', 'people_vaccinated', 'people_fully_vaccinated', 'total_boosters'. Two are considered for the analysis among these eight columns. Name of these two selected columns are 'date', 'people_fully_vaccinated'.

In Dataset-2, initially there were total five columns. Name of these columns are 'date', 'Country', 'Confirmed', 'Recovered', 'Deaths'. Total three columns are considered for the final analysis among these five columns, which are Date, 'Country', 'Confirmed', 'Deaths'.

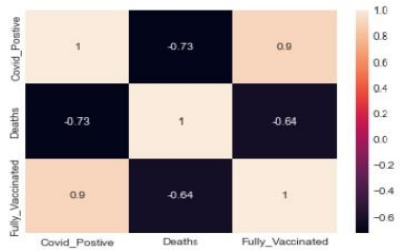
Reason of Selecting Mentioned Datasets and Variables: The purpose is to measure the efficacy of Covid-19 vaccine in the USA. These two datasets have date wise Covid-19 positive cases, vaccination, number of death records. Therefore, hypothesis can be justified by merging these two datasets, and analysing the result. Some variables are dropped from both datasets because they have no strong influence to the analysis.

Correlations of Variables: Variables of the datasets has been selected very carefully to serve the purpose of the project. From the figure, it has been seen that the two variables (‘Death’ and‘Fully_Vaccinated’) are negatively correlated. It indicates that with the increase of vaccination, the death rate of Covid-19 infected is decreased.

Fig- Correlation between ‘Death’ and ‘Fully_Vaccinated’ Variables:

```
In [129]: #finding corelation between Death and Fully_Vaccinated variables
          import pandas as pd
          import seaborn as sn
          import matplotlib.pyplot as plt

          corrMatrix = finaldfVaccCovid19.corr()
          sn.heatmap(corrMatrix, annot=True)
          plt.show()
          print('It has been seen that the variables "Death" and "Fully_Vaccinated" are negetively co-related.\nThat means, if
```




```
It has been seen that the variables "Death" and "Fully_Vaccinated" are negetively co-related.
That means, if vaccination is increased death is decreased.
```

```
In [130]: #finding corelation between Death and Fully_Vaccinated in another method
          finaldfVaccCovid19[['Deaths','Fully_Vaccinated']].corr()
```

Out[130]:

	Deaths	Fully_Vaccinated
Deaths	1.00000	-0.63876
Fully_Vaccinated	-0.63876	1.00000

2. **Clean Data:** Data cleaning activities are given below-

- Fortunately there were no Null values to handle in these two datasets. **After merging these two datasets, intentionally some zero values are taken in a column, which will explain in the data integration part.**

Fig-Data info of Dataset-1

```
In [78]: #to get information about column of dataset
         dfUSVacc.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 460 entries, 0 to 459
Data columns (total 8 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   location                460 non-null    object
 1   date                    460 non-null    object
 2   vaccine                 460 non-null    object
 3   source_url              460 non-null    object
 4   total_vaccinations      460 non-null    int64
 5   people_vaccinated       460 non-null    int64
 6   people_fully_vaccinated 460 non-null    int64
 7   total_boosters          460 non-null    int64
dtypes: int64(4), object(4)
memory usage: 28.9+ KB
```

Fig-Data info of Dataset-2

```
In [116]: #Information of colums and data after extraing USA data
          dfUSCovid19Data.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 776 entries, 144031 to 144806
Data columns (total 5 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   date       776 non-null    object
 1   Country    776 non-null    object
 2   Confirmed  776 non-null    int64
 3   Recovered  776 non-null    int64
 4   Deaths     776 non-null    int64
dtypes: int64(3), object(2)
memory usage: 36.4+ KB
```

- The unnecessary columns are removed from the Dataset-1. These are removed as these have no influence to justify the efficacy of the vaccine.

Fig- Code for Removing Unnecessary Columns from Dataset-1.

```
memory usage: 28.5+ KB

In [341]: #removed unecessary columns (vaccine,source_url)
          dfUSVacc.drop(['location','vaccine','source_url','total_vaccinations','tota

In [342]: #Columns after removed unecessary columns
          dfUSVacc.columns

Out[342]: Index(['date', 'people_fully_vaccinated'], dtype='object')
```

- Another dataset containing information of the Covid-19 (Dataset-2) of many countries has been loaded.

Fig- Loading Dataset-2.

```
In [348]: #load covid19 death,confirmed and recovered data
          urlCovid19='https://raw.githubusercontent.com/datasets/covid-19/master/data/countries-aggregated
          dfUSCovid19Data=pd.read_csv(urlCovid19)
          dfUSCovid19Data.rename(columns={'Date': 'date'}, inplace=True)
          # Print Initial dataset of covid-19
          dfUSCovid19Data
```

Out[348]:

	date	Country	Confirmed	Recovered	Deaths
0	2020-01-22	Afghanistan	0	0	0
1	2020-01-23	Afghanistan	0	0	0
2	2020-01-24	Afghanistan	0	0	0
3	2020-01-25	Afghanistan	0	0	0
4	2020-01-26	Afghanistan	0	0	0

- Later, only the rows of USA has extracted from the dataset by dropping the other countries' rows. These rows are removed because here we are measuring only the efficacy of vaccine in USA.

Fig- Extracting data of USA from Dataset-2.

```
memory usage: 5.9+ MB

In [183]: #filtering data for considering only USA data
          dfUSCovid19Data=dfUSCovid19Data.loc[(dfUSCovid19Data['Country'] == 'US') & (dfUSCovid19Data['date'] >= '2020-02-01')

In [184]: # Printing Covid-19 USA Data
          dfUSCovid19Data
```

Out[184]:

	date	Country	Confirmed	Recovered	Deaths
143482	2020-02-01	US	8	0	0
143483	2020-02-02	US	8	0	0
143484	2020-02-03	US	11	0	0
143485	2020-02-04	US	11	0	0
143486	2020-02-05	US	11	0	0
...	...	...	...	...	...
144236	2022-02-24	US	78812640	0	946099
144237	2022-02-25	US	78887236	0	948130
144238	2022-02-26	US	78934671	0	948826
144239	2022-02-27	US	78950518	0	949018
144240	2022-02-28	US	79047371	0	951114

- - The unnecessary columns are removed from the Dataset-2. These are removed as these have no influence to justify the efficacy of the vaccine. The 'Recovered' variable has removed because the efficacy will be measured by comparing the ratio of death and vaccination of the Covid-19 infected.

Fig- Code for Removing Unnecessary Columns from Dataset-2.

```
memory usage: 35.6+ KB
```

```
n [186]: #removed unnecessary columns, calculating death rate, and grouping
         dfUSCovid19Data.drop(['Recovered'],axis=1,inplace=True)
         #calculating death rate
```

3. Construct Data:

To make the data usable for further analysis and calculation, some tasks have performed. These are selecting data using a time frame, changing data types, summarizing and plotting graphs. In this section, new attributes are derived by percentage calculation and month wise summing variables. Details are as follows:

- In Dataset-2, Covid-19 information were from worldwide. Here only data of USA have extracted from the dataset. While doing this task, specific time limit has been used. The logic is that in the Vaccination dataset (Dataset-1), data is available up to the middle of the March 2022. Therefore, USA data have extracted from Dataset-2 matching with the time limit of Dataset-1 so that no mismatch is happened after merging these two datasets and the analysis become correct.

Fig- Dataset-1 till March 17, 2022

```
In [82]: #Printing Data after removing unnecessary columns
         dfUSVacc
```

Out[82]:

	date	people_fully_vaccinated
0	2020-12-13	5706
1	2020-12-14	5825
2	2020-12-15	6085
3	2020-12-16	6558
4	2020-12-17	7299
...	...	...
455	2022-03-13	216825493
456	2022-03-14	216871383
457	2022-03-15	216916501
458	2022-03-16	216947808
459	2022-03-17	216952347

460 rows × 2 columns

Fig-Code for Selecting Time Limit for Dataset-2

```
In [90]: data
SCovid19Data['Country'] == 'US') & (dfUSCovid19Data['date'] >= '2020-02-01') & (dfUSCovid19Data['date'] <= '2022-03-17')
```

Fig-Data Table of Dataset-2

```
In [118]: dfUSCovid19Data
```

Out[118]:

date	Country	Confirmed	Deaths
2020-02-29	US	25	0.040000
2020-03-31	US	192079	0.027900
2020-04-30	US	1076478	0.061454
2020-05-31	US	1788187	0.060214
2020-06-30	US	2645046	0.048152
2020-07-31	US	4543581	0.033856
2020-08-31	US	6042745	0.030348
2020-09-30	US	7239776	0.028575
2020-10-31	US	9156199	0.025285
2020-11-30	US	13626608	0.019875
2020-12-31	US	20191923	0.017421
2021-01-31	US	26338845	0.017025
2021-02-28	US	28747921	0.017886
2021-03-31	US	30563154	0.018042
2021-04-30	US	32452189	0.017728
2021-05-31	US	33376665	0.017788
2021-06-30	US	33777015	0.017888
2021-07-31	US	35101042	0.017463
2021-08-31	US	39383351	0.016273
2021-09-30	US	43526291	0.016074
2021-10-31	US	46035834	0.016235
2021-11-30	US	48583376	0.016092
2021-12-31	US	54834939	0.015097
2022-01-31	US	75093931	0.011854
2022-02-28	US	79051482	0.012034
2022-03-17	US	79683737	0.012173

- Datatype of a date column is changed to perform sorting and match up with the values of other dataset's column.
  Fig-Code for changing data type.

```
#Changing the format of date for the convenience of grouping data
dfUSCovid19Data['date'] = pd.to_datetime(dfUSCovid19Data['date'])
```

- Percentage of death is calculated by comparing number of covid-19 infected with number of deaths for Covid-19. Deriving death ratio from the two columns have made the analysis easier. For example, if a graph is made with the given number of deaths and vaccinations as these were in the dataset, it would plotted in such way that might hard for us to understand because the number of death is very low compared to the numbers of vaccination. This problem is solved by deriving the death percentage.

Fig- Code for Calculating Death Percentage.

```
#calculating death rate
dfUSCovid19Data['Deaths'] =  (dfUSCovid19Data['Deaths']/dfUSCovid19Data['Confirmed'])*1
```

Fig-Before Calculating the Death Percentage:

```
In [55]: #Graph to show death percentage before and after vaccination
         plt.style.use('seaborn-whitegrid')
         finaldfVaccCovid19.plot.line(x='Fully_Vaccinated', y=['Deaths'])
```

```
Out[55]: <AxesSubplot:xlabel='Fully_Vaccinated'>
```

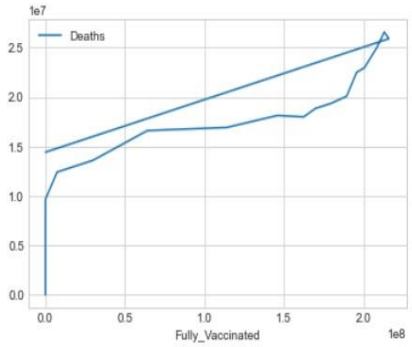
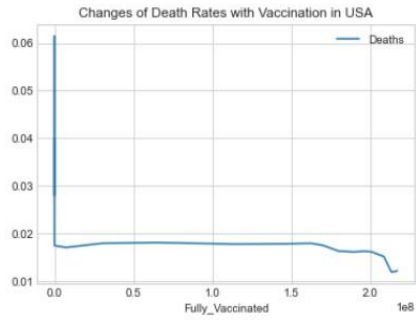


Fig-Before Calculating the Death Percentage:

```
In [121]: #Graph to show death percentage before and after vaccination
          plt.style.use('seaborn-whitegrid')
          finaldfVaccCovid19.plot.line(x='Fully_Vaccinated', y=['Deaths'])
          plt.title('Changes of Death Rates with Vaccination in USA')
          plt.show()
```



- Summarized data month wise in both the datasets for the convenience of analyzing.

Fig-Code and Output Table of Grouping Vaccination Data (Dataset-1):

```
In [55]: # Grouping vaccination data month wise
dfUSVacc['date'] = pd.to_datetime(dfUSVacc['date'])
dfUSVacc = dfUSVacc.sort_values(by='date')
dfUSVacc=dfUSVacc.groupby(pd.DatetimeIndex(dfUSVacc.date).to_period('M')).nth([-1])
dfUSVacc.set_index('date', inplace=True)
dfUSVacc.to_csv('dfUSVacc.csv')
dfUSVacc
```

Out[55]:

date	people_fully_vaccinated
2020-12-31	40720
2021-01-31	7351091
2021-02-28	29867214
2021-03-31	64171965
2021-04-30	114146822
2021-05-31	145971530
2021-06-30	162484459
2021-07-31	169897435
2021-08-31	179658096
2021-09-30	189511379
2021-10-31	195765678

Fig- Code for Grouping Covid-19 data (Dataset-2):

```
# Grouping month wise death rate of covid infected
dfUSCovid19Data=dfUSCovid19Data.groupby(pd.DatetimeIndex(dfUSCovid19Data.date).to_period('M')).nth([-1])
dfUSCovid19Data.set_index('date', inplace=True)
```

Fig-Data table of Dataset-2 after Month wise Summarize.

```
In [77]: dfUSCovid19Data
```

Out[77]:

date	Country	Confirmed	Deaths
2020-02-29	US	25	0.040000
2020-03-31	US	192079	0.027900
2020-04-30	US	1076478	0.061454
2020-05-31	US	1788187	0.060214
2020-06-30	US	2645046	0.048152
2020-07-31	US	4543581	0.033856
2020-08-31	US	6042745	0.030348
2020-09-30	US	7239776	0.028575
2020-10-31	US	9156199	0.025285
2020-11-30	US	13626608	0.019875
2020-12-31	US	20191923	0.017421
2021-01-31	US	26338845	0.017025
2021-02-28	US	28747921	0.017886
2021-03-31	US	30563154	0.018042
2021-04-30	US	32452189	0.017728

4. Integrate Data:

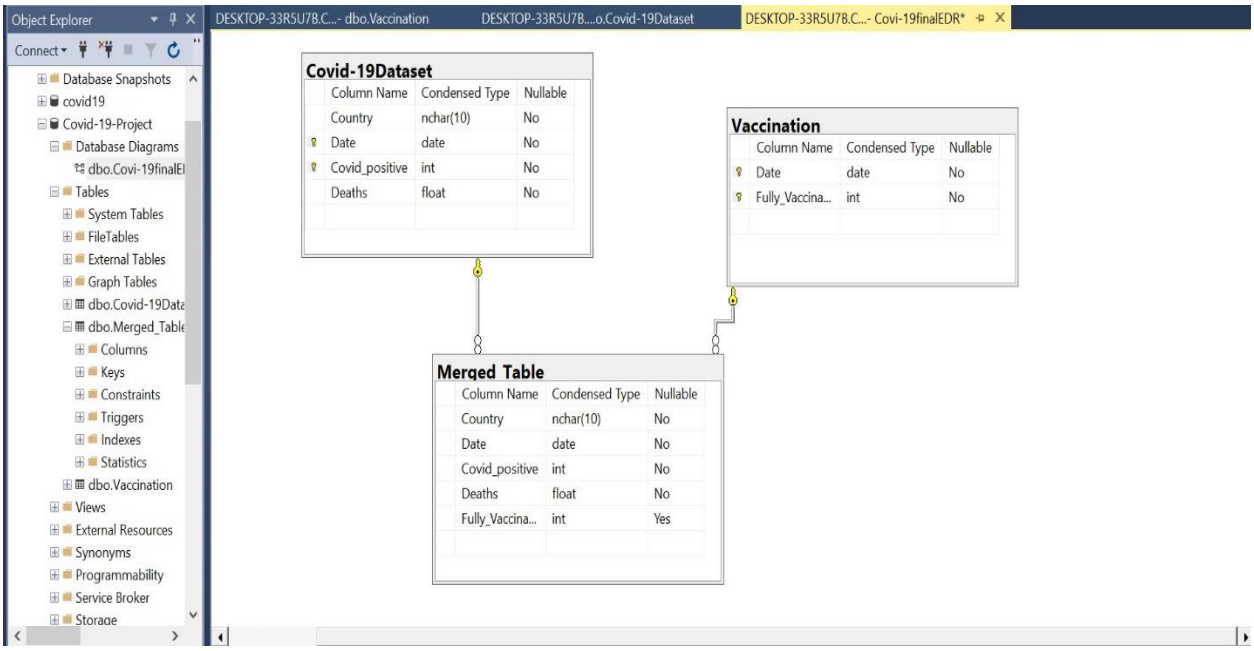
Reason of Merging Datasets:

After data cleaning and construction part, one dataset contains the data of month wise covid-19 vaccination and another dataset contains the data of date wise death percentage of covid-19 infected. It is not possible to draw any conclusion about the research objective from any of the individual dataset. Therefore, combined them to justify the hypothesis and have a meaningful conclusion.

Data Merging Technique:

Merged two datasets based on the date. Through that, a new datasets has been generated , where there are five columns including name of the country, dates, number of covid-19 positive cases, deaths due to covid-19 and number of fully_vaccinated people. In the ERD diagram, it is shown that, in the merged table data will come from two tables date wise and merged, but if the covid-19 dataset has values but vaccination dataset has no value on that date, the fully_vaccination column will take 'NULL' values in the merged dataset. 'Null' values are handles in the python program.

Fig- ERD Diagram for Joining Tables:



Handling Null/Missing Values:

While performing merging, in some rows where people are Covid-19 infected and death data are also available, but fully_vaccination data are missing on the same date. These rows are filled up with 'zero' in fully_vaccination column. This has been done to make the mathematical calculation and compare the death ratio before the vaccination, which are required to prove the efficacy of the vaccine. If the missing value is not replaced with numeric value (0) we would not make calculations, compare values clearly through graphs and make other analysis.

Code of Data Merging (Left Join): In the code 'Left join' is performed because many Covid-19 infected people died before vaccination. If o 'Right join' is performed in this case, the number of deaths before vaccination will lost from the new dataset. After merging, previous fully_vaccination column evolves with new features.

Fig: Code for Merging and Output Table:

```
: #merge covid19 data and vaccination data -- final comparison
  #Left join because many covid-infected people died before vaccination.
  #if we do right join these death before vaccination will lost from new dataset.
  finaldfVaccCovid19=pd.merge(dfUSCovid19Data,dfUSVacc,on='date',how='left')
  #In some rows where people are infected, deaths data are availabe but fully_vaccination data is Null,
  #I fill that rows with with zero in vaccination column.
  finaldfVaccCovid19.fillna(0, inplace=True)
  finaldfVaccCovid19.reset_index(inplace = True)
  print(finaldfVaccCovid19)
```

```
          date Country  Covid_Postive    Deaths  people_fully_vaccinated
0   2020-02-29      US             25  0.040000                      0.0
1   2020-03-31      US         192079  0.027900                      0.0
2   2020-04-30      US        1076478  0.061454                      0.0
3   2020-05-31      US        1788187  0.060214                      0.0
4   2020-06-30      US        2645046  0.048152                      0.0
5   2020-07-31      US        4543581  0.033856                      0.0
6   2020-08-31      US        6042745  0.030348                      0.0
7   2020-09-30      US        7239776  0.028575                      0.0
8   2020-10-31      US        9156199  0.025285                      0.0
9   2020-11-30      US       13626608  0.019875                      0.0
10  2020-12-31      US       20191923  0.017421                  40720.0
11  2021-01-31      US       26338845  0.017025                7351091.0
12  2021-02-28      US       28747921  0.017886               29867214.0
13  2021-03-31      US       30563154  0.018042               64171965.0
14  2021-04-30      US       32452189  0.017728              114146822.0
15  2021-05-31      US       33376665  0.017788              145971530.0
16  2021-06-30      US       33777015  0.017888              162484459.0
17  2021-07-31      US       35101042  0.017463              169897435.0
18  2021-08-31      US       39383351  0.016273              179658096.0
19  2021-09-30      US       43526291  0.016074              189511379.0
20  2021-10-31      US       46035834  0.016235              195765678.0
21  2021-11-30      US       48583376  0.016092              200776827.0
22  2021-12-31      US       54834939  0.015097              208435998.0
23  2022-01-31      US       75093931  0.011854              213156809.0
24  2022-02-28      US       79051482  0.012034              216129963.0
25  2022-03-17      US       79683737  0.012173              216952347.0
```

Fig- Final Table after Change of Column Name:

```
# Changing column Name
finaldfVaccCovid19.rename(columns={'date': 'Date'}, inplace=True)
finaldfVaccCovid19.rename(columns={'people_fully_vaccinated': 'Fully_Vaccinated'}, inplace=True)
#printing the new generated tables
finaldfVaccCovid19
```

	Date	Country	Covid_Postive	Deaths	Fully_Vaccinated
0	2020-02-29	US	25	0.040000	0.0
1	2020-03-31	US	192079	0.027900	0.0
2	2020-04-30	US	1076478	0.061454	0.0
3	2020-05-31	US	1788187	0.060214	0.0
4	2020-06-30	US	2645046	0.048152	0.0
5	2020-07-31	US	4543581	0.033856	0.0
6	2020-08-31	US	6042745	0.030348	0.0
7	2020-09-30	US	7239776	0.028575	0.0
8	2020-10-31	US	9156199	0.025285	0.0
9	2020-11-30	US	13626608	0.019875	0.0
10	2020-12-31	US	20191923	0.017421	40720.0
11	2021-01-31	US	26338845	0.017025	7351091.0
12	2021-02-28	US	28747921	0.017886	29867214.0
13	2021-03-31	US	30563154	0.018042	64171965.0
14	2021-04-30	US	32452189	0.017728	114146822.0
15	2021-05-31	US	33376665	0.017788	145971530.0
16	2021-06-30	US	33777015	0.017888	162484459.0
17	2021-07-31	US	35101042	0.017463	169897435.0
18	2021-08-31	US	39383351	0.016273	179658096.0
19	2021-09-30	US	43526291	0.016074	189511379.0
20	2021-10-31	US	46035834	0.016235	195765678.0
21	2021-11-30	US	48583376	0.016092	200776827.0
22	2021-12-31	US	54834939	0.015097	208435998.0
23	2022-01-31	US	75093931	0.011854	213156809.0
24	2022-02-28	US	79051482	0.012034	216129963.0
25	2022-03-17	US	79683737	0.012173	216952347.0

5. Format Data: In this part of data preparation, data is formatted numeric to text and graph is plotted to perform further analysis.

- After merging two data tables an existing column evolves with new features. Before merging, in Dataset-1 all the rows of vaccination column was containing values but after merging with Dataset-1 in the new table some of the rows of vaccination columns contain zero value. The reason behind that is to keep the Covid-19 data before and after vaccination. Details and pictures of this has given in the data integration part.

- A new column is derived from the merged table to easily understand who are vaccinated and who are not. Adding this feature helps to understand the table better.

Fig-Code for Creating 'Vaccination Status' Column.

```
In [40]: #Deriving new feature to easily understand which people are died before and after vaccination
         finaldfVaccCovid19['Vaccination_Status']=np.where((finaldfVaccCovid19.Deaths > 0.0) & (finaldfVaccCovid19.Fully_Vaccinat
```

```
In [42]: finaldfVaccCovid19['Vaccination_Status']

Out[42]: 0      Not vaccinated
         1      Not vaccinated
         2      Not vaccinated
         3      Not vaccinated
         4      Not vaccinated
         5      Not vaccinated
         6      Not vaccinated
         7      Not vaccinated
```

Fig-Table with New Attribute:

```
In [45]: finaldfVaccCovid19
```

Out[45]:

	Date	Country	Covid_Postive	Deaths	Fully_Vaccinated	Vaccination_Status
0	2020-02-29	US	25	0.040000	0.0	Not vaccinated
1	2020-03-31	US	192079	0.027900	0.0	Not vaccinated
2	2020-04-30	US	1076478	0.061454	0.0	Not vaccinated
3	2020-05-31	US	1788187	0.060214	0.0	Not vaccinated
4	2020-06-30	US	2645046	0.048152	0.0	Not vaccinated
5	2020-07-31	US	4543581	0.033856	0.0	Not vaccinated
6	2020-08-31	US	6042745	0.030348	0.0	Not vaccinated
7	2020-09-30	US	7239776	0.028575	0.0	Not vaccinated
8	2020-10-31	US	9156199	0.025285	0.0	Not vaccinated
9	2020-11-30	US	13626608	0.019875	0.0	Not vaccinated
10	2020-12-31	US	20191923	0.017421	40690.0	Fully Vaccinated
11	2021-01-31	US	26338845	0.017025	7350806.0	Fully Vaccinated
12	2021-02-28	US	28747921	0.017886	29866531.0	Fully Vaccinated

Final Analysis: In the final data table, it has been noticed that from February, 2020 to November, 2020 the death percentage of Covid-19 infected people were higher. During that time, no one was fully vaccinated. After that period, number of fully vaccination gradually increased and proportionately death percentage of Covid-19 infected decreased. This is also clearly visible in the line graph how the death percentage has changed with the change of vaccination over the time.

Fig- Final Data Table for Comparison with hypothesis.

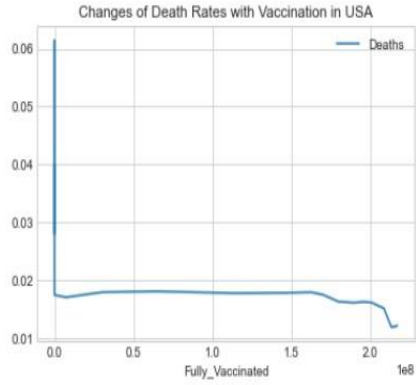
```
In [120]: finaldfVaccCovid19
```

Out[120]:

	Date	Country	Covid_Postive	Deaths	Fully_Vaccinated	Vaccination_Status
0	2020-02-29	US	25	0.040000	0.0	Not vaccinated
1	2020-03-31	US	192079	0.027900	0.0	Not vaccinated
2	2020-04-30	US	1076478	0.061454	0.0	Not vaccinated
3	2020-05-31	US	1788187	0.060214	0.0	Not vaccinated
4	2020-06-30	US	2645046	0.048152	0.0	Not vaccinated
5	2020-07-31	US	4543581	0.033856	0.0	Not vaccinated
6	2020-08-31	US	6042745	0.030348	0.0	Not vaccinated
7	2020-09-30	US	7239776	0.028575	0.0	Not vaccinated
8	2020-10-31	US	9156199	0.025285	0.0	Not vaccinated
9	2020-11-30	US	13626608	0.019875	0.0	Not vaccinated
10	2020-12-31	US	20191923	0.017421	40720.0	Fully Vaccinated
11	2021-01-31	US	26338845	0.017025	7351091.0	Fully Vaccinated
12	2021-02-28	US	28747921	0.017886	29867214.0	Fully Vaccinated
13	2021-03-31	US	30563154	0.018042	64171965.0	Fully Vaccinated
14	2021-04-30	US	32452189	0.017728	114146822.0	Fully Vaccinated
15	2021-05-31	US	33376665	0.017788	145971530.0	Fully Vaccinated
16	2021-06-30	US	33777015	0.017888	162484459.0	Fully Vaccinated
17	2021-07-31	US	35101042	0.017463	169897435.0	Fully Vaccinated
18	2021-08-31	US	39383351	0.016273	179658096.0	Fully Vaccinated
19	2021-09-30	US	43526291	0.016074	189511379.0	Fully Vaccinated
20	2021-10-31	US	46035834	0.016235	195765678.0	Fully Vaccinated
21	2021-11-30	US	48583376	0.016092	200776827.0	Fully Vaccinated
22	2021-12-31	US	54834939	0.015097	208435998.0	Fully Vaccinated
23	2022-01-31	US	75093931	0.011854	213156809.0	Fully Vaccinated
24	2022-02-28	US	79051482	0.012034	216129963.0	Fully Vaccinated
25	2022-03-17	US	79683737	0.012173	216952347.0	Fully Vaccinated

Fig- Changes of Date Rate with Vaccination in USA

```
In [121]: #Graph to show death percentage before and after vaccination
          plt.style.use('seaborn-whitegrid')
          finaldfVaccCovid19.plot.line(x='Fully_Vaccinated', y=['Deaths'])
          plt.title('Changes of Death Rates with Vaccination in USA')
          plt.show()
```



Conclusion of the Analysis: Based on all these analysis, it is justified that Covid-19 vaccine is effective in USA and reducing the death rates.

Advance Visualization of the Whole Project: Pandas_Profiling is used to get an overview of the project. In the coding section it has been mentioned. As the report is too long, few parts of the report are provided to have an overlook.

Fig- Some Pictures of pandas_profiling:

```
#Creating Report in Pandas
from pandas_profiling import ProfileReport

prof = ProfileReport(finaldfVaccCovid19)
prof.to_file(output_file='finaldfVaccCovid19_merge_col_prof.html')
prof
```

Summarize dataset: 100% 28/28 [00:03<00:00, 8.83it/s, Completed]

Generate report structure: 100% 1/1 [00:01<00:00, 1.72s/it]

Render HTML: 100% 1/1 [00:00<00:00, 2.61it/s]

Export report to file: 100% 1/1 [00:00<00:00, 49.42it/s]

Pandas Profiling Report Overview Variables Interactions Correlations Missing values Sample

Pandas Profiling Report Overview Variables Interactions Correlations Missing values Sample

Overview Alerts 21 Reproduction

Dataset statistics

Number of variables	6
Number of observations	26
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.3 KiB
Average record size in memory	52.9 B

Variable types

DateTime	1
Categorical	2
Numeric	3

Pandas Profiling Report | Overview | Variables | Interactions | Correlations | Missing values | Sample

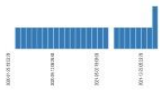
Variables

Date
Date

HIGH CORRELATION
UNIQUE

Distinct	26
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Memory size	336.0 B

Minimum	2020-02-29 00:00:00
Maximum	2022-03-17 00:00:00



Toggle details

Covid_Postive
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
UNIQUE

Distinct	26
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	28963554.58

Minimum	25
Maximum	79683737
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	336.0 B

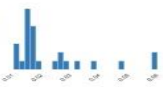
Toggle details

Deaths
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
UNIQUE

Distinct	26
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	0.02433589503

Minimum	0.0118538341
Maximum	0.06145411239
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	336.0 B



Toggle details

Out[421]:

Fully_Vaccinat...
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	17
Distinct (%)	65.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	89012243.58

Minimum	0
Maximum	216952347
Zeros	10
Zeros (%)	38.5%
Negative	0
Negative (%)	0.0%
Memory size	336.0 B

Toggle details

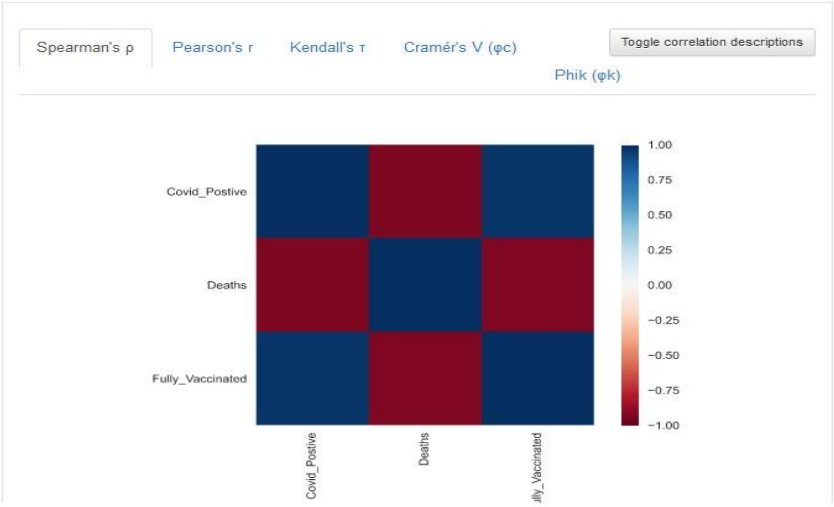
Vaccination_S...
Categorical

HIGH CORRELATION
HIGH CORRELATION

Distinct	2
Distinct (%)	7.7%
Missing	0
Missing (%)	0.0%
Memory size	336.0 B

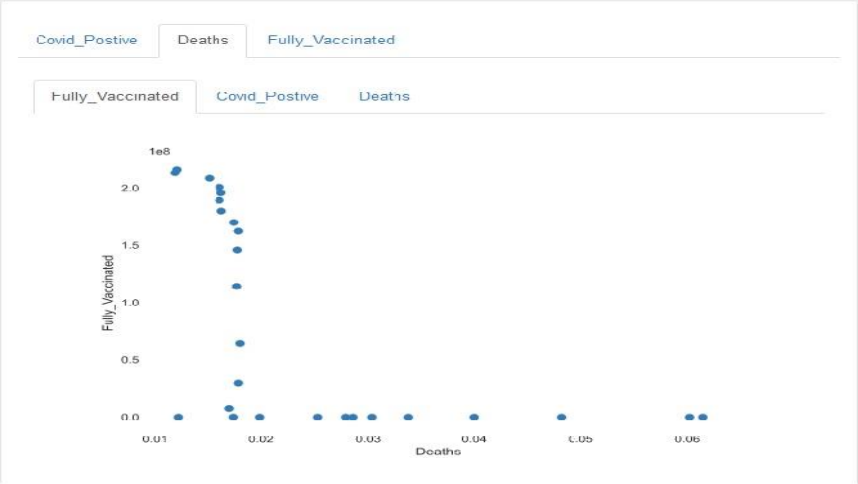
Toggle details

Correlations



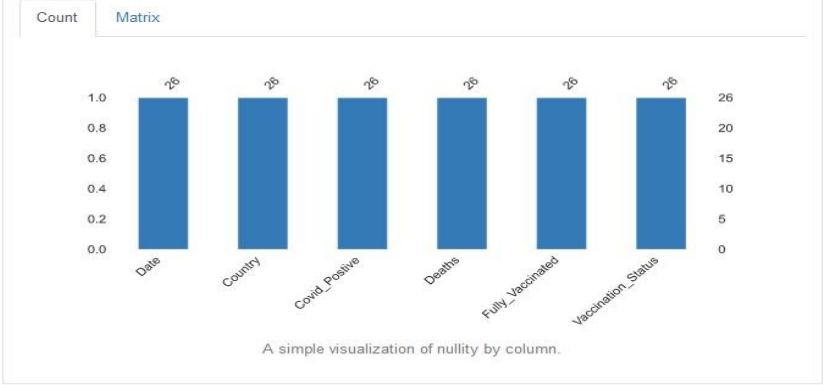
Out[42]:

Interactions



Out[42]:

Missing values



Sample

First rows

	Date	Country	Covid_Postive	Deaths	Fully_Vaccinated	Vaccination_Status
0	2020-02-29	US	25	0.040000	0.0	Not vaccinated
1	2020-03-31	US	192079	0.027900	0.0	Not vaccinated
2	2020-04-30	US	1076478	0.061454	0.0	Not vaccinated
3	2020-05-31	US	1788187	0.060214	0.0	Not vaccinated
4	2020-06-30	US	2645046	0.048152	0.0	Not vaccinated
5	2020-07-31	US	4543581	0.033856	0.0	Not vaccinated
6	2020-08-31	US	6042745	0.030348	0.0	Not vaccinated
7	2020-09-30	US	7239776	0.028575	0.0	Not vaccinated
8	2020-10-31	US	9156199	0.025285	0.0	Not vaccinated
9	2020-11-30	US	13626608	0.019875	0.0	Not vaccinated

Last rows

	Date	Country	Covid_Postive	Deaths	Fully_Vaccinated	Vaccination_Status
16	2021-06-30	US	33777015	0.017888	162484459.0	Fully Vaccinated
17	2021-07-31	US	35101042	0.017463	169897435.0	Fully Vaccinated
18	2021-08-31	US	39383351	0.016273	179658096.0	Fully Vaccinated
19	2021-09-30	US	43526291	0.016074	189511379.0	Fully Vaccinated
20	2021-10-31	US	46035834	0.016235	195765678.0	Fully Vaccinated
21	2021-11-30	US	48583376	0.016092	200776827.0	Fully Vaccinated
22	2021-12-31	US	54834939	0.015097	208435998.0	Fully Vaccinated
23	2022-01-31	US	75093931	0.011854	213156809.0	Fully Vaccinated
24	2022-02-28	US	79051482	0.012034	216129963.0	Fully Vaccinated
25	2022-03-17	US	79683737	0.012173	216952347.0	Fully Vaccinated

Python Code:

Code of the analysis is added with the report. Jupyter Notebook platform has been used for coding and analysis.