Report

Table of Contents

[Introduction to Data set 2](#_Toc153147648)

[Data Analysis Findings 3](#_Toc153147649)

[List of Research Questions 3](#_Toc153147650)

[Data Preprocessing 3](#_Toc153147651)

[Data Visualization 3](#_Toc153147652)

[Literature Review 8](#_Toc153147653)

[Limitations 9](#_Toc153147654)

[Conclusion 10](#_Toc153147655)

[Appendix 10](#_Toc153147656)

[References 13](#_Toc153147657)

# Introduction to Data set

The selected dataset for the analysis is global UNICEF immunization data which is taken from Kaggle website. There are more than 58000 records with 11 attributes in the selected dataset. The dataset includes data from the dates from 1980 until 2019. The dataset contains data regarding percentage of children vaccinated pertaining to 214 countries over a period of 40 years for 14 vaccines grouped by type/dose information. It can be seen that these columns contains the same information. It can also be noted that the data is not available for all countries for all vaccines. Hence for any particular use case, the sufficiency of available data needs to be checked. The dataset consists of the data of various vaccines administered to the children geographical area wise. The attributes of the dataset are as follows

* Geographic area – This attribute contains the geographic area of the world i.e. which country the record is belong to.
* Indicator – This attribute contains the value of the indicator for which the data has been collected such as Percentage of surviving infants who received the first dose of DTP-containing vaccine etc.
* Vaccine – This attribute contains the name of the vaccine for which the data has been collected.
* Current age – This attribute contains the value of the age of the kid to whom the vaccine will be administered.
* Unit multiplier – In this the value of unit multiplier has been captured like units etc.
* Unit of measure – This attribute contains the value of unit of measure of the collected data such as per cent etc.
* TIME\_PERIOD – In this field the time period of the collected record would be captured such as 1980, 2019 etc. This field contains records from 1980 to 2019.
* OBS\_VALUE – This attribute contains the observation value of the collected record of vaccine.
* Observation Status – This attribute contains the status value of the collected observation regarding the vaccine.
* Observation confidentiality – This attribute contains whether the collected observation is free, private or confidential etc.
* DATA\_SOURCE – This attribute contains the value from where the data has been collected.

# Data Analysis Findings

## List of Research Questions

By the analysis of the selected dataset we would like to find following research questions, and I am sure we could find the same easily:

How many vaccines are there across the world and which ones are more effective and popular?

In which country the vaccines are getting administered to the infants or kids in large numbers and effectively?

What is the share of vaccines across the world, i.e. which one is the highest and which one is the lowest?

What is the indicator being followed in various countries?

There are many more analysis questions which could be derived from the analysis of the given dataset.

## Data Preprocessing

Null values were removed from the dataset and fresh dataset was created using the same, the data frame newdata was created.

Data was checked using data.describe.T method. Following result was found

TIME\_PERIOD OBS\_VALUE

Count 58317.000000 58317.000000

Mean 2003.730593 80.390795

Std 10.927390 22.855291

Min 1980.000000 1.000000

25% 1995.000000 73.000000

50% 2005.000000 90.000000

75% 2013.000000 96.000000

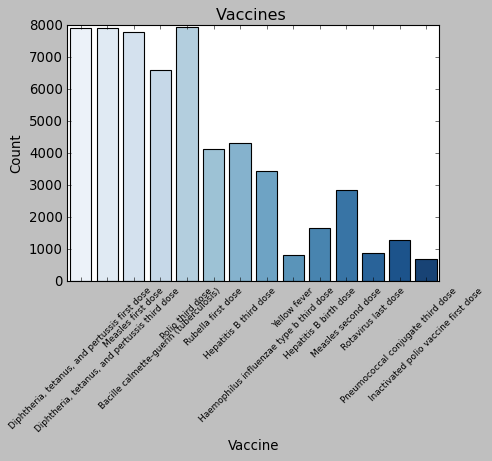
Max 2019.000000 99.000000

As per the above correlation matrix the dataset contains data regarding percentage of children vaccinated pertaining to 214 countries over a period of 40 years for 14 vaccines grouped by type/dose information.

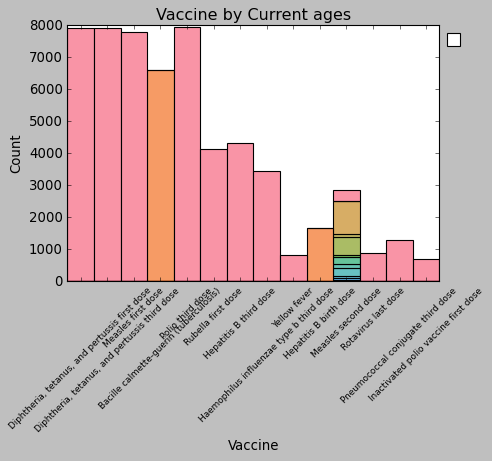
## Data Visualization

Following are some of the data visualizations:

Below plot shows the count of various available vaccines across the globe

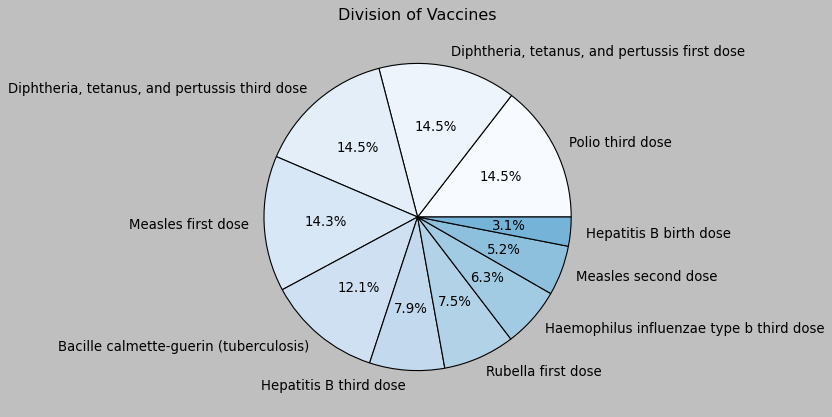


Above chart illustrates the number of vaccines administered to the kids across the globe. In this lighter color bar is the most number of vaccines and darker color bar is the least.



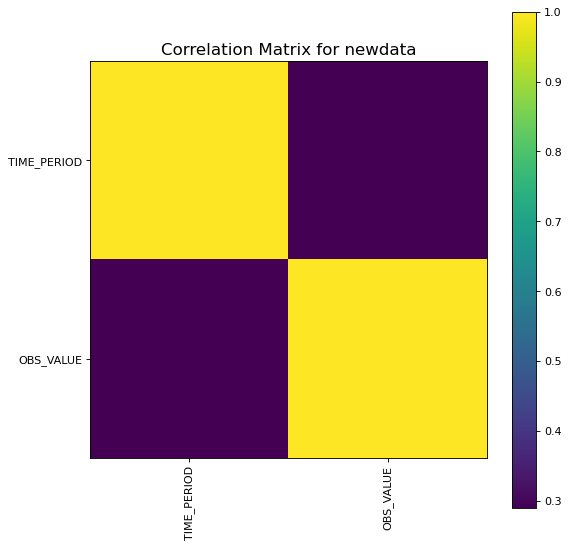
Above graph has used two categorical variables Vaccine and Current Age to build the plot.

Below pie chart is the percentage of top 10 vaccines administered across the globe.

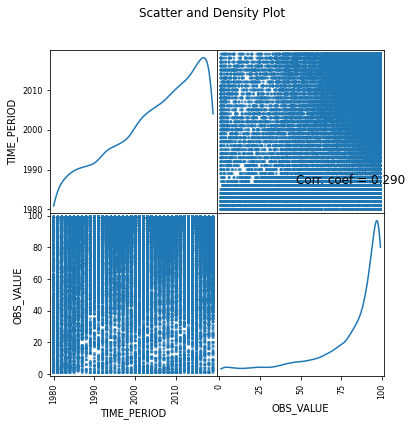


The above graph shows highest share is captured by polio third dose, diphtheria, tetanus and pertussis first and third dose each having 14.5%.

Correlation matrix between the time period and OBS value has been created

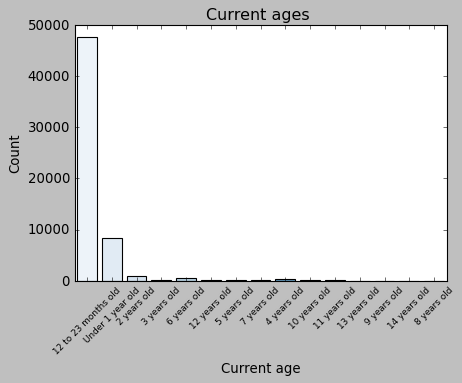
****

Scatter and density plot between the time period and OBS value has been created

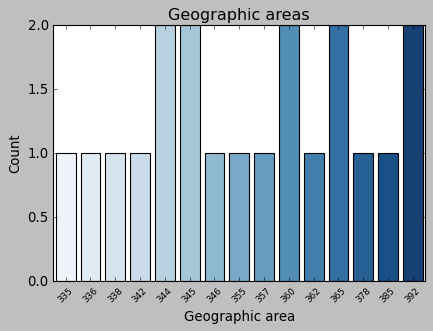
****

Below plot shows the count of various current ages for the vaccine across the globe.

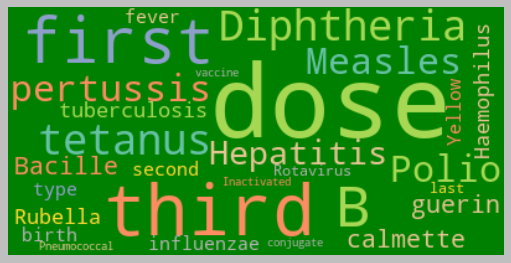
It illustrates the count of current age for the vaccine across the globe. In this lighter color bar is the most number of vaccines and darker color bar is the least.

****

Following chart illustrates the top 20 count of vaccine records geographical area wise of various administered vaccine across the globe to the infants and kids.

****

Following chart illustrates the word cloud of various administered vaccine across the globe to the infants and kids.

****

## Literature Review

According to the sources the vaccines are among the supreme developments in world health and advancement. Since very long time almost like more than 200 years, vaccines have effectively abridged the plague of many deadly diseases like diphtheria, polio, measles and smallpox, aiding infants and kids come of age healthier and happier. The success ratio of these vaccines are tremendous, almost saving 5 or 6 lives per minute which sums to saving up to 3 million casualties each year, and that too when this dataset does not contains records of latest pandemic of COVID-19.

This was possible only because of immunization drive which is conducting with the help of UNICEF globally kids are walking properly, playing, dancing, learning and doing their all daily chores without any help. The children who are vaccinated performing well at school, with financial welfares that flow across their societies. Today, vaccines are estimated to be one of the most cost-effective means of advancing global welfare.

Although the data of vaccinated infants and kids are very high and success rate also tremendous, still the immunization level is not as per the expected or how it should be. A very large number of infants and kids roughly around 21 million do not get the vaccine they supposed to. The infants and kids from the financially backward background and from under developed countries are the most affected. As per the data suggests major sector of such children live in such countries, in far-flung places, or where polio is still an endemic.

Poor state of immunization drive too is major concern for immunization drive in many places of mother and infants’ health and well-being.

According to the WHO sources the hesitant behavior to get the vaccine is the most common concern for the immunization drive. This concern is not new it was there too when the first vaccine was developed, it’s basically human nature to accept any new thing, only the global places are shifted from well-educated or well sourced countries to the poor ones.

With the help of its associates across the globe, the UNICEF now delivers vaccines to stretch nearly half of the world’s infants and kids below or equal to five years’ age. Many organizations, governments, private companies, NGOs, and of course the United Nations are working very hard to achieve the utmost target somehow in more than hundred countries globally. They are taking help from local bodies to safeguard the deliveries, reach the deliveries on time to the needy ones and most importantly they are trying hard to keep it affordable or free if possible to each and every infant or kid in any far-flung area.

On the dataset of historical hydrologic for the purpose of visualizations many charts or graphs have been prepared such as bar chart, pie chart, correlation matrix, scatter plots etc.

# Limitations

The limitations of this dataset are that as the number of attributes is not very much so more number of critical analysis is bit tricky, if we could have more elaborate data then we can analyze the situation better. It will also not able to handle the dynamic changes of vaccines’ data. In this dataset user level results cannot be presented or any visualizations or analysis cannot be done as user level as the data is at only broad country or geographical area level. As we all know many countries are not providing any data or proving insufficient data at the area level within the country or from all member states. According to sources roughly more than 50 countries are not providing any data at all regarding the immunization, so it’s impossible to make any global consensus regarding the drive.

The quality of data which countries are providing is not good as they are providing substantial average of data regarding immunization.

# Conclusion

The given data is bit old as it dates back to year 2019. If we consider some latest scenario for analysis, the global coverage of the third dose of diphtheria tetanus pertussis fell from eighty six per cent in year 2019 to eighty three per cent in last year. The latest WHO and UNICEF vaccination estimates of national immunization coverage also illustrates that ninety per cent of countries across the world that reported 2020 data experienced stagnant or declining coverage of third dose of diphtheria tetanus pertussis compared to 2019 with nearly thirty per cent viewing a degeneration of at least five percentage points. As a result, 23 million children were not vaccinated or not fully vaccinated (some children have not received all three doses of diphtheria tetanus pertussis) in 2020. Of those 23 million children, more than sixty per cent live in just ten countries (India, Nigeria, Democratic Republic of the Congo, Pakistan, Indonesia, Ethiopia, Brazil, Philippines, Angola and Mexico) and 17 million of them did not receive any vaccines (zero-dose children).

# Appendix

import warnings

warnings.filterwarnings("ignore")

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

data = pd.read\_csv('fusion\_IMMUNISATION\_UNICEF\_1.0\_all.csv')

newdata = data.dropna(axis=1)

newdata.head().T

# describe the data

newdata.describe()

# The selected data belongs to the period from 1980-2019.

# per column unique values

newdata.nunique()

newdata.shape

newdata.head()

# To check the count of values by Current Age

newdata['Current age'].value\_counts()

# Plot tp show frequency chart for Current Age

plt.style.use("classic")

plt.figure(figsize=(6, 4))

sns.countplot(newdata['Current age'], palette='Blues', \*\*{'hatch':'','linewidth':1})

plt.xlabel("Current age")

plt.ylabel("Count")

plt.title("Current ages ")

plt.xticks(rotation=45, fontsize=8)

plt.show()

# Plot to show number of administered vaccine by vaccine type and current age

plt.style.use("classic")

plt.figure(figsize=(6, 4))

sns.histplot(binwidth=0.5, x="Vaccine", hue="Current age", data=newdata, stat="count", multiple="stack")

plt.legend(bbox\_to\_anchor=(1, 1), loc=2)

plt.xlabel("Vaccine")

plt.ylabel("Count")

plt.title("Vaccine by Current ages")

plt.xticks(rotation=45, fontsize=8)

plt.show()

# Plot to show number of administered vaccine by vaccine type

import seaborn as sns

plt.style.use("classic")

plt.figure(figsize=(6, 4))

sns.countplot(newdata['Vaccine'], palette='Blues', \*\*{'hatch':'','linewidth':1})

plt.xlabel("Vaccine")

plt.ylabel("Count")

plt.title("Vaccines ")

plt.xticks(rotation=45, fontsize=8)

plt.show()

# Plot to show number of administered vaccine by OBS Values

plt.style.use("classic")

plt.figure(figsize=(6, 4))

sns.countplot(newdata['OBS\_VALUE'], palette='Blues', \*\*{'hatch':'','linewidth':1})

plt.xlabel("Observation VAlue")

plt.ylabel("Count")

plt.title("Observation Values ")

plt.xticks(rotation=45, fontsize=8)

plt.show()

# Plot to show number of administered vaccine by Time Period

plt.style.use("classic")

plt.figure(figsize=(8, 6))

sns.countplot(newdata['TIME\_PERIOD'], palette='Blues', \*\*{'hatch':'','linewidth':1})

plt.xlabel("TIME PERIOD")

plt.ylabel("Count")

plt.title("Time Periods ")

plt.xticks(rotation=45, fontsize=8)

plt.show()

# Plot to show shared division of vaccines

color = plt.cm.Blues(np.linspace(0,1,20))

newdata["Vaccine"].value\_counts().sort\_values(ascending=False).head(10).plot.pie(y="Indicator",colors=color,autopct="%0.1f%%")

plt.title("Division of Vaccines")

plt.axis("off")

plt.show()

# Correlation matrix

import matplotlib.pyplot as plt

newdata = newdata[[col for col in newdata if newdata[col].nunique() > 1]]

corr = newdata.corr()

plt.figure(num=None, figsize=(8, 8), dpi=80, facecolor='w', edgecolor='k')

corrMat = plt.matshow(corr, fignum = 1)

plt.xticks(range(len(corr.columns)), corr.columns, rotation=90)

plt.yticks(range(len(corr.columns)), corr.columns)

plt.gca().xaxis.tick\_bottom()

plt.colorbar(corrMat)

plt.title(f'Correlation Matrix for newdata', fontsize=15)

plt.show()

# Scatter matrix

df = newdata.select\_dtypes(include =[np.number]) # keep only numerical columns

df = df[[col for col in df if df[col].nunique() > 1]]

columnNames = list(df)

if len(columnNames) > 10: # reduce the number of columns for matrix inversion of kernel density plots

columnNames = columnNames[:10]

df = df[columnNames]

ax = pd.plotting.scatter\_matrix(df, alpha=0.75, figsize=[6, 6], diagonal='kde')

corrs = df.corr().values

for i, j in zip(\*plt.np.triu\_indices\_from(ax, k = 1)):

ax[i, j].annotate('Corr. coef = %.3f' % corrs[i, j], (0.8, 0.2), xycoords='axes fraction', ha='center', va='center', size=12)

plt.suptitle('Scatter and Density Plot')

plt.show()

# plot for count geographical area

import seaborn as sns

plt.style.use("classic")

plt.figure(figsize=(6, 4))

sns.countplot(newdata['Geographic area'].value\_counts().head(20), palette='Blues', \*\*{'hatch':'','linewidth':1})

plt.xlabel("Geographic area")

plt.ylabel("Count")

plt.title("Geographic areas")

plt.xticks(rotation=45, fontsize=8)

plt.show()

#word cloud

from wordcloud import WordCloud, ImageColorGenerator

text = " ".join(str(each) for each in newdata.Vaccine.unique())

wordcloud = WordCloud(max\_words=200,colormap='Set2', background\_color="green").generate(text)

plt.figure(figsize=(8,5))

plt.figure(figsize=(8,5))

plt.imshow(wordcloud, interpolation='Bilinear')

plt.axis("off")

plt.figure(1,figsize=(12, 12))

plt.show()

# Plot to show number of administered vaccine by vaccine type and Time Period

plt.style.use("classic")

plt.figure(figsize=(6, 4))

sns.histplot(binwidth=0.5, x="TIME\_PERIOD", hue="Vaccine", data=newdata, stat="count", multiple="stack")

plt.legend(bbox\_to\_anchor=(1, 1), loc=2)

plt.xlabel("Vaccine")

plt.ylabel("Count")

plt.title("Vaccine by TIME\_PERIOD and type")

plt.xticks(rotation=45, fontsize=8)

plt.show()

# References

Carol Liu, Yi-Chun, et al. “Sensorineural Hearing Loss (SNHL) as an Adverse Event Following Immunization (AEFI): Case Definition & Guidelines for Data Collection, Analysis, and Presentation of Immunization Safety Data.” *Vaccine*, vol. 38, no. 30, June 2020, pp. 4717–4731, 10.1016/j.vaccine.2020.05.019. Accessed 21 Jan. 2021.

Thompson, Kimberly M., and Grace E. Logan. “Characterization of Heterogeneity in Childhood Immunization Coverage in Central Florida Using Immunization Registry Data.” *Risk Analysis*, vol. 36, no. 7, 29 May 2015, pp. 1418–1426, 10.1111/risa.12424.

“Vaccination and Immunization Statistics - UNICEF DATA.” *UNICEF DATA*, 2018, data.unicef.org/topic/child-health/immunization/.

Chieu, YC., Chang, CW., & Lea, TY. (2016). modelling of bio-chemical O2 demand at the up-stream catchment of Feiitsui Reservoir, Taiwan. *Hydrology Research*, *47*(5), 1069–1085.

Heapşen, A., & Vaatansever, M. (2011). Forecasting upcoming tendencies in Dubai housing market by using BoxJenkins auto-regressive integrated moving mean. *International Journal of Housing Markets and Analysis*, *4*(3), 210–223.

MaC., Waong Smit, H., ChuCk., & Taiera Juuarez, D. (2015). Big-data in medicine practice: current use, challenges, and the future. *Integrated Pharmacy Research and Practice*, 91.