KLE Technological University
Hubballi

A Course Project Report on

"Covid-19 Statewise Vaccine Analysis"

A Course Project Report Submitted in Partial Fulfillment of the Requirement for the Course of

Exploratory Data Analysis

in

4th Semester of Computer Science and Engineering

by

AKHILESH JOSHI	02FE22BCS013
DILIPSINGH RAJPUROHIT	02FE22BCS035
SANA MULLA	02FE22BCS099
SEEBA DODDMANI	02FE22BCS116

Under the guidance of

Dr. Prema T. Akkasaligar
Professor,
Department of Computer Science and Engineering,
KLE Technological University's Dr. MSSCET, Belagavi.

KLE Technological University's
Dr. M. S. Sheshgiri College of Engineering and Technology,
Belagavi – 590 008.

June 2024

DECLARATION

We hereby declare that the matter embodied in this report entitled "**Covid-19 Statewise Vaccinie Analysis**" submitted to KLE Technological University for the course completion of Exploratory Data Analysis (21ECSC210) in the 4th Semester of Computer Science and Engineering is the result of the work done by us in the Department of Computer Science and Engineering, KLE Dr. M. S. Sheshgiri College of Engineering, Belagavi under the guidance of Dr.Prema T. Akkasaligar, Professor, Department of Computer Science and Engineering. We further declare that to the best of our knowledge and belief, the work reported here in doesn't form part of any other project on the basis of which a course or award was conferred on an earlier occasion on this by any other student(s), also the results of the work are not submitted for the award of any course, degree or diploma within this or in any other University or Institute. We hereby also confirm that all of the experimental work in this report has been done by us.

Belagavi – 590 008
Date : 10th June 2024

AKHILESH JOSHI (02FE22BCS013)	DILIPSHINGH RAJPUROHIT (02FE22BCS035)
SANA MULLA (02FE22BCS099)	SEEBA DODDMANI (02FE22BCS116)

KLE TECH KLE
TECHNOLOGICAL UNIVERSITY
Creating Value, Leveraging Knowledge
Belagavi Campus

Dr.M.S.Sheshgiri College of Engineering & Technology

Department of Computer Science & Engineering

CERTIFICATE

This is to certify that the project entitled "Covid-19 Statewise Vaccine Analysis" submitted to KLE Technological University's Dr. MSSCET, Belagavi for the partial fulfillment of the requirement for the course - Exploratory Data Analysis (21ECSC210) by Akhilesh Joshi,Dilipsingh Rajpurohit,Sana Mulla,Seeba Doddmani, students in the Department of Computer Science and Engineering, KLE Technological University's Dr. MSSCET, Belagavi, is a bonafide record of the work carried out by them under my supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any other course completion.

Belagavi – 590 008
Date : 10th June 2024

Dr. Prema T Akkasaligar	Prof. Priyanka Gavade
(Course Teacher)	(Course Coordinator)

Dr. Rajashri Khanai
(Head of the Department)

Abstract

Coronaviruses, a diverse group of viruses, can lead to illnesses in both animals and humans. In humans, these viruses can cause a range of respiratory infections, from mild colds to more serious conditions like Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The newest addition to this family of viruses is the one responsible for the COVID-19 disease, as identified by the World Health Organization

Problem Statement: Analyze the distribution and administration of COVID-19 vaccines across different states in India

Solution: We aim to uncover regional disparities, assess vaccination rates, understand demographic trends, address challenges, and predict future vaccine needs. By doing so, the project aims to provide actionable insights to optimize vaccine deployment strategies and combat the pandemic effectively.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Coronaviruses, a diverse group of viruses, can lead to illnesses in both animals and humans. In humans, these viruses can cause a range of respiratory infections, from mild colds to more serious conditions like Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The newest addition to this family of viruses is the one responsible for the COVID-19 disease, as identified by the World Health Organization.

1.2 Problem Statement

Analyze the distribution and administration of COVID-19 vaccines across different states in India

1.2.1 Objectives

We aim to uncover regional disparities, assess vaccination rates, under- stand demographic trends, address challenges, and predict future vaccine needs. By doing so, the project aims to provide actionable insights to optimize vaccine deployment strategies and combat the pandemic effectively.

Chapter 2

Knowing the Dataset

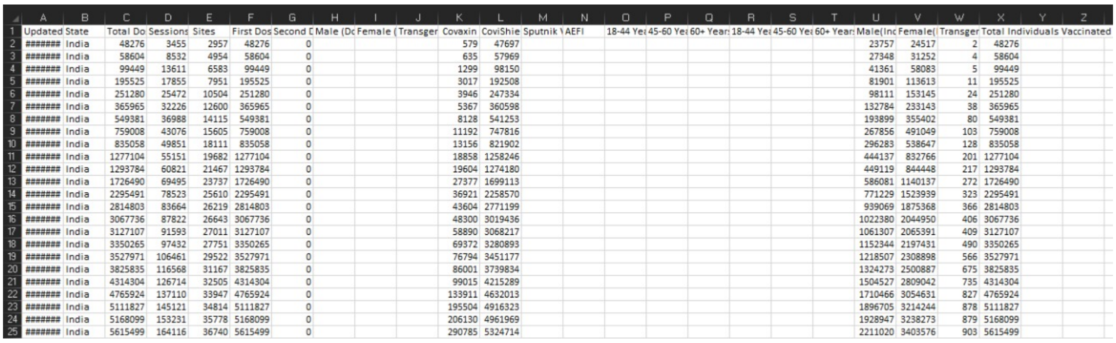
2.1 Dataset

- The Dataset has 7846 entries, 0 to 7845 and total 24 columns.
- The data was recorded from 16th January 2021 to 12th August 2021
- Source URL :https://www.kaggle.com/datasets/sudalairajkumar/covid19-in-india

2.2 Features of the Dataset

2.3 Observations

List your observations from the dataset here.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Updated	State	Total Do	Sessions	Sites	First Dos	Second D	Male (Dc	Female (	Transger	Covaxin	CoviShie	Sputnik \	AEFI	18-44 Yea	45-60 Yea	60+ Year	18-44 Yea	45-60 Yea	60+ Year	Male(Inc	Female(I	Transger	Total Individuals Vaccinated		
2	#######	India	48276	3455	2957	48276	0				579	47697									23757	24517	2	48276		
3	#######	India	58604	8532	4954	58604	0				635	57969									27348	31252	4	58604		
4	#######	India	99449	13611	6583	99449	0				1299	98150									41361	58083	5	99449		
5	#######	India	195525	17855	7951	195525	0				3017	192508									81901	113613	11	195525		
6	#######	India	251280	25472	10504	251280	0				3946	247334									98111	153145	24	251280		
7	#######	India	365965	32226	12600	365965	0				5367	360598									132784	233143	38	365965		
8	#######	India	549381	36988	14115	549381	0				8128	541253									193899	355402	80	549381		
9	#######	India	759008	43076	15605	759008	0				11192	747816									267856	491049	103	759008		
10	#######	India	835058	49851	18111	835058	0				13156	821902									296283	538647	128	835058		
11	#######	India	1277104	55151	19682	1277104	0				18858	1258246									444137	832766	201	1277104		
12	#######	India	1293784	60821	21467	1293784	0				19604	1274180									449119	844448	217	1293784		
13	#######	India	1726490	69495	23737	1726490	0				27377	1699113									586081	1140137	272	1726490		
14	#######	India	2295491	78523	25610	2295491	0				36921	2258570									771229	1523939	323	2295491		
15	#######	India	2814803	83664	26219	2814803	0				43604	2771199									939069	1875368	366	2814803		
16	#######	India	3067736	87822	26643	3067736	0				48300	3019436									1022380	2044950	406	3067736		
17	#######	India	3127107	91593	27011	3127107	0				58890	3068217									1061307	2065391	409	3127107		
18	#######	India	3350265	97432	27751	3350265	0				69372	3280893									1152344	2197431	490	3350265		
19	#######	India	3527971	106461	29522	3527971	0				76794	3451177									1218507	2308898	566	3527971		
20	#######	India	3825835	116568	31167	3825835	0				86001	3739834									1324273	2500887	675	3825835		
21	#######	India	4314304	126714	32505	4314304	0				99015	4215289									1504527	2809042	735	4314304		
22	#######	India	4765924	137110	33947	4765924	0				133911	4632013									1710466	3054631	827	4765924		
23	#######	India	5111827	145121	34814	5111827	0				195504	4916323									1896705	3214244	878	5111827		
24	#######	India	5168099	153231	35778	5168099	0				206130	4961969									1928947	3238273	879	5168099		
25	#######	India	5615499	164116	36740	5615499	0				290785	5324714									2211020	3403576	903	5615499		

Figure 2.1: Snapshot of Dataset

Characteristics	**Values**
Name	Covid-19 Statewise Vaccine Ananlysis
Type	
Source	https://www.kaggle.com/datasets/sudalairajkumar/covid19-in-india
Number of Instance	7845
Number of Features	24
Dataaset Format	CSV (Comma Separated Values)

Table 2.1: Properties of the dataset

- How are the features? All categorical? Mix?

  - We have mixed features in our dataset

- Are there any missing values? If yes, are they large or small?

  -Yes, there are large number of missing values in the dataset

- What is the range of data items? How are they distributed?

  - Data items in the dataset are Categorical and Numerical

- Are there any outliers?

  - Yes, there are outliers in the dataset

- Are any of the features skewed?

  - Yes, there are skewed features in the dataset

- Does any of the features require normalization, scaling?

  - No

- Overall what are the characteristics of your dataset?

  - overall, there is more numerical data in our dataset and only 2 categorical features exist, i,e, Updated On and States

SL No	Feature Name	Description
1.	Updated On	Date on which the data is recorded
2.	State	Name of the region
3.	Total Doses Administered	Total number of vaccine doses administrated
4.	Sessions	Total number of Sessions conducted
5.	Sites	Total number of Sites in respective region
6.	First Dose Administered	No of First Doses Administered
7.	Second Dose Administered	No of Second Doses Administered
8.	Male (Doses Administered)	No of Male (Doses Administered)
9.	Female (Doses Administered)	No of Female (Doses Administered)
10.	Transgender (Doses Administered)	No of Transgender (Doses Administered)
11.	Covaxin (Doses Administered)	No of Covaxin (Doses Administered)
12.	CoviSheild (Doses Administered)	No of CoviShield (Doses Administered)
13.	Sputnik V (Doses Administered)	No of Sputnik-V (Doses Administered)
14.	AEFI	No of people having Adverse Effects Following Immunization
15.	18-44 Years (Doses Administered)	No of doses administrated for people of age group of 18-44 years
16.	45-60 Years (Doses Administered)	No of doses administrated for people of age group of 45-60 years
17.	60+ Years (Doses Administered)	No of doses administrated for people aged 60+
18.	18-44 Years(Individuals Vaccinated)	No of people vaccinated of age group of 18-44 years
19.	45-60 Years(Individuals Vaccinated)	No of people vaccinated of age group of 45-60 years
20.	60+ Years(Individuals Vaccinated)	No of people vaccinated of age more than 60 years
21.	Male(Individuals Vaccinated)	No of Male people vaccinated
22.	Female(Individuals Vaccinated)	No of Female people vaccinated
23.	Transgender(Individuals Vaccinated)	No of Transgenders vaccinated
24.	Total Individuals Vaccinated	Total No of people vaccinated

Table 2.2: Feature Set Description

TABLE 2.3: Details of the Features in the Dataset.

Feature Name	**Data Type**	**Distinct Values**	**Missing Values**
Updated On	object	213	0
State	object	37	0
Total Doses Administered	float64	7376	224
Sessions	float64	6464	224
Sites	float64	3044	224
First Dose Administered	float64	3767	224
Second Dose Administered	float64	6275	224
Male (Doses Administered)	float64	7170	384
Female (Doses Administered)	float64	7155	384
Transgender (Doses Administered)	float64	2117	384
Covaxin (Doses Administered)	float64	4353	224
CoviSheild (Doses Administered)	float64	7375	224
Sputnik - V (Doses Administered)	float64	1040	4850
AEFI	float64	15488	2407
18-44 Years (Doses Administered)	float64	1694	6143
45-60 Years (Doses Administered)	float64	1693	6143
60+ Years (Doses Administered)	float64	1692	6143
18-44 Years(Individuals Vaccinated)	float64	3696	4112
45-60 Years(Individuals Vaccinated)	float64	3700	4111
60+ Years(Individuals Vaccinated)	float64	3684	4111
Male(Individuals Vaccinated)	float64	159	7685
Female(Individuals Vaccinated)	float64	159	7685
Transgender(Individuals Vaccinated)	float64	156	7685
Total Individuals Vaccinated	float64	5676	1926

2.4 Statistical Data Analysis

The mean,maximum,minimum,standard deviation and quartile scores of all features are given below:

1. Total Doses Administered:

The mean of Total Doses Administered is 9188170.544023095

The max of Total Doses Administered is 513228400.0

The min of Total Doses Administered is 7.0

The standard deviation of Total Doses Administered is 37461801.16976709

The 25th percentile of Total Doses Administered is 135657.0

The 50th percentile of Total Doses Administered is 818202.0

The 75th percentile of Total Doses Administered is 6625243.0

2. Sessions:

The mean of Sessions is 479235.7989765123

The max of Sessions is 35010311.0

The min of Sessions is 0.0

The standard deviation of Sessions is 1911511.1925942916

The 25th percentile of Sessions is 6004.0

The 50th percentile of Sessions is 45470.0

The 75th percentile of Sessions is 342869.0

3. Sites:

The mean of Sites is 2282.8720640335914

The max of Sites is 73933.0

The min of Sites is 0.0

The standard deviation of Sites is 7275.973729842437

The 25th percentile of Sites is 69.0

The 50th percentile of Sites is 597.0

The 75th percentile of Sites is 1708.0

4. First Dose Administered:

The mean of First Dose Administered is 7414415.300354284

The max of First Dose Administered is 400150406.0

The min of First Dose Administered is 7.0

The standard deviation of First Dose Administered is 29952087.780029744

The 25th percentile of First Dose Administered is 116632.0

The 50th percentile of First Dose Administered is 661459.0

The 75th percentile of First Dose Administered is 5387805.0

5. Second Dose Administered:

The mean of Second Dose Administered is 1773755.2436688098

The max of Second Dose Administered is 113077994.0

The min of Second Dose Administered is 0.0

The standard deviation of Second Dose Administered is 7570382.401890757

The 25th percentile of Second Dose Administered is 12831.0

The 50th percentile of Second Dose Administered is 138818.0

The 75th percentile of Second Dose Administered is 1166434.0

6. Male (Doses Administered):

The mean of Male Doses Administered is 3620156.0107224234

The max of Male Doses Administered is 270163622.0

The min of Male Doses Administered is 0.0

The standard deviation of Male Doses Administered is 17379382.71749272

The 25th percentile of Male Doses Administered is 56555.0

The 50th percentile of Male Doses Administered is 389785.0

The 75th percentile of Male Doses Administered is 2735777.0

7. Female (Doses Administered):

The mean of Female Doses Administered is 3168416.360407452

The max of Female Doses Administered is 239518609.0

The min of Female Doses Administered is 2.0

The standard deviation of Female Doses Administered is 15153103.67224777

The 25th percentile of Female Doses Administered is 52107.0

The 50th percentile of Female Doses Administered is 334238.0

The 75th percentile of Female Doses Administered is 2561513.0

8. Transgender (Doses Administered):

The mean of Transgender (Doses Administered) is 1162.9780190323013

The max of Transgender (Doses Administered) is 98275.0

The min of Transgender (Doses Administered) is 0.0

The standard deviation of Transgender (Doses Administered) is 5931.353995292976

The 25th percentile of Transgender (Doses Administered) is 8.0

The 50th percentile of Transgender (Doses Administered) is 113.0

The 75th percentile of Transgender (Doses Administered) is 800.0

9. Covaxin (Doses Administered):

The mean of Covaxin (Doses Administered) is 1044669.3220049862

The max of Covaxin (Doses Administered) is 62367416.0

The min of Covaxin (Doses Administered) is 0.0

The standard deviation of Covaxin (Doses Administered) is 4452258.870168522

The 25th percentile of Covaxin (Doses Administered) is 0.0

The 50th percentile of Covaxin (Doses Administered) is 11851.0

The 75th percentile of Covaxin (Doses Administered) is 757930.0

10. CoviShield (Doses Administered):

The mean of CoviShield (Doses Administered) is 8126552.9236320695

The max of CoviShield (Doses Administered) is 446825051.0

The min of CoviShield (Doses Administered) is 7.0

The standard deviation of CoviShield (Doses Administered) is 32984142.543515686

The 25th percentile of CoviShield (Doses Administered) is 133134.0

The 50th percentile of CoviShield (Doses Administered) is 756736.0

The 75th percentile of CoviShield (Doses Administered) is 6007817.0

11. Sputnik V (Doses Administered):

The mean of Sputnik V (Doses Administered) is 9655.57061769616

The max of Sputnik V (Doses Administered) is 588039.0

The min of Sputnik V (Doses Administered) is 0.0

The standard deviation of Sputnik V (Doses Administered) is 43882.53617733601

The 25th percentile of Sputnik V (Doses Administered) is 0.0

The 50th percentile of Sputnik V (Doses Administered) is 0.0

The 75th percentile of Sputnik V (Doses Administered) is 2519.0

12. AEFI:

The mean of AEFI is 1139.4025376976829

The max of AEFI is 26542.0

The min of AEFI is 0.0

The standard deviation of AEFI is 3454.608045574231

The 25th percentile of AEFI is 109.25

The 50th percentile of AEFI is 294.0

The 75th percentile of AEFI is 808.0

13. 18-44 Years (Doses Administered):

The mean of 18-44 Years (Doses Administered) is 8773958.21386604

The max of 18-44 Years (Doses Administered) is 224330364.0

The min of 18-44 Years (Doses Administered) is 26624.0

The standard deviation of 18-44 Years (Doses Administered) is 26608287.5871943

The 25th percentile of 18-44 Years (Doses Administered) is 434484.25

The 50th percentile of 18-44 Years (Doses Administered) is 3095970.0

The 75th percentile of 18-44 Years (Doses Administered) is 7366240.75

14. 45-60 Years (Doses Administered):

The mean of 45-60 Years (Doses Administered) is 7442161.202115159

The max of 45-60 Years (Doses Administered) is 166757453.0

The min of 45-60 Years (Doses Administered) is 16815.0

The standard deviation of 45-60 Years (Doses Administered) is 22259992.510206062

The 25th percentile of 45-60 Years (Doses Administered) is 0.0

The 50th percentile of 45-60 Years (Doses Administered) is 232627.5

The 75th percentile of 45-60 Years (Doses Administered) is 6969726.5

15. 60+ Years (Doses Administered):

The mean of 60+ Years (Doses Administered) is 5641605.495299648

The max of 60+ Years (Doses Administered) is 118692689.0

The min of 60+ Years (Doses Administered) is 9994.0

The standard deviation of 60+ Years (Doses Administered) is 16816496.62130506

The 25th percentile of 60+ Years (Doses Administered) is 128560.5

The 50th percentile of 60+ Years (Doses Administered) is 1805696.5

The 75th percentile of 60+ Years (Doses Administered) is 5294762.75

16. 18-44 Years(Individuals Vaccinated):

The mean of 18-44 Years(Individuals Vaccinated) is 1395894.5357621217

The max of 18-44 Years(Individuals Vaccinated) is 92243148.0

The min of 18-44 Years(Individuals Vaccinated) is 1059.0

The standard deviation of 18-44 Years(Individuals Vaccinated)is 5501454.261410572

The 25th percentile of 18-44 Years(Individuals Vaccinated) is 56554.0

The 50th percentile of 18-44 Years(Individuals Vaccinated) is 294727.0

The 75th percentile of 18-44 Years(Individuals Vaccinated) is 910516.0

17. 45-60 Years (Individuals Vaccinated):

The mean of 45-60 Years (Individuals Vaccinated) is 2916514.789769684

The max of 45-60 Years (Individuals Vaccinated) is 90968877.0

The min of 45-60 Years (Individuals Vaccinated)) is 1136.0

The standard deviation of 45-60 Years (Individuals Vaccinated) is 9567607.054644108

The 25th percentile of 45-60 Years (Individuals Vaccinated) is 92482.25

The 50th percentile of 45-60 Years (Individuals Vaccinated) is 833039.5

The 75th percentile of 45-60 Years (Individuals Vaccinated) is 2499280.5

18. 60+ Years(Individuals Vaccinated):

The mean of 60+ Years(Individuals Vaccinated) is 2627444.0565077662

The max of 60+ Years(Individuals Vaccinated) is 67310981.0

The min of 60+ Years(Individuals Vaccinated) is 558.0

The standard deviation of 60+ Years(Individuals Vaccinated) is 8192225.180721852

The 25th percentile of 60+ Years(Individuals Vaccinated) is 56159.75

The 50th percentile of 60+ Years(Individuals Vaccinated) is 788742.5

The 75th percentile of 60+ Years(Individuals Vaccinated) is 2337874.0

19. Male (Individuals Vaccinated):

The mean of Male (Individuals Vaccinated) is 44616867.8625

The max of Male (Individuals Vaccinated) is 134941971.0

The min of Male (Individuals Vaccinated) is 23757.0

The standard deviation of Male (Individuals Vaccinated) is 39507492.96552456

The 25th percentile of Male (Individuals Vaccinated) is 5739350.0

The 50th percentile of Male (Individuals Vaccinated) is 37165905.0

The 75th percentile of Male (Individuals Vaccinated) is 74416634.5

20. Female(Individuals Vaccinated):

The mean of Female(Individuals Vaccinated) is 39510179.6

The max of Female(Individuals Vaccinated) is 115668447.0

The min of Female(Individuals Vaccinated) is 24517.0

The standard deviation of Female(Individuals Vaccinated) is 34176840.95163574

The 25th percentile of Female(Individuals Vaccinated) is 5023407.25

The 50th percentile of Female(Individuals Vaccinated) is 33654024.5

The 75th percentile of Female(Individuals Vaccinated) is 66853682.25

21. Transgender (Individuals Vaccinated):

The mean of Transgender (Individuals Vaccinated) is 12370.54375

The max of Transgender (Individuals Vaccinated) is 46462.0

The min of Transgender (Individuals Vaccinated) is 2.0

The standard deviation of Transgender (Individuals Vaccinated) is 12485.026752752348

The 25th percentile of Transgender (Individuals Vaccinated) is 1278.75

The 50th percentile of Transgender (Individuals Vaccinated) is 8007.5

The 75th percentile of Transgender (Individuals Vaccinated)is 19851.0

22. Total Individuals Vaccinated:

The mean of Total Individuals Vaccinated is 4547841.557357661

The max of Total Individuals Vaccinated is 250656880.0

The min of Total Individuals Vaccinated is 7.0

The standard deviation of Total Individuals Vaccinated is 18341821.27664394

The 25th percentile of Total Individuals Vaccinated is 74275.5

The 50th percentile of Total Individuals Vaccinated is 402288.0

The 75th percentile of Total Individuals Vaccinated is 3501562.0

Chapter 3

Implement Framework

To perform exploratory data analysis on the Covid-19 dataset, we have performed the following implementation framework.

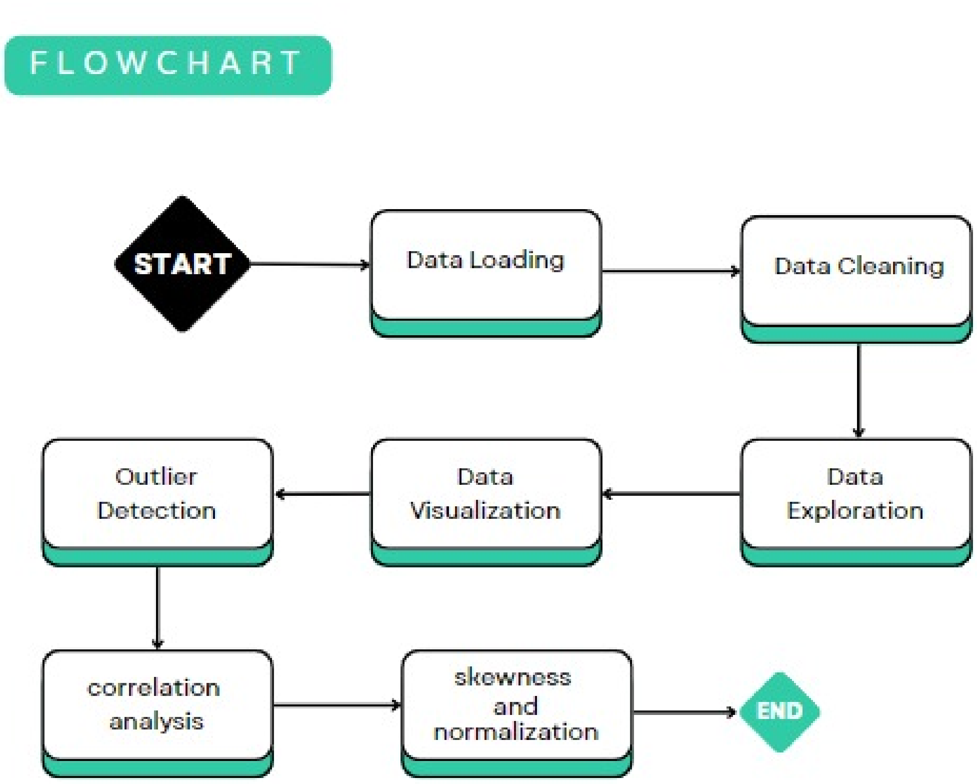


Figure 3.1: Implementation Framework flowchart

1. Data Loading: Load the dataset into your preferred data analysis environment, such as Python with Pandas. Make sure to read the data correctly, including the correct data types for each feature.

2. Data Cleaning: Handle missing values appropriately. Depending on the amount and nature of the missing data, you might decide to impute missing values, remove rows or columns with missing values, or use other techniques to handle the missing data.

3. Data Exploration: Perform summary statistics to gain initial insights into the dataset. This may include calculating mean, median, standard deviation, minimum, and maximum values for numerical features. For categorical features like ”State” you can count unique values and explore their distribution.

4. Outlier Detection: Identify and handle outliers in the data. Outliers can significantly impact statistical analyses and model performance. Consider using box plots or other methods to detect and decide how to handle outliers.

5. Encoding Technique: Converting categorical attributes such as ”Updated on” and ”States” to numerical attributes .

6. Skewness and Normalization: Check the skewness of numerical features. If needed, apply transformations (e.g., log transformation) to make the data more normally distributed. Normalization or scaling might also be necessary for certain analyses or modeling techniques.

7. Correlation Analysis: Investigate the correlations between different numerical features using correlation matrices or heat map. This will help you understand the relationships between variables and identify potential multicollinearity.

8. Data Visualization: Create visualizations to better understand the distribution and relationships between different features. You can use histograms, box plots, scatter plots, line plots, and other types of plots to visualize the data.

9. Ridge Regression: Ridge Regression, is a type of linear regression that includes a regularization term. The regularization term helps to prevent overfitting by penalizing large coefficients in the model.

Chapter 4

Data Pre-processing

Data pre-processing is a crucial step in any data analysis or machine learn- ing project. It involves cleaning and transforming the raw data to make it suitable for analysis and modeling. Here are the data pre-processing steps that can be applied to the Covid-19 dataset.

1. Handling Missing Values: Check for missing values in the dataset and decide how to handle them. Depending on the extent of missing values and the nature of the data, you can either remove rows or columns with missing values or impute them with appropriate methods such as mean, median, or regression imputation.
In our Dataset 3 columns (”Male (Individuals Vaccinated)” ,”Female (Individuals Vaccinated)”,”Transgender (Individuals Vaccinated)”) have 97.96 percent null values, hence we have removed those attributes. We have dropped first 200 rows due to inconsistent data.

2. Dealing with Duplicates: Look for duplicate entries in the dataset and remove them if necessary. Duplicate records can skew analysis results and model performance.
There are no Duplicate values in the Covid-19 Dataset.

3. Encoding Categorical Variables: If there are categorical variables in the dataset, convert them into numerical format using techniques like one- hot encoding or label encoding.

We have applied Label Encoding on ”Updated On” and ”State” columns.

4. Outlier Detection and Handling: Identify and handle outliers in the data. Outliers can significantly impact the performance of certain machine learning models. You can choose to remove outliers or apply transformations to mitigate their effects.

We have used box plot to detect the outliers. The Dataset has extreme outliers therefore we have used IQR method to cap the outliers. Some attributes (”Sputnik V”, ”18-44 Years(Doses Administered)”, ”45-60 Years(Doses Administered)”,”60+ Years(Doses Administered)”,”18-44 Years(Individuals Vaccinated)”,”45-60 Years (Individuals Vaccinated)”,”60+ Years(Individuals Vaccinated)”, have significant data in the outliers hence we have not treated them.

Results of Data Pre-processing After applying the data pre-processing steps, you will have a cleaned and transformed dataset that is ready for analysis and modeling. This pre-processed dataset can now be used for exploratory data analysis (EDA) to gain insights and visualize patterns in the data. Additionally, it can serve as input to various machine learning algorithms for training and testing to build predictive models for forecasting covid-19 vacine status based on the available features.

Chapter 5

Exploratory Data Analysis

5.1 Hypothesis on the Problem Statement

1. How is the distribution of vaccine types represented across all states based on the total doses administered?

2. How is the distribution of Covaxin, CoviShield and Sputnik V vaccine doses administered across all states of India ?

3. Which age group has received the highest number of vaccine doses, and how does vaccination coverage vary across different age groups?

4. Which gender group has received the highest number of vaccine doses across all States of India, and how does vaccination coverage vary across different gender groups?

5. What is the state-wise distribution of different age group individuals vaccinated?

6. Which state has the highest of total individuals vaccinated?

7. Is there a significant difference between the number of first doses and second doses administered within each state?

8. Do states with higher numbers of sessions have lower rates of AEFI cases?

9. What is the relationship between the number of doses administered and the total number of individuals vaccinated across different states?

10. Which state has the highest covid-19 cases?

11. Is there a significant difference between the Total Doses Administered and Total Individuals Vaccinated within each state?

12. What is the loss of vaccine doses administered?

13. What is covid-19 cases threat across different states?

14. What is the average number of doses administered per session?

15. What is the Distribution of Doses Administered by Vaccine Type and Gender ?

16. How are Doses Administered by Gender and Age Group ?

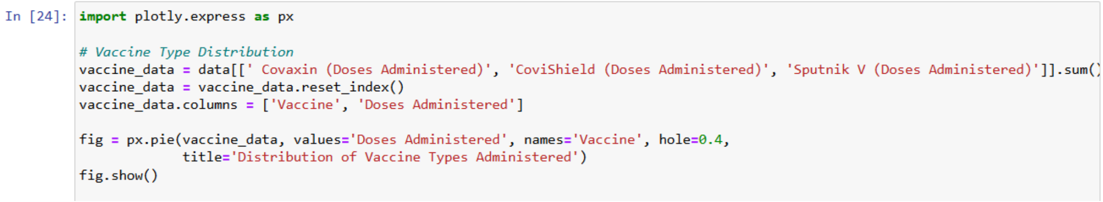
17. How many sessions were conducted in the whole vaccination process ?

18. Comparison of First and Second Dose Administration.

19. Is there any relation between adverse events following immunization (AEFI) and the total number of vaccine doses administered ?

5.2 Analysis

1. How is the distribution of vaccine types represented across all states based on the total doses administered?



```
In [24]: import plotly.express as px

         # Vaccine Type Distribution
         vaccine_data = data[[' Covaxin (Doses Administered)', 'CoviShield (Doses Administered)', 'Sputnik V (Doses Administered)']].sum(
         vaccine_data = vaccine_data.reset_index()
         vaccine_data.columns = ['Vaccine', 'Doses Administered']

         fig = px.pie(vaccine_data, values='Doses Administered', names='Vaccine', hole=0.4,
                      title='Distribution of Vaccine Types Administered')
         fig.show()
```

FIGURE 5.1: Vaccine Type Distribution Code

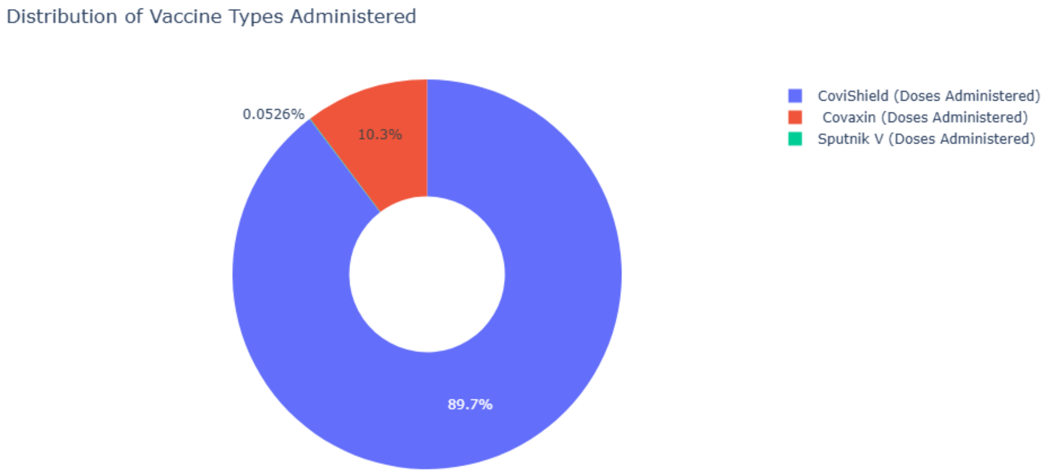


FIGURE 5.2: Vaccine Type Distribution

Inference:

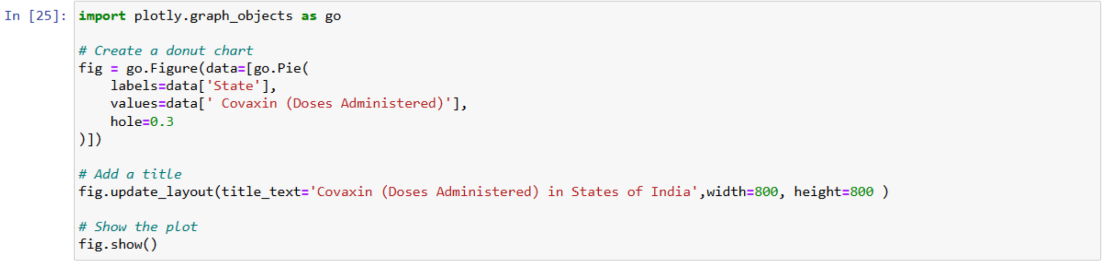
- CoviShield Vaccine has been administered in more number - 89.7
- Covaxin - 9.66
- Sputnik V Vaccines have been administered in very low number as compared to Covaxin and CoviShield vaccine - 0.15

Possible Reasons:

- Covishield has been produced in larger quantities and made available earlier in India compared to Covaxin and Sputnik V.

- The Serum Institute of India, one of the world's largest vaccine manufacturers, has been able to scale up production of Covishield rapidly to meet the demand.

2. How is the distribution of Covaxin, CoviShield and Sputnik V vaccine doses administered across all states of India ?



```
In [25]: import plotly.graph_objects as go

         # Create a donut chart
         fig = go.Figure(data=[go.Pie(
             labels=data['State'],
             values=data[' Covaxin (Doses Administered)'],
             hole=0.3
         )])

         # Add a title
         fig.update_layout(title_text='Covaxin (Doses Administered) in States of India',width=800, height=800 )

         # Show the plot
         fig.show()
```

FIGURE 5.3: State-wise Covaxin Doses Administered Code

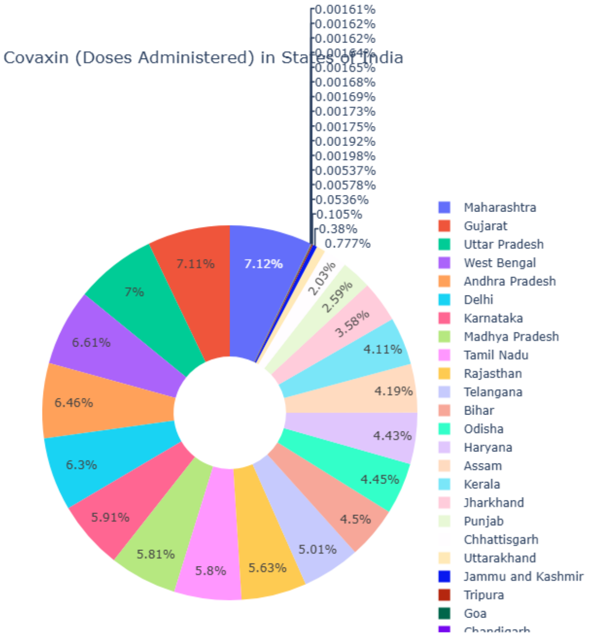


FIGURE 5.4: State-wise Covaxin Doses Administered

Inference:

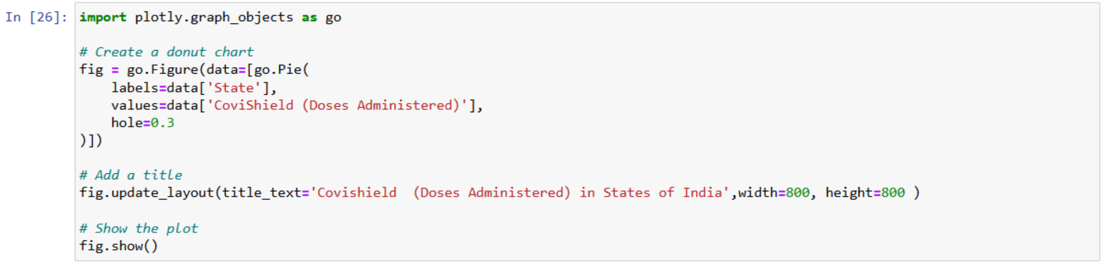
- State with highest number of Covaxin Vaccine Administered is Maharashtra and Gujarat and a few states have no to least vaccinces administered as in Andaman and Nicobar islands, Mizoram, Manipur etc

Possible Reasons:

- Maharashtra is one of the most populous states in India, with densely populated cities like Mumbai and Pune. Higher population density may result in greater demand for vaccines and higher vaccination rates.

- Urban areas often have better healthcare infrastructure, including vaccination centers and distribution networks. Maharashtra's urban centers may have facilitated easier access to vaccines for its residents.

- Maharashtra was one of the earliest states in India to experience significant COVID-19 outbreaks. This may have led to heightened awareness and a proactive approach towards vaccination among its population.



```
In [26]: import plotly.graph_objects as go

         # Create a donut chart
         fig = go.Figure(data=[go.Pie(
             labels=data['State'],
             values=data['CoviShield (Doses Administered)'],
             hole=0.3
         )])

         # Add a title
         fig.update_layout(title_text='Covishield  (Doses Administered) in States of India',width=800, height=800 )

         # Show the plot
         fig.show()
```

Figure 5.5: State-wise CoviShield Doses Administered Code

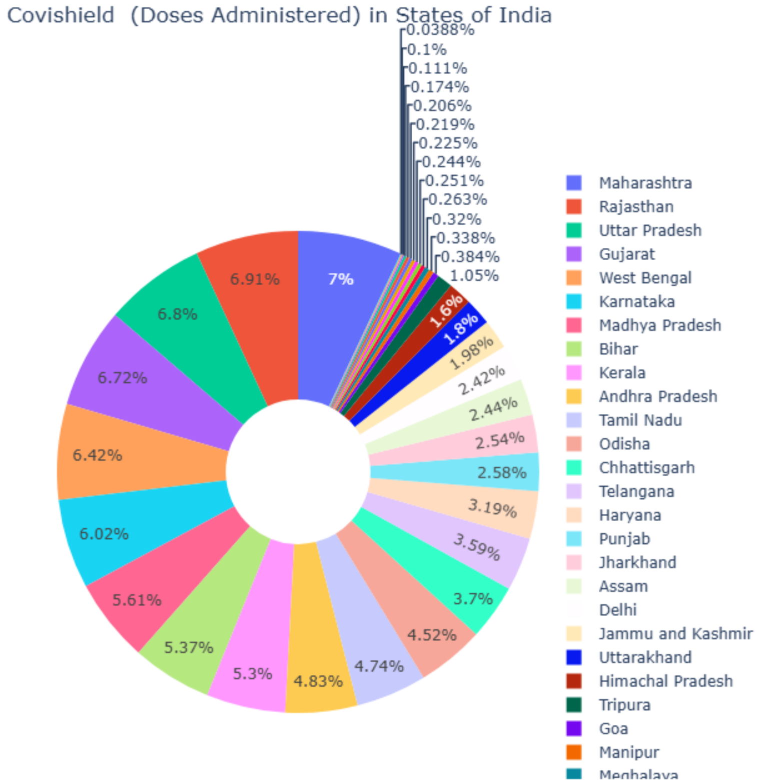
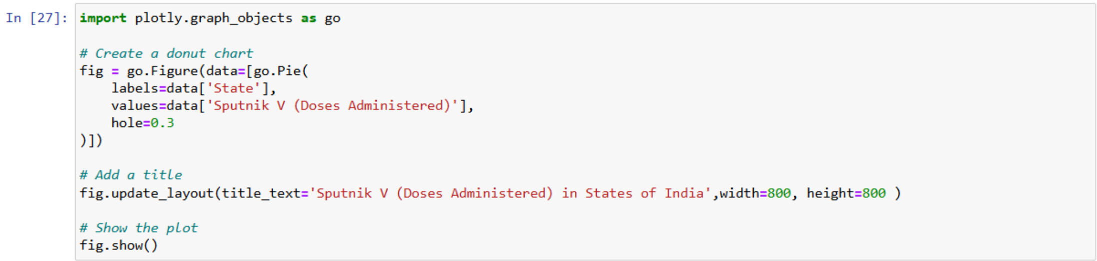


FIGURE 5.6: State-wise CoviShield Doses Administered

Inference:

- State with highest number of CoviShield Vaccine Administered is Maharashtra and lakshadweep has the least number of doses adminitered



```
In [27]: import plotly.graph_objects as go

         # Create a donut chart
         fig = go.Figure(data=[go.Pie(
             labels=data['State'],
             values=data['Sputnik V (Doses Administered)'],
             hole=0.3
         )])

         # Add a title
         fig.update_layout(title_text='Sputnik V (Doses Administered) in States of India',width=800, height=800 )

         # Show the plot
         fig.show()
```

FIGURE 5.7: State-wise Sputnik V Doses Administered Code

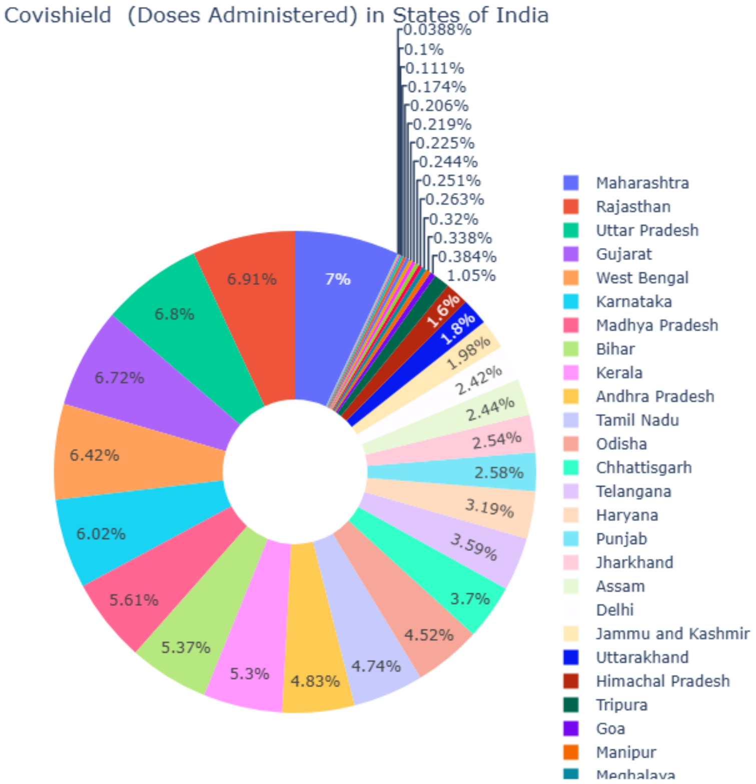


FIGURE 5.8: State-wise Sputnik V Doses Administered

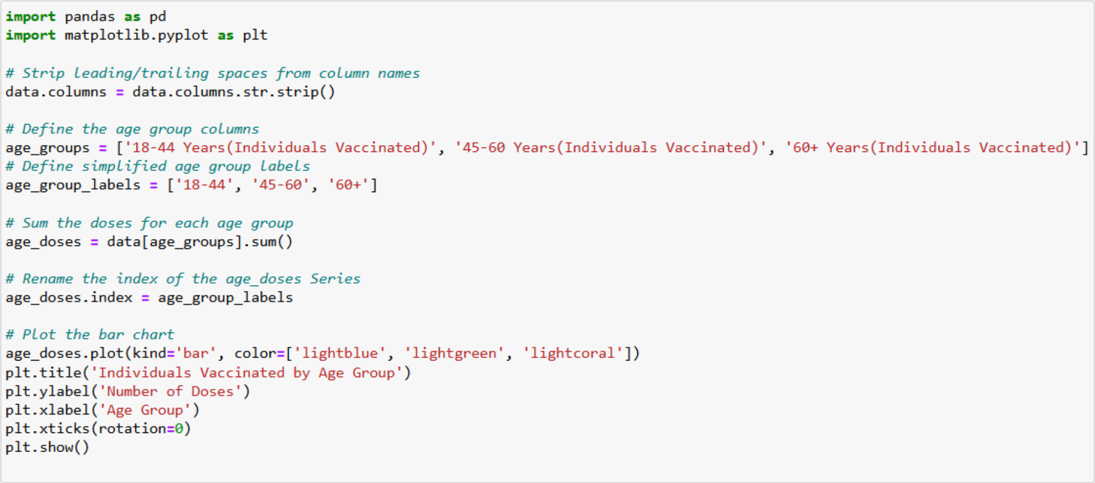
Inference:

- State with highest number of Sputnik-V Vaccine Administered is Telangana and a many states have no vaccinces administered as in Uttrakhand, Tripura, Sikkim, Mizoram etc

Possible Reasons:

- Variations in public confidence in the Sputnik V vaccine and vaccine hesitancy could impact vaccination rates. States with higher levels of vaccine hesitancy may experience slower uptake of the vaccine, resulting in fewer doses administered.

3. Which age group has received the highest number of vaccine doses, and how does vaccination coverage vary across different age groups?



```
import pandas as pd
import matplotlib.pyplot as plt

# Strip leading/trailing spaces from column names
data.columns = data.columns.str.strip()

# Define the age group columns
age_groups = ['18-44 Years(Individuals Vaccinated)', '45-60 Years(Individuals Vaccinated)', '60+ Years(Individuals Vaccinated)']
# Define simplified age group labels
age_group_labels = ['18-44', '45-60', '60+']

# Sum the doses for each age group
age_doses = data[age_groups].sum()

# Rename the index of the age_doses Series
age_doses.index = age_group_labels

# Plot the bar chart
age_doses.plot(kind='bar', color=['lightblue', 'lightgreen', 'lightcoral'])
plt.title('Individuals Vaccinated by Age Group')
plt.ylabel('Number of Doses')
plt.xlabel('Age Group')
plt.xticks(rotation=0)
plt.show()
```

Figure 5.9: Individuals Vaccinated by Age Group Code

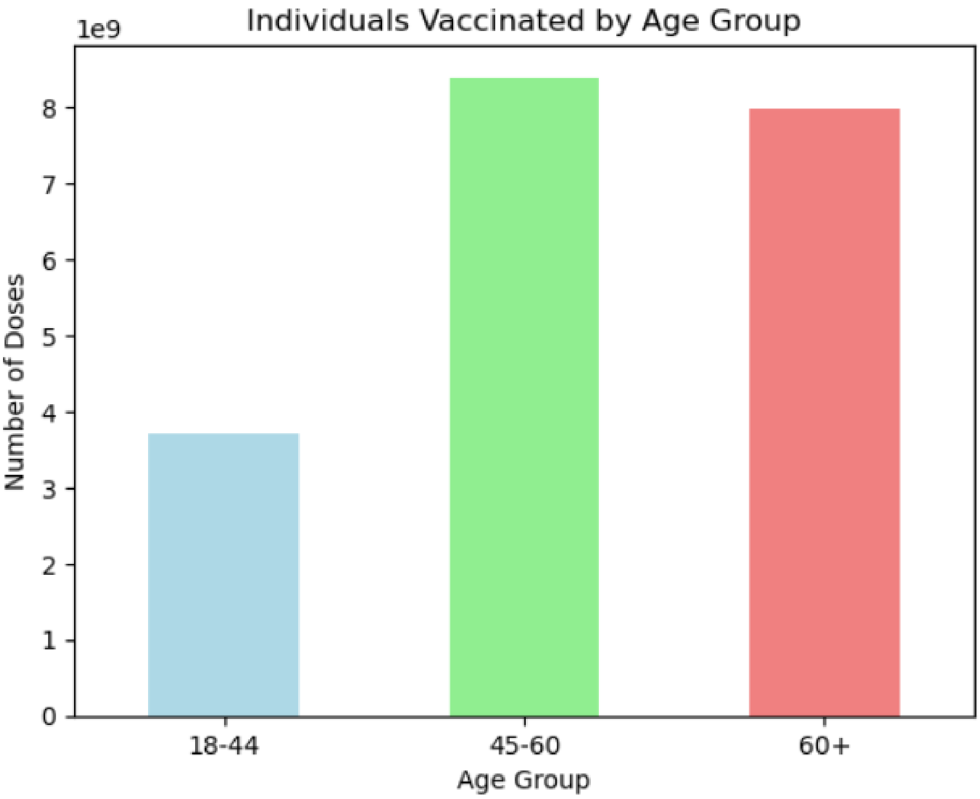


Figure 5.10: Individuals Vaccinated by Age Group

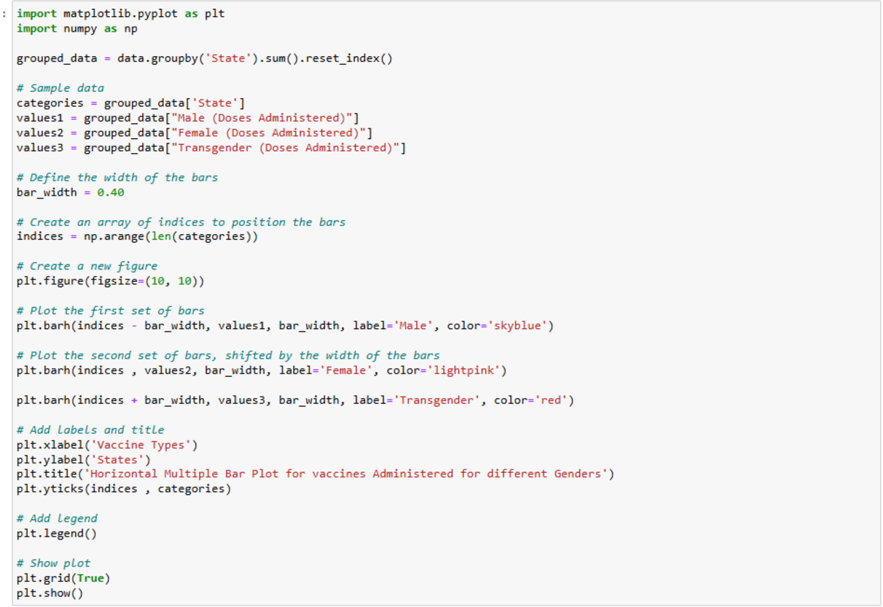
Inference:

- The Individuals of age group 45-60 have been vaccinated in good numbers
- The age group of 18-44 have relatively lesser number of vaccinated individuals

Possible Reasons:

- Initially, the Indian government prioritized vaccination for certain age groups based on vulnerability to severe COVID-19 illness and exposure risk. The 45-60 age group may have been prioritized earlier in the vaccination rollout, leading to higher coverage among this demographic.
- Eligibility criteria for vaccination may have initially favored older age groups due to their higher vulnerability to COVID-19. This could have resulted in more individuals aged.

4. Which gender group has received the highest number of vaccine doses across all States of India, and how does vaccination coverage vary across different gender groups?



```
import matplotlib.pyplot as plt
import numpy as np

grouped_data = data.groupby('State').sum().reset_index()

# Sample data
categories = grouped_data['State']
values1 = grouped_data["Male (Doses Administered)"]
values2 = grouped_data["Female (Doses Administered)"]
values3 = grouped_data["Transgender (Doses Administered)"]

# Define the width of the bars
bar_width = 0.40

# Create an array of indices to position the bars
indices = np.arange(len(categories))

# Create a new figure
plt.figure(figsize=(10, 10))

# Plot the first set of bars
plt.barh(indices - bar_width, values1, bar_width, label='Male', color='skyblue')

# Plot the second set of bars, shifted by the width of the bars
plt.barh(indices , values2, bar_width, label='Female', color='lightpink')

plt.barh(indices + bar_width, values3, bar_width, label='Transgender', color='red')

# Add labels and title
plt.xlabel('Vaccine Types')
plt.ylabel('States')
plt.title('Horizontal Multiple Bar Plot for vaccines Administered for different Genders')
plt.yticks(indices , categories)

# Add legend
plt.legend()

# Show plot
plt.grid(True)
plt.show()
```

Figure 5.11: Vaccine Administered for different genders code

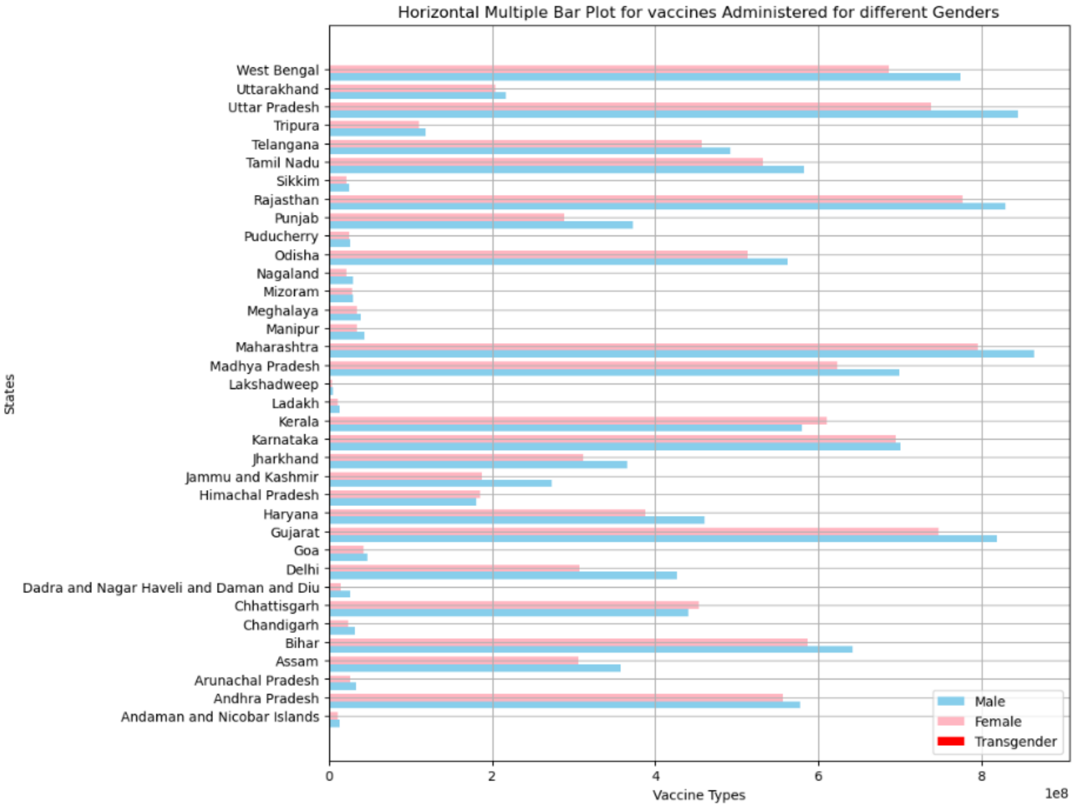
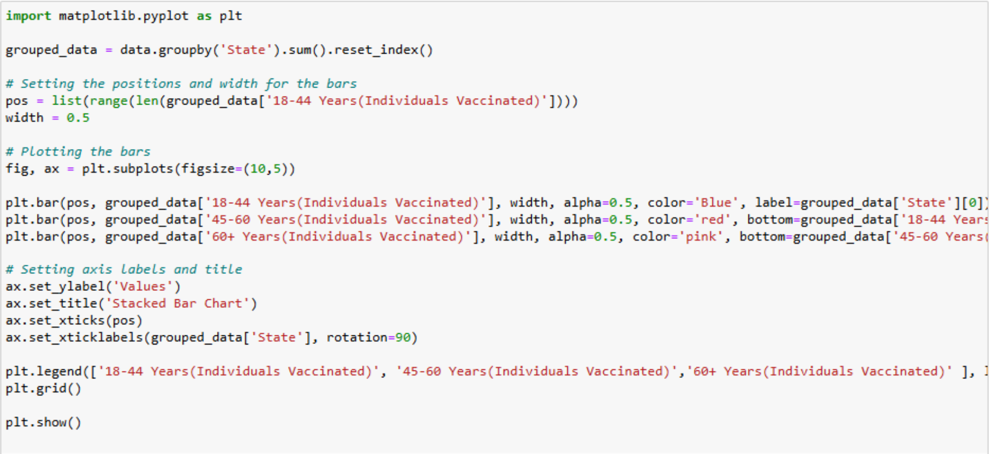


FIGURE 5.12: Vaccine Administered for different genders

Inference:

- Vaccines administered for male and female are relatively equal
- Vaccines administered for Transgender are very less as compared to male and female

5. What is the state-wise distribution of different age group individuals vaccinated?



```
import matplotlib.pyplot as plt

grouped_data = data.groupby('State').sum().reset_index()

# Setting the positions and width for the bars
pos = list(range(len(grouped_data['18-44 Years(Individuals Vaccinated)'])))
width = 0.5

# Plotting the bars
fig, ax = plt.subplots(figsize=(10,5))

plt.bar(pos, grouped_data['18-44 Years(Individuals Vaccinated)'], width, alpha=0.5, color='Blue', label=grouped_data['State'][0]
plt.bar(pos, grouped_data['45-60 Years(Individuals Vaccinated)'], width, alpha=0.5, color='red', bottom=grouped_data['18-44 Year
plt.bar(pos, grouped_data['60+ Years(Individuals Vaccinated)'], width, alpha=0.5, color='pink', bottom=grouped_data['45-60 Years

# Setting axis labels and title
ax.set_ylabel('Values')
ax.set_title('Stacked Bar Chart')
ax.set_xticks(pos)
ax.set_xticklabels(grouped_data['State'], rotation=90)

plt.legend(['18-44 Years(Individuals Vaccinated)', '45-60 Years(Individuals Vaccinated)','60+ Years(Individuals Vaccinated)' ],
plt.grid()

plt.show()
```

FIGURE 5.13: State-wise distribution of different age group individuals vaccinated code

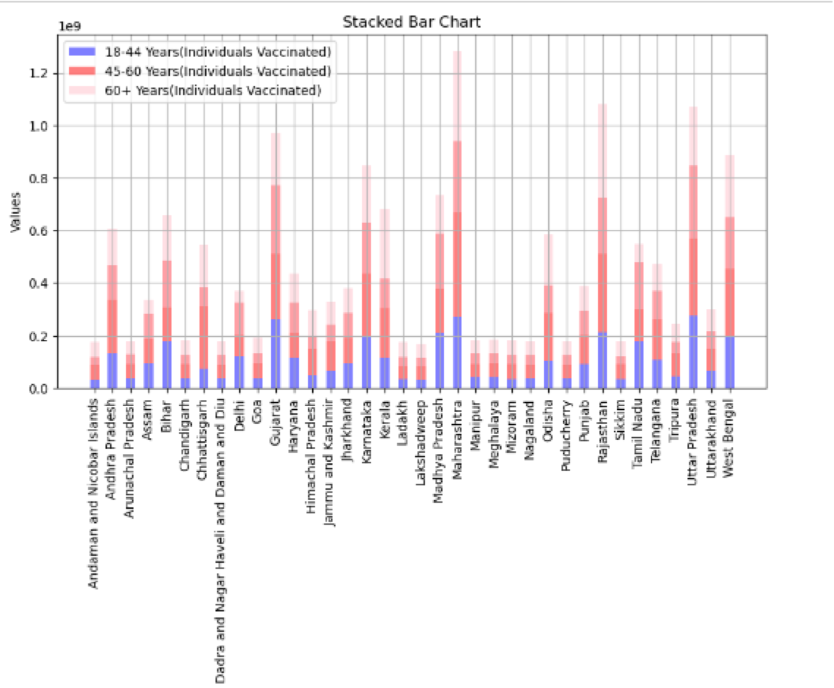


FIGURE 5.14: State-wise distribution of different age group individuals vaccinated

Inference:

- Maharashtra has the highest doses administered.

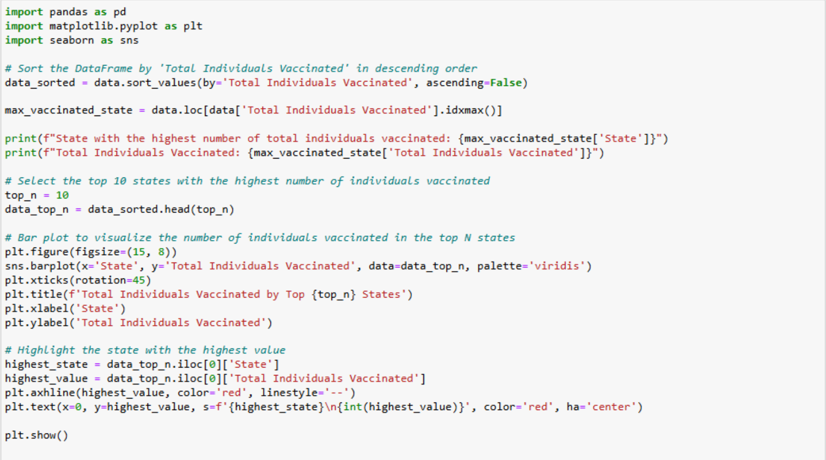
Possible Reasons:

- Maharashtra is one of the most populous states in India, with densely populated cities like Mumbai and Pune. Higher population density may result in greater demand for vaccines and higher vaccination rates.

- Urban areas often have better healthcare infrastructure, including vaccination centers and distribution networks. Maharashtra's urban centers may have facilitated easier access to vaccines for its residents.

- Maharashtra was one of the earliest states in India to experience significant COVID-19 outbreaks. This may have led to heightened awareness and a proactive approach towards vaccination among its population.

6. Which state has the highest of total individuals vaccinated?



```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Sort the DataFrame by 'Total Individuals Vaccinated' in descending order
data_sorted = data.sort_values(by='Total Individuals Vaccinated', ascending=False)

max_vaccinated_state = data.loc[data['Total Individuals Vaccinated'].idxmax()]

print(f"State with the highest number of total individuals vaccinated: {max_vaccinated_state['State']}")
print(f"Total Individuals Vaccinated: {max_vaccinated_state['Total Individuals Vaccinated']}")

# Select the top 10 states with the highest number of individuals vaccinated
top_n = 10
data_top_n = data_sorted.head(top_n)

# Bar plot to visualize the number of individuals vaccinated in the top N states
plt.figure(figsize=(15, 8))
sns.barplot(x='State', y='Total Individuals Vaccinated', data=data_top_n, palette='viridis')
plt.xticks(rotation=45)
plt.title(f'Total Individuals Vaccinated by Top {top_n} States')
plt.xlabel('State')
plt.ylabel('Total Individuals Vaccinated')

# Highlight the state with the highest value
highest_state = data_top_n.iloc[0]['State']
highest_value = data_top_n.iloc[0]['Total Individuals Vaccinated']
plt.axhline(highest_value, color='red', linestyle='--')
plt.text(x=0, y=highest_value, s=f'{highest_state}\n{int(highest_value)}', color='red', ha='center')

plt.show()
```

Figure 5.15: State-wise Total Individuals Vaccinated code

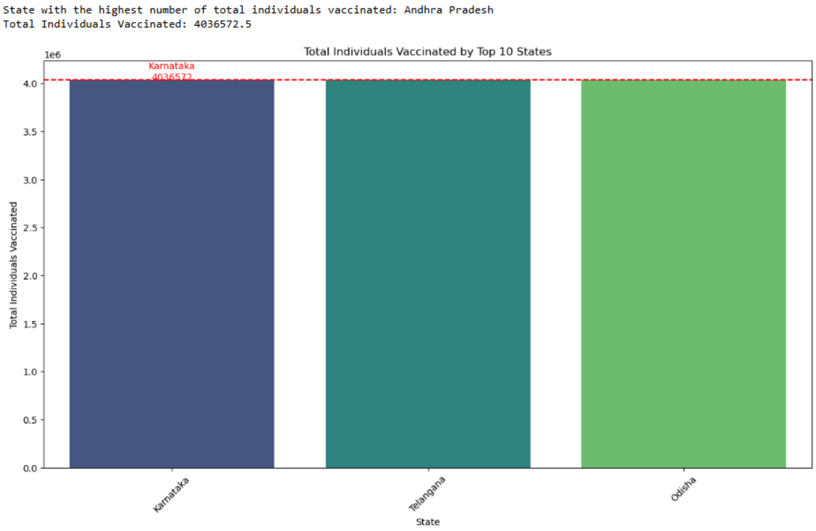
Inference:

- State with the highest number of total individuals vaccinated: Karnatak

- Total Individuals Vaccinated: 4036572.5

Possible Reasons:

- Karnataka is one of the most populous states in India, with a large number of residents eligible for vaccination. The sheer size of the population could contribute to higher vaccination numbers.

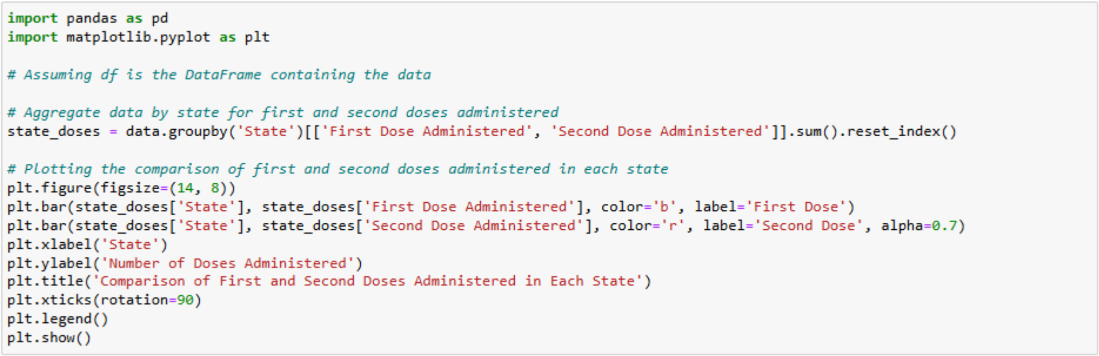


```
State with the highest number of total individuals vaccinated: Andhra Pradesh
Total Individuals Vaccinated: 4036572.5
```

FIGURE 5.16: State-wise Total Individuals Vaccinated

- Karnataka is home to major urban centers such as Bangalore (Bengaluru), which have densely populated areas and a higher concentration of vaccination centers. This may result in more efficient vaccination campaigns and higher coverage rates.

7. Is there a significant difference between the number of first doses and second doses administered within each state?



```python
import pandas as pd
import matplotlib.pyplot as plt

# Assuming df is the DataFrame containing the data

# Aggregate data by state for first and second doses administered
state_doses = data.groupby('State')[['First Dose Administered', 'Second Dose Administered']].sum().reset_index()

# Plotting the comparison of first and second doses administered in each state
plt.figure(figsize=(14, 8))
plt.bar(state_doses['State'], state_doses['First Dose Administered'], color='b', label='First Dose')
plt.bar(state_doses['State'], state_doses['Second Dose Administered'], color='r', label='Second Dose', alpha=0.7)
plt.xlabel('State')
plt.ylabel('Number of Doses Administered')
plt.title('Comparison of First and Second Doses Administered in Each State')
plt.xticks(rotation=90)
plt.legend()
plt.show()
```

FIGURE 5.17: State-wise First and Second Dose administered code

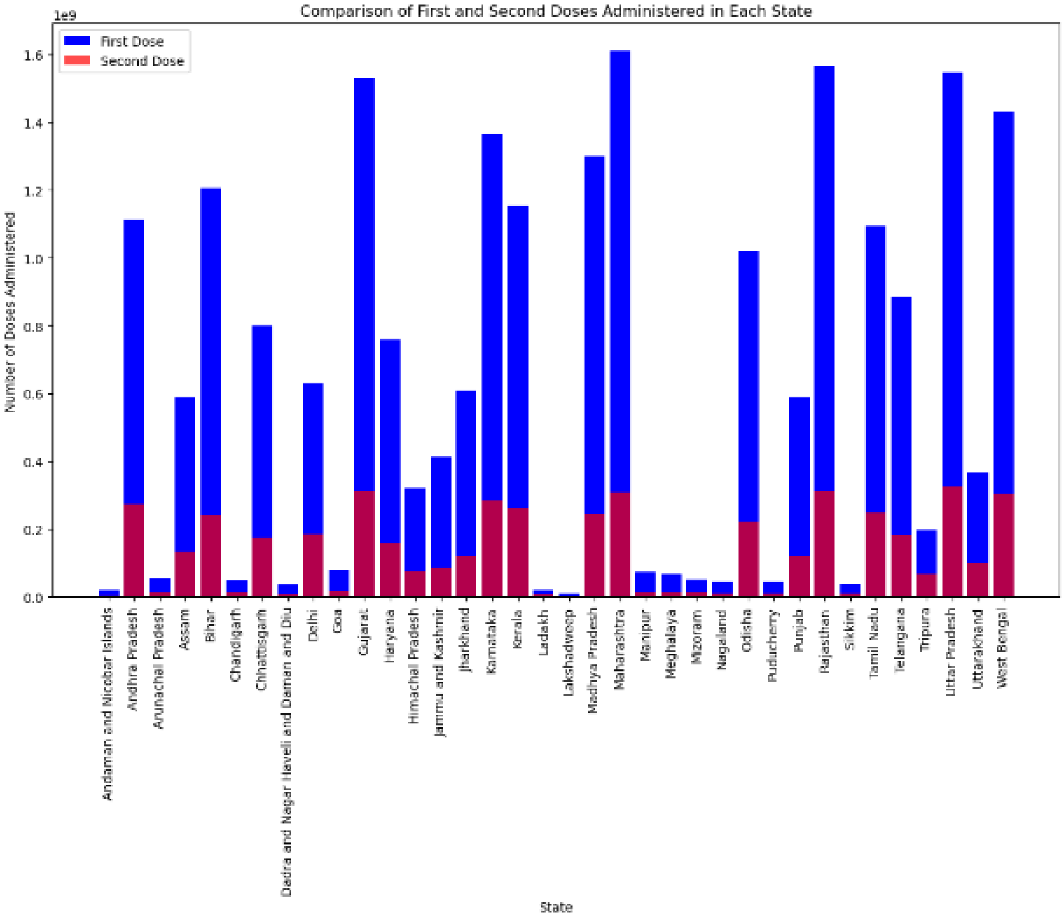


FIGURE 5.18: State-wise First and Second Dose administered

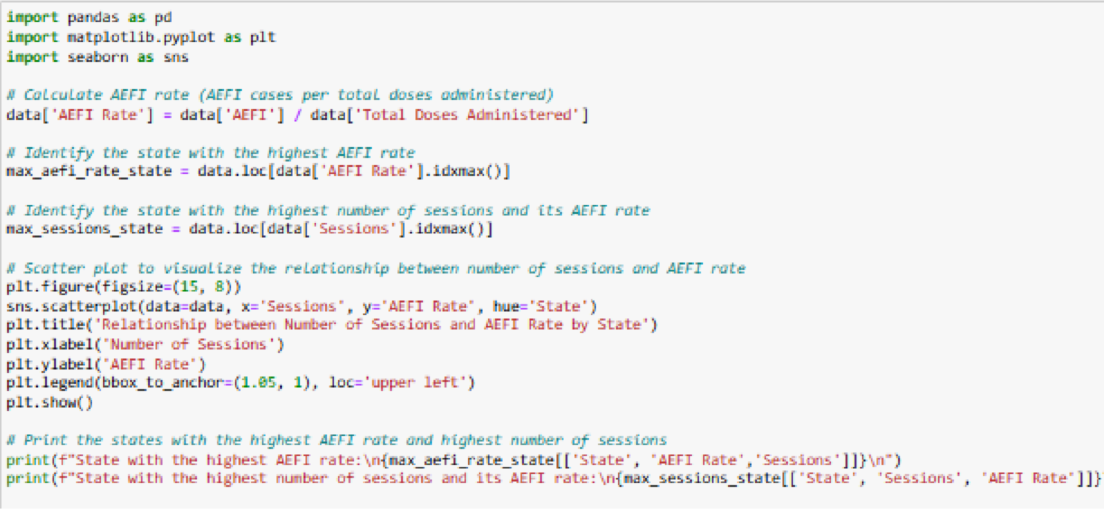
Inference:

- In the every state the people were taken first dose more than the second dose...
- In Maharastra the number of doses taken by people were high in first and also second dose

Possible Reasons:

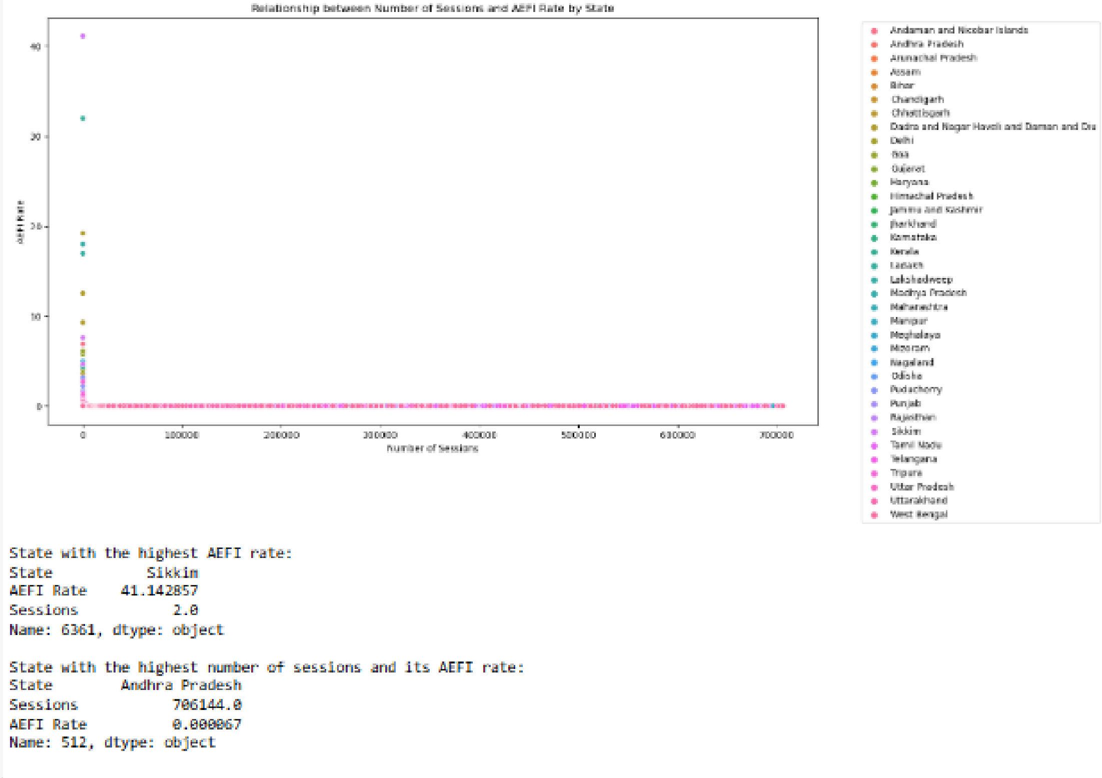
- The gap between the first dose and second dose was extended in many cases.
- And also some individuals might have been hesitant or faced challenges in getting the second dose due to misinformation.

8. Do states with higher numbers of sessions have lower rates of AEFI cases?



```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Calculate AEFI rate (AEFI cases per total doses administered)
data['AEFI Rate'] = data['AEFI'] / data['Total Doses Administered']

# Identify the state with the highest AEFI rate
max_aefi_rate_state = data.loc[data['AEFI Rate'].idxmax()]

# Identify the state with the highest number of sessions and its AEFI rate
max_sessions_state = data.loc[data['Sessions'].idxmax()]

# Scatter plot to visualize the relationship between number of sessions and AEFI rate
plt.figure(figsize=(15, 8))
sns.scatterplot(data=data, x='Sessions', y='AEFI Rate', hue='State')
plt.title('Relationship between Number of Sessions and AEFI Rate by State')
plt.xlabel('Number of Sessions')
plt.ylabel('AEFI Rate')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()

# Print the states with the highest AEFI rate and highest number of sessions
print(f"State with the highest AEFI rate:\n{max_aefi_rate_state[['State', 'AEFI Rate','Sessions']]}\n")
print(f"State with the highest number of sessions and its AEFI rate:\n{max_sessions_state[['State', 'Sessions', 'AEFI Rate']]}
```

FIGURE 5.19: Sessions vs AEFI cases code



```
State with the highest AEFI rate:
State             Sikkim
AEFI Rate      41.142857
Sessions             2.0
Name: 6361, dtype: object

State with the highest number of sessions and its AEFI rate:
State        Andhra Pradesh
Sessions           706144.0
AEFI Rate          0.000067
Name: 512, dtype: object
```

FIGURE 5.20: Sessions vs AEFI cases

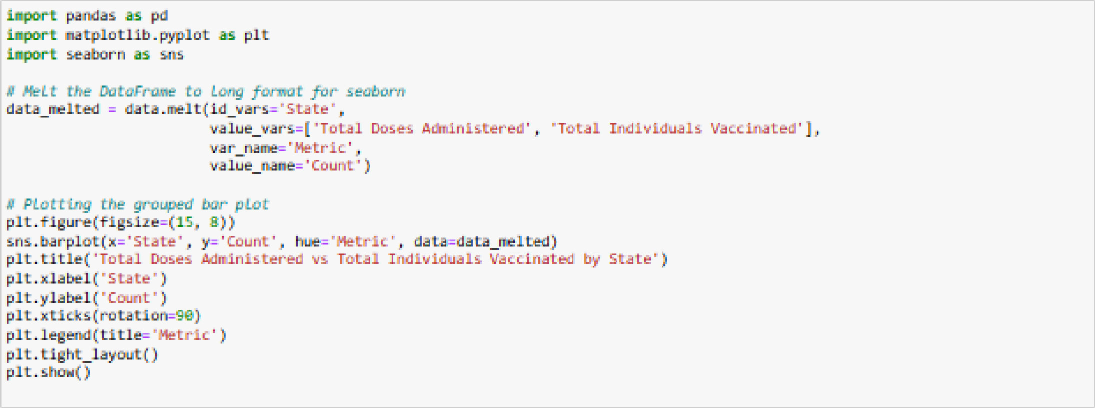
Inference:

- State with the highest AEFI rate: State Sikkim AEFI Rate 41.142857 Sessions 2.0
- State with the highest number of sessions and its AEFI rate: State Andhra Pradesh Sessions 706144.0 AEFI Rate 0.000067

Possible Reasons:

- States with higher numbers of vaccination sessions likely have better training protocols for healthcare workers. Regular training sessions can improve the management of vaccinations and reduce the occurrence of AEFI.
- States with high vaccination rates often have better access to medical resources and emergency response mechanisms to quickly address any adverse events.

9. What is the relationship between the number of doses administered and the total number of individuals vaccinated across different states?



```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Melt the DataFrame to long format for seaborn
data_melted = data.melt(id_vars='State',
                        value_vars=['Total Doses Administered', 'Total Individuals Vaccinated'],
                        var_name='Metric',
                        value_name='Count')

# Plotting the grouped bar plot
plt.figure(figsize=(15, 8))
sns.barplot(x='State', y='Count', hue='Metric', data=data_melted)
plt.title('Total Doses Administered vs Total Individuals Vaccinated by State')
plt.xlabel('State')
plt.ylabel('Count')
plt.xticks(rotation=90)
plt.legend(title='Metric')
plt.tight_layout()
plt.show()
```

FIGURE 5.21: Total Doses Administered vs total number of individuals Vaccinated Code

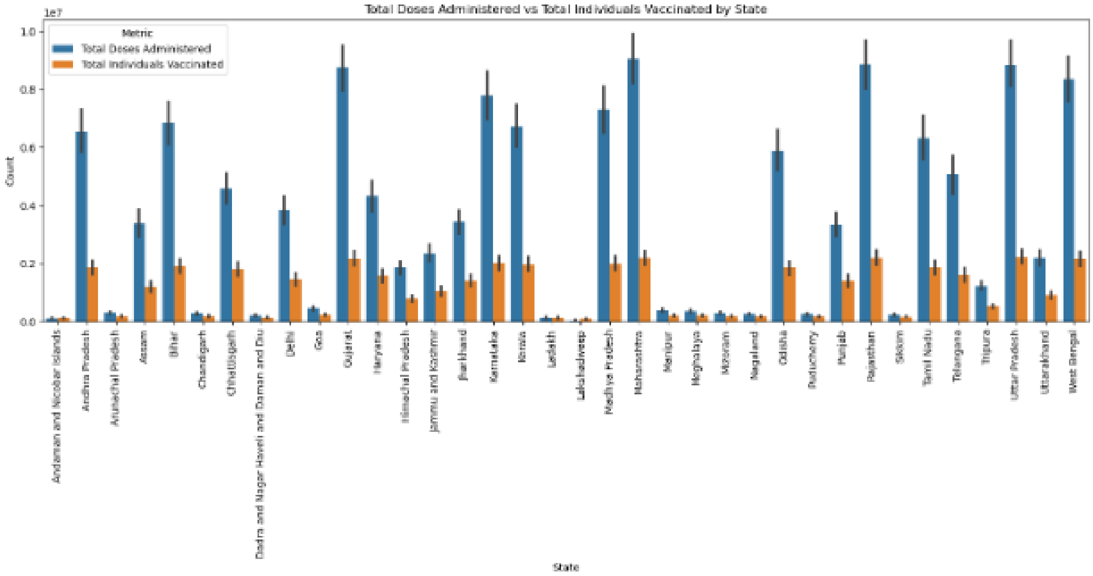


Figure 5.22: Total Doses Administered vs total number of individuals Vaccinated

Inference:

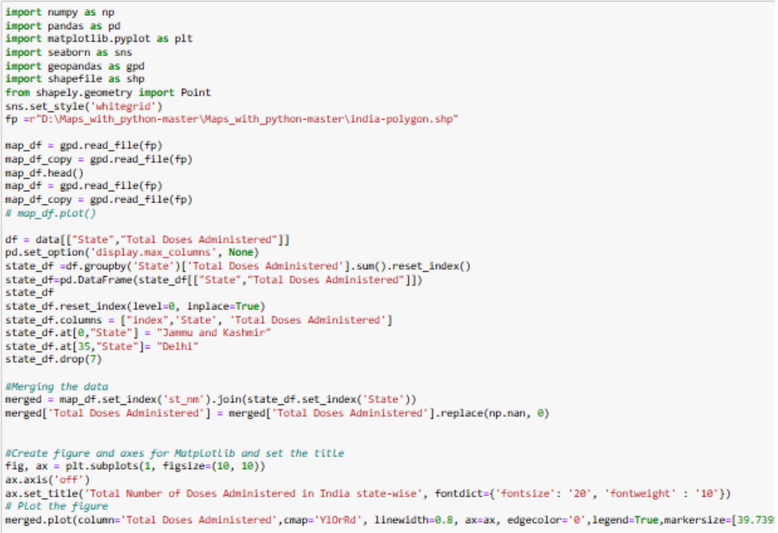
- By comparing total doses administered and total individuals vaccinated we concluded that Maharastra has the highest total doses administered and highest total individuals vaccinated

Possible Reasons:

- Urban areas in Maharashtra have a well-developed healthcare infrastructure, which supports extensive vaccination campaigns.

- Effective public awareness campaigns have been conducted to educate the population about the importance of vaccination, helping to reduce vaccine hesitancy and increase uptake.

10. Which state has the highest covid-19 cases?



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import geopandas as gpd
import shapefile as shp
from shapely.geometry import Point
sns.set_style('whitegrid')
fp =r"D:\Maps_with_python-master\Maps_with_python-master\india-polygon.shp"

map_df = gpd.read_file(fp)
map_df_copy = gpd.read_file(fp)
map_df.head()
map_df = gpd.read_file(fp)
map_df_copy = gpd.read_file(fp)
# map_df.plot()

df = data[["State","Total Doses Administered"]]
pd.set_option('display.max_columns', None)
state_df =df.groupby('State')['Total Doses Administered'].sum().reset_index()
state_df=pd.DataFrame(state_df[["State","Total Doses Administered"]])
state_df
state_df.reset_index(level=0, inplace=True)
state_df.columns = ["index",'State', 'Total Doses Administered']
state_df.at[0,"State"] = "Jammu and Kashmir"
state_df.at[35,"State"]= "Delhi"
state_df.drop(7)

#Merging the data
merged = map_df.set_index('st_nm').join(state_df.set_index('State'))
merged['Total Doses Administered'] = merged['Total Doses Administered'].replace(np.nan, 0)


#Create figure and axes for Matplotlib and set the title
fig, ax = plt.subplots(1, figsize=(10, 10))
ax.axis('off')
ax.set_title('Total Number of Doses Administered in India state-wise', fontdict={'fontsize': '20', 'fontweight' : '10'})
# Plot the figure
merged.plot(column='Total Doses Administered',cmap='YlOrRd', linewidth=0.8, ax=ax, edgecolor='0',legend=True,markersize=[39.739
```

FIGURE 5.23: Highest Covid-19 case code for geo map

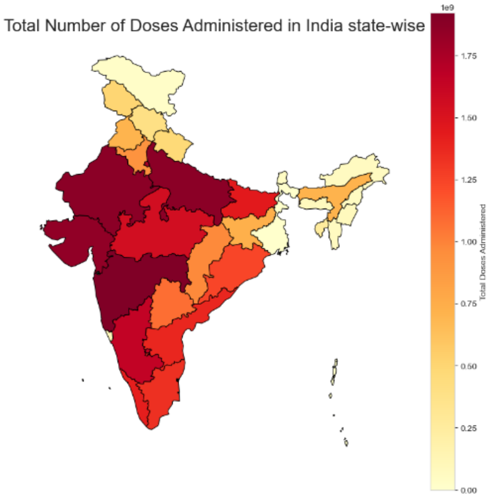


FIGURE 5.24: Highest Covid-19 case geo map

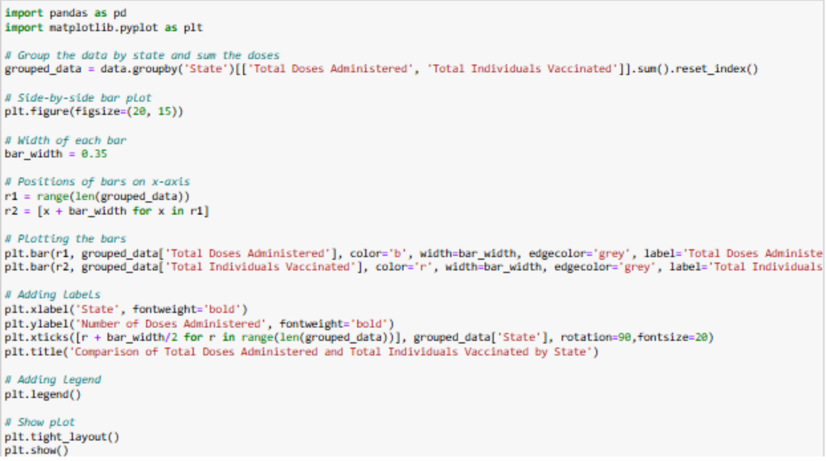
Inference:

- State with most COVID-19 Cases prediction by (State v/s Total Doses Administered)’ is Maharashtra.

Possible Reasons:

- Due to high pouplation density and less care in precautions.
- Effective public awareness campaigns have been launched to educate the population about the benefits of vaccination, thereby increasing acceptance and participation rates.

11. Is there a significant difference between the Total Doses Administered and Total Individuals Vaccinated within each state?



```
import pandas as pd
import matplotlib.pyplot as plt

# Group the data by state and sum the doses
grouped_data = data.groupby('State')[['Total Doses Administered', 'Total Individuals Vaccinated']].sum().reset_index()

# Side-by-side bar plot
plt.figure(figsize=(20, 15))

# Width of each bar
bar_width = 0.35

# Positions of bars on x-axis
r1 = range(len(grouped_data))
r2 = [x + bar_width for x in r1]

# Plotting the bars
plt.bar(r1, grouped_data['Total Doses Administered'], color='b', width=bar_width, edgecolor='grey', label='Total Doses Administe
plt.bar(r2, grouped_data['Total Individuals Vaccinated'], color='r', width=bar_width, edgecolor='grey', label='Total Individuals

# Adding labels
plt.xlabel('State', fontweight='bold')
plt.ylabel('Number of Doses Administered', fontweight='bold')
plt.xticks([r + bar_width/2 for r in range(len(grouped_data))], grouped_data['State'], rotation=90,fontsize=20)
plt.title('Comparison of Total Doses Administered and Total Individuals Vaccinated by State')

# Adding legend
plt.legend()

# Show plot
plt.tight_layout()
plt.show()
```

FIGURE 5.25: Total Doses Administered and Total Individuals Vaccinated within each state code

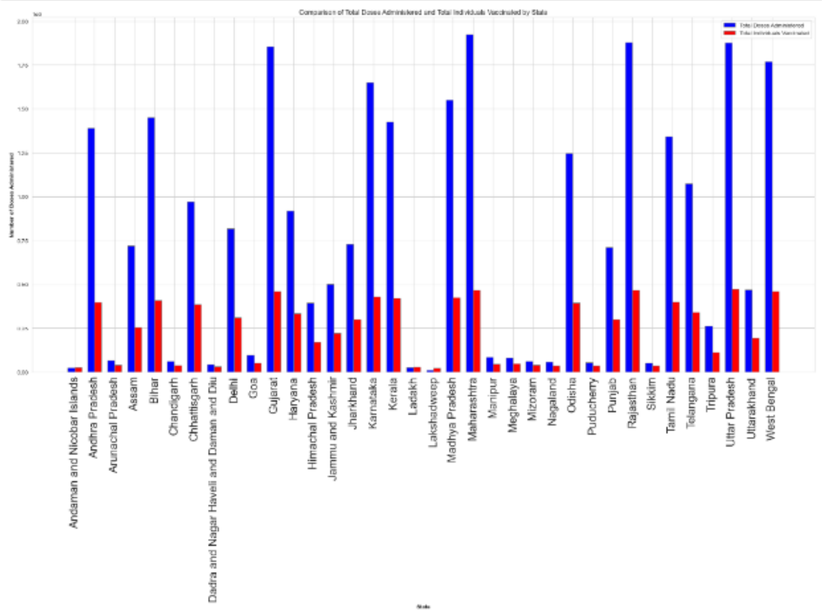
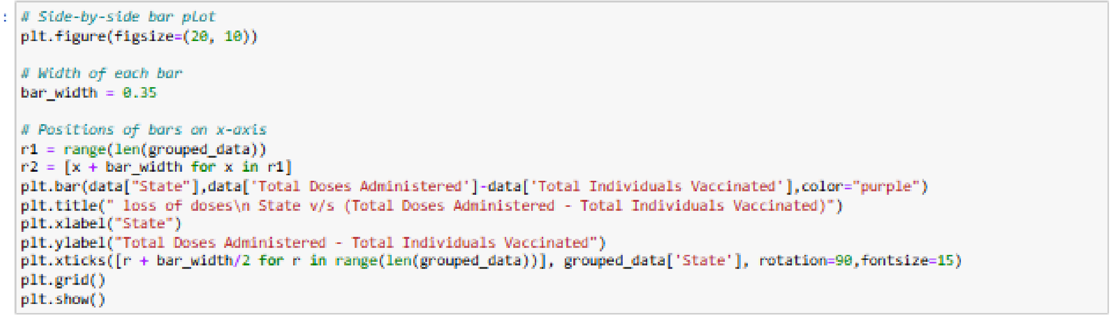


FIGURE 5.26: Total Doses Administered and Total Individuals Vaccinated within each state

12. What is the loss of vaccine doses administered?



```
# Side-by-side bar plot
plt.figure(figsize=(20, 10))

# Width of each bar
bar_width = 0.35

# Positions of bars on x-axis
r1 = range(len(grouped_data))
r2 = [x + bar_width for x in r1]
plt.bar(data["State"],data['Total Doses Administered']-data['Total Individuals Vaccinated'],color="purple")
plt.title(" loss of doses\n State v/s (Total Doses Administered - Total Individuals Vaccinated)")
plt.xlabel("State")
plt.ylabel("Total Doses Administered - Total Individuals Vaccinated")
plt.xticks([r + bar_width/2 for r in range(len(grouped_data))], grouped_data['State'], rotation=90,fontsize=15)
plt.grid()
plt.show()
```

FIGURE 5.27: Loss of vaccine doses administered code

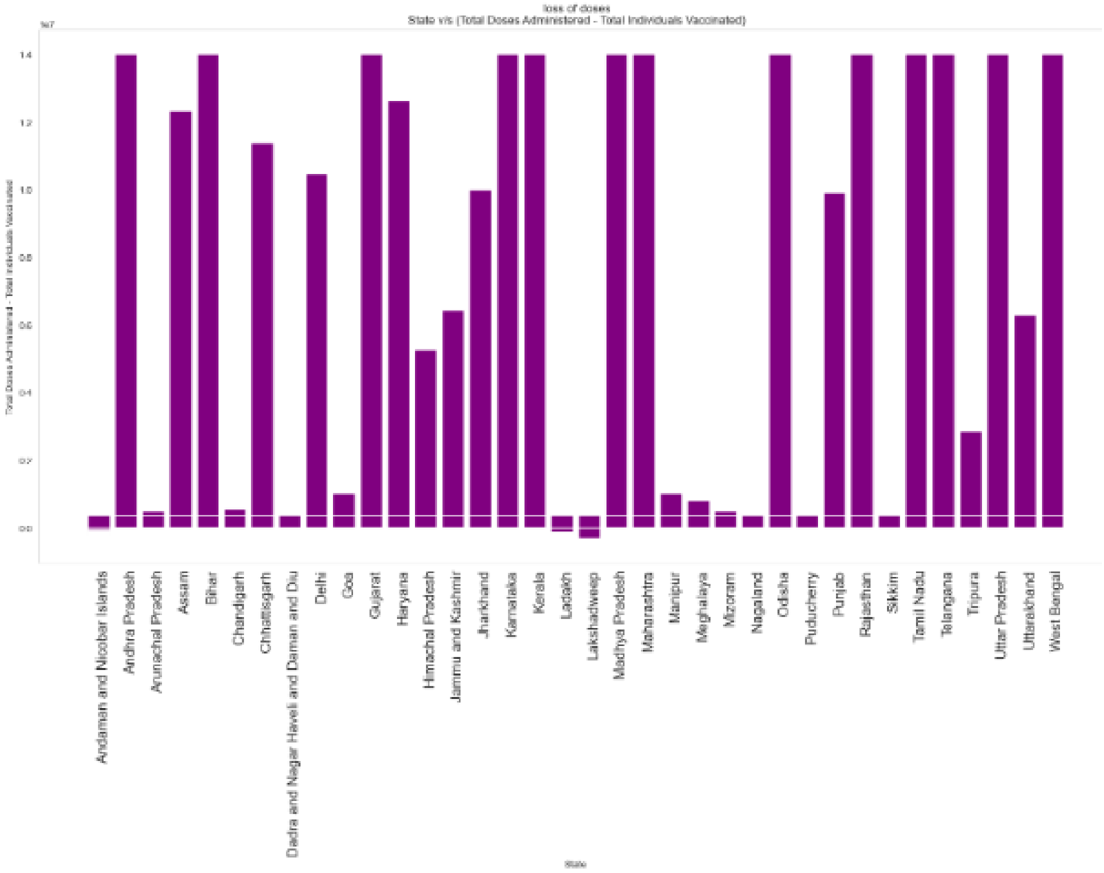
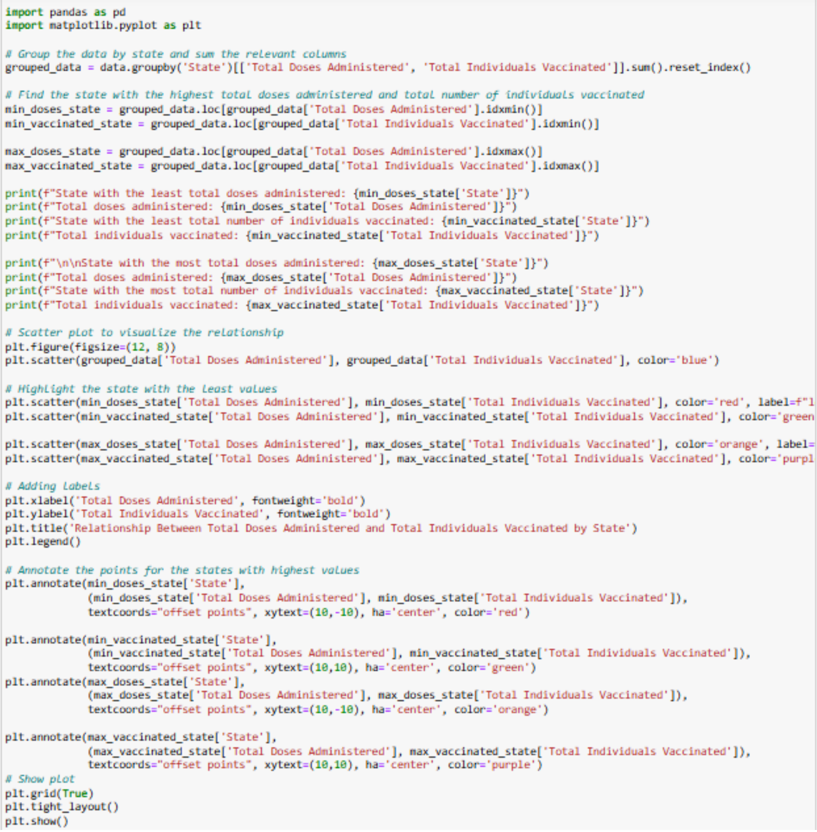


FIGURE 5.28: Loss of vaccine doses administered

Inference for 11 and 12:

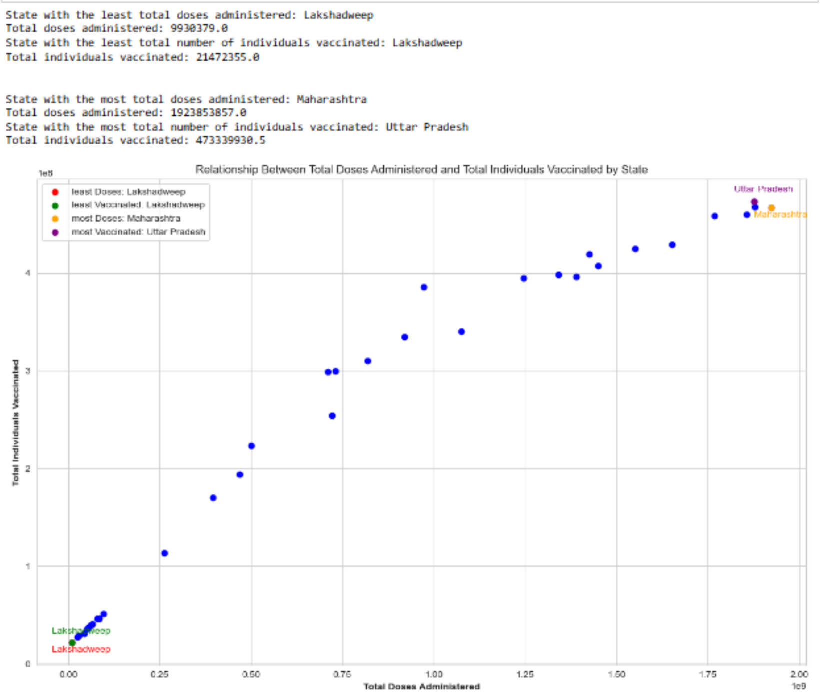
- Some of the states has a loss of covid doses in negative which idicates that there was sortage of doses in that region (goa,chandigarh,andaman and nicobar,manipur,mizoram)etc. or thae people have migrated to different nearest state to get vaccinated due to lock down rules or due to sharing of doses between two nearest states there might be some other reasons for loss of doses.

13. What is covid-19 cases threat across different states?



```
import pandas as pd
import matplotlib.pyplot as plt

# Group the data by state and sum the relevant columns
grouped_data = data.groupby('State')[['Total Doses Administered', 'Total Individuals Vaccinated']].sum().reset_index()

# Find the state with the highest total doses administered and total number of individuals vaccinated
min_doses_state = grouped_data.loc[grouped_data['Total Doses Administered'].idxmin()]
min_vaccinated_state = grouped_data.loc[grouped_data['Total Individuals Vaccinated'].idxmin()]

max_doses_state = grouped_data.loc[grouped_data['Total Doses Administered'].idxmax()]
max_vaccinated_state = grouped_data.loc[grouped_data['Total Individuals Vaccinated'].idxmax()]

print(f"State with the least total doses administered: {min_doses_state['State']}")
print(f"Total doses administered: {min_doses_state['Total Doses Administered']}")
print(f"State with the least total number of individuals vaccinated: {min_vaccinated_state['State']}")
print(f"Total individuals vaccinated: {min_vaccinated_state['Total Individuals Vaccinated']}")

print(f"\n\nState with the most total doses administered: {max_doses_state['State']}")
print(f"Total doses administered: {max_doses_state['Total Doses Administered']}")
print(f"State with the most total number of individuals vaccinated: {max_vaccinated_state['State']}")
print(f"Total individuals vaccinated: {max_vaccinated_state['Total Individuals Vaccinated']}")

# Scatter plot to visualize the relationship
plt.figure(figsize=(12, 8))
plt.scatter(grouped_data['Total Doses Administered'], grouped_data['Total Individuals Vaccinated'], color='blue')

# Highlight the state with the least values
plt.scatter(min_doses_state['Total Doses Administered'], min_doses_state['Total Individuals Vaccinated'], color='red', label=f"l
plt.scatter(min_vaccinated_state['Total Doses Administered'], min_vaccinated_state['Total Individuals Vaccinated'], color='green

plt.scatter(max_doses_state['Total Doses Administered'], max_doses_state['Total Individuals Vaccinated'], color='orange', label=
plt.scatter(max_vaccinated_state['Total Doses Administered'], max_vaccinated_state['Total Individuals Vaccinated'], color='purpl

# Adding labels
plt.xlabel('Total Doses Administered', fontweight='bold')
plt.ylabel('Total Individuals Vaccinated', fontweight='bold')
plt.title('Relationship Between Total Doses Administered and Total Individuals Vaccinated by State')
plt.legend()

# Annotate the points for the states with highest values
plt.annotate(min_doses_state['State'],
             (min_doses_state['Total Doses Administered'], min_doses_state['Total Individuals Vaccinated']),
             textcoords="offset points", xytext=(10,-10), ha='center', color='red')

plt.annotate(min_vaccinated_state['State'],
             (min_vaccinated_state['Total Doses Administered'], min_vaccinated_state['Total Individuals Vaccinated']),
             textcoords="offset points", xytext=(10,10), ha='center', color='green')
plt.annotate(max_doses_state['State'],
             (max_doses_state['Total Doses Administered'], max_doses_state['Total Individuals Vaccinated']),
             textcoords="offset points", xytext=(10,-10), ha='center', color='orange')

plt.annotate(max_vaccinated_state['State'],
             (max_vaccinated_state['Total Doses Administered'], max_vaccinated_state['Total Individuals Vaccinated']),
             textcoords="offset points", xytext=(10,10), ha='center', color='purple')
# Show plot
plt.grid(True)
plt.tight_layout()
plt.show()
```

Figure 5.29: Covid-19 cases threat code



```
State with the least total doses administered: Lakshadweep
Total doses administered: 9930379.0
State with the least total number of individuals vaccinated: Lakshadweep
Total individuals vaccinated: 21472355.0


State with the most total doses administered: Maharashtra
Total doses administered: 1923853857.0
State with the most total number of individuals vaccinated: Uttar Pradesh
Total individuals vaccinated: 473339930.5
```

Relationship Between Total Doses Administered and Total Individuals Vaccinated by State
1e8
least Doses: Lakshadweep
least Vaccinated: Lakshadweep
most Doses: Maharashtra
most Vaccinated: Uttar Pradesh
Uttar Pradesh
Maharashtra
Lakshadweep
Lakshadweep
Total Individuals Vaccinated
0
1
2
3
4
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e9
Total Doses Administered

FIGURE 5.30: Covid-19 cases threat

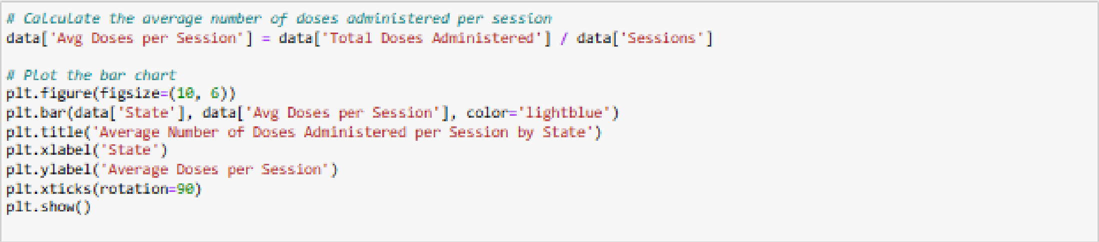
Inference:

- State with the least total doses administered: Lakshadweep Total doses administered: 9930379.0 State with the least total number of individuals vaccinated: Lakshadweep Total individuals vaccinated: 21472355.0
- State with the most total doses administered: Maharashtra Total doses administered: 1923853857.0 State with the most total number of individuals vaccinated: Uttar Pradesh Total individuals vaccinated: 473339930.5

Possible Reasons:

- State with the most total number of individuals vaccinated is uttar pradesh as uttra pradesh is heighest populated state. The proportion of Communicable, Maternal, Neonatal, and Nutritional Diseases [CMNND] contribute to 40.5 percent of total disease burden, in which Diarrheal Diseases, Lower Respiratory Infection, and Tuberculosis remain the major causes of death in Uttar Pradesh(Annexure 2, Figure 6). As per QPR reports, the annualized total case notification rate for TB is 193total number of least individuals vaccinated is lakshadweep as it is isolate state island and as the tourim was suuspended the caues was reduced.

14. What is the average number of doses administered per session?



```
# Calculate the average number of doses administered per session
data['Avg Doses per Session'] = data['Total Doses Administered'] / data['Sessions']

# Plot the bar chart
plt.figure(figsize=(10, 6))
plt.bar(data['State'], data['Avg Doses per Session'], color='lightblue')
plt.title('Average Number of Doses Administered per Session by State')
plt.xlabel('State')
plt.ylabel('Average Doses per Session')
plt.xticks(rotation=90)
plt.show()
```

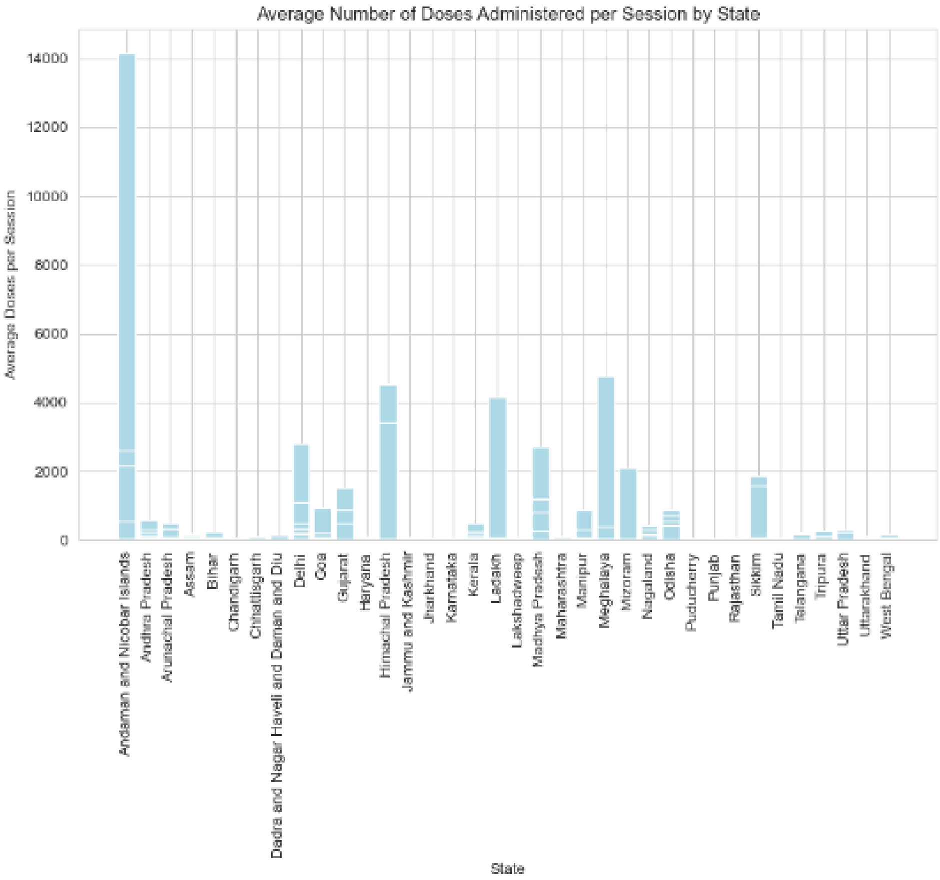


FIGURE 5.31: Number of doses administered per session

Inference:

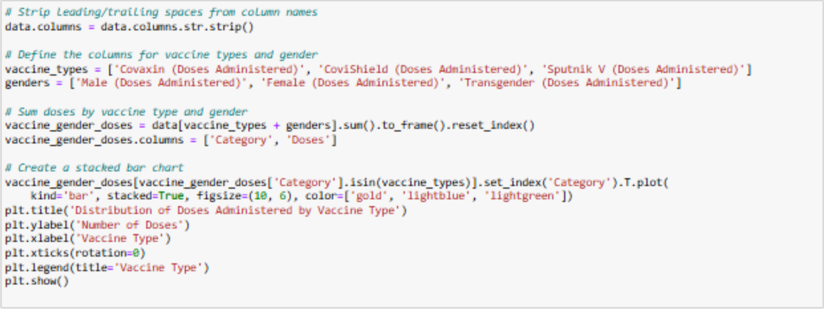
- There is a large variation in the average number of doses administered per session across different states. Some states have a much higher average than others.

- Andra pradesh appears to have a relatively high average number of doses administered per session, exceeding 10,000 doses.
- The majority of states seem to have an average number of doses administered per session between 2,000 and 6,000.
- A few states appear to have a low average number of doses administered per session, falling below 2,000 doses.

Possible Reasons:

- Urban areas with large populations and better accessibility can conduct mass vaccination drives more efficiently, leading to higher averages per session.

15. What is the Distribution of Doses Administered by Vaccine Type and Gender ?



```python
# Strip leading/trailing spaces from column names
data.columns = data.columns.str.strip()

# Define the columns for vaccine types and gender
vaccine_types = ['Covaxin (Doses Administered)', 'CoviShield (Doses Administered)', 'Sputnik V (Doses Administered)']
genders = ['Male (Doses Administered)', 'Female (Doses Administered)', 'Transgender (Doses Administered)']

# Sum doses by vaccine type and gender
vaccine_gender_doses = data[vaccine_types + genders].sum().to_frame().reset_index()
vaccine_gender_doses.columns = ['Category', 'Doses']

# Create a stacked bar chart
vaccine_gender_doses[vaccine_gender_doses['Category'].isin(vaccine_types)].set_index('Category').T.plot(
    kind='bar', stacked=True, figsize=(10, 6), color=['gold', 'lightblue', 'lightgreen'])
plt.title('Distribution of Doses Administered by Vaccine Type')
plt.ylabel('Number of Doses')
plt.xlabel('Vaccine Type')
plt.xticks(rotation=0)
plt.legend(title='Vaccine Type')
plt.show()
```

Figure 5.32: Distribution based on gender and vaccine type code

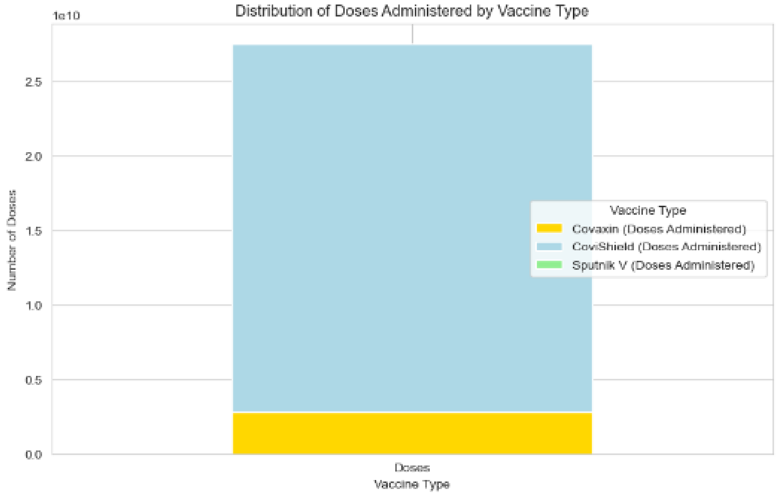


FIGURE 5.33: Distribution based on gender and vaccine type

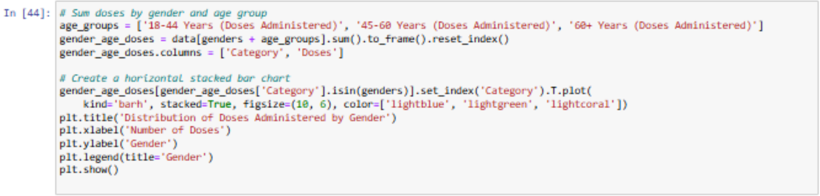
Inference:

- Covishield: This vaccine type has been administered the most widely, with a significant lead over the other two types.
- Covaxin: This vaccine type has been administered a moderate number of doses, but considerably less than Covishield.
- Sputnik V: This vaccine type has seen the least amount of administration compared to the other two.

Possible Reasons:

- Availability: Covishield might have been produced or imported in larger quantities than the other vaccines.
- Approval Timeline: Covishield might have received emergency use authorization earlier, allowing for vaccinations to start sooner.
- Public Preference: There could be a public preference for Covishield due to factors like familiarity or prior experience.

16. How are Doses Administered by Gender and Age Group ?



```
In [44]: # Sum doses by gender and age group
         age_groups = ['18-44 Years (Doses Administered)', '45-60 Years (Doses Administered)', '60+ Years (Doses Administered)']
         gender_age_doses = data[genders + age_groups].sum().to_frame().reset_index()
         gender_age_doses.columns = ['Category', 'Doses']

         # Create a horizontal stacked bar chart
         gender_age_doses[gender_age_doses['Category'].isin(genders)].set_index('Category').T.plot(
             kind='barh', stacked=True, figsize=(10, 6), color=['lightblue', 'lightgreen', 'lightcoral'])
         plt.title('Distribution of Doses Administered by Gender')
         plt.xlabel('Number of Doses')
         plt.ylabel('Gender')
         plt.legend(title='Gender')
         plt.show()
```

FIGURE 5.34: Doses Administered by Gender and Age Group Code

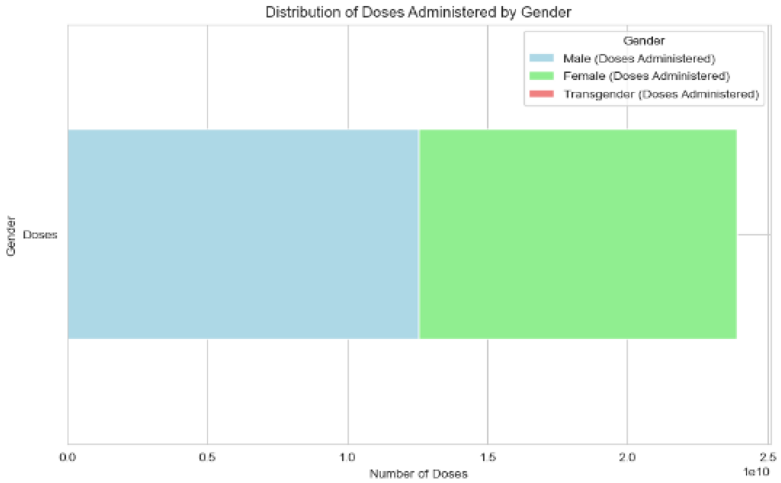


FIGURE 5.35: Doses Administered by Gender and Age Group

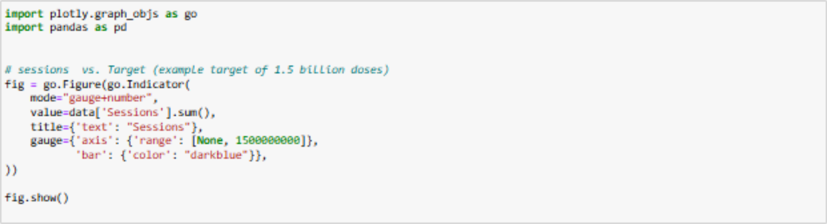
Inference:

- A larger portion of the graph is colored green, which represents doses administered to males.
- A smaller portion of the graph is colored blue, which represents doses administered to females.

Possible Reasons:

- Registration Differences: There could be differences in registration rates between men and women.
- Work Constraints: Women might have faced greater challenges in taking time off work to get vaccinated.
- Vaccine Hesitancy: There could be gender-based differences in vaccine hesitancy.

17. How many sessions were conducted in the whole vaccination process ?



```
import plotly.graph_objs as go
import pandas as pd


# sessions  vs. Target (example target of 1.5 billion doses)
fig = go.Figure(go.Indicator(
    mode="gauge+number",
    value=data['Sessions'].sum(),
    title={'text': "Sessions"},
    gauge={'axis': {'range': [None, 1500000000]},
           'bar': {'color': "darkblue"}},
))

fig.show()
```

FIGURE 5.36: Total Session code

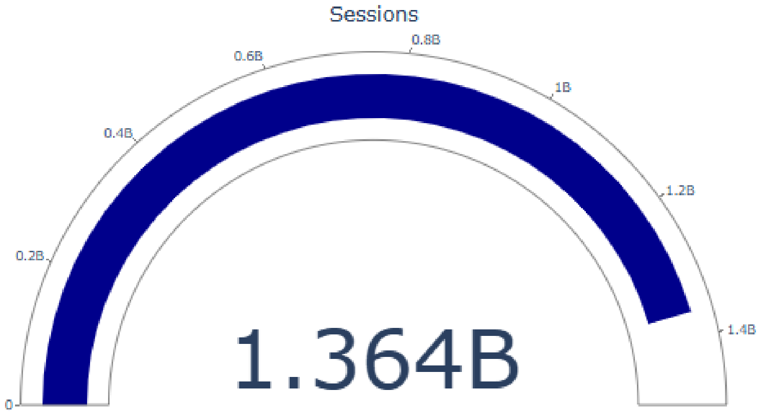
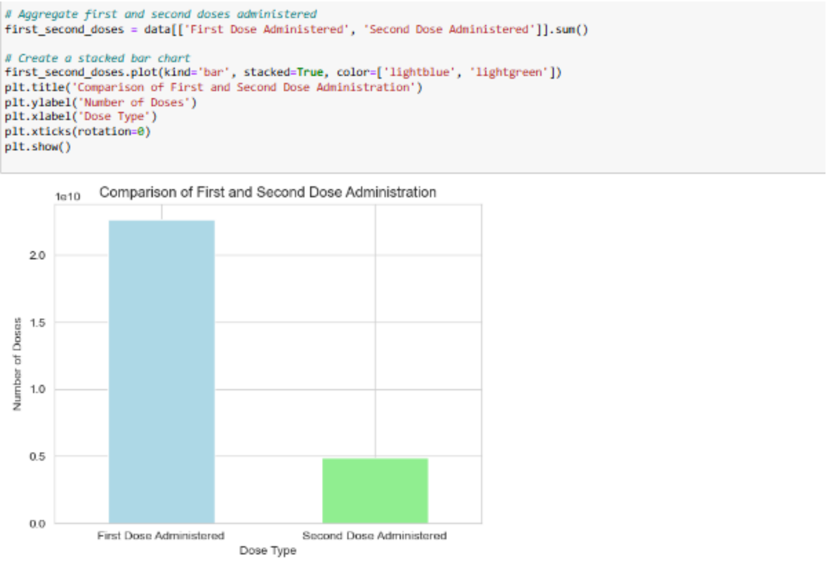


FIGURE 5.37: Total Session

Inference:

- 1.366 B sessions were conducted that was less than the target i.e.,1.5B
- it indicates progress towards a goal of 100 million sessions

18. Comparison of First and Second Dose Administration.



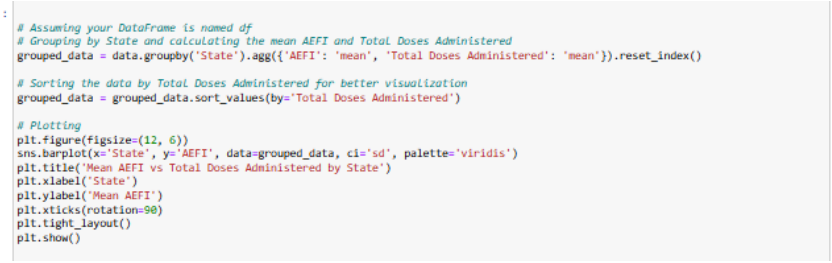
```
# Aggregate first and second doses administered
first_second_doses = data[['First Dose Administered', 'Second Dose Administered']].sum()

# Create a stacked bar chart
first_second_doses.plot(kind='bar', stacked=True, color=['lightblue', 'lightgreen'])
plt.title('Comparison of First and Second Dose Administration')
plt.ylabel('Number of Doses')
plt.xlabel('Dose Type')
plt.xticks(rotation=0)
plt.show()
```

Figure 5.38: Total 1st and 2nd Doses Administered

Inference:

- More than half of the people who were administrated with 1st dose , Didnt recieve the second dose
- The data might only represent a specific timeframe. If there was a lag between administering first and second doses, the graph might show a temporary difference that evens out over time.

19. Is there any relation between adverse events following immunization (AEFI) and the total number of vaccine doses administered ?



```
# Assuming your DataFrame is named df
# Grouping by State and calculating the mean AEFI and Total Doses Administered
grouped_data = data.groupby('State').agg({'AEFI': 'mean', 'Total Doses Administered': 'mean'}).reset_index()

# Sorting the data by Total Doses Administered for better visualization
grouped_data = grouped_data.sort_values(by='Total Doses Administered')

# Plotting
plt.figure(figsize=(12, 6))
sns.barplot(x='State', y='AEFI', data=grouped_data, ci='sd', palette='viridis')
plt.title('Mean AEFI vs Total Doses Administered by State')
plt.xlabel('State')
plt.ylabel('Mean AEFI')
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()
```

Figure 5.39: AEFI vs Total number of Doses Administered Code

Inference:

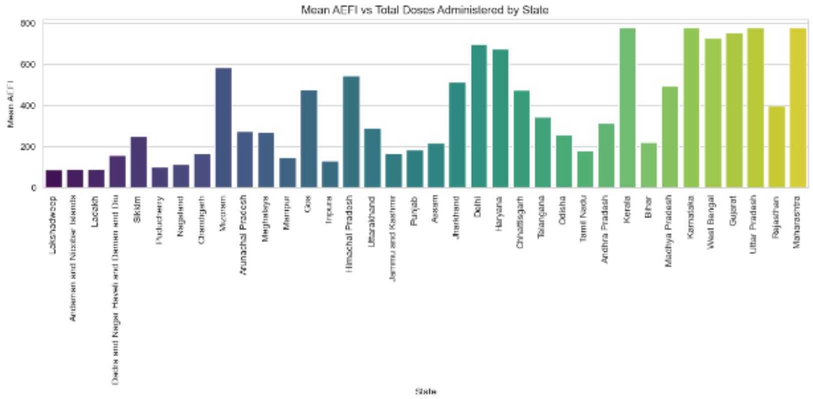


FIGURE 5.40: AEFI vs Total number of Doses Administered

- Maharashtra reports High AEFI Rate compared to other states in india
- Lakshadweep and Puducherry reports High AEFI Rate compared to other states in india

Possible Reasons:

- Pre-existing health conditions: India has a high burden of chronic diseases like diabetes and heart conditions. While these don't necessarily preclude vaccination, they can increase susceptibility to mild AEFIs like fever or pain, which are more commonly reported.
- Vaccine hesitancy and anxiety: Vaccine hesitancy can lead to anxiety around vaccination, which can in some cases manifest as physical symptoms following vaccination. This is a psychosomatic effect and not a direct result of the vaccine itself.

Chapter 6

Results and Outcomes

The outcomes of the course project include:

1. Identification of the state-wise distribution of different vaccine types administered across India.

2. Analysis of the age group that received the highest number of vaccine doses.

3. Examination of gender-based vaccination coverage across different states.

4. Evaluation of the state-wise distribution of vaccinated individuals across different age groups.

5. Determination of the state with the highest number of vaccinated individuals.

6. Comparison between the number of first doses and second doses administered within each state.

7. Investigation of the relationship between the number of doses administered and the total number of vaccinated individuals across different states.

8. Assessment of the correlation between the number of vaccination sessions and the rates of adverse events following immunization (AEFI).

9. Estimation of vaccine wastage by analyzing the loss of vaccine doses administered.

Conclusions

Problem Statement and Importance :

The problem statement was to analyze the distribution and administration of COVID-19 vaccines across different states in India. This analysis was crucial to uncover regional disparities, assess vaccination rates, understand demographic trends, address challenges, and predict future vaccine needs. Solving this problem was important to provide actionable insights for optimizing vaccine deployment strategies and effectively combating the pandemic.

Solving the Problem Statement: .

To solve the problem statement, a comprehensive data analysis approach was employed. This included loading the COVID-19 vaccination dataset, performing data cleaning to handle missing values and outliers, and encoding categorical variables for better analysis. Exploratory data analysis (EDA) was then conducted to visualize patterns and gain insights into the data. Various statistical methods and visualization techniques were used to analyze the distribution of vaccine types, vaccination coverage across age and gender groups, and the state-wise distribution of vaccinated individuals.

Results and Their Outcomes:

The analysis revealed several key outcomes:

1. CoviShield was the most administered vaccine, followed by Covaxin and Sputnik V.

2. Maharashtra had the highest number of vaccine doses administered, while some states had significantly lower vaccination rates.

3. The age group 18-44 received the highest number of vaccine doses.

4. Males had a higher vaccination rate compared to females and transgender individuals.

5. A significant difference was observed between the number of first doses and second doses administered within each state.

6. States with higher numbers of vaccination sessions tended to have lower rates of AEFI cases.

7. The study provided insights into vaccine wastage and suggested ways to optimize vaccine distribution to minimize loss.

These outcomes highlight the effectiveness and challenges of the vaccination campaign in India, providing valuable insights for future public health strategies.

Bibliography