**Missing Data**

* 2013-- QUARTERWISE PERCENTAGE DISTRIBUTION OF DESTINATION COUNTRY WISE DEPARTURES OF INDIAN NATIONALS
* 2014-- DISTRIBUTION OF NATIONALITY-WISE FTAs IN INDIA ACCORDING TO PURPOSE
* 2018-- QUARTERWISE PERCENTAGE DISTRIBUTION OF DESTINATION COUNTRY WISE DEPARTURES OF INDIAN NATIONALS
* Data for Question 2 is incomplete. The information on fees received from each country or continent for specific months is missing. Additionally, it’s unclear whether the Excel file detailing tourism fees is regarding Arrivals or Departures.
* No Data for Question 5. Same as question 2, but for Departing tourists.

**Wrong Data**

* 2014 – The data in QUARTERWISE PERCENTAGE DISTRIBUTION OF DESTINATION COUNTRY WISE DEPARTURES OF INDIAN NATIONALS is the same as DISTRIBUTION OF NATIONALITY-WISE FTAs IN INDIA ACCORDING TO AGE GROUP.
* 2017 – Same as above, not a complete copy, but 95% copy.
* 2019 – China and Taiwan in QUARTERWISE PERCENTAGE DISTRIBUTION OF DESTINATION COUNTRY WISE DEPARTURES OF INDIAN NATIONALS is in West Asia, everywhere else it's in East Asia. This might affect calculation of Q7 and 8.
* 2021 – Some entries in PORT-WISE DEPARTURES OF INDIAN NATIONALS FROM INDIA is percentage and not the actual count. Since the total count is not given, we cannot convert the percentage to actual count.

| **Port/land check post** | **Count** |
| --- | --- |
| Delhi (Airport) | 2278502 |
| Mumbai (Airport) | 1246851 |
| Chennai (Airport) | 6.20% |
| Cochin (Airport) | 960821 |
| Bengaluru (Airport) | 5.00% |
| Hyderabad (Airport) | 551170 |
| Calicut (Airport) | 638778 |
| Trivandrum (Airport) | 4.80% |
| Kolkata (Airport) | 2.90% |
| Ahmedabad (Airport) | 2.50% |
| Others | 1046819 |

* 2022 – STATE UT-WISE DOMESTIC AND FOREIGN TOURIST VISITS has decimal point data for Domestic and Foreign Arrival counts.

**Inconsistencies**

* China and Taiwan appear under both East and West Asia. – Slight problem, may affect the calculations a bit.
* There are some cell merging issues. Like in some cases multiple columns of the same row are merged together and the values are given in the merged cell, separated by spaces.
* The total number of FTAs for some years is not the same in Age, purpose, gender files.

2013

Total Arrivals from Age file: 7648432.0

Total Arrivals from Purpose file: 3317974.0

Total Arrivals from Gender file: 7648432.0

2016

Total Arrivals from Age file: 8789423.0

Total Arrivals from Purpose file: 8789423.0

Total Arrivals from Gender file: 8789423.0

2017

Total Arrivals from Age file: 10012347.0

Total Arrivals from Purpose file: 6088369.0

Total Arrivals from Gender file: 14512347.0

2018

Total Arrivals from Age file: 10547939.0

Total Arrivals from Purpose file: 10547939.0

Total Arrivals from Gender file: 10547939.0

2019

Total Arrivals from Age file: 10919514.0

Total Arrivals from Purpose file: 6088369.0

Total Arrivals from Gender file: 10919514.0

2020

Total Arrivals from Age file: 2741748.0

Total Arrivals from Purpose file: 2741748.0

Total Arrivals from Gender file: 2741748.0

2021

Total Arrivals from Age file: 1525972.0

Total Arrivals from Purpose file: 1525972.0

Total Arrivals from Gender file: 1525972.0

2022

Total Arrivals from Age file: 6924909.0

Total Arrivals from Purpose file: 6430141.0

Total Arrivals from Gender file: 6430141.0

* Purpose of Visit options are not the same in all years. **Is comparing the entropies of different years fair, if some years have more categories.** Also, some options with the same meaning have different names. Ex: 'OCI' and 'Indian Diaspora'

'Medical Visa' and 'Medical'

2013 - ['Tourist Visa', 'OCI', 'Business Visa', 'Medical Visa', 'Student Visa', 'Others']

2016 - ['Business and Professional', 'Leisure Holiday and Recreation', 'Medical', 'Indian Diaspora', 'Others']

2017 - ['Business and Professional', 'Leisure Holiday and Recreation', 'Medical', 'Indian Diaspora', 'Others']

2018 - ['Business and Professional', 'Leisure Holiday and Recreation', 'Medical', 'Indian Diaspora', 'Others']

2019 - ['Business and Professional', 'Leisure Holiday and Recreation', 'Medical', 'Indian Diaspora', 'Others']

2020 - ['Business and Professional', 'Leisure Holiday and Recreation', 'Medical', 'Indian Diaspora', 'Others']

2021 - ['Business and Professional', 'Leisure Holiday and Recreation', 'Medical', 'Indian Diaspora', 'Student', 'Others']

2022 - ['Business and Professional', 'Leisure Holiday and Recreation', 'Medical', 'Indian Diaspora', 'Student', 'Unknown', 'Others']

**Naming Errors**

* Same excel files are named slightly differently in some years, which makes loading them programmatically difficult.

Ex: 2013 – DISTRIBUTION OF NATIONALITY-WISE FTAs IN INDIA ACCORDING TO PURPOSE, 2013.xlsx

2021 – DISTRIBUTION OF NATIONALITY-WISE FTAs IN INDIA ACCORDING TO PURPOSE, 2021 (in percentage).xlsx

* Same countries are named differently in different Excel files.

Ex: UK or United Kingdom

USA or United States or United States of America

China or Republic of China

* Some States/UT in the STATE UT-WISE DOMESTIC AND FOREIGN TOURIST VISITS.xlsx files are named differently across the years.
  + Ex: Andaman & Nicobar Islands vs A&N Island.
  + Jammu & Kashmir vs J&K.

**Question Wise**

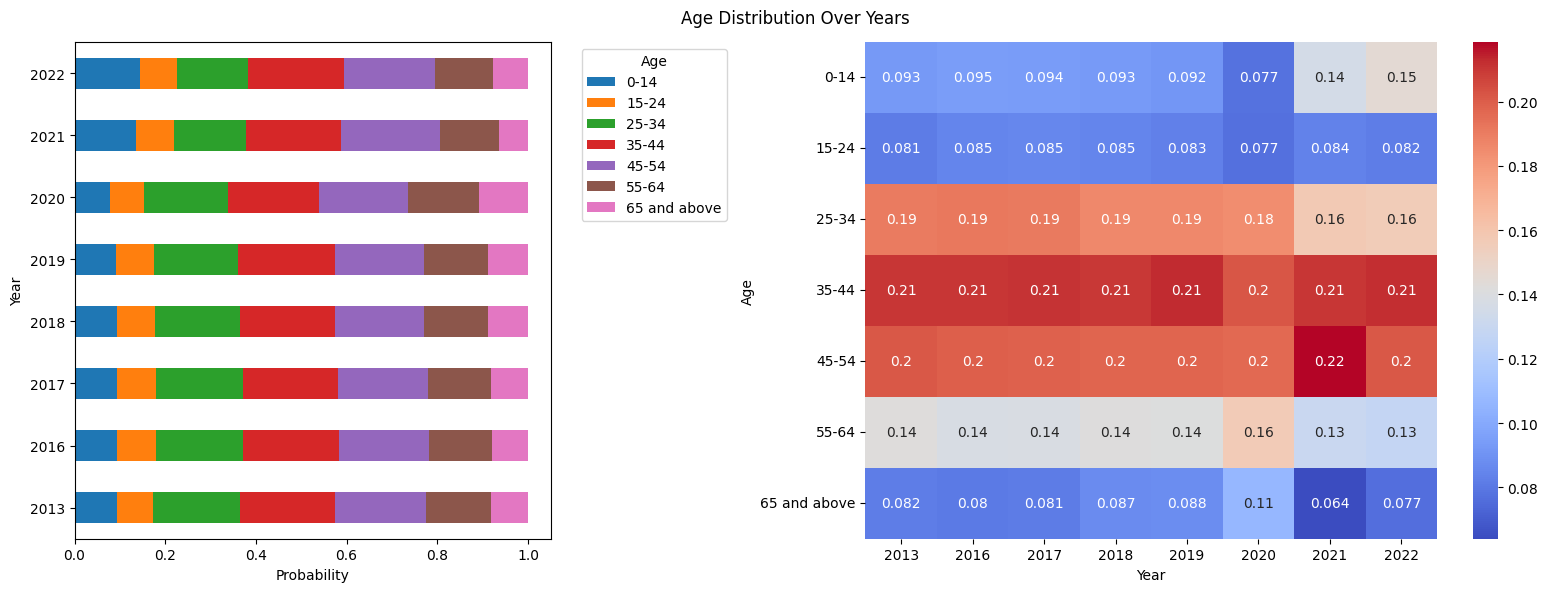
**Question 1**

What does  *“For each month, find the changes* ***across the entire data*** *from the previous month.”*  mean?

**Question 2**

Data is incomplete. The information on fees received from each country or continent for specific months is missing. Additionally, it’s unclear whether the Excel file detailing tourism fees is regarding Arrivals or Departures.

**Question 3**

* Purpose of Visit options are not the same in all years.
* 2022 has a ‘Third Gender’ category, which no other year has.
* **Is comparing the entropies of different years fair, if some years have more categories.**
* Also, some options with the same meaning have different names. This is creating redundant categories. **Can we rename them?**
* For the multinomial probability distribution part, is plotting a graph showing the probability of each category for each feature enough?
* 

**Question 5**

No data on fees received from outgoing tourists.

**Question 6**

* This question requires repeating the analysis from Question 3 for outgoing tourists. However, Question 3's analysis includes a breakdown by country and continent, but the data for this question is not categorized by destination country or continent.
* 2019 has Rail as an option for Mode of Departure, which no other year has. All other years only have 3 categories while 2019 has 4. Again, is comparing entropies fair if the number of categories differ?
* For the third feature in this question - state-wise visit of domestic versus foreign tourists , are the states/UTs the features and Domestic/Foreign the categories or reverse?
* 2022 – STATE UT-WISE DOMESTIC AND FOREIGN TOURIST VISITS has decimal point data for Domestic and Foreign Arrival counts.

**Question 9**

* Will interpolation work properly if the number and names of purpose of visit categories are different in different years.