



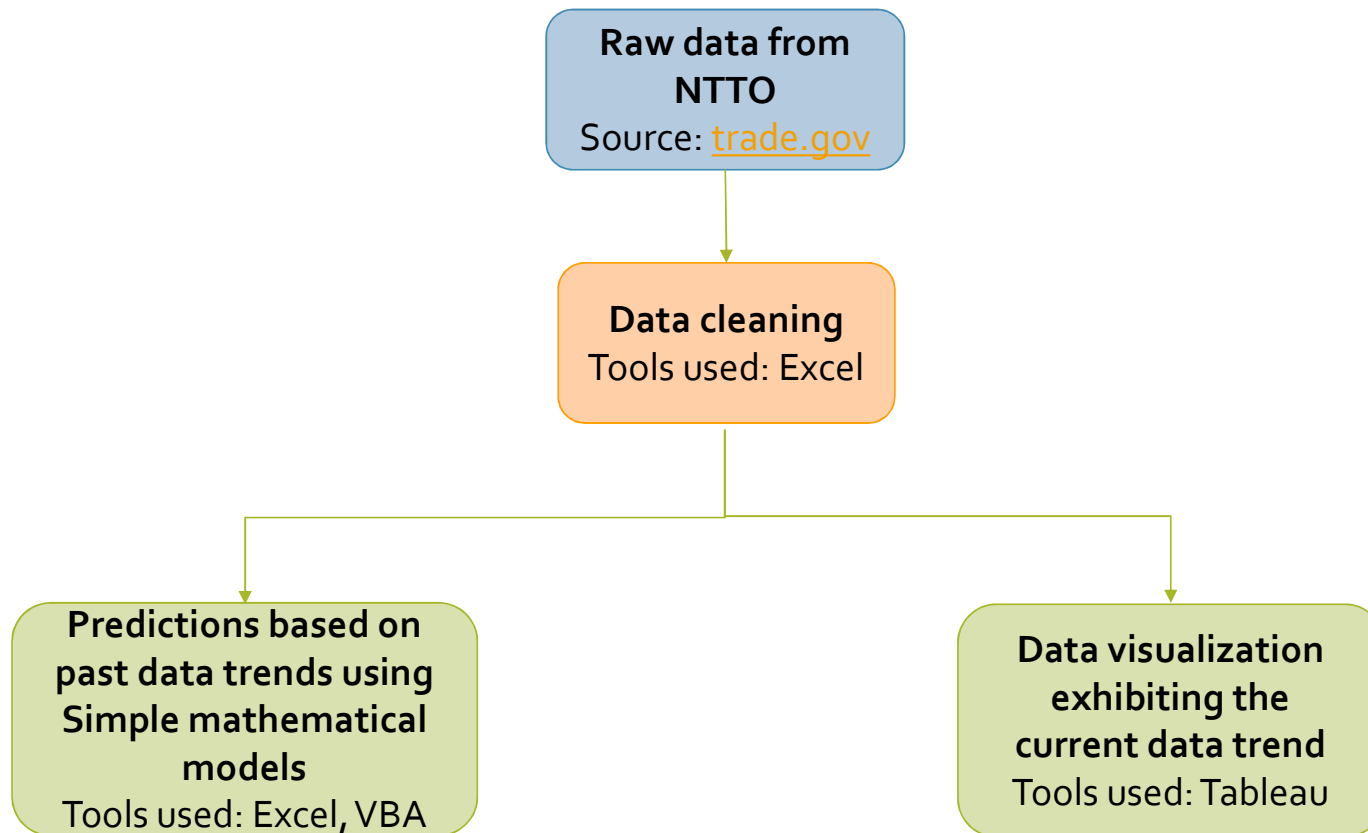
ANALYSIS AND PREDICTIONS OF INTERNATIONAL TOURIST ARRIVALS IN THE UNITED STATES

By:
Jalpaben Patel

Problem

- **Travel industry on halt due to pandemic.**
 - As the covid restrictions emerged, there is a decline in travel plans.
 - Impacting the country's revenue in terms of pleasure and business travels put on hold.
- **High level of uncertainty in predicting travel forecast.**
 - Due to the huge drop in 2020, it will be challenging to forecast what the future travel will look like.
 - Identifying areas of focus and target market for improving tourism is taxing

Flow Diagram



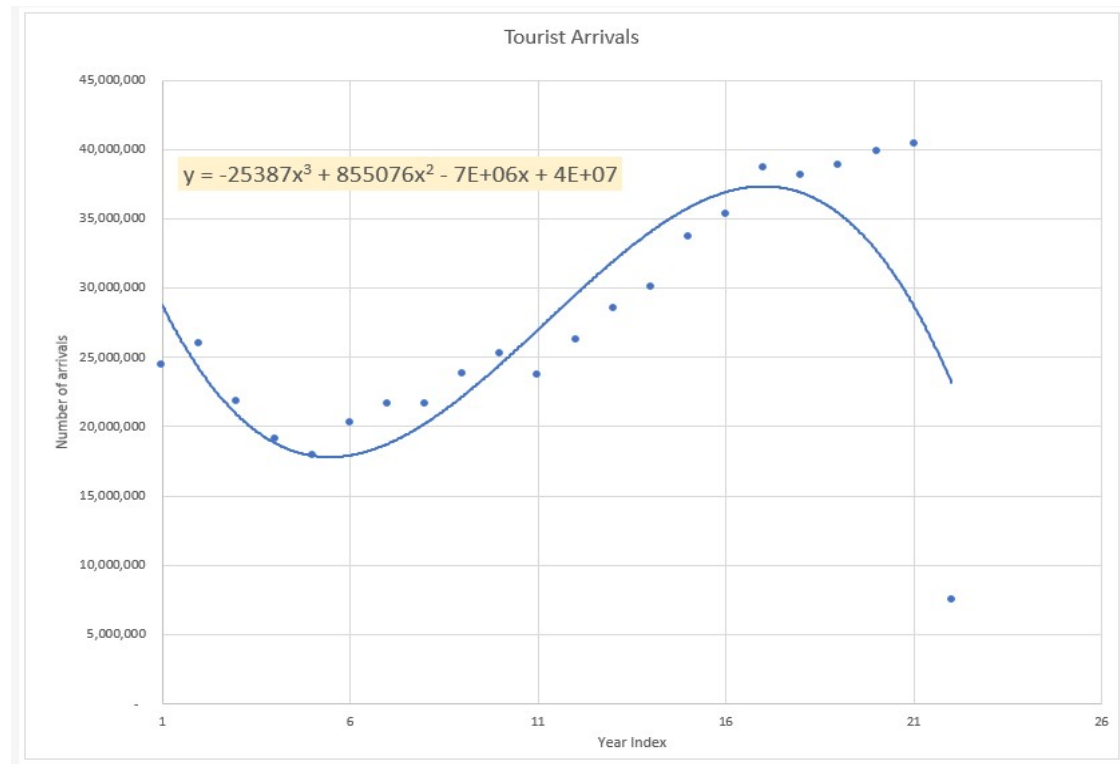
Data Cleaning

- The input raw data obtained from the trade.gov website had information about number of arrivals to the United States based on various categories such as Region, Country, Port of Entry, Visa type for years 1999-2020.
- Since these data was in wide format and was present in multiple files separated for each year, it had to be grouped and cleansed for further analysis and visualizations.
- Primarily Excel's Get and Transform tools and some other excel functions like index and match, transpose, etc. was used in data cleaning. The usage of the functions is described as follows.
 - The index and match functions have been used to find the appropriate region for each country.
 - Get and Transform Tool has been used to convert the raw data from wide to long format (As it was suitable for the visualization tool).
 - Range names have been used in the formulas instead of cell references.
 - Transpose function has also been used in the data clean-up process.
 - Some of the input fields have extra spaces which made it difficult to analyze. To clean up the raw data, the *trim()* function has been used.

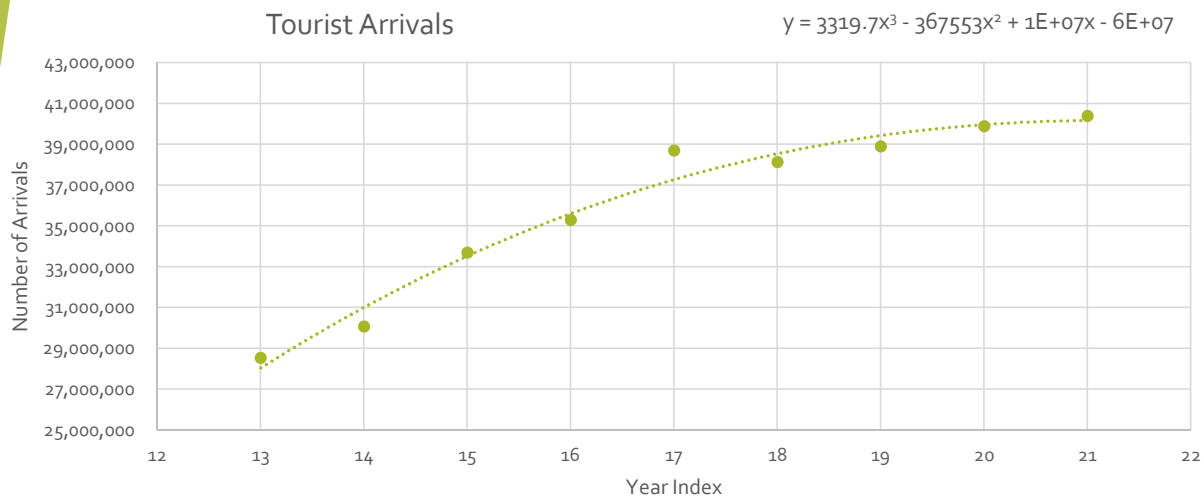
Creating the model

Total Tourist Arrivals in US		
Year	YearIndex	Tourist Arrivals
1999	1	24,466,187
2000	2	25,974,701
2001	3	21,832,868
2002	4	19,116,707
2003	5	18,026,213
2004	6	20,322,257
2005	7	21,678,528
2006	8	21,668,290
2007	9	23,892,277
2008	10	25,341,451
2009	11	23,756,184
2010	12	26,362,616
2011	13	28,539,631
2012	14	30,070,908
2013	15	33,688,677
2014	16	35,295,842
2015	17	38,700,150
2016	18	38,129,404
2017	19	38,905,524
2018	20	39,883,361
2019	21	40,393,346
2020	22	7,594,470

	Squared Error					
	Linear	Power	Exponential	Logarithmic	Quadratic	Cubic
Sum of Errors	1.12E+15	1.30E+15	1.22E+15	1.24E+15	1.10E+15	5.45E+14
Mean Squared Error	5.07E+13	5.90E+13	5.54E+13	5.62E+13	5.01E+13	2.48E+13
RMSE	7,120,217	7,680,388	7,440,381	7,493,698	7,081,102	4,979,389

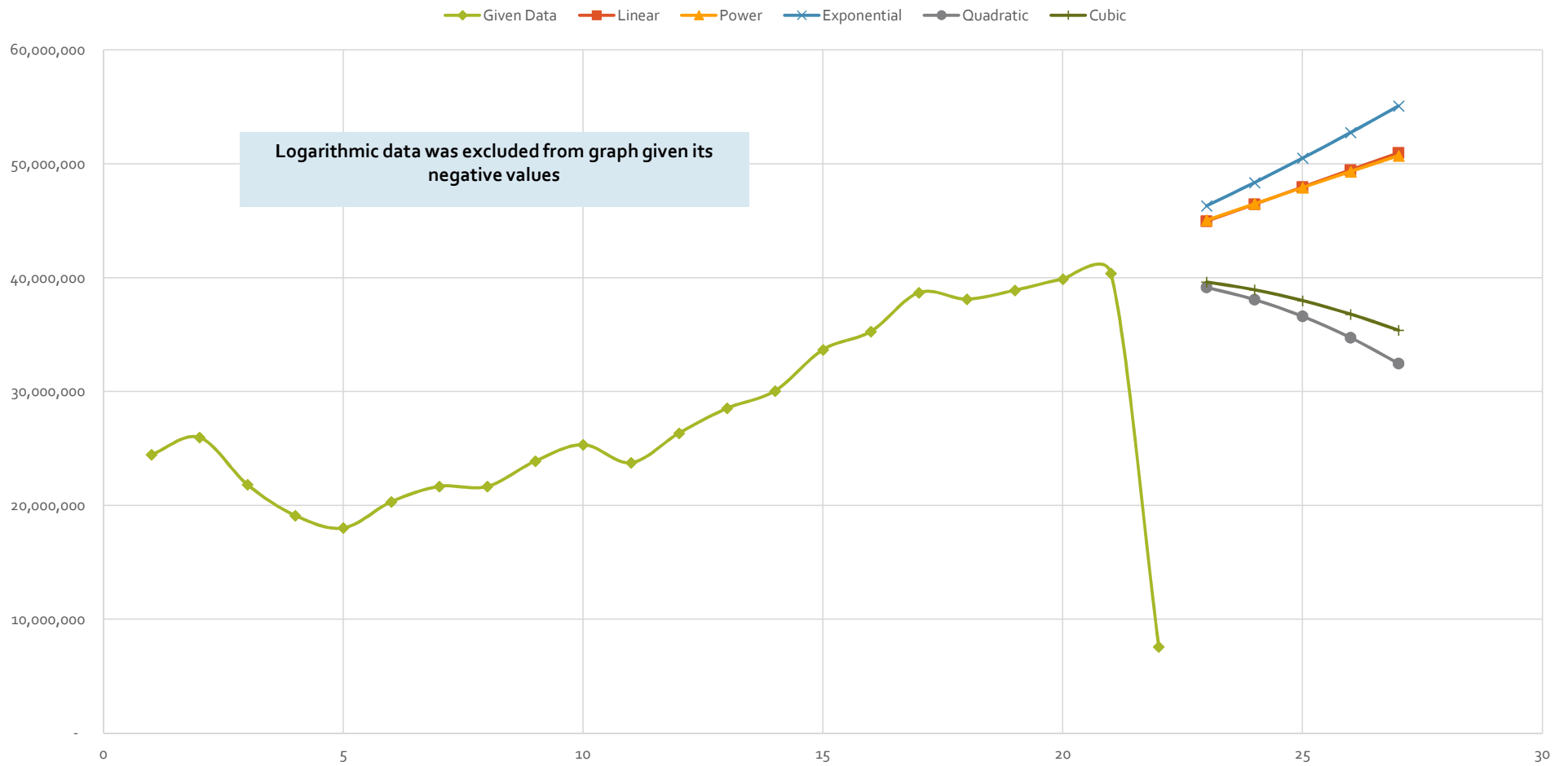


Making Predictions – The Cubic Model



	Squared Error					
	Linear	Power	Exponential	Logarithmic	Quadratic	Cubic
Sum of Errors	1.59E+13	1.48E+13	2.05E+13	1.15E+17	3.82E+12	3.80E+12
Mean Squared Error	1.77E+12	1.64E+12	2.28E+12	1.27E+16	4.24E+11	4.22E+11
RMSE	1,330,108	1,281,621	1,510,913	112,914,685	651,272	649,930

FUTURE TOURIST ARRIVALS



Making Predictions - Result

Total Tourist Arrivals in US - Future Years						
	Predictions					
Year	Linear	Power	Exponential	Logarithmic	Quadratic	Cubic
2021	44,968,263.39	45,055,294.05	46,319,363.81	(87,917,004.35)	39,152,944.31	39,634,960.64
2022	46,470,254.64	46,497,564.81	48,372,158.01	(89,457,454.82)	38,077,691.88	38,942,134.39
2023	47,972,245.89	47,924,311.67	50,515,928.50	(90,935,011.93)	36,605,940.42	37,992,235.32
2024	49,474,237.14	49,336,313.18	52,754,707.20	(92,354,610.47)	34,737,689.93	36,805,181.47
2025	50,976,228.39	50,734,280.52	55,092,704.70	(93,720,626.27)	32,472,940.41	35,400,890.86

- Model was made using years 2011-2019 as the base
- Cubic function was the best fit, so we based our final prediction on that mathematical model

Factors Affecting Tourism

Between 2019 and 2020, there was a decrease of 32,798,876 tourist arrivals, almost entirely due to the impacts of COVID-19 on the tourism industry. This corresponds to an 81.2% decrease in tourist arrivals. As the tourism industry begins to build itself back up, people like hotel owners, airline companies, and those investing in tourism-based companies may be wondering what the future of travel looks like.

In previous years, Quarter 3 (July, August, September) tends to have the highest rates of tourism. As restrictions due to COVID begin to be lifted, it is likely that tourism will gradually grow in the early summer of 2021, and expand to its peak in July/August of 2021 as people become more comfortable with the idea of traveling, and as more vaccines are administered. As the years progress and we get into 2022 and 2023, the world will begin to return to normalcy. Our predictions show a relatively consistent rate of tourism, as there are still unknown factors affecting tourism related to COVID. However, we feel as though our model is consistent with what is to be expected.

Some other factors that influence tourism rates are: economic stability, seasonal trends, attractions, and the weather/climate. Beyond the fact that COVID resulted in quarantine, and a generalized discouragement of traveling, it also left some people without jobs, or with a reduced income. This is another factor that is going to affect tourism as borders begin to reopen, as some families are not able to afford to travel. Seasonal trends and the weather/climate will likely result in tourism picking up more in the summer, as seen in trends of the prior years.

Find Best Fit Model using VBA Automation

- The calculation of RMSE and finding best fit model for future predictions has been automated using VBA.
- The formulas involved in the calculation of RMSE has been formulated using range of year index values as reference.
- This allows the application users to find the best fit model easily without having to add any formulas.
- This program also enables the users to compute the RMSE for any additional year and arrival information.
- The VBA automation is available for both historical and recent year prediction data.
- The program also allows the user to delete the existing prediction information (in case of any changes in the number of arrivals) and recalculate RMSE for new data

Find Best Fit Model For Historical Data

Copy Year Index to
ModelParams

Clear Model Parameters

Calculate Model Paramaters

Clear Model data

Calculate RMSE

Find Best Fit Model For Recent Data

Copy Year Index to
ModelParamsRecentYears

Clear Model Parameters

Calculate Model Paramaters

Clear Model data

Calculate RMSE for Recent
Year Model

Data Visualization

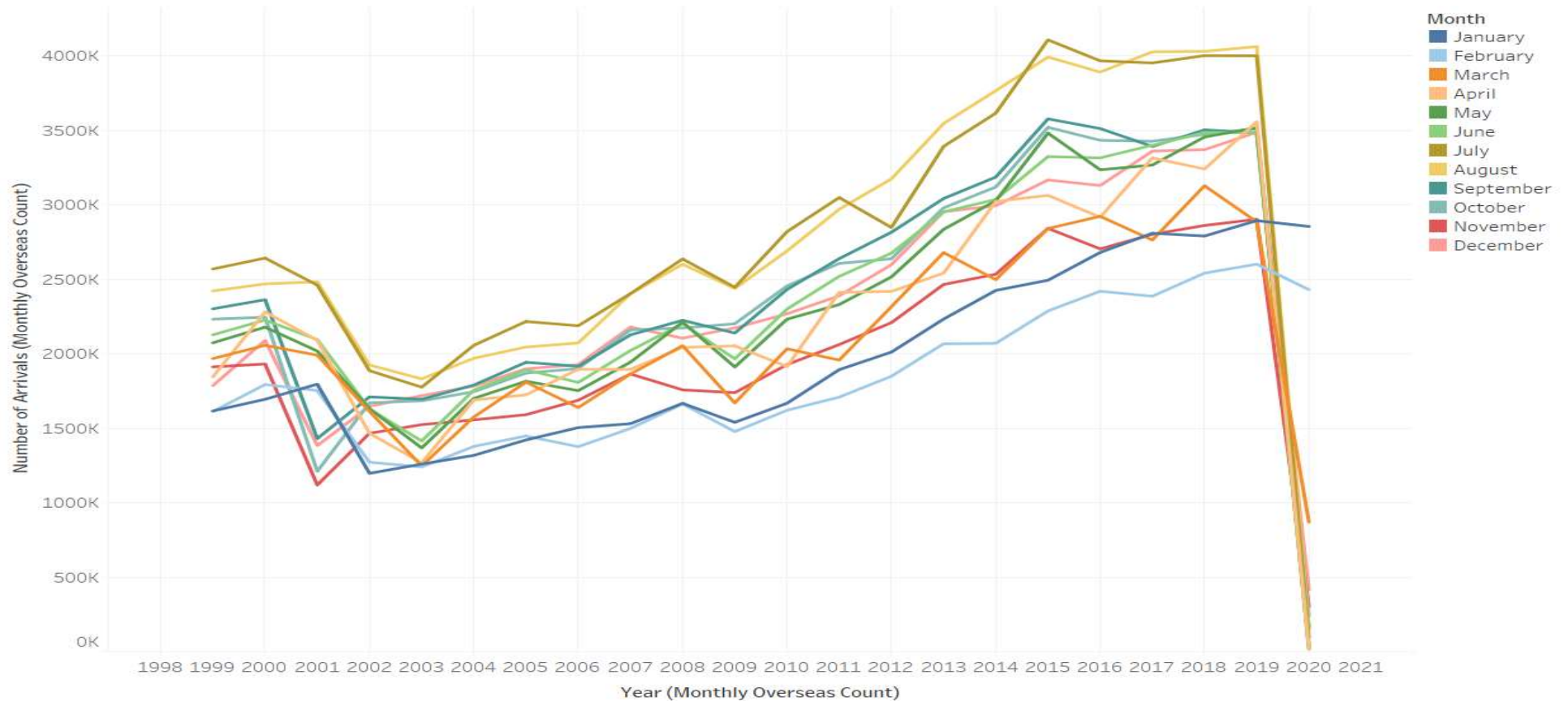


Interactive dashboards have been created to address the following queries using Tableau

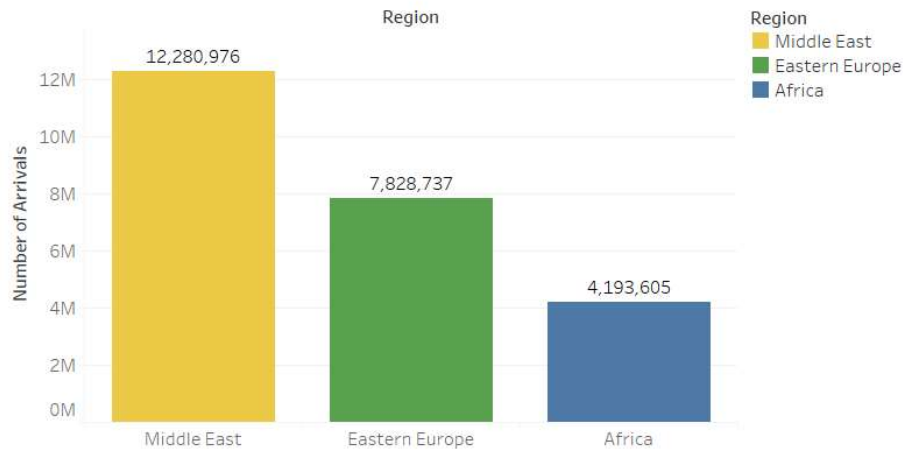
- Overall arrival trend (1999-2020)
- Top regions that had the greatest number of arrivals
- Regions that contribute to least number of arrivals
- Top Countries contributing to most arrivals and least arrivals
- Overall country wise distribution of number of arrivals
- Busiest Ports of Entry in the United States
- Overall arrival trend based on port of entry
- Visa type contributing to highest number of arrivals
- Visa Type contributing to highest number of arrivals in each region
- Percent Change in number of arrivals by Month

Month of arrivals with percent change

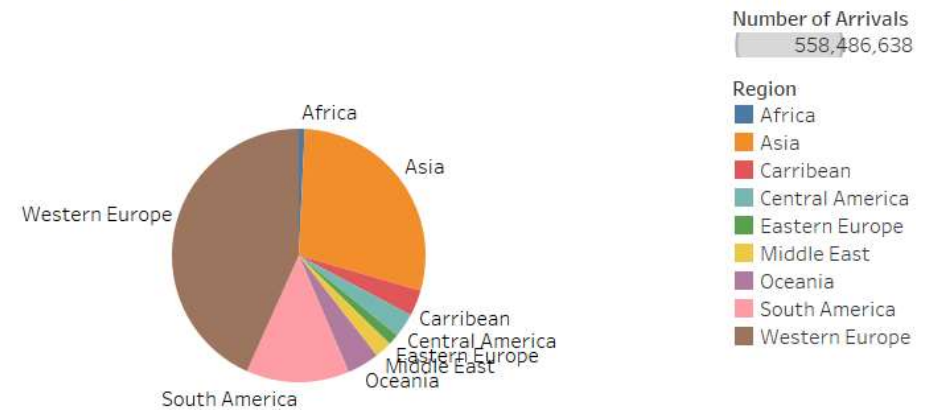
- Here, We found that there is a pattern between number of arrival in each month. Like July and August month share maximum arrival whereas January and February have least number of arrival.



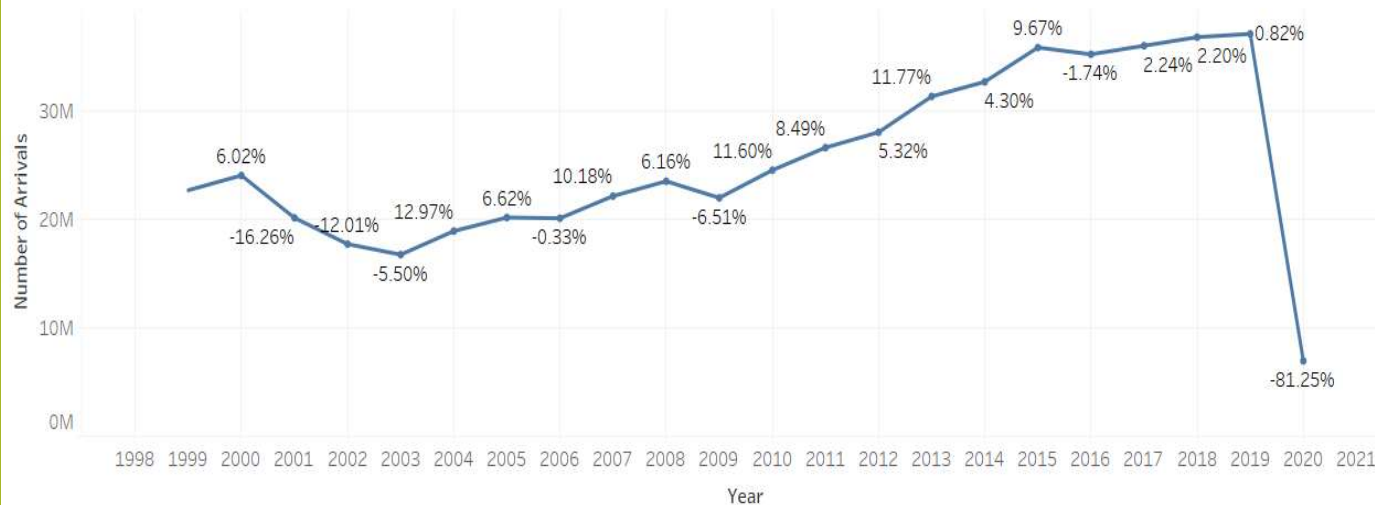
Bottom 3 Regions with least number of arrival



Region wise top arrivals



Overall arrival trend



- Western Europe, Asia and South America are region, which contribute more tourist entry in United States.
- Whereas Middle East, Eastern Europe and Africa have least tourist count.
- Overall arrival trend increase over the period of time. But in 2020 due to pandemic and travel restriction, there was huge fall in arrival.

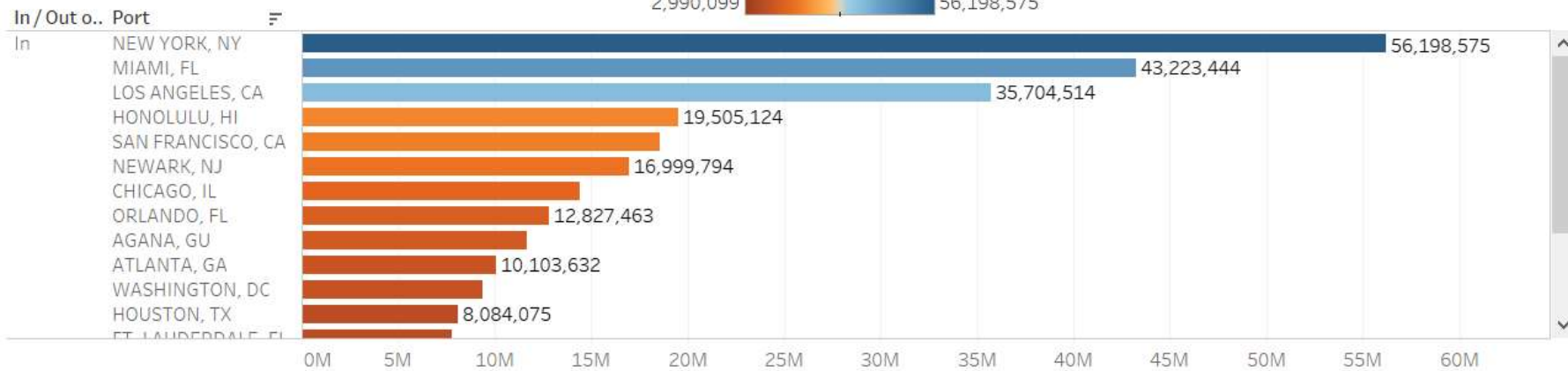
Busiest Ports of Entry in the United States

Number of Arrivals (Portwise Arrival Data)

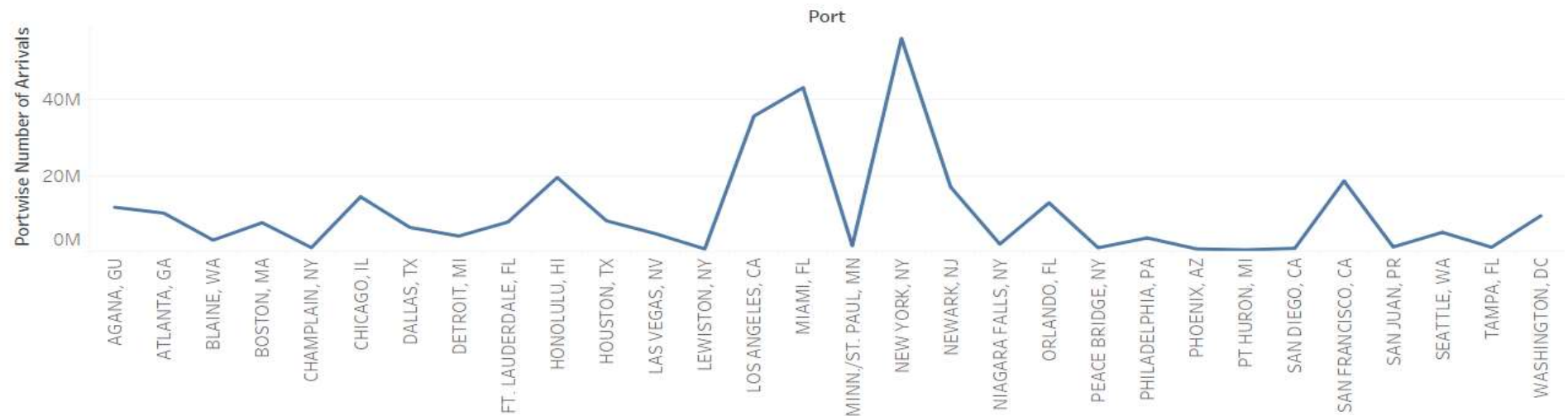
2,990,099 56,198,575

Top N Port

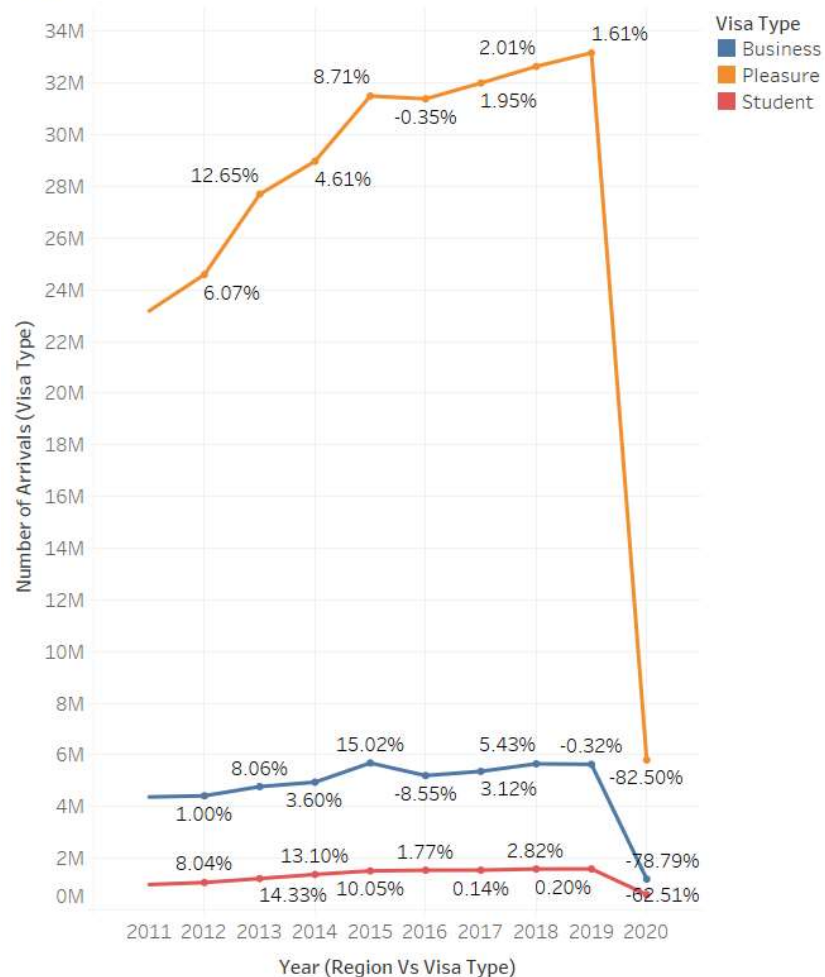
20



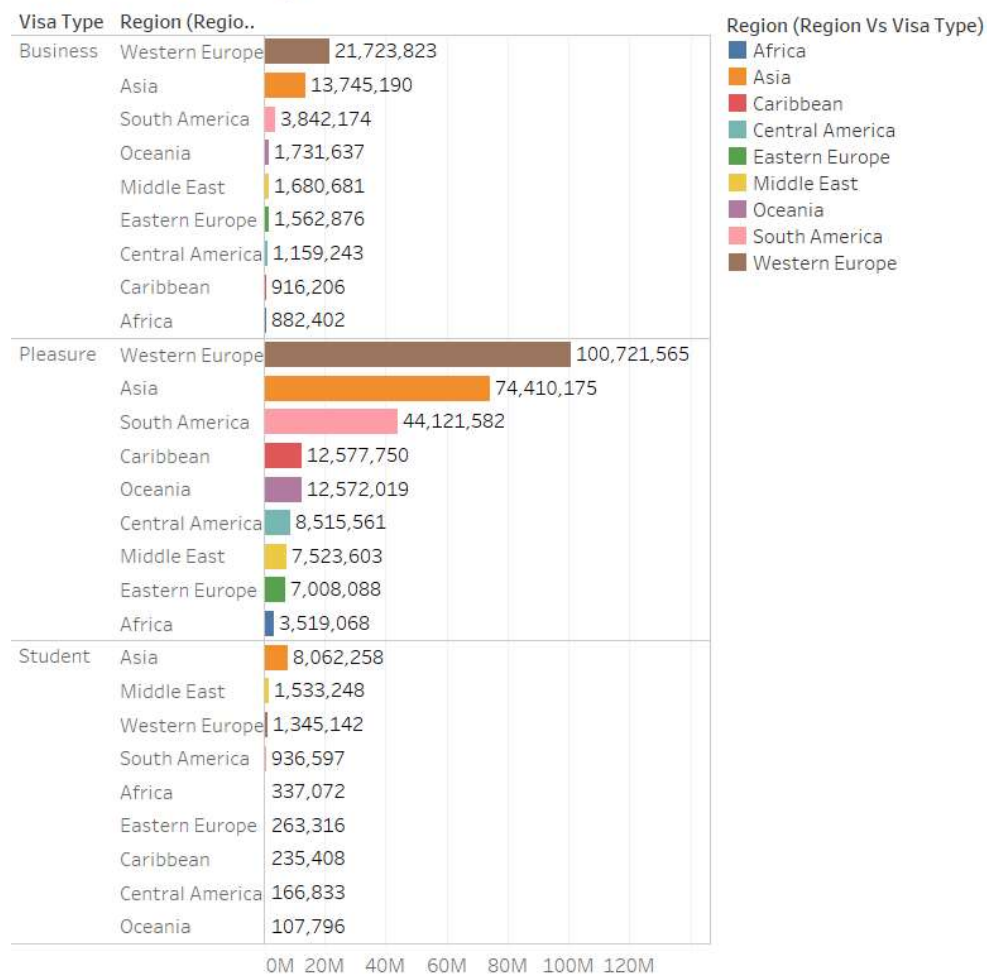
Overall arrival trend based on port of entry



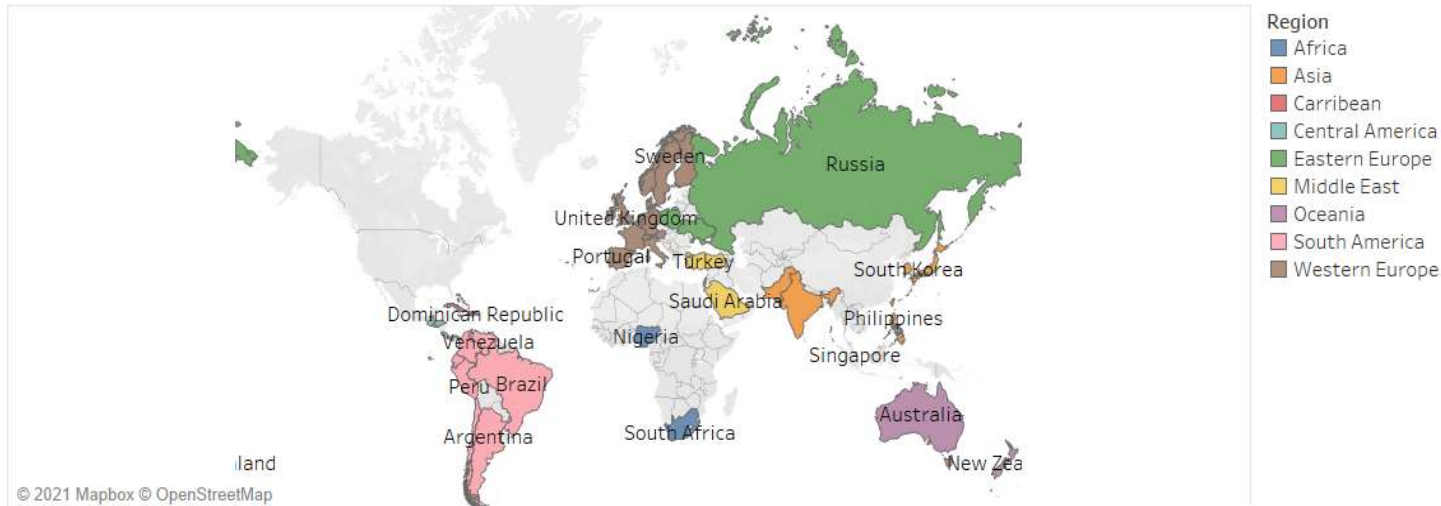
Visa type contributing to highest number of arrivals



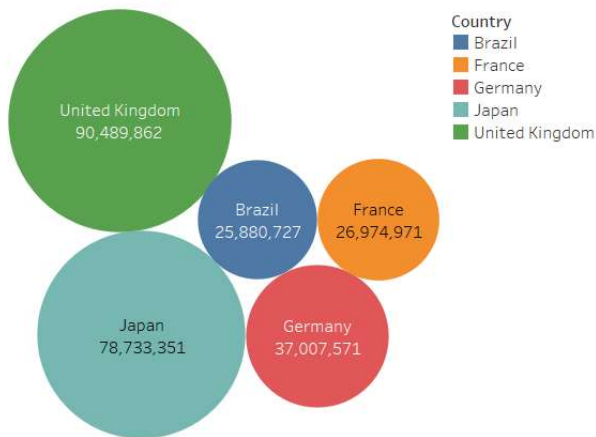
Visa Type contributing to highest number of arrivals in each region



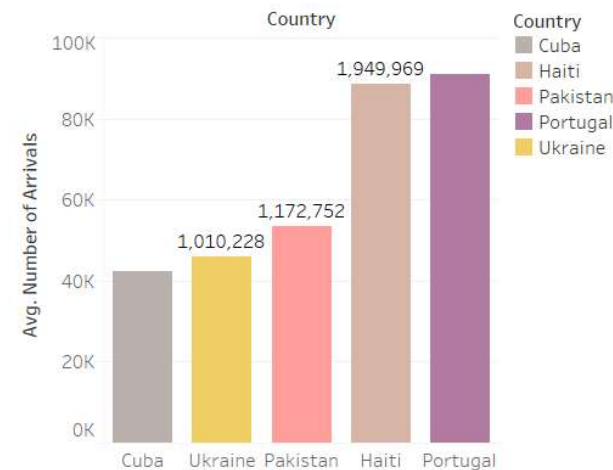
Overall country wise distribution of number of arrivals



Top 5 Countries contributing to most arrivals



Countries with least arrival Entry



- Any traveler taking a trip to a main destination outside their residence, for business, pleasure, study purpose. Previous slide shows that Maximum tourist arrive with Pleasure visa type and mainly from Western Europe and Asia region.
- Map based on Longitude and latitude. Colors shows details about region. The marks are labeled by County.
- United Kingdom, Japan, Brazil, Germany and France contribute more tourist entry in United States.
- Whereas Cuba, Ukraine, Pakistan, Haiti and Portugal have least tourist count.
- New York, NY and Miami, FL are the busiest port.

Capabilities:

- Tableau helped us visualize raw data in form of multiple dashboards.
- Identify trends and patterns from analyzed data which can help Tourism companies to yield important decisions.
- Through this project we would be aiding the US tourism industry including airlines, hotels and ground transportation businesses by predicting international tourist arrivals trends in the US which they can use for making investment related decisions.
- With the help of our prediction model, additional data can be analyzed for forecasting number of arrivals for future years with reduced manual effort (As we have automated the prediction for best fit model using VBA).

Limitations:

- We did not have country wise data for different visa types which would have helped us to provide a data comparison based on the number of arrivals from different countries and visa types.
- Also, the I-94 data from the website have only the Port of entry details but do not have any information about the destination. If we have those details, it will be more useful for the people from tourism industry as there could be cases where port of entry is different from the destination of travelers.
- Our project does not provide a user-friendly form to add additional data in the model automatically and search for regional and country wise data for their own analysis.
- Our prediction is based on simple mathematical model, and it does not consider any external factors impacting the number of arrivals.

Development and Analysis Challenges

- In developing our model for predictions, we faced a challenge due to the dramatic decrease in tourists from 2019 to 2020 which made us to eliminate the 2020 data for making predictions. When forming the predictions based on 2016 to 2020, we were getting negative values for the future year predictions, though that is impossible.
- The input data had extra spaces which made it difficult to analyze. The index and match functions failed due to this reason
- During automation of RMSE calculation, we ran into different errors such as subscript out of range error while calculating the prediction values for quadratic and cubic functions as the data was populated using array function.
- During initial data compilation we tried to combine the raw data based on year as it was common for all categories. But that structure proved to be incompatible for visualization purposes
- We tried creating interactive drop-down list for year wise arrival data based on Port of entry but instead used filter option.
- During VBA automation we came across errors such as Method range of object - 'Global' failed, Sub for Function not defined.

Thank You

