FLIGHT PRICES ANALYSIS

California State University, Los Angeles

CIS 5200 – System Design and Analysis

Dr. Jongwook Woo

Team Members: Ragi Dave, An Mach,

Ankita Hasmukhbhai Savaliya, Bhumika Suvagia

AGENDA

- Goals
- Dataset Information
- Hardware Specification
- Architectural Workflow
- Analysis
- Conclusion

GOALS

- Provide a high-level, general and aggregated approach to hive analysis.
- Performed analysis could be valuable for travelers who are looking to plan their trips in advance and find the best deals on flights. Such as,

- What are the top 10 most popular routes?
- Which airport has the highest traffic during August and September?

FLIGHT DATA INTRODUCTION

- This dataset contains flight prices and includes information on flight itineraries from April to November 2022. The size of the dataset is 31.09 GB. In the dataset, there are 27 columns, we utilized only 17 of them.
- Flight information to and from 16 airports, including ATL, DFW, DEN, ORD, LAX, CLT, MIA, JFK, EWR, SFO, DTW, BOS, PHL, LGA, IAD, and OAK.

FLIGHTDATA TABLE

col_name	data_type
# col_name	data_type
flightid	string
searchdate	date
flightdate	date
startingairport	string
destinationairport	string
farebasiscode	string
travelduration	string
elapseddays	int
isbasiceconomy	boolean
isrefundable	boolean
isnonstop	boolean
basefare	double
totalfare	double
seatsremaining	int
totaltraveldistance	int
segmentsdepartureti meepochseconds	string
segmentsdepartureti meraw	string
segmentsarrivalti meepochseconds	string
segmentsarrivalti meraw	string
segmentsarrivalairportcode	string
segmentsdepartureairportcode	string
segmentsairlinename	string
segmentsairlinecode	string
segmentsequipmentdescription	string
segmentsdurationinseconds	string
segmentsdistance	int
segmentscabincode	string

AIRPORT TABLE

col_name	data_type	comment
# col_name	data_type	comment
name	string	
city	string	
country	string	
iata	string	
icao	string	
latitude	double	
longitude	double	
	NULL	NULL
# Detailed Table Information	NULL	NULL
Database:	amach3	NULL
OwnerType:	USER	NULL
Owner:	amach3	NULL
CreateTime:	Fri Apr 21 04:06:36 GMT 2023	NULL
LastAccessTime:	UNKNOWN	NULL
Retention:	0	NULL
Location:	hdfs://bigdaimn0.sub03291929060.trainingvcn.oraclevcn.com:8020/user/amach3/Airport	NULL
Table Type:	EXTERNAL_TABLE	NULL
Table Parameters:	NULL	NULL
	EXTERNAL	TRUE
	bucketing_version	2
	numFiles	1
	skip.header.line.count	1
	totalSize	580468
	transient_lastDdlTime	1682049996
	NULL	NULL

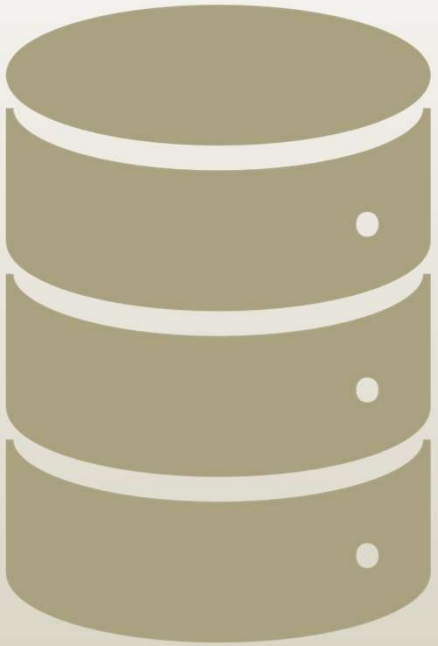
DATASET INFORMATION

- **DATASET NAME:** Flight Prices
- **DATASET URL:**

https://www.kaggle.com/datasets/dilwong/flightprices

https://www.kaggle.com/datasets/mike90/airport-codes

- **TOTAL SIZE:** 31.09 GB
- **COUNTRIES CONSIDERED:** USA
- **NUMBER OF FILES:** 2
- **FILE FORMAT:** CSV
- **GITHUB LINK:** https://github.com/amach3/Flight-Price-Analysis.git



HADOOP CLUSTER SPECIFICATION

- Cluster Version: Hadoop 3.1.2
- CPU Speed: 1995.312 MHz
- Number of CPU cores:

  8 cores x 5 nodes = 40 cores
- Number of nodes: 5

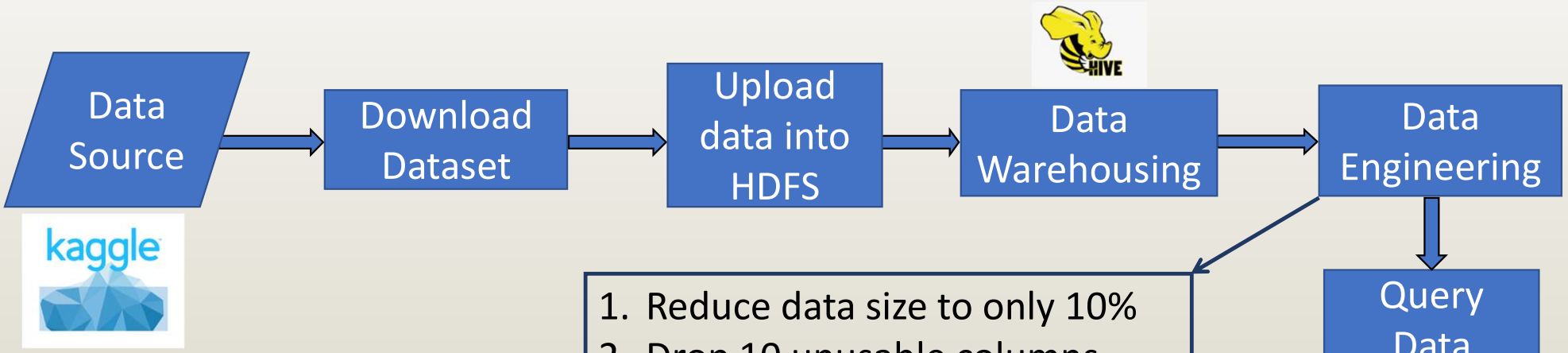
  (2 Master and 3 Worker)
- Total Memory Size: 390.71 GB

```
-bash-4.2$ lscpu
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                8
On-line CPU(s) list:   0-7
Thread(s) per core:    2
Core(s) per socket:    4
Socket(s):             1
NUMA node(s):          1
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 85
Model name:            Intel(R) Xeon(R) Platinum 8167M CPU @ 2.00GHz
Stepping:              4
CPU MHz:               1995.312
BogoMIPS:              3990.62
Virtualization:        VT-x
Hypervisor vendor:     KVM
Virtualization type:   full
L1d cache:             32K
L1i cache:             32K
L2 cache:              4096K
L3 cache:              16384K
NUMA node0 CPU(s):     0-7
```

```
-bash-4.2$ hdfs dfsadmin -report
Configured Capacity: 419520548352 (390.71 GB)
Present Capacity: 417859894557 (389.16 GB)
DFS Remaining: 42330897693 (39.42 GB)
DFS Used: 375528996864 (349.74 GB)
DFS Used%: 89.87%
```

```
-bash-4.2$ hdfs version
Hadoop 3.1.2
Source code repository ssh://git@bitbucket.oci.oraclecorp.com:7999/bdcs/apache_bigtop.git -r 4100eb8d8581c4328601079ff5af522f95e9977f
Compiled by root on 2023-02-27T08:26Z
Compiled with protoc 2.5.0
From source with checksum b367ca15864aef16725a3035859c9ece
This command was run using /usr/odh/1.1.5/hadoop/hadoop-common-3.1.2.jar
```

ARCHITECTURE WORKFLOW



1. Reduce data size to only 10%
2. Drop 10 unusable columns
3. Remove duplicate records
4. Remove null values
5. Add columns by extracting the data we need

Roles:

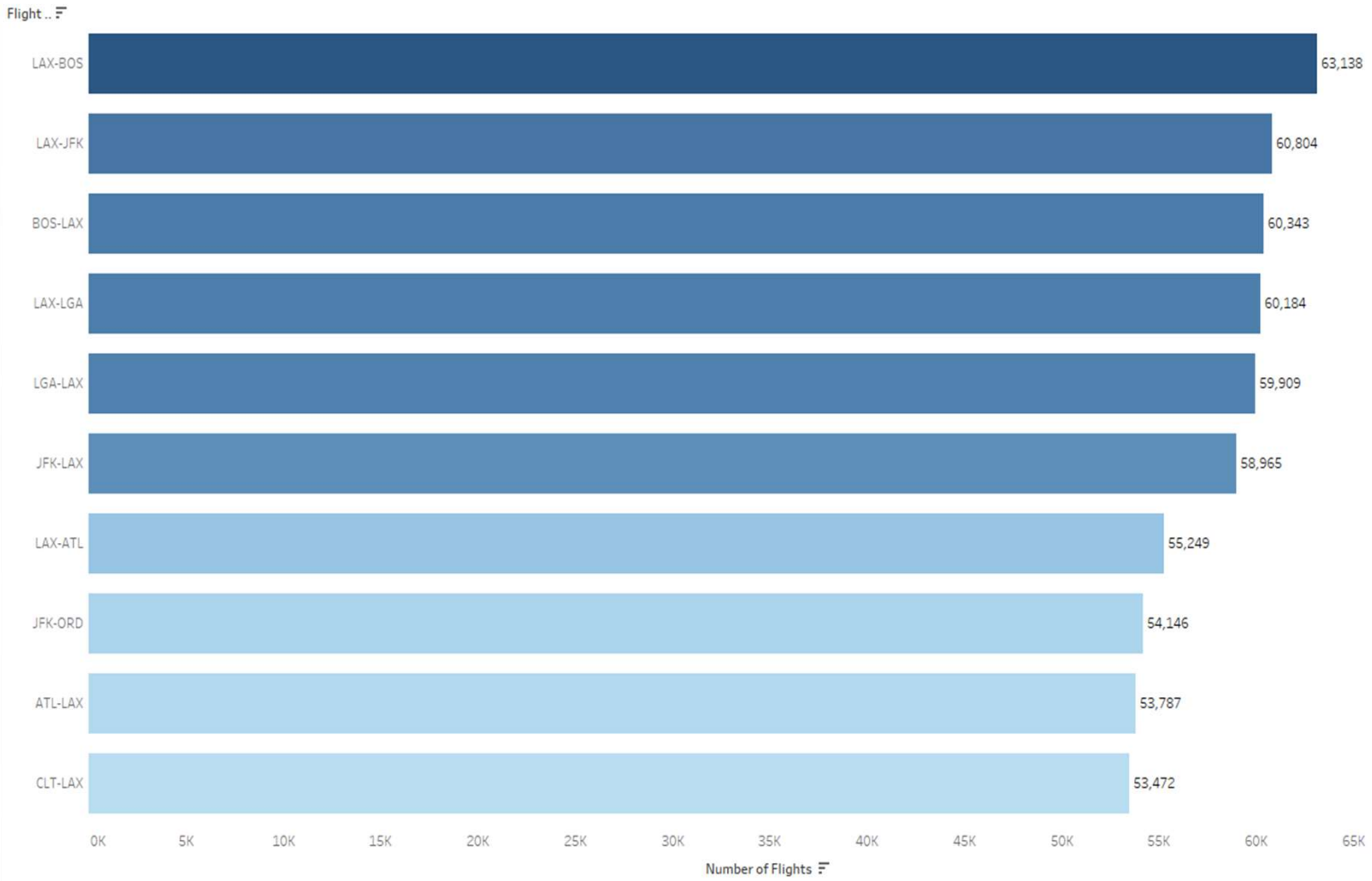
- An: Download & upload data into HDFS, create tables in Hive
- All: Perform data cleaning, write queries, create visualizations together during Teams meetings
- Bhumika, Ankita, Ragi: write analysis on visualizations

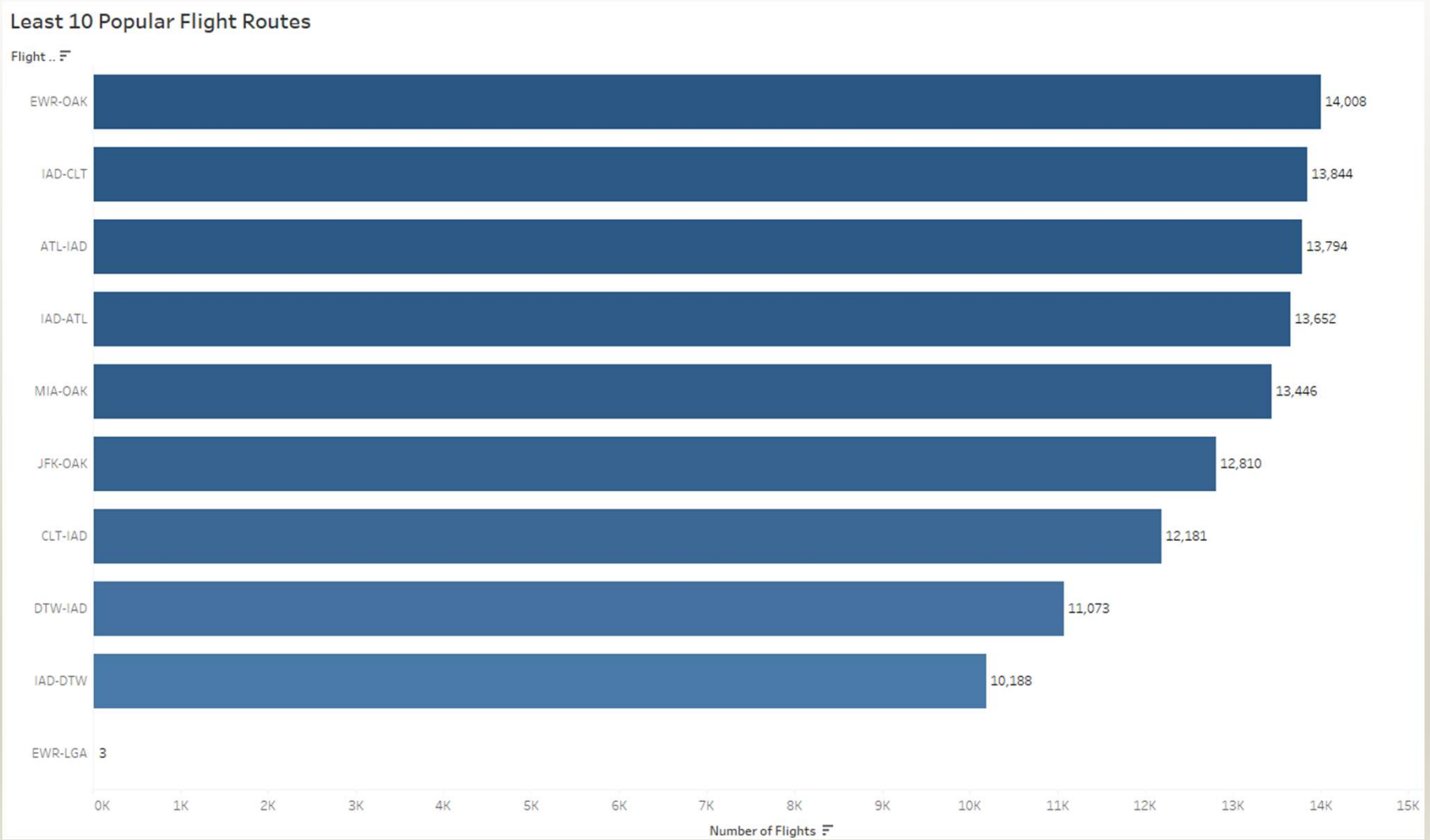
FLIGHTDATA2 TABLE (AFTER DATA CLEANING)

col_name	data_type
# col_name	data_type
flightid	string
searchdate	date
flightdate	date
startingairport	string
destinationairport	string
farebasiscode	string
travelduration	string
elapseddays	int
isbasiceconomy	boolean
isrefundable	boolean
isnonstop	boolean
basefare	double
totalfare	double
seatsremaining	int
totaltraveldistance	int
segmentsairlinename	string
segmentsequipmentdescription	string
flightmonth	int
flightroute	string

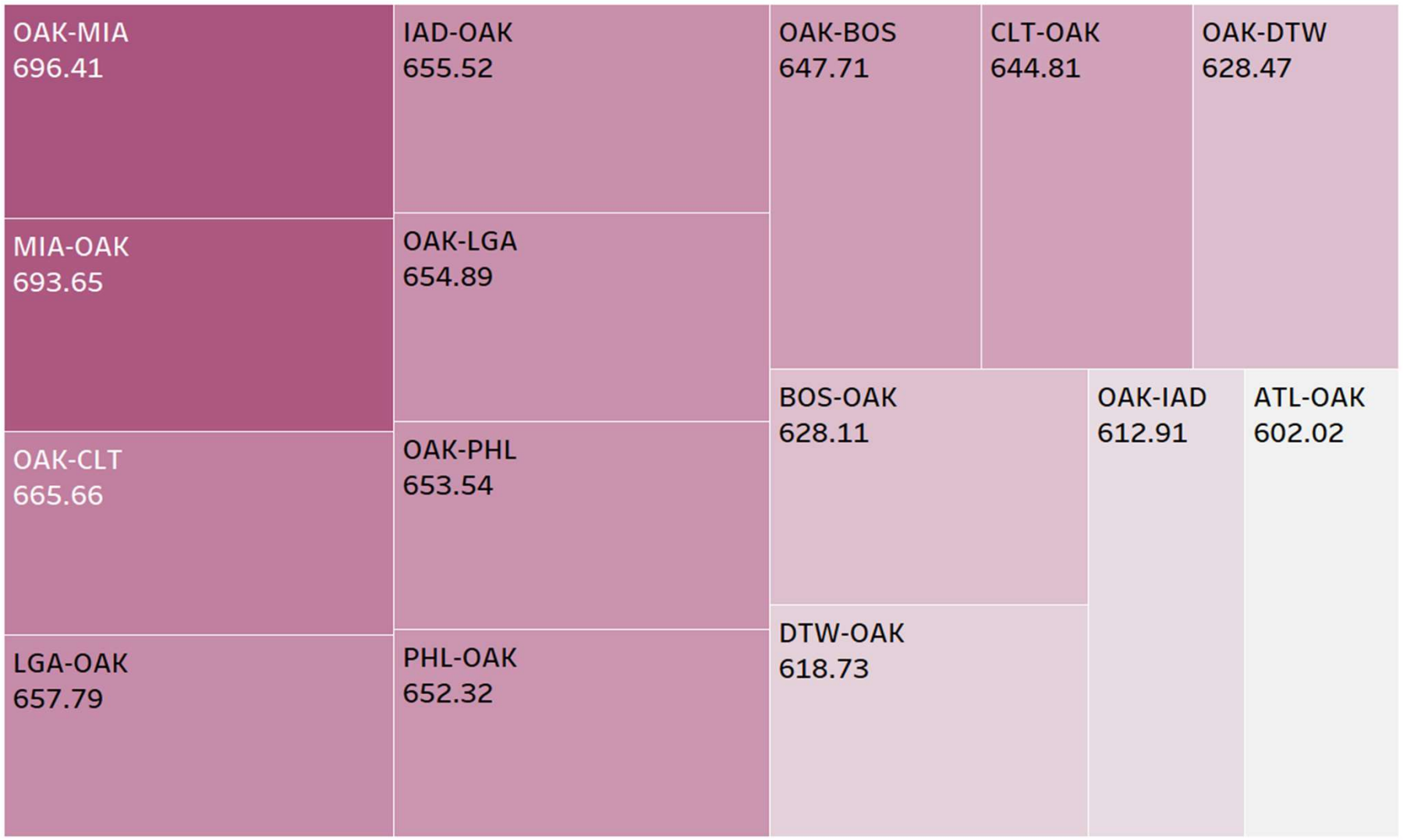

ANALYSIS
Q1
Q2
Q3
1,000
Q2
Q3
Q4

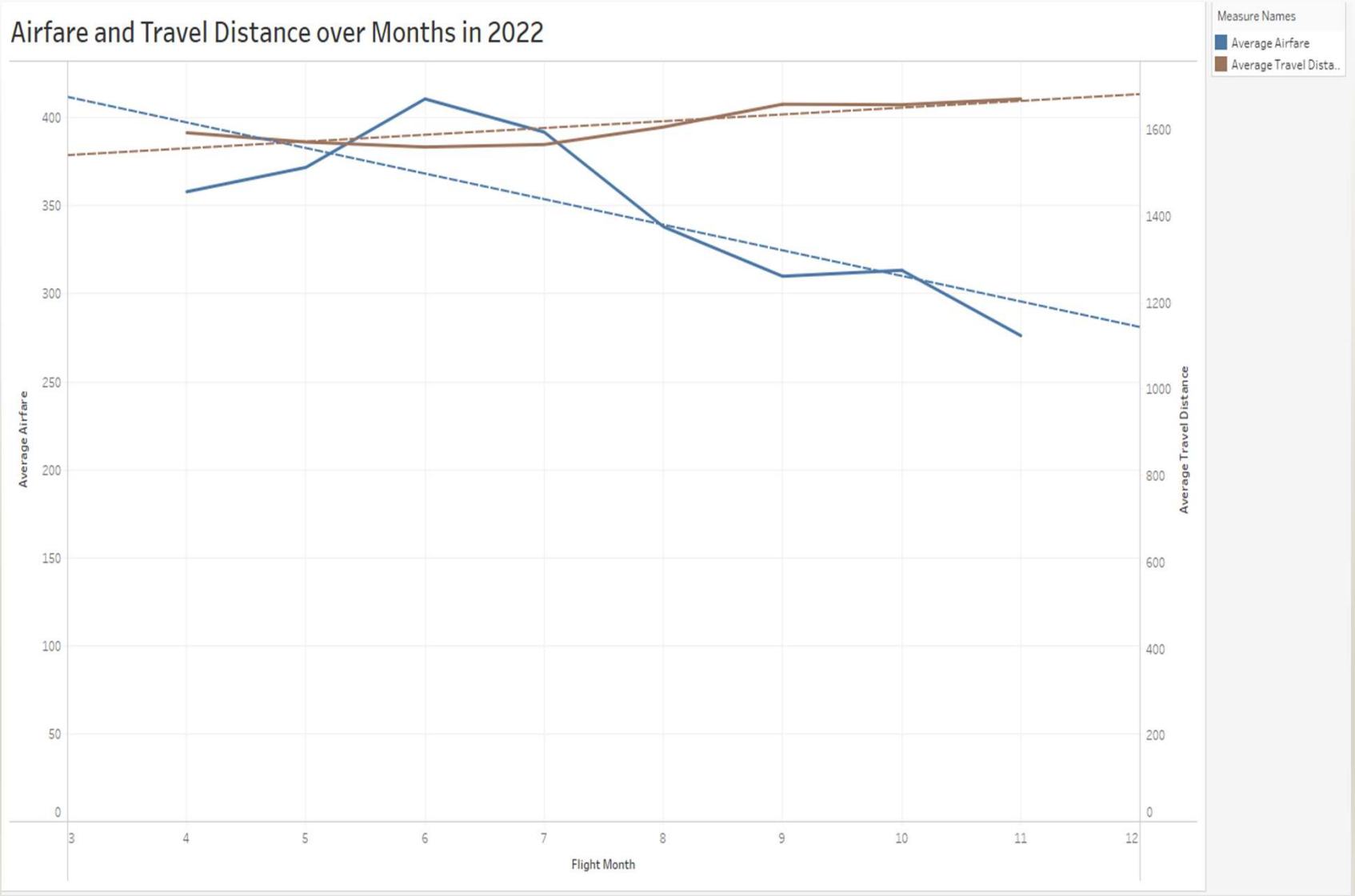
Top 10 Most Popular Flight Routes

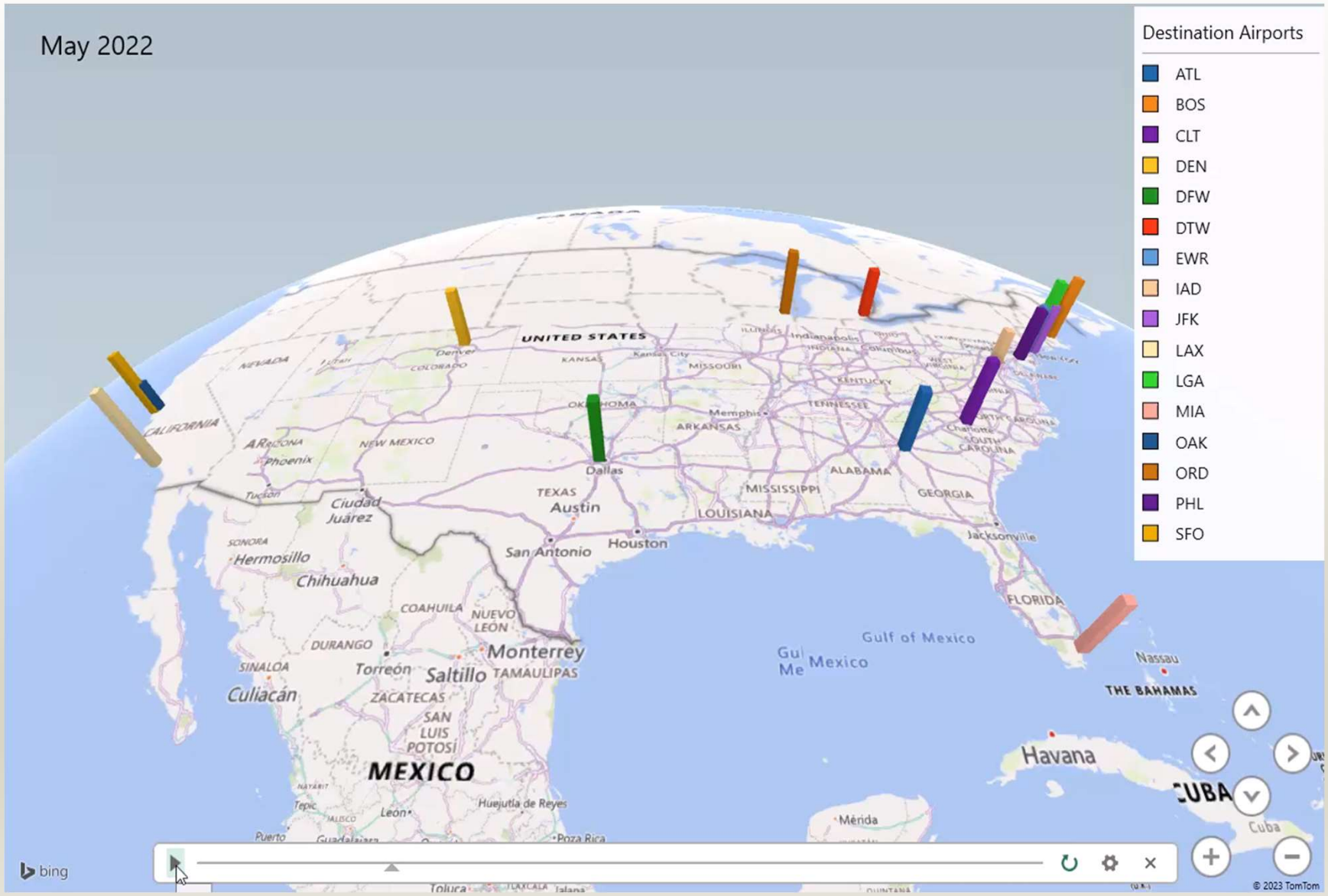


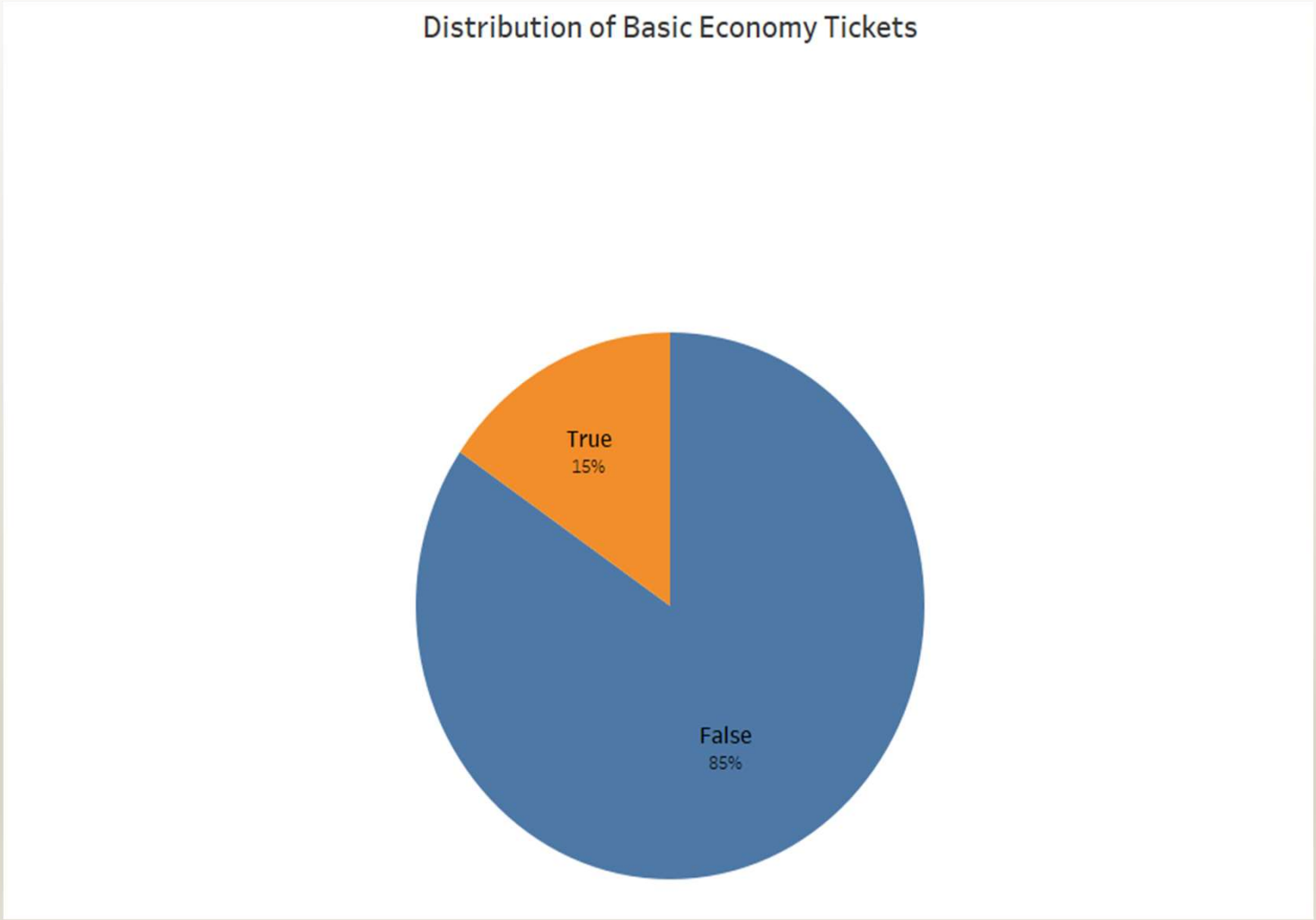

Least 10 Popular Flight Routes
Flight ..
EWR-OAK
IAD-CLT
ATL-IAD
IAD-ATL
MIA-OAK
JFK-OAK
CLT-IAD
DTW-IAD
IAD-DTW
EWR-LGA
14,008
13,844
13,794
13,652
13,446
12,810
12,181
11,073
10,188
3
0K
1K
2K
3K
4K
5K
6K
7K
8K
9K
10K
11K
12K
13K
14K
15K
Number of Flights

Top 15 Most Expensive Flight Routes




Airfare and Travel Distance over Months in 2022
Measure Names
Average Airfare
Average Travel Dista..
Average Airfare
Average Travel Distance
Flight Month
0
50
100
150
200
250
300
350
400
0
200
400
600
800
1000
1200
1400
1600
3
4
5
6
7
8
9
10
11
12


May 2022
Destination Airports
ATL
BOS
CLT
DEN
DFW
DTW
EWR
IAD
JFK
LAX
LGA
MIA
OAK
ORD
PHL
SFO
UNITED STATES
Denver
Kansas City
Indianapolis
Columbus
Memphis
Charlotte
Phoenix
Tucson
Dallas
Austin
Houston
San Antonio
Jacksonville
Ciudad Juárez
Hermosillo
Chihuahua
Monterrey
Torreón
Saltillo
Culiacán
Mazatlán
MEXICO
Gulf of Mexico
Nassau
THE BAHAMAS
Havana
CUBA
Cuba
Mérida
Huejutla de Reyes
Poza Rica
Toluca
bing
© 2023 TomTom


Distribution of Basic Economy Tickets
True
15%
False
85%

CONCLUSION

Data Analysis Summary:

i. The report highlights the top and least 10 flight routes based on the number of tickets sold.
ii. The report also identifies the top 15 most expensive flight routes, based on the ticket price.
iii. In June, airfare was found to be higher in comparison to the traveled distance.
iv. In August and September, the majority of destination airports experienced higher traffic.
v. It was observed that airlines offer only a small proportion of basic economy tickets.

Further research could be conducted by analyzing data from additional airports to gain a more comprehensive understanding of the trends in the airline industry.

THANK YOU