WEATHER IMPOSED FLIGHT DELAY ANALYSIS

Group 22 Project

AGENDA

- **Group 22 Team**
- **Background**
  - US Aviation Industry
  - Impacts of Flight Delay
- **Potential Business Questions**
- **Data Processing**
  - ETL Process
  - BDDS and DI Tools
  - Raw Data Extraction
  - Selected Columns
  - Procedure
  - Data Quality Check
  - Data Merging
- **Star Schema Design and FACTs Constellation**
- **Visualisations**
- **Insights and Recommendations**
- **References**
- **Appendix**

THE GROUP 22 TEAM

➢ Project Manager: Kabilan Rajendiran

He was chosen as the PM because of his leadership skill and previous experience in managing projects. He also had respective scores of 14 and 12 as a Coordinator and Completer Finisher on Belvin's Self-Perception Inventory. He oversaw the project, maintained project files, called for meetings, and ensured timelines were respected and adhered to.

➢ Data Analyst: Oyebanji Olusanya

He has experience working as a wireless network data analyst. He was appointed based on his experience and his self-evaluation result using Belvin's self-Perception Inventory. He extracted the raw datasets, cleaned and analysed the data.

➢ Data Engineer: Funmilayo Oyawoye

She was appointed based on her score on Belbin's Self-Perception which were 16 as a Team Worker and 15 as a shaper. She was supported to build and maintain the data infrastructure and ensure data quality.

➢ Visualisation Expert: Dhakirah Salahudin-Mukeen

She was appointed based on her experience as a data visualisation expert. She has worked with Tableau and PowerBI. She supported the project with the visualisation part by translating the analysed data to pictorial views and assisted in drawing out the underlining insights.

BACKGROUND

➢ US Aviation Industry:

- Carries 29.2% of the world's total air traffic
- 666.15 million passengers in 2021
- Commercial aviation drives 5% of the total US GDP
- Valued at $1.25 trillion
- According to Financial Times, over 300 flights were delayed in July 2021 due to smoke caused by fire from extreme heat in US Pacific Northwest alone.

➢ Impact of Flight Delay:

- Cost average $101.18 per minute for aircraft block time
- Need for extra gates and ground personnel
- Lost productivity, wages, and goodwill
- Cost estimated $28B in 2018 to airlines and passengers when their time is monetized

POTENTIAL BUSINESS QUESTIONS

POTENTIAL BUSINESS QUESTIONS

- What is the monthly breakdown of flights by airline and departure airport from 2013-2019?

Justification: This aids in identifying the busiest airlines and airports, which is helpful for allocating resources, developing infrastructure, and comprehending travel trends.

- What is the monthly breakdown of departure flight delays by airline, airport, and weather event from 2013-2019?

Justification: With this information, authorities may better understand how weather conditions affect departure delays and devise plans to reduce disruptions, increase operational effectiveness, and improve passenger experience in inclement weather.

- What are the top 10 popular destinations by flight and year between 2013 and 2019?

Justification: Airlines, airports, and tourism authorities can use this information to better understand demand trends, plan the best flight paths, and allocate resources.

- What is the monthly breakdown of arrival delay by airport and airline between 2013 and 2019?

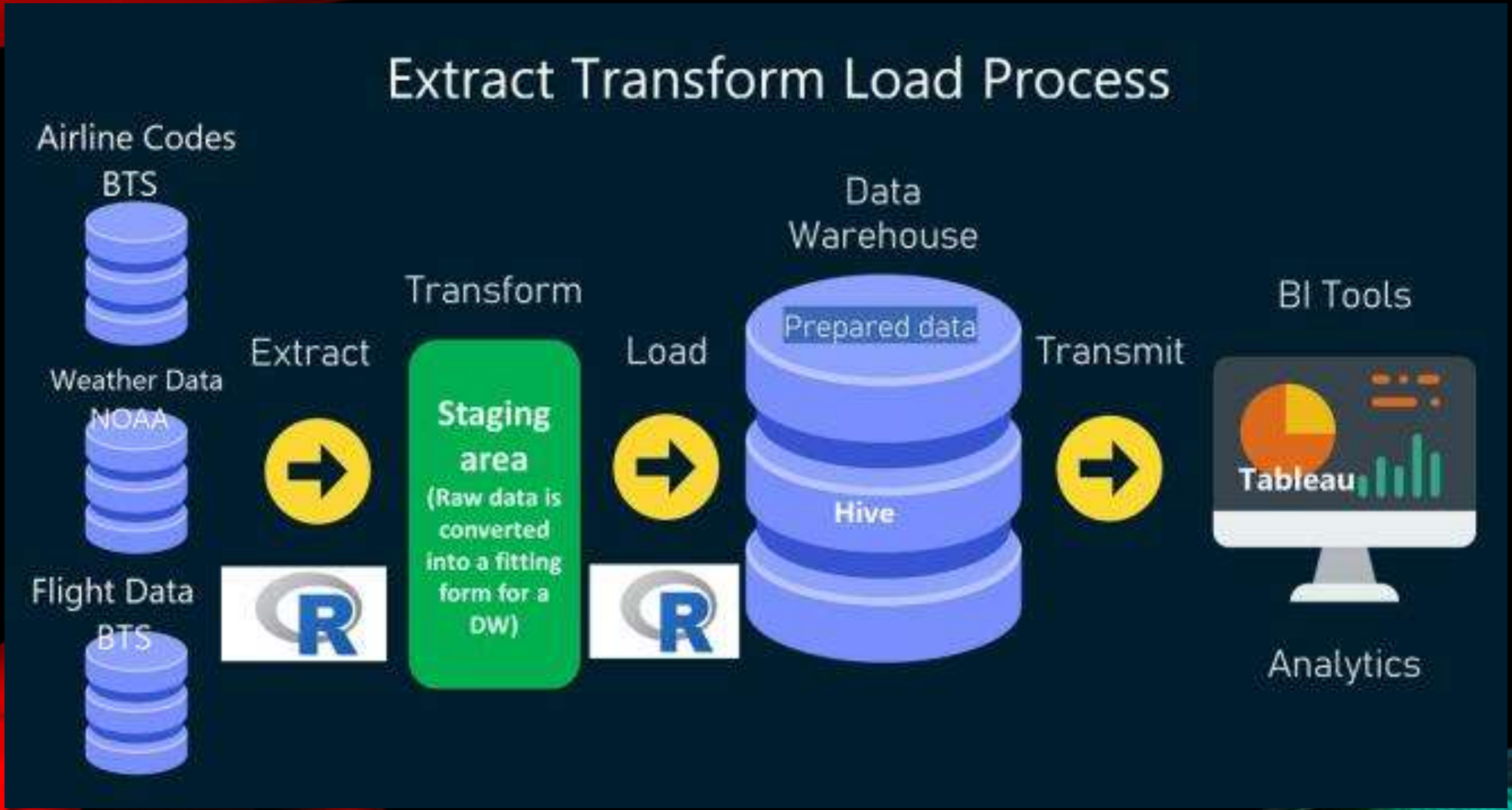
Justification: Aviation authorities may better manage resources, reduce delays, and improve passenger satisfaction by using this information to identify problem areas, optimise timetables, and boost operational efficiency.

- Which airport had the most arrival delay by weather events in 2019?

Justification: With the use of this information, authorities can put plans in place to lessen the effects of bad weather and create methods to increase safety, reduce delays, and enhance the entire travel experience.

DATA PROCESSING

ETL PROCESS



ETL was chosen over ELT to translate the data from the OLTP to OLAP because of the following:

- ✓ Supports fast analysis once the data has been structured
- ✓ Supports both on-premise and cloud-based environments
- ✓ Supports encryption of data before loading on the warehouse

BDDS AND DI TOOLS

Tool	Function in ETL	Similar Tools	Choice Reason
R/Rstudio	Data Extraction, Data cleaning, Data aggregation, Data Statistics and analysis, Data visualisation, Integrated programming environment, etc	Jupyter Notebook with Python, Spark with Scala	Familiarity, Ability to adequately handle the data size, Easy data manipulation and visualization, User friendly coding environment
Hadoop DFS	Distributed processing capability which allows for parallel processing	Google Cloud Big Query, Cloudera, Microsoft SQL Server, Apache Spark, Apache Flink, Databricks, Amazon EMR, Vertica, Snowflakes	Ability to quickly process massive volumes of data, It is an open source platform, It is scalable, resilient and flexible, Allows storage of any data format, both structured and unstructured data
Hive	Apache Hive data warehouse software facilitates SQL querying managing of large datasets residing in distributed storage	Amazon Redshift, Apache Impala, PrestoDB	It is scalable, flexible and cost efficient, It is user friendly, Fault tolerant, SQL query capability
Tableau	For data visualisation, manipulation, exploration, structuring, etc	Power BI, GoodData, SAS, Zebra BIInsightSquared	It is user-friendly, Ability to handle large dataset, Outstanding visualization, etc

CHOICE OF DATA WAREHOUSING APPROACH

Kimball	Inmon
Utilises dimensional model for data	Utilises dimensional model for data mart
Bottom-up approach	Top-down approach
Uses key business approach to solution	Uses corporate model approach
Analytical system can access data directly from the data warehouse	Analytical system can only access data from the data warehouse via the data mart
Less time to implement	Takes time to implement
Less expensive to implement	Highly expensive to implement
Requires a generalist team to implement	Requires a specialist team to implement
Focuses on individual the business area	Requires enterprise-wide data integration

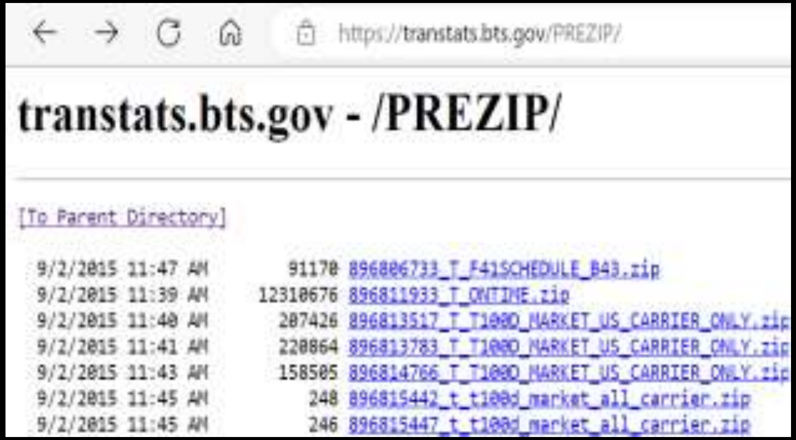
The team decided to use Kimball approach because of highlighted advantages over Inmon

RAW DATA EXTRACTION

- **Flight data extracted from the Bureau of Transportation Statistics (BTS)**
- **Weather data was extracted from the National Centre for Environmental Information**
- **7 years data from 2013 to 2019 was collected**

Flight Data

- ✓ Contains 109 columns
- ✓ Contains 43,928,883 rows



Weather Data

- ✓ Contains 51 columns
- ✓ Contains 420, 957 rows

Airline Code

- ✓ Contains 2 columns
- ✓ Contains 1653 rows

Sample_airlineCode

Justification for Choosing the Data Sources

The two referenced sources are government data open sources which makes them reliable and up to date

SELECTED COLUMNS

Flight Data (Columns: 14 Rows: 43,928,883)

Raw Data Column Name	Refined Data Column Name
Year	year
DestAirportID	destAirportID
OriginAirportID	originAirportID
ArrDelay	arrDelay
Month	month
DestCityName	destCity_name
OriginCityName	originCity_name
ArrDelayMinutes	arrDelay_minutes
DayofMonth	dayofMonth
DepDelay	depDelay
WeatherDelay	weatherDelay
IATA_CODE_Reporting_Airline	airlineCode
DepDelayMinutes	depDelay_minutes
Flights	flights

Weather Data (Columns: 4 Rows: 420,957)

Raw Data Column Name	Refined Data Column Name
BEGIN_YEARMONTH	yearMonth
BEGIN_DAY	day
EVENT_ID	eventID
EVENT_TYPE	eventType

Airline Code (Columns: 2 Rows: 1652)

Raw Data Column Name	Refined Data Column Name
Code	airlineCode
Airline	airlineName

DATA VALIDITY & QUALITY CHECK :: FLIGHT DATASET

Check for Missing data

- Number of missing values: 38,635,529 (total missing values in flight dataframe)
- Number of missing values: 687,574 (total missing values in DepDelay)
- Number of missing values: 687,574 (total missing values in DepDelayMinutes)
- Number of missing values: 819,936 (total missing values in ArrDelay)
- Number of missing values: 819,936 (total missing values in ArrDelayMinutes)
- Number of missing values: 35,620,509 (total missing values in weatherDelay)

➢ **Rows with missing values for weather delay were removed because they belong to other delay causes and the data reduces to:**

Columns: 14

Rows: 8,308,374

➢ **Then, the data frame was again checked for missing data :**

- Number of missing values: 44 (DepDelay column)
- Number of missing values: 44 (DepDelayMinutes column)

➢ **After removing rows with no values for DepDelay and DepDelayMinutes columns:**

Columns: 14

Rows: 8,308,330

➢ **Again, missing data was checked across the dataset and on each column. None was found**

```
> # Check for missing values in the dataset
> num_missing <- sum(is.na(df_flight))
> print(paste("Number of missing values:", num_missing))
[1] "Number of missing values: 38635529"
```

```
> # Check for missing values in the  ArrDelay column
> num_missing_ArrDelay <- sum(is.na(df_flight$ArrDelay))
> print(paste("Number of missing values:", num_missing_ArrDelay  ))
[1] "Number of missing values: 819936"
>
> # Check for missing values in the  ArrDelayMinutes column
> num_missing_ArrDelayMinutes <- sum(is.na(df_flight$ArrDelayMinutes))
> print(paste("Number of missing values:", num_missing_ArrDelayMinutes  ))
[1] "Number of missing values: 819936"
>
> # Check for missing values in the  DepDelay column
> num_missing_DelDelay <- sum(is.na(df_flight$DepDelay))
> print(paste("Number of missing values:", num_missing_DelDelay  ))
[1] "Number of missing values: 687574"
>
> # Check for missing values in the  DepDelayMiniutes column
> num_missing_DelDelayMinutes <- sum(is.na(df_flight$DepDelayMinutes))
> print(paste("Number of missing values:", num_missing_DelDelayMinutes  ))
[1] "Number of missing values: 687574"
>
> # Check for missing values in the  weatherDelay column
> num_missing_weatherDelay <- sum(is.na(df_flight$WeatherDelay))
> print(paste("Number of missing values:", num_missing_weatherDelay  ))
[1] "Number of missing values: 35620509"
```

```
> cat("Number of columns:", num_columns2, "\n")
Number of columns: 14
> cat("Number of rows:", num_rows2, "\n")
Number of rows: 8308374
> #Check the data for missing values again
>
> num_missing_tot2 <- sum(is.na(df_flight2))
> print(paste("Number of missing values:", num_missing_tot2))
[1] "Number of missing values: 88"
```

```
> cat("Number of columns:", num_columns3, "\n")
Number of columns: 14
> cat("Number of rows:", num_rows3, "\n")
Number of rows: 8308330
>
>
> #Check the data for missing values again
>
> num_missing_tot4 <- sum(is.na(df_flight3))
> print(paste("Number of missing values:", num_missing_tot4))
[1] "Number of missing values: 0"
```

DATA VALIDITY & QUALITY CHECK :: FLIGHT DATASET

Check for Duplicated data

- Number of duplicate rows: 2242

```
> # Check for duplicate rows
> num_duplicates <- sum(duplicated(df_flight3))
> print(paste("Number of duplicate rows:", num_duplicates))
[1] "Number of duplicate rows: 2242"
> # Check the number of rows after removing duplicates
> num_rows_unique <- nrow(df_flight_unique)
> print(paste("Number of rows after removing duplicates:", num_rows_unique))
[1] "Number of rows after removing duplicates: 8306088"
> # Check for duplicated rows again
> num_duplicates2 <- sum(duplicated(df_flight_unique))
> print(paste("Number of duplicate rows:", num_duplicates2))
[1] "Number of duplicate rows: 0"
>
```

➢ After removing duplicated rows, the dataset reduces

Columns: 14

Rows: 8,306,088

➢ Duplicated data was again checked across the dataset, and none was found

Check for incorrect data types

- Each column was checked for incorrect data types

```
+   print(paste("Column:", names(column_types)[i], "- Data Type:"
+ }
[1] "Column: Year - Data Type: integer"
[1] "Column: Month - Data Type: integer"
[1] "Column: DayofMonth - Data Type: integer"
[1] "Column: IATA_CODE_Reporting_Airline - Data Type: character"
[1] "Column: OriginCityName - Data Type: character"
[1] "Column: DestAirportID - Data Type: integer"
[1] "Column: DestCityName - Data Type: character"
[1] "Column: DepDelay - Data Type: numeric"
[1] "Column: DepDelayMinutes - Data Type: numeric"
[1] "Column: ArrDelay - Data Type: numeric"
[1] "Column: ArrDelayMinutes - Data Type: numeric"
[1] "Column: WeatherDelay - Data Type: numeric"
[1] "Column: Flights - Data Type: numeric"
[1] "Column: OriginAirportID - Data Type: integer"
```

DATA VALIDITY & QUALITY CHECK :: FLIGHT DATASET

Flight Dataset Summary

```
> summary(df_flight_unique)
      Year          Month          DayofMonth     IATA_CODE_Reporting_Airline OriginCityName     DestAirportID   DestCityName
 Min.   :2013   Min.   : 1.000   Min.   : 1.00   Length:8306088              Length:8306088     Min.   :10135   Length:8306088
 1st Qu.:2014   1st Qu.: 4.000   1st Qu.: 8.00   Class :character            Class :character   1st Qu.:11292   Class :character
 Median :2016   Median : 6.000   Median :16.00   Mode  :character            Mode  :character   Median :12892   Mode  :character
 Mean   :2016   Mean   : 6.451   Mean   :15.68                                                  Mean   :12703
 3rd Qu.:2018   3rd Qu.: 9.000   3rd Qu.:23.00                                                  3rd Qu.:14025
 Max.   :2019   Max.   :12.000   Max.   :31.00                                                  Max.   :16869
    DepDelay       DepDelayMinutes      ArrDelay        ArrDelayMinutes     WeatherDelay        Flights   OriginAirportID
 Min.   : -56.00   Min.   :   0.00   Min.   :  15.00   Min.   :  15.00   Min.   :   0.00   Min.   :1   Min.   :10135
 1st Qu.:  17.00   1st Qu.:  17.00   1st Qu.:  23.00   1st Qu.:  23.00   1st Qu.:   0.00   1st Qu.:1   1st Qu.:11292
 Median :  37.00   Median :  37.00   Median :  38.00   Median :  38.00   Median :   0.00   Median :1   Median :12889
 Mean   :  57.44   Mean   :  57.85   Mean   :  61.87   Mean   :  61.87   Mean   :   2.97   Mean   :1   Mean   :12651
 3rd Qu.:  72.00   3rd Qu.:  72.00   3rd Qu.:  73.00   3rd Qu.:  73.00   3rd Qu.:   0.00   3rd Qu.:1   3rd Qu.:13930
 Max.   :2710.00   Max.   :2710.00   Max.   :2695.00   Max.   :2695.00   Max.   :2692.00   Max.   :1   Max.   :16869
```

- Some extreme values are observed in DepDelay, DepDelayminutes, ArrDelay, ArrDelayMinutes and weatherDelay columns. These values are possible as flights can be delayed for more than 48 hours (2880 minutes).
- Other values are okay

DATA VALIDITY & QUALITY CHECK :: FLIGHT DATASET

Check for Outliers

- Each column was checked for outliers
- Depdelay, DepDelayMinutes, ArrDelay , ArrDelayMinutes, WeatherDelay contain large values

Check for Inappropriate Values

- Each column was checked for inappropriate values
- Column DepDelay has 791,636 negative values which means the flight departed earlier than scheduled
- These rows and those equal to zero are the early and the on-time flights which are not our concern. Hence, we removed rows whose DepDelay and ArrDelay are zero and less than zero and checked for missing values. Then, the data reduces to:

Columns: 14 Rows: 7,383,251

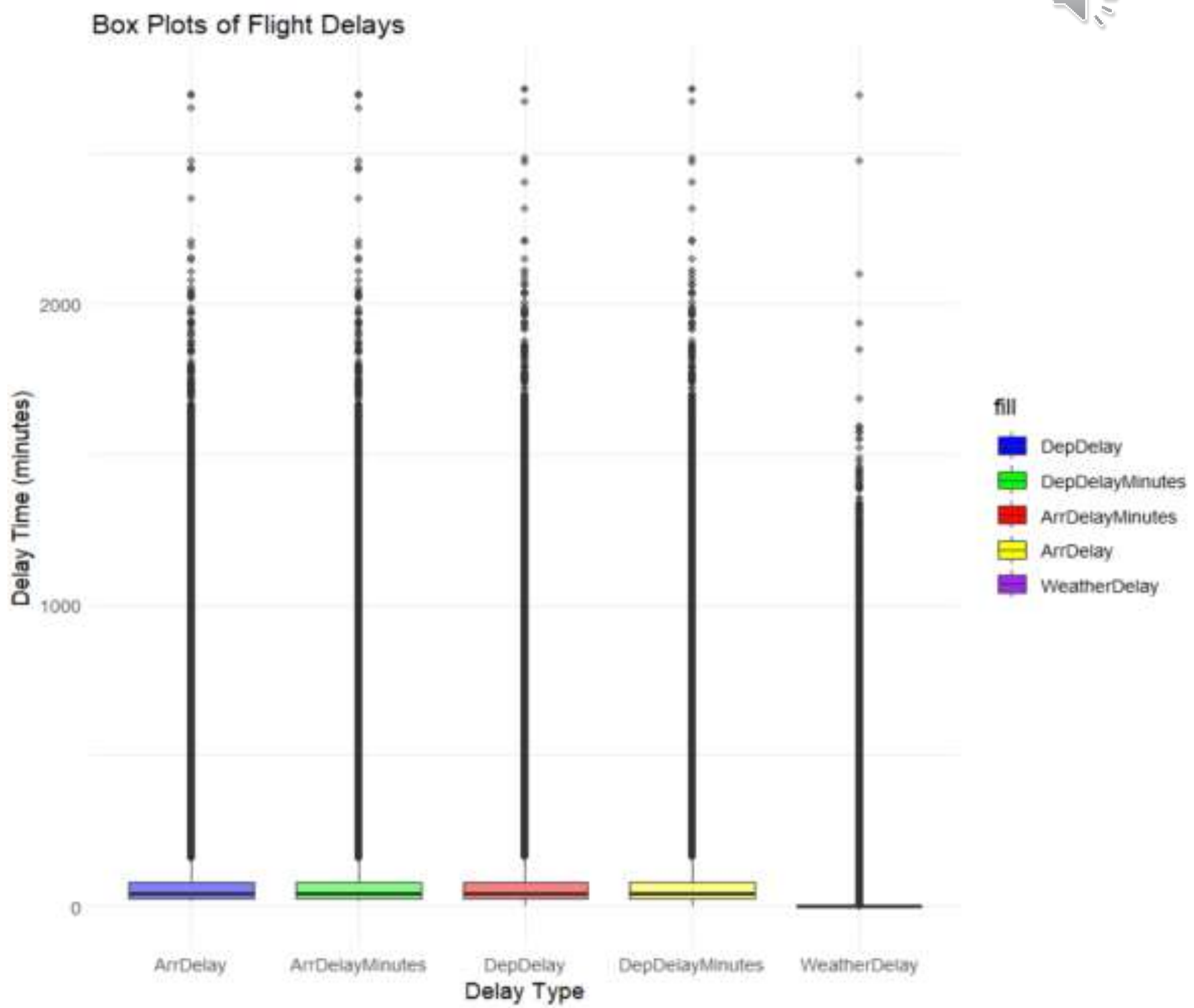
```
>
> # Call the function to check the inappropriate values in the flight data
> check_inappropriate_values(df_flight)
Column DepDelay contains negative values.
> # Checking for the total negative values in the column DepDelay
> negative_dep_delay <- sum(df_flight_unique$DepDelay < 0, na.rm = TRUE)
>
> # Print the result
> cat("Number of negative values in DepDelay column:", negative_dep_delay, "\n")
Number of negative values in DepDelay column: 791636
```

```
> print(paste("Number of missing values:", num_missing_df_flight_delay))
[1] "Number of missing values: 0"
> num_columns_delay <- ncol(df_flight_delay)
> num_rows_delay <- nrow(df_flight_delay)
>
> # Print the results
> cat("Number of columns:", num_columns_delay, "\n")
Number of columns: 14
> cat("Number of rows:", num_rows_delay, "\n")
Number of rows: 7383251
```

DATA VALIDITY QUALITY CHECK:: FLIGHT DATASET

Check for Outliers

- Each column was checked for outliers
- Depdelay, DepDelayMinutes, ArrDelay , ArrDelayMinutes, WeatherDelay contain large values
- Eradicating these extreme values means that we will lose representation for flight delays in this category. Hence, analysis was carried with them



DATA VALIDITY QUALITY CHECK:: FLIGHT DATASET

Flight dataset aggregated by month

```
> head(df_flight_delay_aggregated)
# A tibble: 6 × 14
# Groups:   year [1]
   year month airlineCode originCity_name       destAirportID destCity_name   depDelay depDelay_minutes arrDelay
  <int> <int> <chr>       <chr>                         <int> <chr>              <dbl>            <dbl>    <dbl>
1  2013     1 9E          Dallas/Fort Worth, TX         12478 New York, NY     4318003          4318003  4483222
2  2013     2 DL          Cincinnati, OH                10397 Atlanta, GA      4037376          4037376  4212325
3  2013     3 DL          Atlanta, GA                   14057 Portland, OR     5283047          5283047  5425699
4  2013     4 WN          Manchester, NH                10821 Baltimore, MD    6058777          6058777  6335080
5  2013     5 DL          Phoenix, AZ                   10397 Atlanta, GA      5599059          5599059  5814848
6  2013     6 9E          Detroit, MI                   11278 Washington, DC   8694910          8694910  9033144
# ℹ 5 more variables: arrDelay_minutes <dbl>, weatherDelay <dbl>, flights <dbl>, originAirportID <int>,
#   year_month <chr>
```

```
> #Check the aggregated flight data for missing values again
> num_missing_df_flight_delay <- sum(is.na(df_flight_delay))
> print(paste("Number of missing values:", num_missing_df_flight_delay))
[1] "Number of missing values: 0"
```

```
> num_duplicates_aggregated <- df_flight_delay[duplicated(df_flight_delay) | duplicate
> print(paste("Number of duplicate rows:", num_duplicates_aggregated))
 [1] "Number of duplicate rows: integer(0)"   "Number of duplicate rows: integer(0)"
 [3] "Number of duplicate rows: integer(0)"   "Number of duplicate rows: character(0)"
 [5] "Number of duplicate rows: character(0)" "Number of duplicate rows: integer(0)"
 [7] "Number of duplicate rows: character(0)" "Number of duplicate rows: numeric(0)"
 [9] "Number of duplicate rows: numeric(0)"   "Number of duplicate rows: numeric(0)"
[11] "Number of duplicate rows: numeric(0)"   "Number of duplicate rows: numeric(0)"
[13] "Number of duplicate rows: numeric(0)"   "Number of duplicate rows: integer(0)"
```

No missing or duplicated values were found in the aggregated Flight dataset

DATA VALIDITY & QUALITY CHECK :: WEATHER DATASET

Check for missing values

- No missing values were found

```
> # Check for missing values
> num_missing_weather <- sum(is.na(df_weather))
> print(paste("Number of missing values:", num_missing_weather))
[1] "Number of missing values: 0"
>
> #Check for missing values in BEGIN_YEARMONTH column
> num_missing_beginYear <- sum(is.na(df_weather$BEGIN_YEARMONTH))
> print(paste("Number of missing values:", num_missing_beginYear))
[1] "Number of missing values: 0"
>
> #Check for missing values in BEGIN_DAY column
> num_missing_beginDay <- sum(is.na(df_weather$BEGIN_DAY))
> print(paste("Number of missing values:", num_missing_beginDay))
[1] "Number of missing values: 0"
>
> #Check for missing values in EVENT_ID column
> num_missing_eventID <- sum(is.na(df_weather$EVENT_ID))
> print(paste("Number of missing values:", num_missing_eventID))
[1] "Number of missing values: 0"
>
> #Check for missing values in EVENT_TYPE column
> num_missing_eventType <- sum(is.na(df_weather$EVENT_TYPE))
> print(paste("Number of missing values:", num_missing_eventType))
[1] "Number of missing values: 0"
```

DATA VALIDITY QUALITY CHECK:: WEATHER DATASET

Check for duplicated rows

- ➢ No duplicated rows were found

```
> # Check for duplicate rows
> num_duplicates_weather <- sum(duplicated(df_weather))
> print(paste("Number of duplicate rows:", num_duplicates_weather))
[1] "Number of duplicate rows: 0"
```

Check for data types across the columns

- ➢ Appropriate data type found for each column

```
> # Check the data type for each column
> column_types_weather <- sapply(df_weather, class)
>
> # Print the column names and their corresponding data types
>   for (i in seq_along(column_types_weather)) {
+   print(paste("Column:", names(column_types_weather)[i], "- Data Type:", column_types_weather[i]))
+   }
[1] "Column: BEGIN_YEARMONTH - Data Type: integer"
[1] "Column: BEGIN_DAY - Data Type: integer"
[1] "Column: EVENT_ID - Data Type: integer"
[1] "Column: EVENT_TYPE - Data Type: character"
```

DATA VALIDITY QUALITY CHECK:: WEATHER DATASET

Summary of the Weather dataset

- All columns have appropriate values

```
> summary(df_weather)
   BEGIN_DAY          EVENT_ID          EVENT_TYPE              Year          Month
 Min.   : 1.00    Min.   : 418371    Length:420957        Min.   :2013   Min.   : 1.000
 1st Qu.: 7.00    1st Qu.: 536476    Class :character     1st Qu.:2014   1st Qu.: 4.000
 Median :15.00    Median : 646813    Mode  :character     Median :2016   Median : 6.000
 Mean   :15.17    Mean   : 646633                         Mean   :2016   Mean   : 5.919
 3rd Qu.:23.00    3rd Qu.: 756692                         3rd Qu.:2018   3rd Qu.: 8.000
 Max.   :31.00    Max.   :1092656                         Max.   :2019   Max.   :12.000
```

Weather dataset aggregated by month

```
> head(df_weather_aggregated)
# A tibble: 6 x 4
# Groups:   year [1]
   year month eventID eventType
  <dbl> <dbl>   <int> <chr>
1  2013     1  427840 High Wind
2  2013     2  436163 Winter Weather
3  2013     3  440167 Heavy Snow
4  2013     4  455688 Flood
5  2013     5  456844 Drought
6  2013     6  453536 Tropical Storm
>
```

DATA VALIDITY & QUALITY CHECK:: AIRLINE CODE DATASET

No issue found on this dataset

```
>
> #Quality data check on the airline code
>
> # Check for missing values
> num_missing_airlineCode <- sum(is.na(df_airlineCode))
> print(paste("Number of missing values:", num_missing_airlineCode))
[1] "Number of missing values: 0"
>
> #Check for missing values in code column
> num_missing_Code <- sum(is.na(df_airlineCode$Code))
> print(paste("Number of missing values:", num_missing_Code))
[1] "Number of missing values: 0"
>
> #Check for missing values in code column
> num_missing_Airline <- sum(is.na(df_airlineCode$Airline))
> print(paste("Number of missing values:", num_missing_Airline))
[1] "Number of missing values: 0"
>
>
> # Check for duplicated rows
> num_duplicates_airlinecode <- sum(duplicated(df_airlineCode))
> print(paste("Number of duplicate rows:", num_duplicates_airlinecode))
[1] "Number of duplicate rows: 0"
>
```

MERGING OF DATASETS

Combined dataset

```
> head(cleaned_data)
  timeID year month airlineID       airlineName       originCityname destAirportID                    destCityname
1 201301 2013     1        9E Endeavor Air Inc. Dallas/Fort Worth, TX         12478                    New York, NY
2 201306 2013     6        9E Endeavor Air Inc.           Detroit, MI         11278                  Washington, DC
3 201307 2013     7        9E Endeavor Air Inc.           Memphis, TN         11433                     Detroit, MI
4 201308 2013     8        9E Endeavor Air Inc.           Memphis, TN         11433                     Detroit, MI
5 201810 2018    10        9E Endeavor Air Inc.           Atlanta, GA         11617 New Bern/Morehead/Beaufort, NC
6 201310 2013    10        9E Endeavor Air Inc.          New York, NY         11057                   Charlotte, NC
  originAirportID depDelay depDelayminutes arrDelay arrDelayminutes weatherDelay flights eventID          eventType
1           11298  4318003         4318003  4483222         4483222       232681   75964  427840          High Wind
2           11433  8694910         8694910  9033144         9033144       507791  131832  453536     Tropical Storm
3           13244  8107008         8107008  8382190         8382190       341823  130193  468057     Excessive Heat
4           13244  5708567         5708567  5841468         5841468       205699  102448  474194          Dense Fog
5           10397  5505389         5505389  5652103         5652103       274188   86193  785047 Thunderstorm Wind
6           12953  3655390         3655390  3687378         3687378        77180   71640  473492        Strong Wind
```

No missing or duplicated values were found. No inappropriate data types were found

```
>
> # Print the results
> cat("Number of columns:", num_columns_cleaned_data, "\n")
Number of columns: 17
> cat("Number of rows:", num_rows_w_cleaned_data, "\n")
Number of rows: 84
>
```

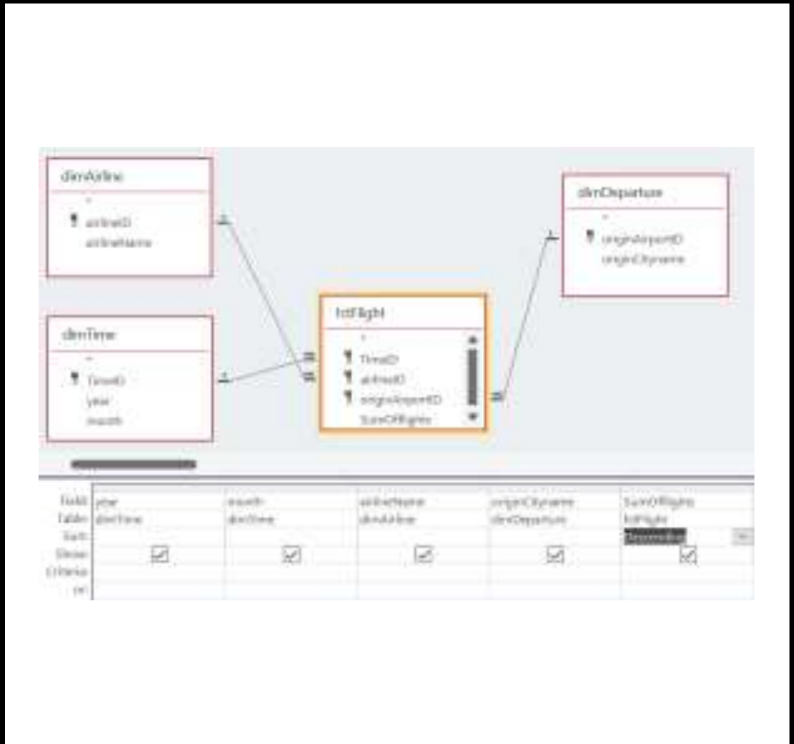
Validity & Quality Check on the Combined Dataset

```
> #Carry out quality data check on the new cleaned data
>
> # Check for missing values
> num_missing_cleaned_data <- sum(is.na(cleaned_data))
> print(paste("Number of missing values:", num_missing_cleaned_data))
[1] "Number of missing values: 0"
>
>
> # Check for duplicate rows
> num_duplicates_cleaned_data <- sum(duplicated(cleaned_data))
> print(paste("Number of duplicate rows:", num_duplicates_cleaned_data))
[1] "Number of duplicate rows: 0"
```

```
> # Check the data type for each column
> column_types_cleaned_data <- sapply(cleaned_data, class)
> column_types_cleaned_data
         timeID            year           month       airlineID     airlineName  originCityname   destAirportID
    "character"       "integer"       "integer"     "character"     "character"     "character"       "integer"
   destCityname originAirportID        depDelay depDelayminutes        arrDelay arrDelayminutes    weatherDelay
    "character"       "integer"       "numeric"       "numeric"       "numeric"       "numeric"       "numeric"
        flights         eventID       eventType
      "numeric"       "integer"     "character"
```

STAR SCHEMA DESIGN & FACTS CONSTELLATION

BUSINESS QUESTION STAR SCHEMA

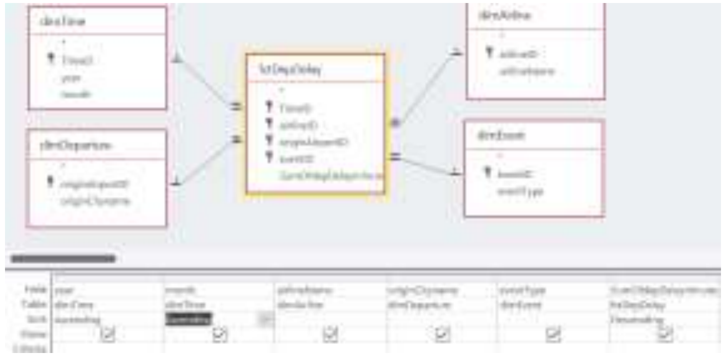


year	month	airlineName	originCityname	SumOfFlights
2019	6	Frontier Airlines Inc.	Orlando, FL	137934
2013	8	Endeavor Air Inc.	Detroit, MI	131832
2013	12	Delta Air Lines Inc.	Atlanta, GA	131457
2013	7	Endeavor Air Inc.	Memphis, TN	130193
2018	7	American Airlines Inc.	Miami, FL	127712
2018	8	United Air Lines Inc.	San Francisco, CA	127126
2019	7	Allegiant Air	St. Petersburg, FL	121787
2018	6	American Airlines Inc.	Los Angeles, CA	110777
2014	6	American Airlines Inc.	New York, NY	118759
2019	8	Delta Air Lines Inc.	Atlanta, GA	118547
2019	12	Delta Air Lines Inc.	Atlanta, GA	112574
2019	5	Republic Airways	Minneapolis, MN	112532
2014	1	American Airlines Inc.	Dallas/Fort Worth, TX	109586
2014	7	American Airlines Inc.	New York, NY	107077
2015	6	Spirit Air Lines	Detroit, MI	105901
2018	5	American Airlines Inc.	Portland, OR	105432
2016	7	American Airlines Inc.	Boston, MA	103801
2019	2	PSA Airlines Inc.	Dayton, OH	103186
2013	8	Endeavor Air Inc.	Memphis, TN	102448

BQ1:

What is the monthly breakdown of flights by airline and departure airport from 2013-2019?

BUSINESS QUESTION STAR SCHEMA

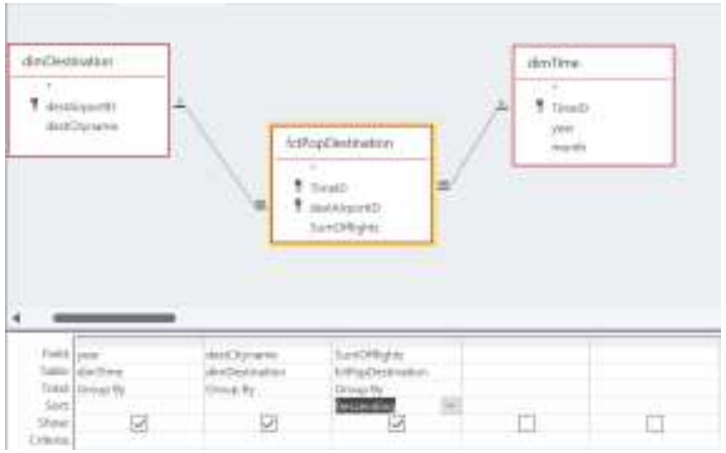


year	month	airlineName	originCityNar	eventType	SumOfDepDelar
2013	1	Endeavor Air In	Dallas/Fort Wo	High Wind	4318003
2013	2	Delta Air Lines I	Cincinnati, OH	Winter Weathe	4037176
2013	3	Delta Air Lines I	Atlanta, GA	Heavy Snow	5293047
2013	4	Southwest Airli	Manchester, NI	Flood	6058777
2013	5	Delta Air Lines I	Phoenix, AZ	Drought	5599056
2013	6	Endeavor Air In	Detroit, MI	Tropical Storm	8694810
2013	7	Endeavor Air In	Memphis, TN	Excessive Heat	8107008
2013	8	Endeavor Air In	Memphis, TN	Dense Fog	5708567
2013	9	Endeavor Air In	Buffalo, NY	Flash Flood	3914528
2013	10	Endeavor Air In	New York, NY	Strong Wind	3655390
2013	11	Endeavor Air In	New York, NY	High Wind	3697866
2013	12	Delta Air Lines I	Atlanta, GA	Heavy Snow	7261821
2014	1	American Airlin	Dallas/Fort Wo	Heavy Snow	7125922
2014	2	American Airlin	New York, NY	Heavy Snow	5334934
2014	3	Southwest Airli	Salt Lake City, U	Strong Wind	4973122
2014	4	American Airlin	New York, NY	Strong Wind	4514499
2014	5	American Airlin	Dallas/Fort Wo	Hail	5716537
2014	6	American Airlin	New York, NY	Hail	7150009

BQ2:

What is the monthly breakdown of departure flight delays by airline, airport and weather event from 2013-2019?

BUSINESS QUESTION STAR SCHEMA

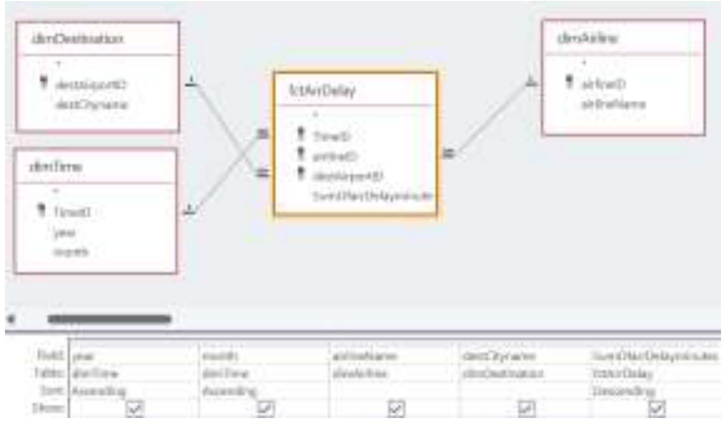


year	destCityname	SumOfflights
2019	Nashville, TN	137934
2013	Washington, DC	131832
2013	Savannah, GA	131457
2013	Detroit, MI	130193
2018	Philadelphia, PA	127712
2018	Newark, NJ	127126
2019	Asheville, NC	121787
2018	Indianapolis, IN	119777
2014	Los Angeles, CA	118759
2019	Dallas/Fort Worth, TX	118547

BQ3:

What are the top 10 popular destination by flight and year between 2013 and 2019?

BUSINESS QUESTION STAR SCHEMA

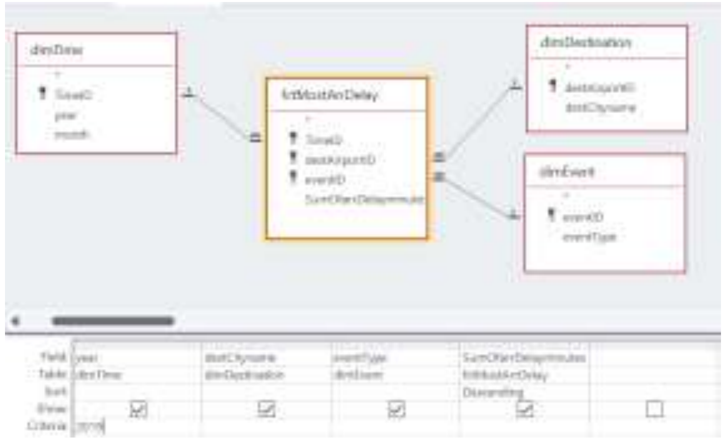


year	month	airlineName	destCitynam	SumOfarrDelayminutes
2013	1	Endeavor Air In	New York, NY	4483222
2013	2	Delta Air Lines I	Atlanta, GA	4212325
2013	3	Delta Air Lines I	Portland, OR	5425699
2013	4	Southwest Airli	Baltimore, MD	6335080
2013	5	Delta Air Lines I	Atlanta, GA	5814848
2013	6	Endeavor Air In	Washington, DC	9033144
2013	7	Endeavor Air In	Detroit, MI	8382190
2013	8	Endeavor Air In	Detroit, MI	5841468
2013	9	Endeavor Air In	New York, NY	3988215
2013	10	Endeavor Air In	Charlotte, NC	3687378
2013	11	Endeavor Air In	Nashville, TN	3720166
2013	12	Delta Air Lines I	Savannah, GA	7577086
2014	1	American Airlin	Wichita, KS	7400972
2014	2	American Airlin	Los Angeles, CA	5547594
2014	3	Southwest Airli	Denver, CO	5097404
2014	4	American Airlin	Los Angeles, CA	4627587
2014	5	American Airlin	Salt Lake City, U	5850418
2014	6	American Airlin	Los Angeles, CA	7452107
2014	7	American Airlin	Los Angeles, CA	6592463
2014	8	American Airlin	Miami, FL	5776184
2014	9	SkyWest Airline	Los Angeles, CA	4090069

BQ4:

What is the monthly breakdown of arrival delay by airport and airline between 2013 and 2019?

BUSINESS QUESTION STAR SCHEMA

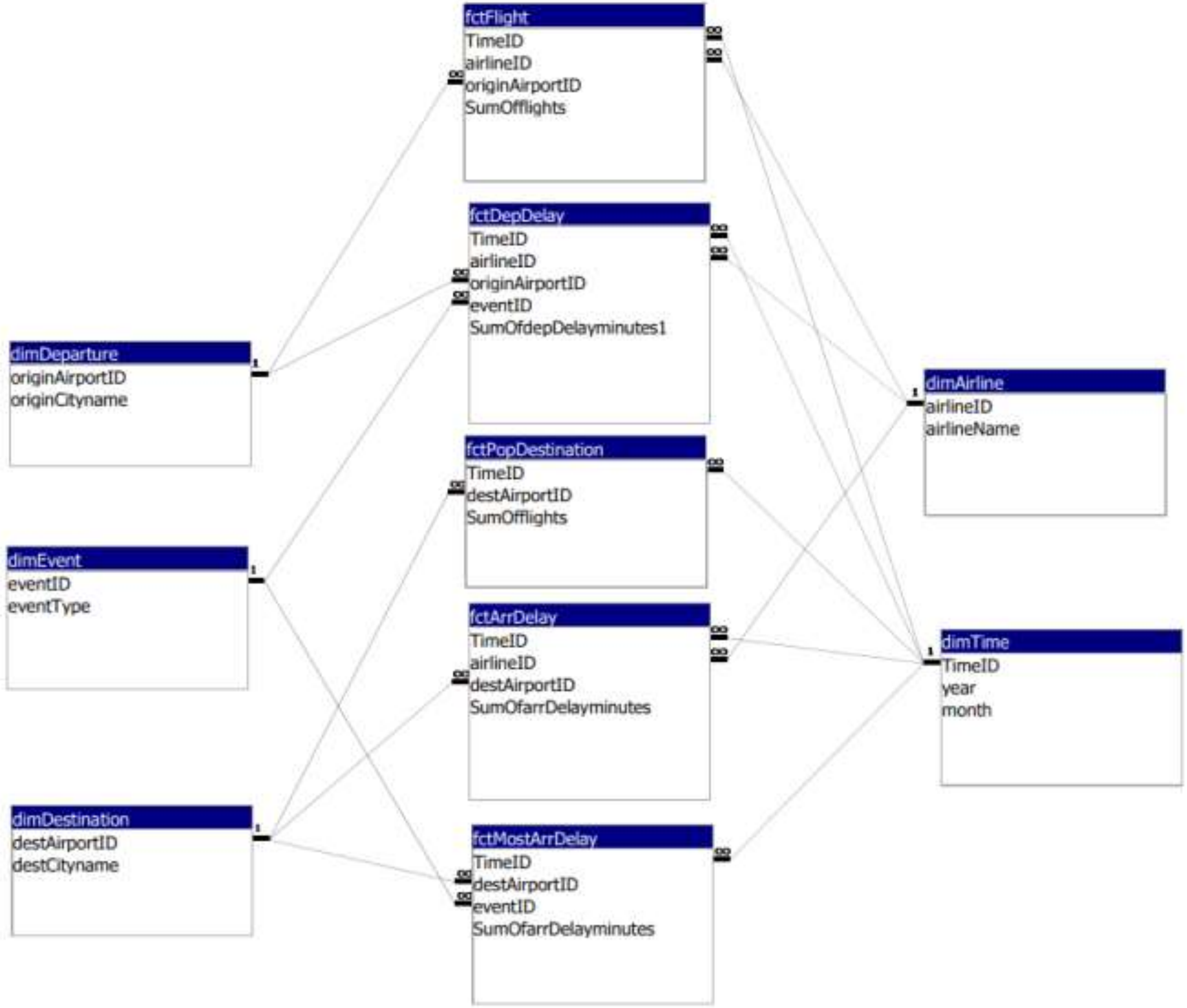


year	destCitynam	eventType	SumOfarrDe
2019	Nashville, TN	Thunderstorm	10640070
2019	Asheville, NC	Hail	9912987
2019	Dallas/Fort Wo	Flood	9129018
2019	Newark, NJ	Flash Flood	8473309
2019	Newark, NJ	Heavy Rain	8191076
2019	Philadelphia, PA	Ice Storm	7797958
2019	Chicago, IL	Hail	7346393
2019	Richmond, VA	Winter Weathe	6752647
2019	Charlotte, NC	Thunderstorm	6561634
2019	Houston, TX	Heavy Rain	6177794
2019	Philadelphia, PA	High Wind	5230550
2019	Charlotte, NC	High Wind	5026435

BQ5:

Which airport had the most arrival delay by weather events in 2019?

FACTS CONSTELLATION



DATA MART AND R/HIVE CONNECTION

```
+----------------------+-------------------------+----------------------------+-------------------------------+-----------------------------------+---------------------------------------+-----
| cleaned_data.timeid  | cleaned_data.year  | cleaned_data.month  | cleaned_data.airlineid  | cleaned_data.airlinename  | cleaned_data.origincityname  | clea
ned_data.destairportid  |    cleaned_data.destcityname    | cleaned_data.originairportid  | cleaned_data.depdelay  | cleaned_data.depdelayminutes  | cleaned_
data.arrdelay  | cleaned_data.arrdelayminutes  | cleaned_data.weatherdelay  | cleaned_data.flights  | cleaned_data.eventid  | cleaned_data.eventtype  |
+----------------------+-------------------------+----------------------------+-------------------------------+-----------------------------------+---------------------------------------+-----
| 201301               | 2013                    | 1                          | 9E                            | Endeavor Air Inc.                 | Dallas/Fort Worth. TX                 | 1247
8                        | New York. NY          | 11298                       | 4318003                       | 4318003                           | 4483222
          | 4483222                  | 232681                     | 75964                         | 100                               | High Wind                             |
| 201306               | 2013                    | 6                          | 9E                            | Endeavor Air Inc.                 | Detroit. MI                           | 1127
8                        | Washington. DC        | 11433                       | 8694910                       | 8694910                           | 9033144
          | 9033144                  | 507791                     | 131832                        | 101                               | Tropical Storm                        |
| 201307               | 2013                    | 7                          | 9E                            | Endeavor Air Inc.                 | Memphis. TN                           | 1143
3                        | Detroit. MI           | 13244                       | 8107008                       | 8107008                           | 8382190
          | 8382190                  | 341823                     | 130193                        | 102                               | Excessive Heat                        |
| 201308               | 2013                    | 8                          | 9E                            | Endeavor Air Inc.                 | Memphis. TN                           | 1143
3                        | Detroit. MI           | 13244                       | 5708567                       | 5708567                           | 5841468
          | 5841468                  | 205699                     | 102448                        | 103                               | Dense Fog                             |
| 201810                                                                                                                                                                            | 1161
7                                                                                                                                                                                 5652103
+-------
```

```
39-8a9e-e98d8ae87e0e): Time taken: 0.03 seconds
INFO  : OK
+------------------------------+
|           tab_name           |
+------------------------------+
| cleaned_data                 |
| dimairline                   |
| dimdeparture                 |
| dimdestination               |
| dimevent                     |
| dimtime                      |
| fctarrdelay                  |
| fctdeparrdelay               |
| fctdepdelay                  |
| fctflight                    |
| fctmostarrdelay              |
| fctpopdestination            |
| monthly_arrival_delay_breakdown |
| monthly_dep_delay_breakdown  |
| monthly_flight_breakdown     |
| most_arr__dep_delay_airline  |
| most_arrival_delay_2019      |
| tbl_airline                  |
| tbl_destairport              |
| tbl_flight                   |
| tbl_originairport            |
| tbl_weather                  |
| top_10_destination           |
+------------------------------+
23 rows selected (0.559 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

```
50-8358-7f41e5566
INFO  : OK
+------------+
| row_count  |
+------------+
| 84         |
```

```
+----------------------+
|        field         |
+----------------------+
| airlineid            |
| airlinename          |
| arrdelay             |
| arrdelayminutes      |
| depdelay             |
| depdelayminutes      |
| destairportid        |
| destcityname         |
| eventid              |
| eventtype            |
| flights              |
| month                |
| originairportid      |
| origincityname       |
| timeid               |
| weatherdelay         |
| year                 |
```

Environment | History | Connections | Tutorial

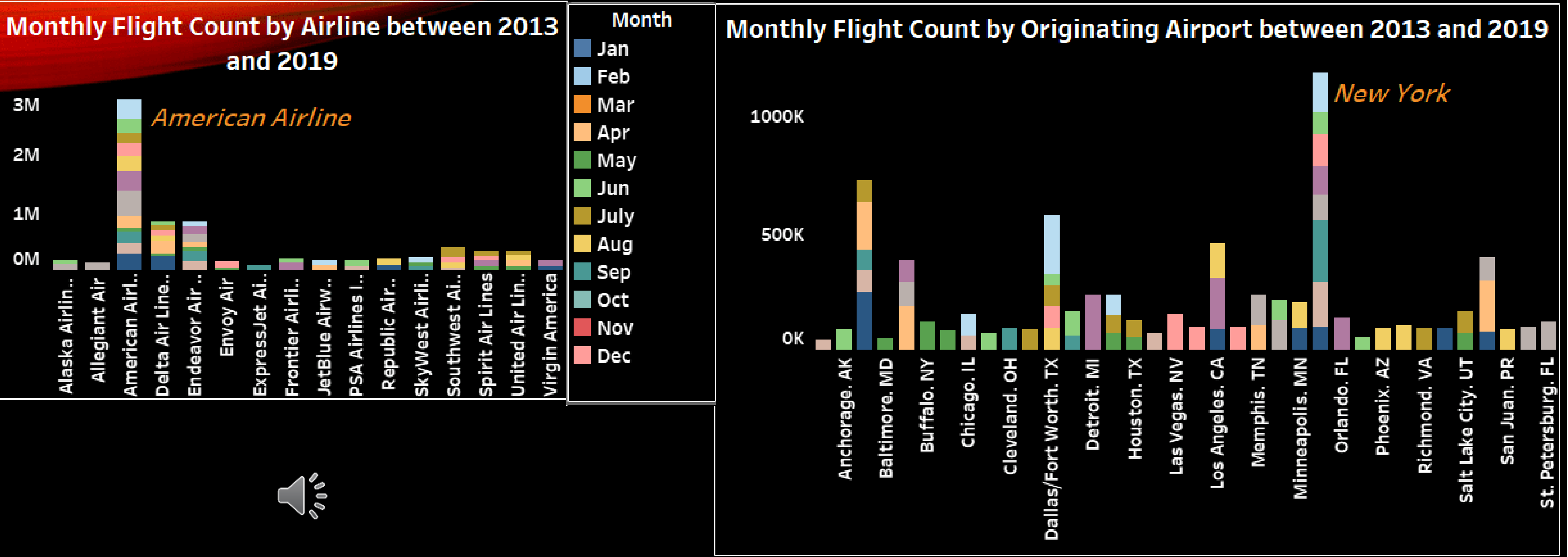
SQL | Help | Refresh Connection Data

ADMP Project Gr22 Hive DNS — Hive

- HIVE
  - admp_grp22
    - cleaned_data
    - dimairline
    - dimdeparture
    - dimdestination
    - dimevent
    - dimtime
    - fctarrdelay
    - fctdeparrdelay
    - fctdepdelay
    - fctflight
    - fctmostarrdelay
    - fctpopdestination
    - monthly_arrival_delay_breakdown
    - monthly_dep_delay_breakdown
    - monthly_flight_breakdown
    - most_arr__dep_delay_airline

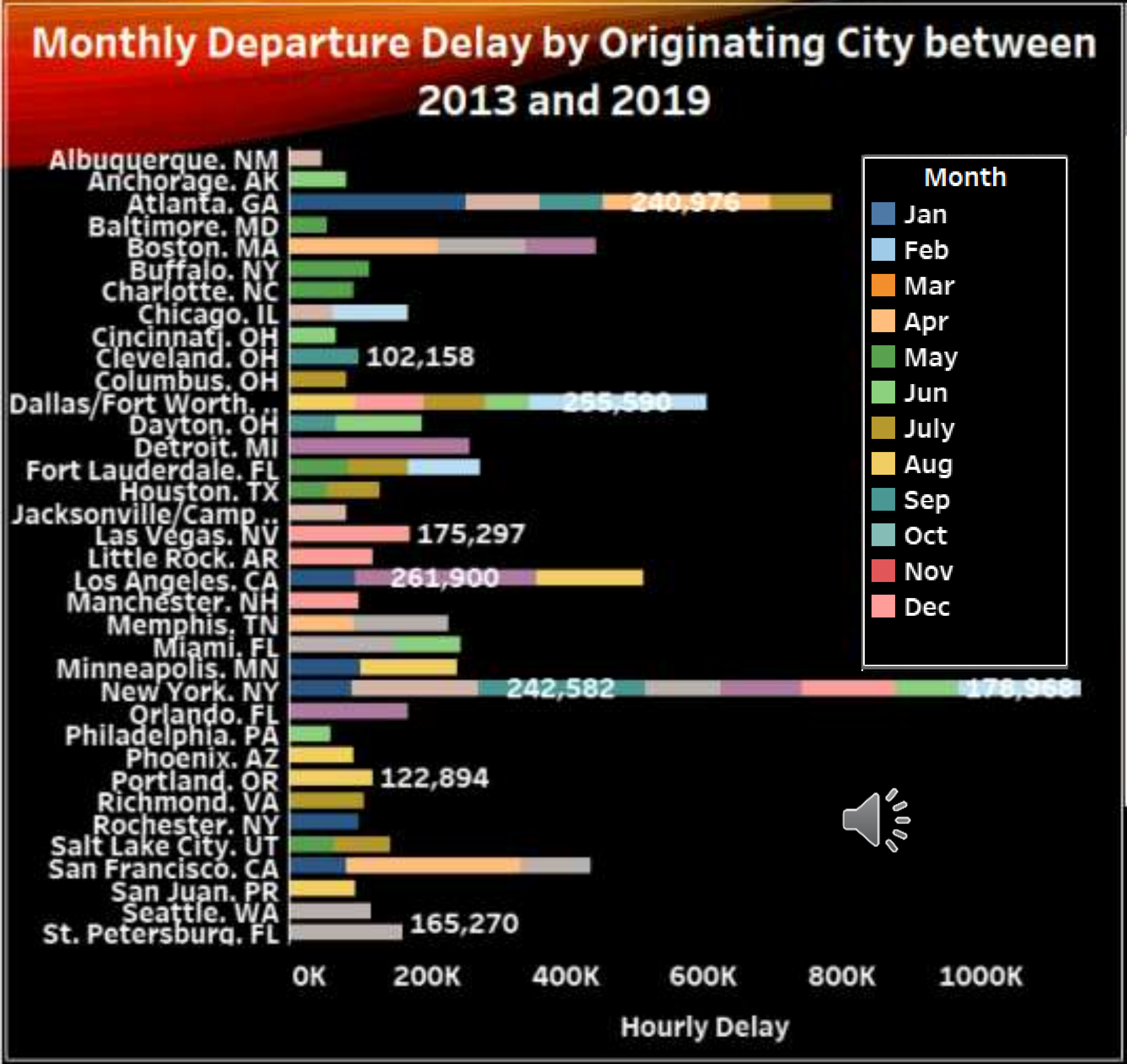
VISUALIZATIONS

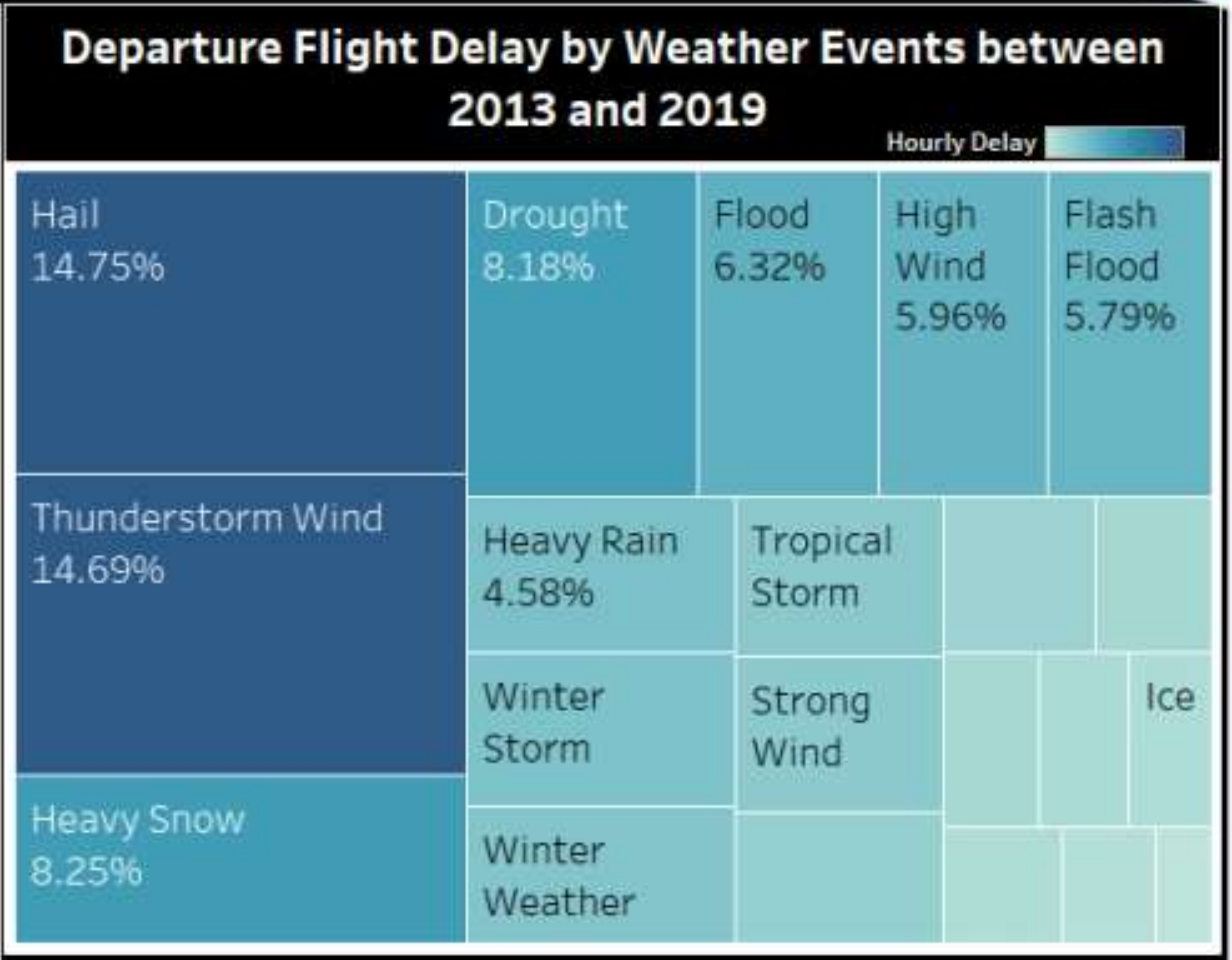
Monthly Flight Count by Airline and Originating City/Airport



The monthly flight breakdown between 2013 and 2019 shows that American Airline was the most patronized airline with almost 3M flights while the New York airport is the airport with most originating flight

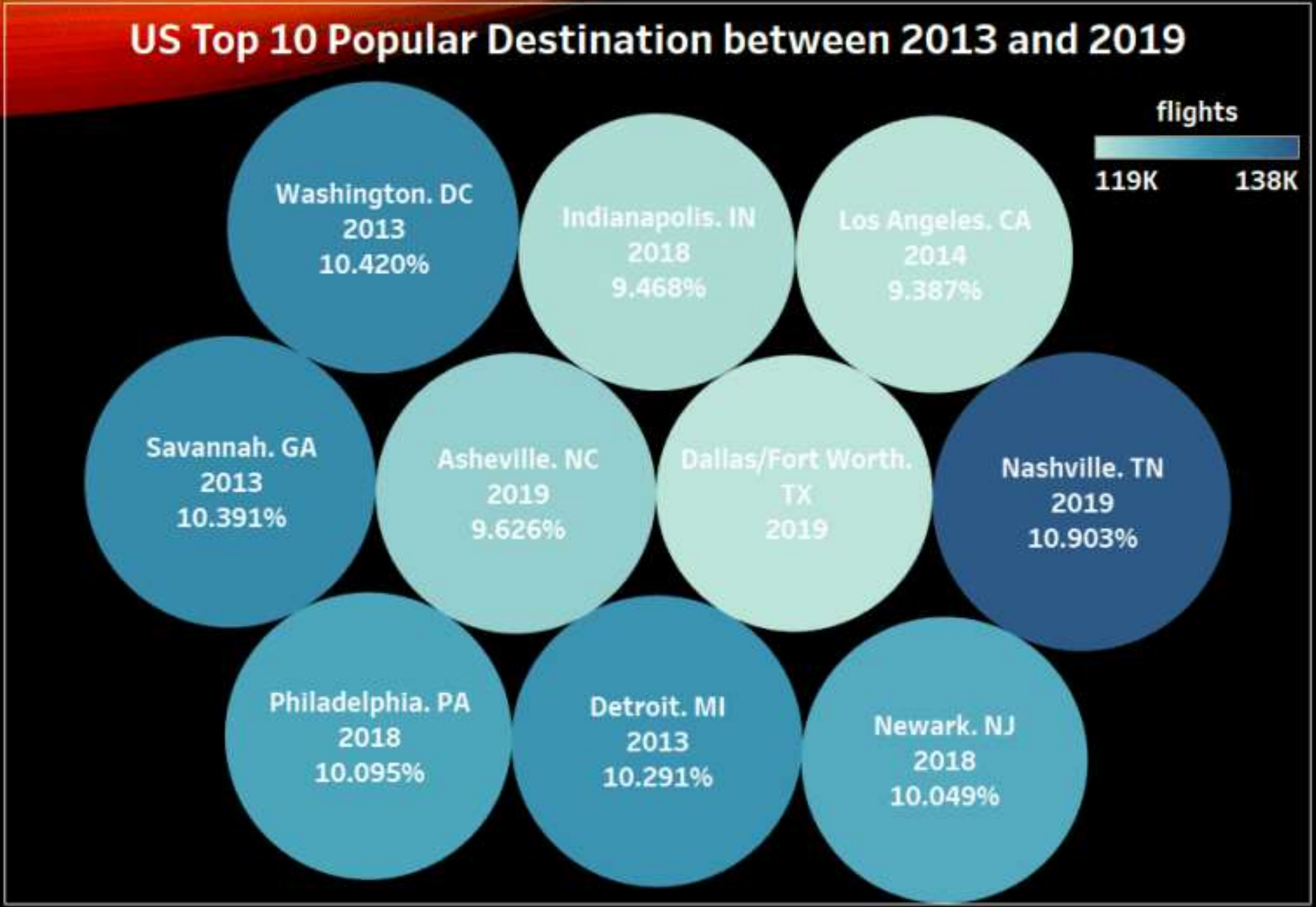
Monthly Departure Delay by Originating City and Weather Events





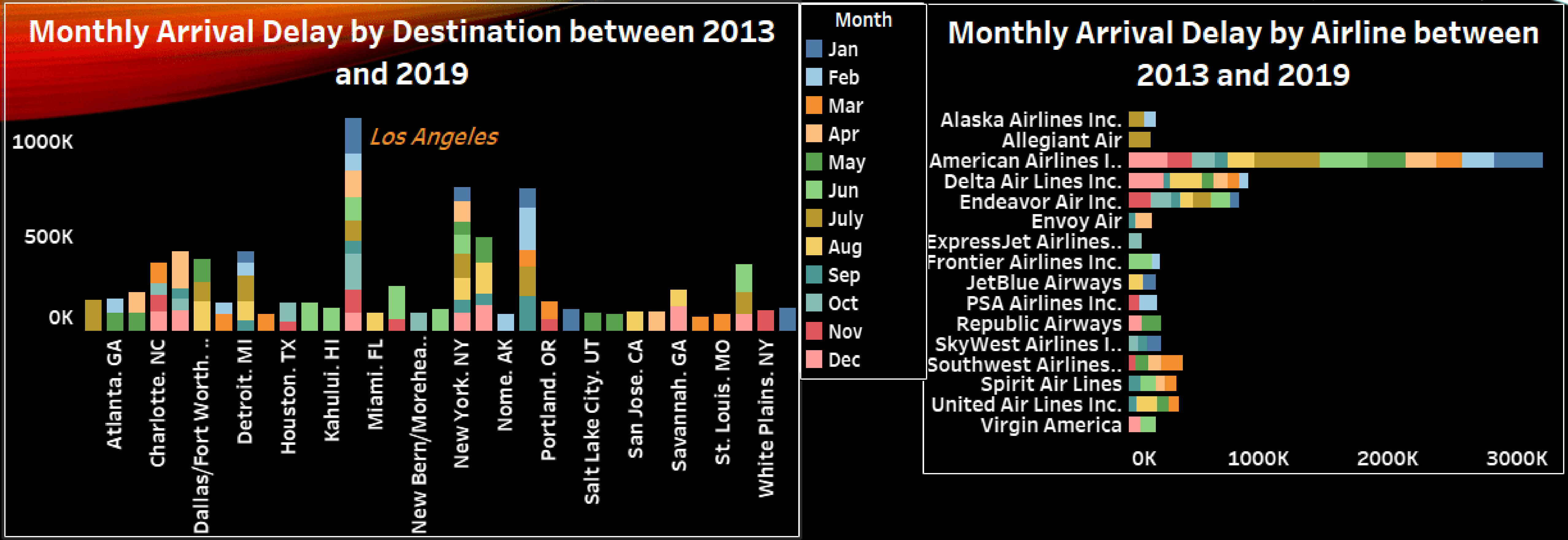
Most originating flight delays as a result of unfavourable weather events came from the New York as the most flights. However, majority of the adverse weather conditions are due to Hail and Thunderstorm wind

Top 10 Popular Destination



The top 10 destinations for airplane passengers during this period are shown with Nashville being the most favourite destination

MONTHLY ARRIVAL DELAY BY DESTINATION AND AIRLINE BETWEEN 2013 AND 2019



Unlike the departure delay, Los Angeles was the city with the most arrival delay as a result of unfavourable weather events. Expectedly, American Airlines had the most arrival flight delay since it was the airline of choice for departure by most passengers within this period.

Most Arrival Delay Destination by Weather Event in 2019

The most arrival delay was caused by Thunderstorm Wind in Nashville in 2019. It was also the cause of flight arrival delay in Charlotte.NC in the same year. However, hail, heavy rain, ice storm and winter weather were high contributors.

INSIGHTS & RECOMMENDATIONS

INSIGHTS & RECOMMENDATIONS

Below insights were drawn from the analysis:

- Majority of the weather-event-related flight delays on arrival in 2019 were caused by thunderstorms, hail, heavy rain, ice storm, and winter weather.
- On arrival between 2013 and 2019, the most impacted city was Los Angeles
- Top ten destinations among which is Los Angeles

Recommendations:

- Weather forecast (using machine learning algorithm) to predict the occurrence of the devastating weather events mentioned above should be prioritized and adequate measures taken to re-route or reschedule flights.
- Flight destinations involving cities like Los Angeles should be placed on a red alert and alternative destinations prepared in collaboration with traffic controllers and other airlines.
- The popular destination airports should be upgraded in terms of technologies and infrastructures to accommodate the traffic and make provisions for possible delays. This will at least reduce the delay significantly when it happens.

REFERENCES

https://data.worldbank.org/indicator/IS.AIR.PSGR

www.bts.gov/explore-topics-and-geography/topics/airline-time-performance-and-causes-flight-delays

www.scientificamerican.com/article/airlines-grapple-with-flights-delayed-by-climate-fueled-heat/

https://www.airlines.org/impact/

transtats.bts.gov/PREZIP

https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles

https://blog.hubspot.com/website/etl-vs-elt

https://www.ft.com/content/55f6c210-00c8-46d1-9165-c99f3dacb938

https://www.gartner.com/reviews/market/product-roadmapping-tools-for-software-engineering/vendor/hive/product/hive/alternatives