2015 US Domestic Flights: Delays and Cancellations

Iulia Tomescu - Data Analytics Career Track

(Capstone 2 - a Python project, October 2021)

SCHOOL OF DATA
by Springboard

OUTLINE

I. Introduction

II. Exploratory data analysis: Data examination & cleaning (APPENDIX)

III. Exploratory data analysis: visualizations, statistics (descriptive & inferential), trends & relationships

IV. Predictive Analytics: Modeling & Prediction

V. Conclusions & Key Insights

INTRODUCTION

- ❑ Why I chose this data
- ❑ Context
- ❑ Business goals
- ❑ Personal goals
- ❑ Sources

Why I chose this data

- Rich in dimensions
- Prediction potential
- High potential business impact
- Interesting subject

Context

- The U.S. Dept. of Transportation's (DOT) Bureau of Transportation Statistics **tracks the on-time performance, and the causes of delays and cancellation of domestic flights operated by large air carriers**; it covers nonstop scheduled-service flights between points within the United States (including territories) by the fourteen U.S. air carriers that have at least one percent of total domestic scheduled-service passenger revenues;
- The current analysis focuses on the **2015 historical daily data of domestic flights operated by large air carriers**, which includes specs on airlines, flights, airports, dates, time, and causes of delays and cancellations.

Business GOALS

- Identify major drivers for delays and cancellations
- Build a model to predict overall flight delays with respect to magnitude

Personal GOALS

- Incorporate technical & analytical skills acquired in the Data Analytics Career Track course
- Demonstrate & strengthen technical skills and computer literacy such as Python, Tableau, statistics, predictive analytics & modeling
- Generate a basic tutorial for people that are new to Python & Exploratory Data Analysis

Sources

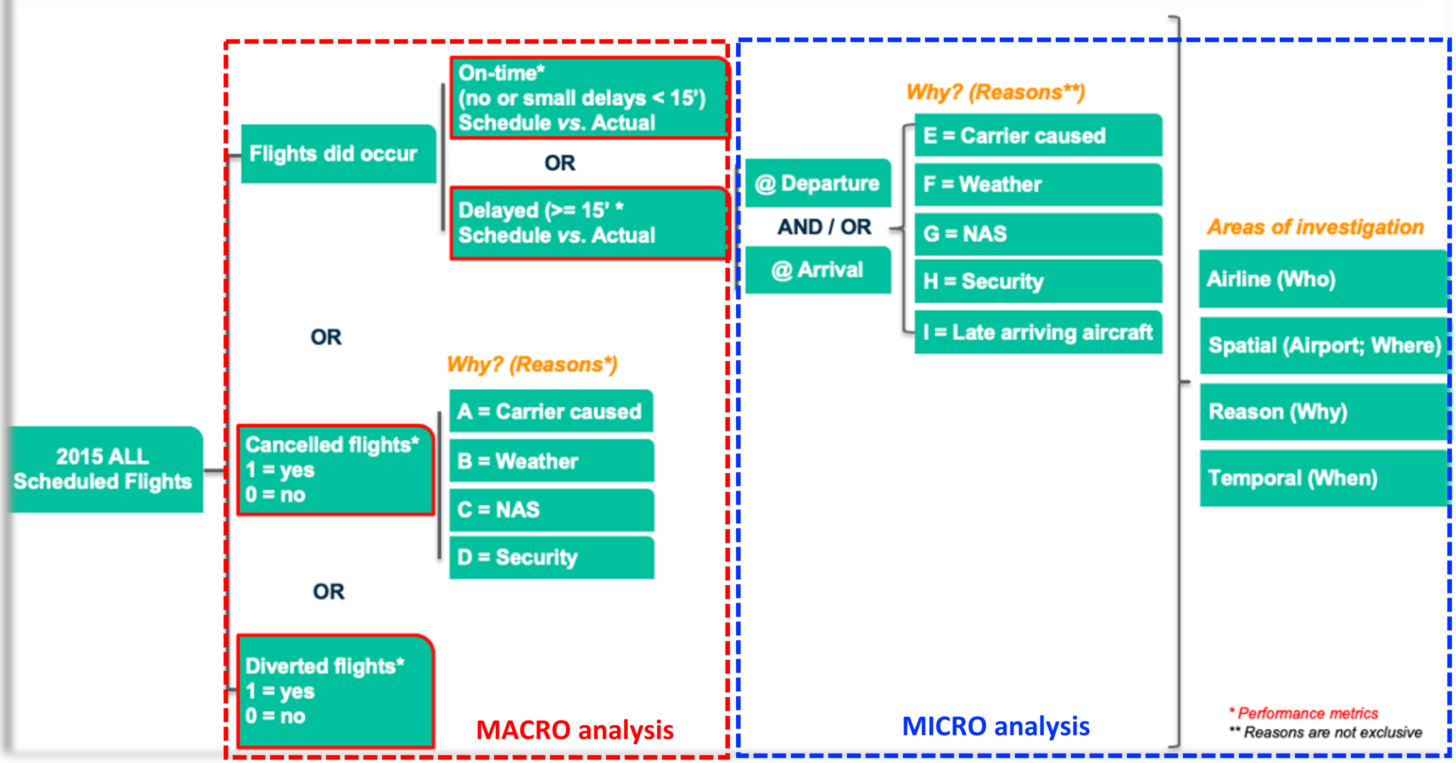
- *Kaggle: https://www.kaggle.com/usdot/flight-delays?select=flights.csv*
- *US Department of Transportation: https://www.transportation.gov/airconsumer*
- *Bureau of Transportation Statistics: https://www.transtats.bts.gov*

ANALYSIS STRUCTURE APPROACH

MAIN file

#	File Name	Description	No. of columns
1.	flights_raw	• the main data set to be used in analysis	29
2.	airports_raw	• reference file for the airports; • includes details for all airports such as name, city, state, & geographic coordinates; • to be merged with flights	7
3.	airlines	• reference file for the airlines; • includes the full names for all the airlines	2
4.	October_2015_Flights	• details of all flights for the month of October 2015 (see below why); • to be merged with flights	65

2015 Domestic Flights: Exploratory Data Analysis (EDA) Approach



OUTLINE

I. Introduction

II. Exploratory data analysis: Data examination & cleaning (APPENDIX)

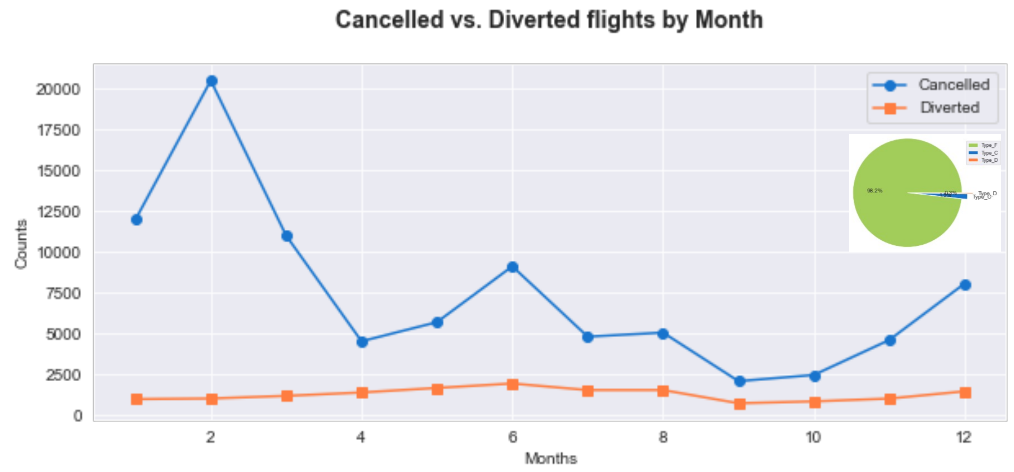
III. **Exploratory data analysis: visualizations, statistics (descriptive & inferential), trends & relationships**
- **Analysis of cancellation, delayed flights, and frequency of flights (MACRO)**
- Analysis of flight delays: frequency, magnitude, reasons, temporal, spatial & carrier-based analysis (MICRO)

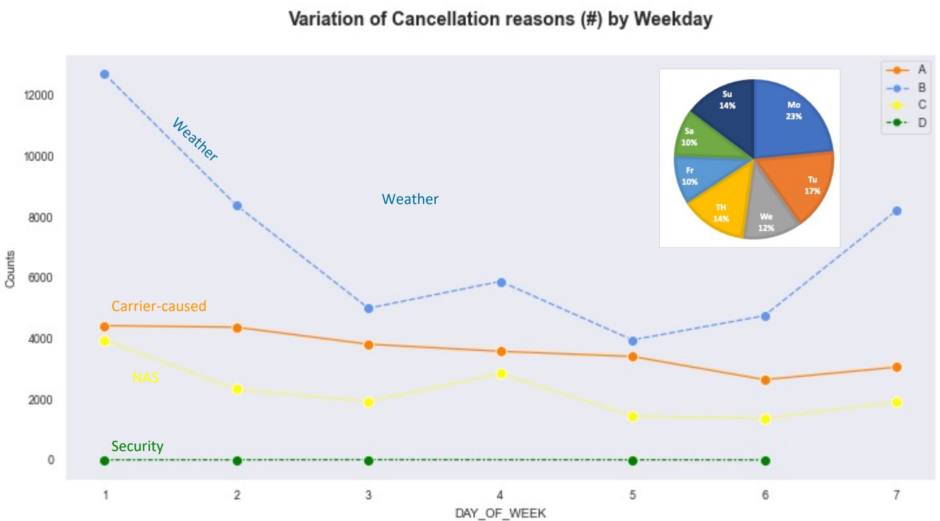
IV. Predictive Analytics: Modeling & Prediction

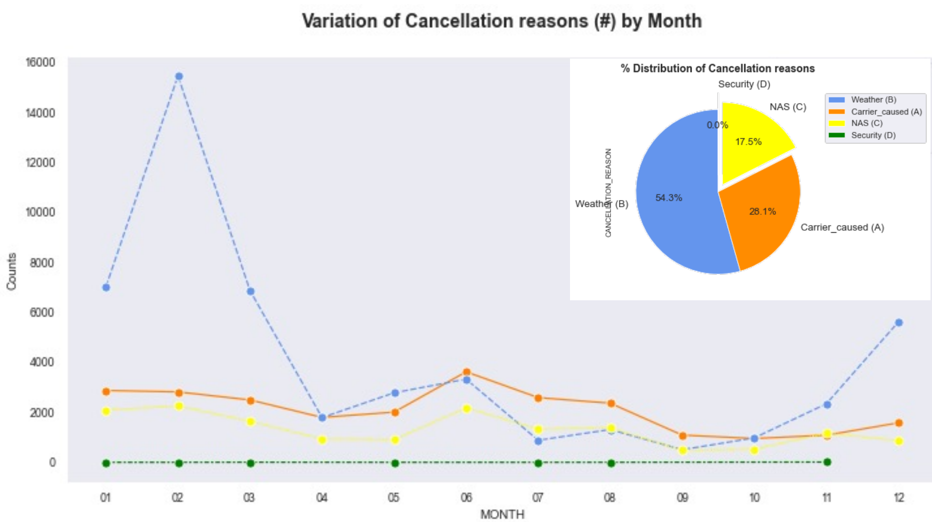
V. Conclusions & Key Insights

WEATHER IS NUMBER ONE CAUSE FOR CANCELLATIONS (54%)

- ❑ **Out of 5.8 MM domestic flights**
  - **1.5%** of flights were **Cancelled (C): 246 cancellation / day**
  - **0.3%** of flights were **Diverted (D): 42 diversions / day**
  - **98.2%** of flights of **did occur (F)**: **15,655 flights / day**
- ❑ **Monthly:** cancellations are most common during Winter & least frequent during Fall
- ❑ **Weekday:** cancellations are most frequent on Mondays & Tuesdays, & least frequent on Fridays & Saturdays
- ❑ **Cancellation reasons:** Weather (54%) & Carrier (28%) related reasons are responsible for 82% of the cancellations
  - Winter and Spring flights greatly affected by weather
  - Spring & Summer greatly affected by the carrier when flights are most frequent

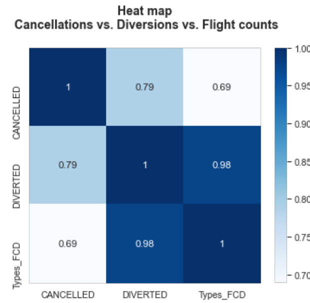


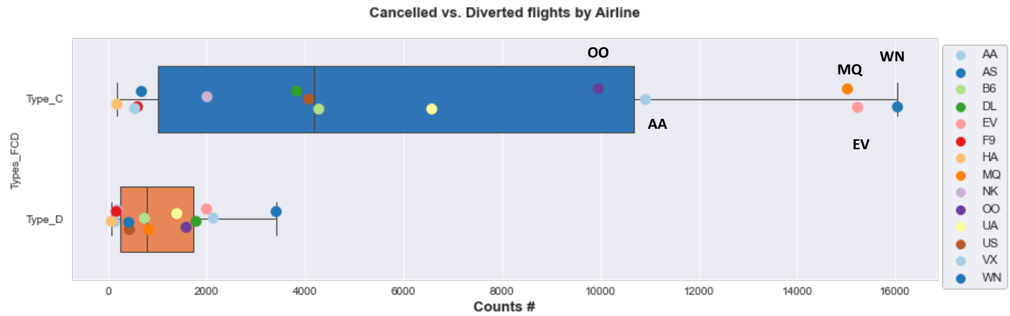


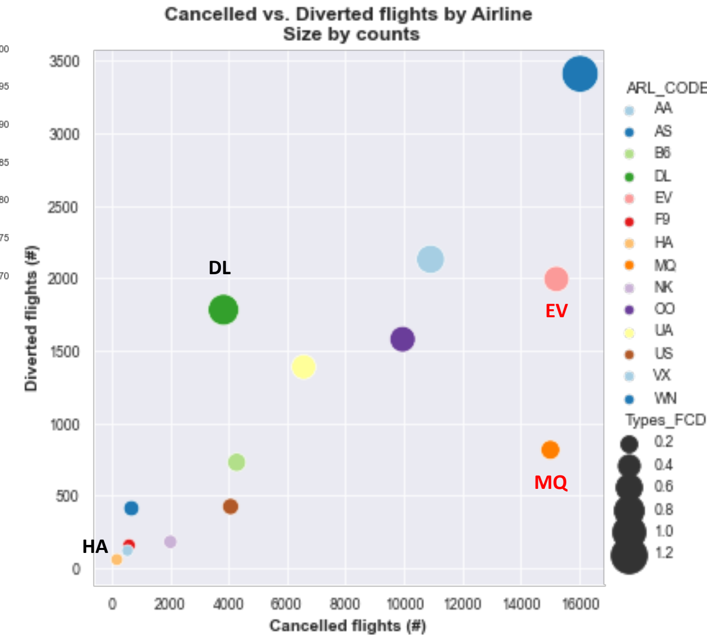


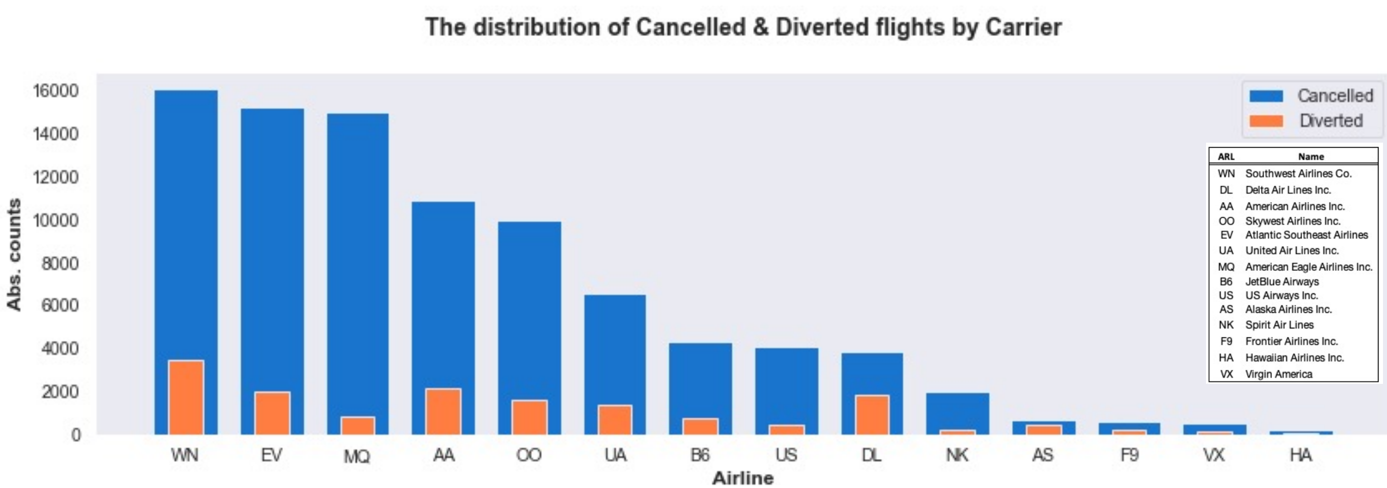
AMERICAN EAGLE & ATLANTIC SOUTHEAST Airlines have high rates of Cancellations DELTA & HAWAIIAN Airlines are top performers

- ❑ 14 total airlines
- ❑ Top 5 airlines account for 75% of the cancellations
- ❑ Cancellations & Diversion frequency vary with the number of flights

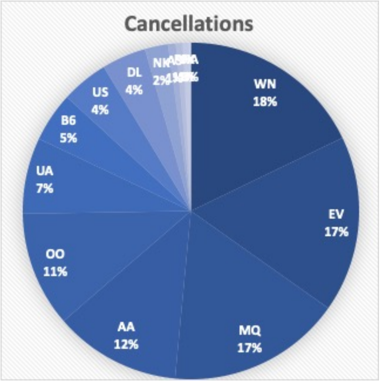






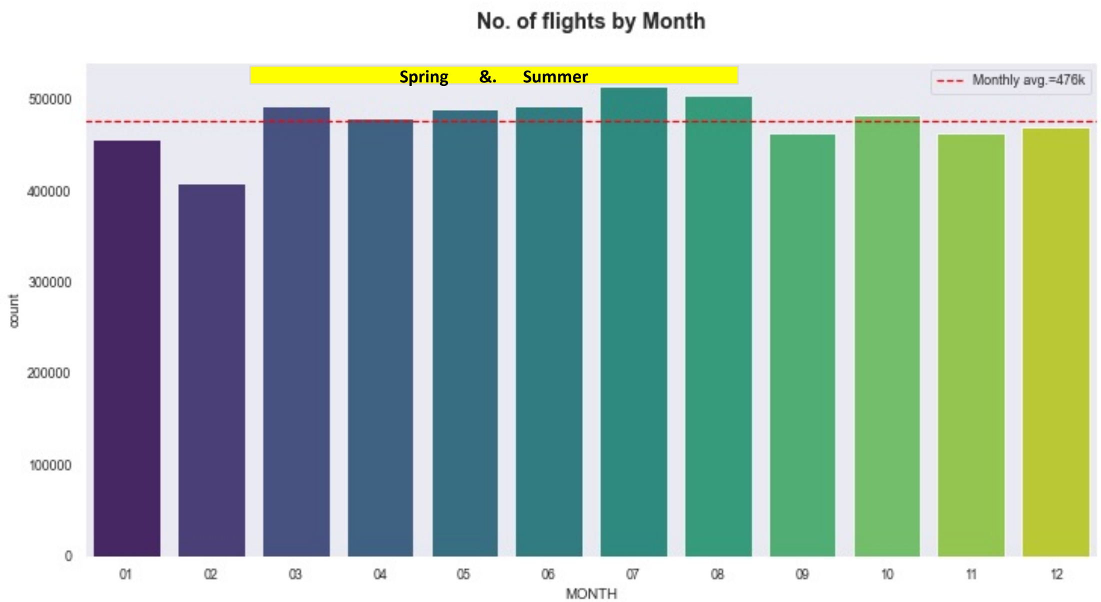


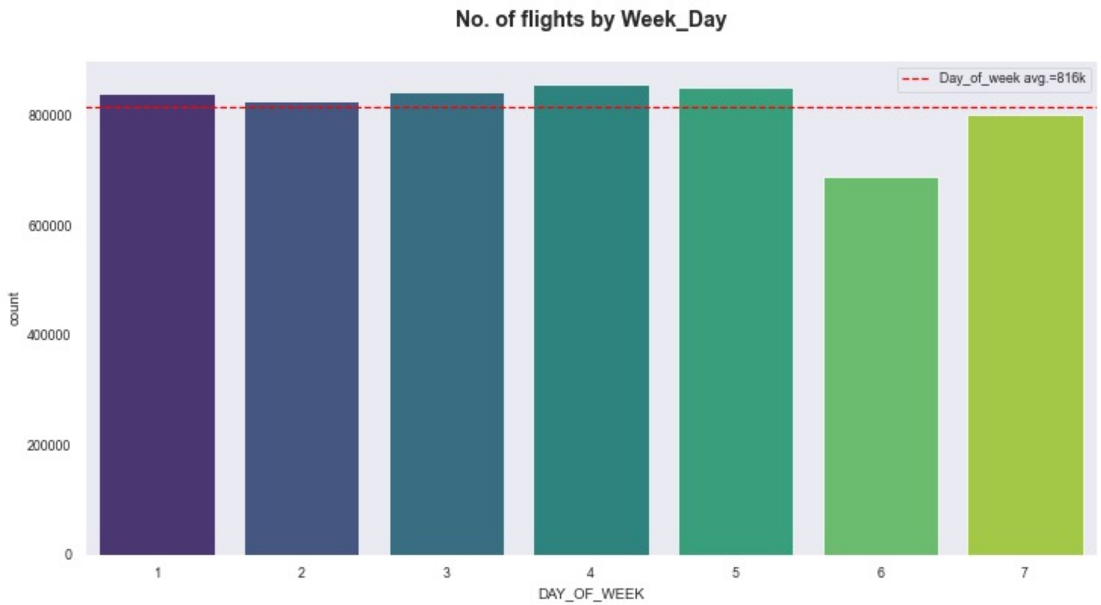
ARL	Name
WN	Southwest Airlines Co.
DL	Delta Air Lines Inc.
AA	American Airlines Inc.
OO	Skywest Airlines Inc.
EV	Atlantic Southeast Airlines
UA	United Air Lines Inc.
MQ	American Eagle Airlines Inc.
B6	JetBlue Airways
US	US Airways Inc.
AS	Alaska Airlines Inc.
NK	Spirit Air Lines
F9	Frontier Airlines Inc.
HA	Hawaiian Airlines Inc.
VX	Virgin America



Temporal insights of flight frequency

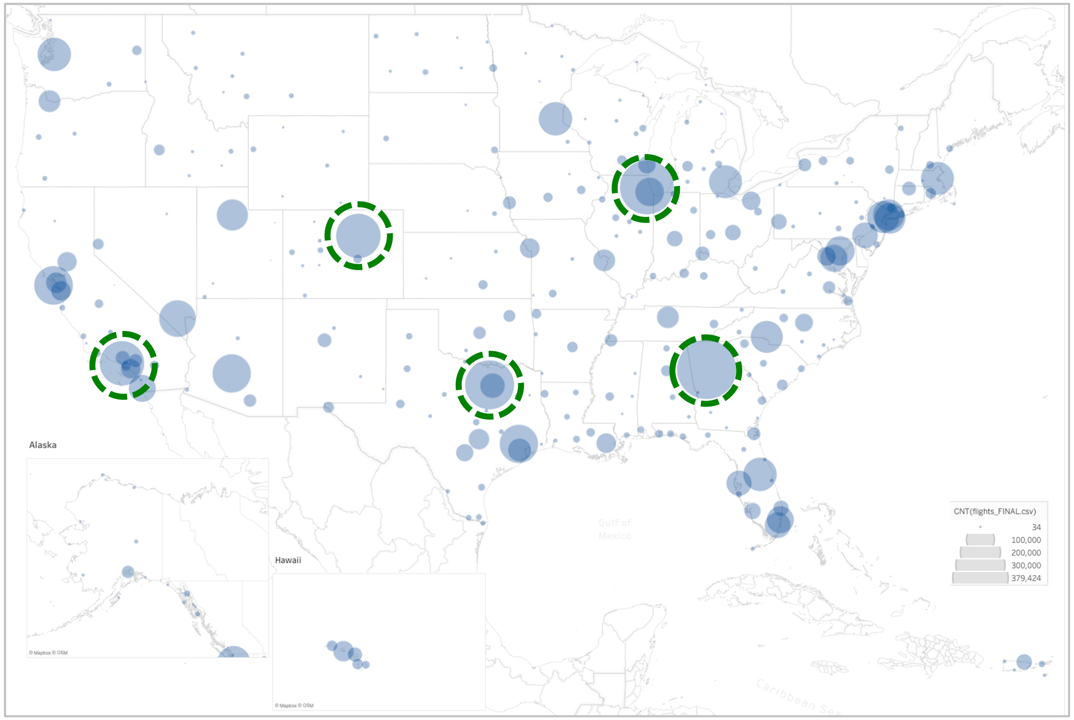
PEOPLE TRAVEL MOST DURING THE SPRING AND SUMMER & AVOID WEEKENDS

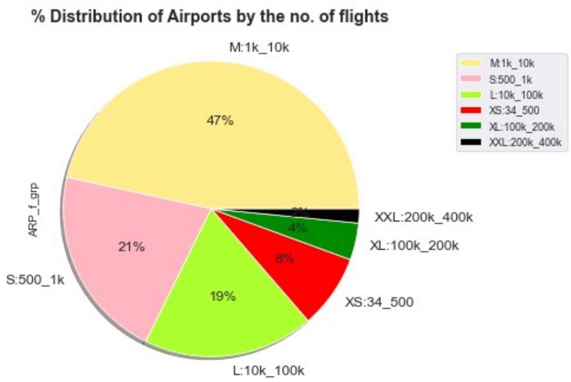


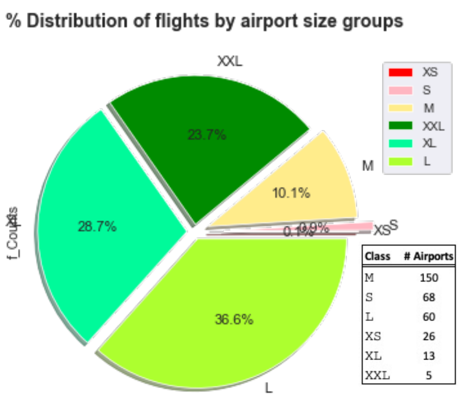


The top 24% busiest airports contributed almost 90% to the total of 2015 volume of US domestic flights

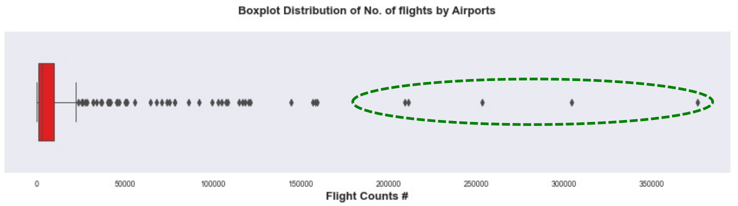
- **322 total airports**
- **Top 5 airports contributed by 24% to the total flight volume:**
    - Atlanta Intl. Airport, Chicago O'Hare Intl. Airport, Dallas/Fort Worth International Airport, Denver Intl. Airport, & Los Angeles Intl. Airport





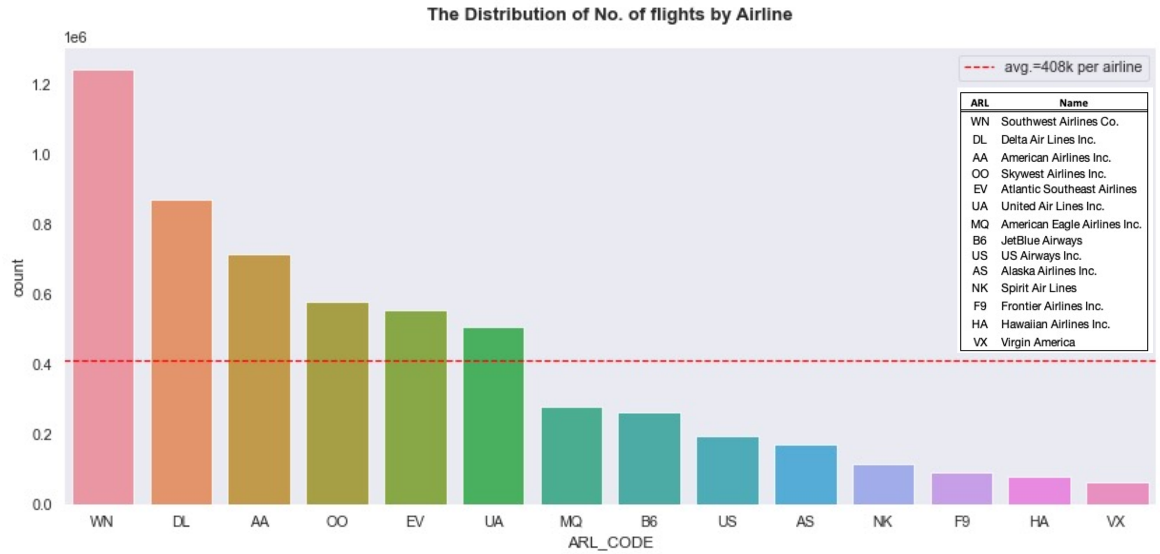


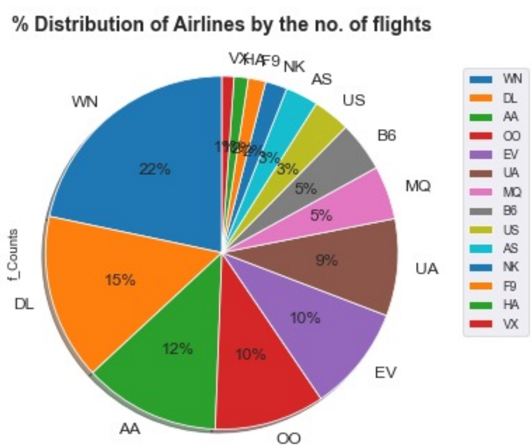
Class	# Airports
M	150
S	68
L	60
XS	26
XL	13
XXL	5

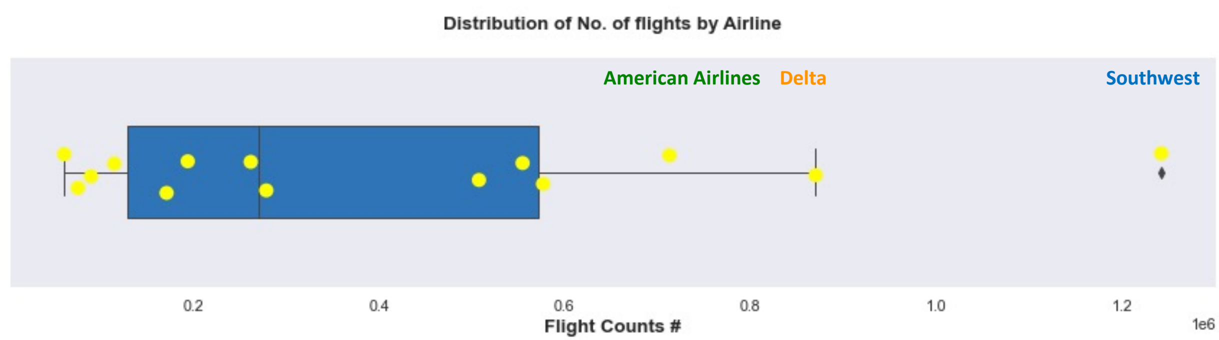


Top 3 airlines dominated the US domestic travel industry (50%) with Southwest Airlines Co. alone controlling 22% of the air travel volume

- **14 total airlines**
- **Top 3 airlines:**
  - Southwest, Delta, American Airlines







OUTLINE

I. Introduction

II. Exploratory data analysis: Data examination & cleaning (APPENDIX)

III. Exploratory data analysis: visualizations, statistics (descriptive & inferential), trends & relationships

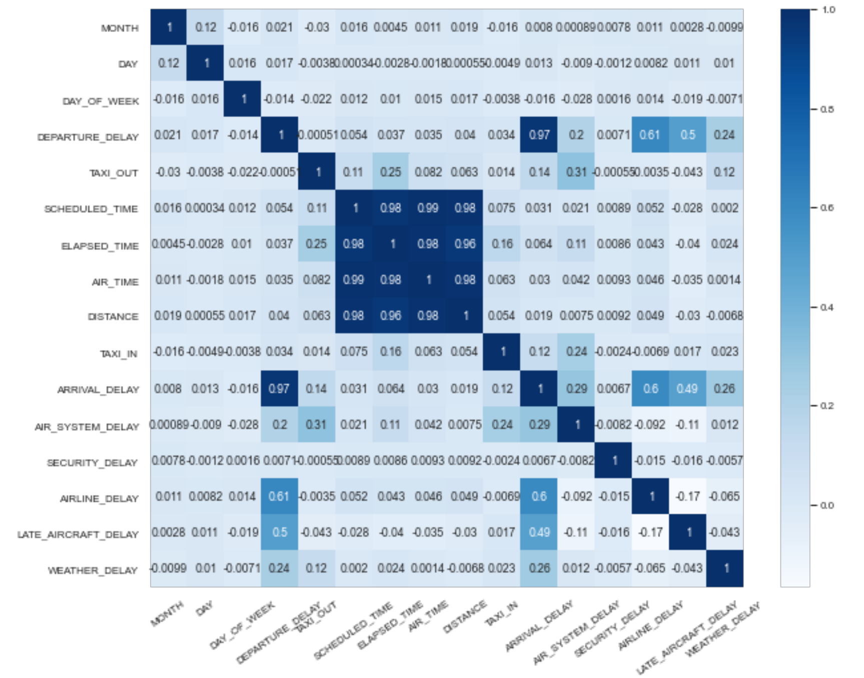
- Analysis of cancellation, delayed flights, and frequency of flights (MACRO)
- **Analysis of flight delays: frequency, magnitude, reasons, temporal, spatial & carrier-based analysis (MICRO)**

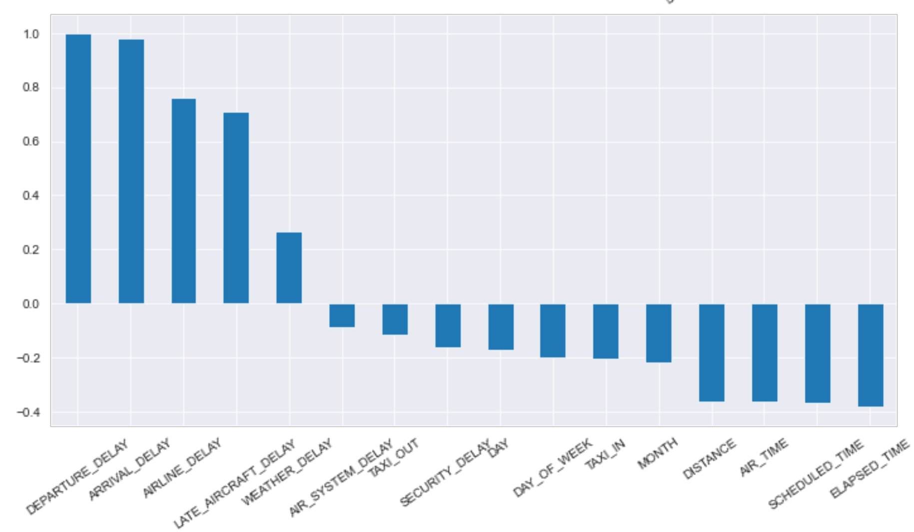
IV. Predictive Analytics: Modeling & Prediction

V. Conclusions & Key Insights

UNDERSTANDING THE RELATIONSHIP BETWEEN VARIABLES

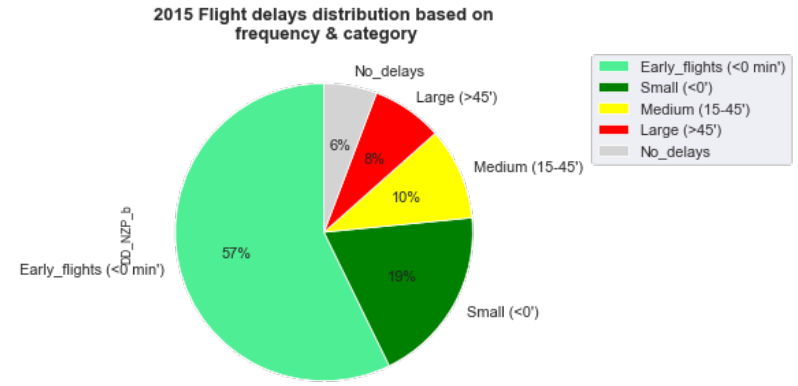
- A flight is considered delayed if it arrived at (or departed) the gate 15 minute or more after the scheduled arrival (departure) time
- Delays at departure strongly correlate with delays at arrival

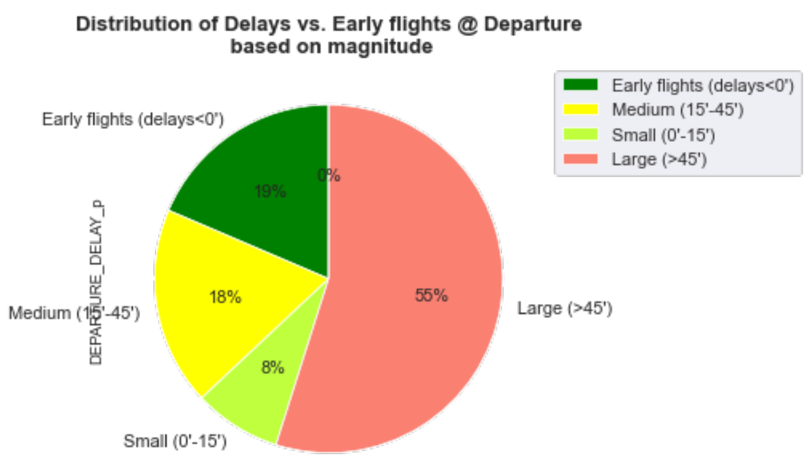


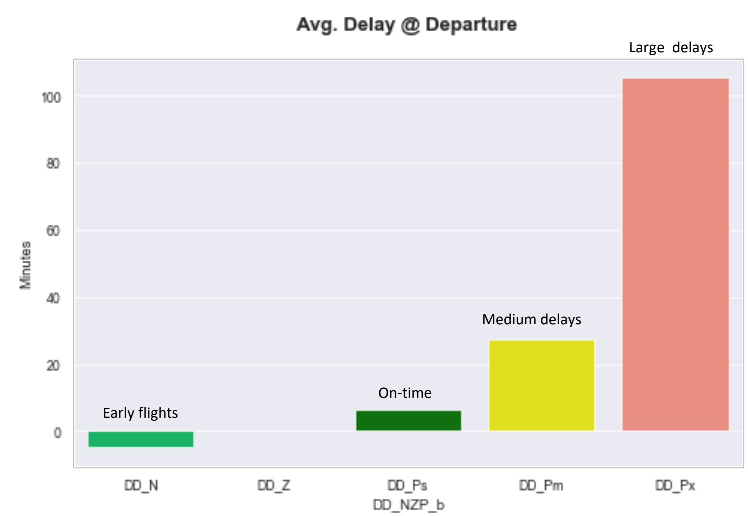


18% OF THE DELAYS ARE LONGER THAN 15 MINUTES (ACTUAL DELAYS)

- **Frequency:**
  - flights are more frequent early (negative delays) and on-time (0 - 15’) than delayed
- **Magnitude:**
  - Medium delays (15’ – 45’) are in average 27’
  - Large delays (>45’) are in average 105’
  - Large delays account for 55% of the delay time

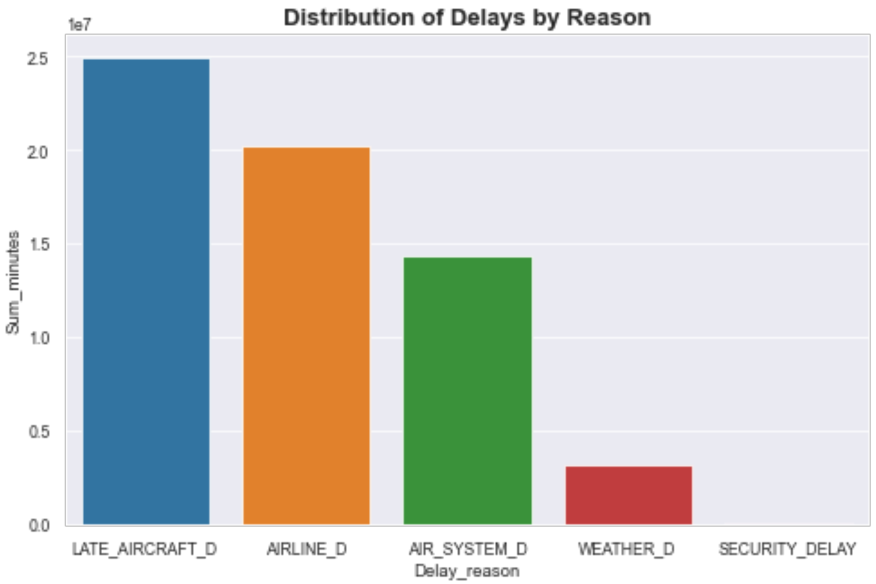


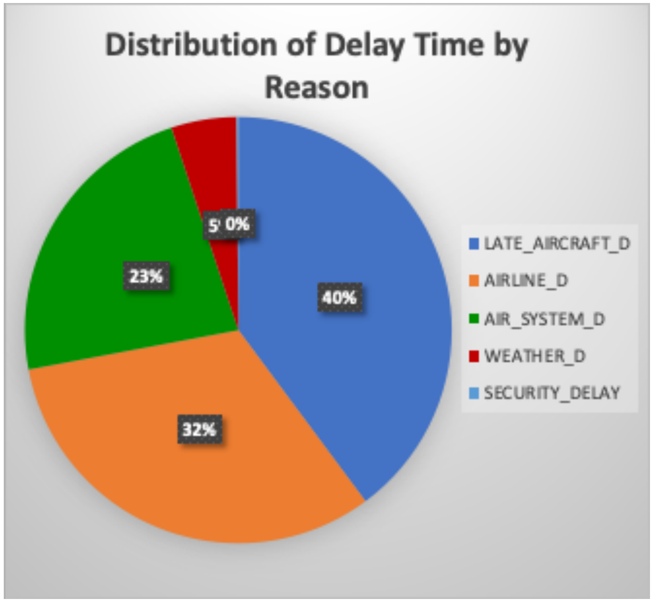




MOST TIME WAS LOST DUE TO DELAY PROPAGATION

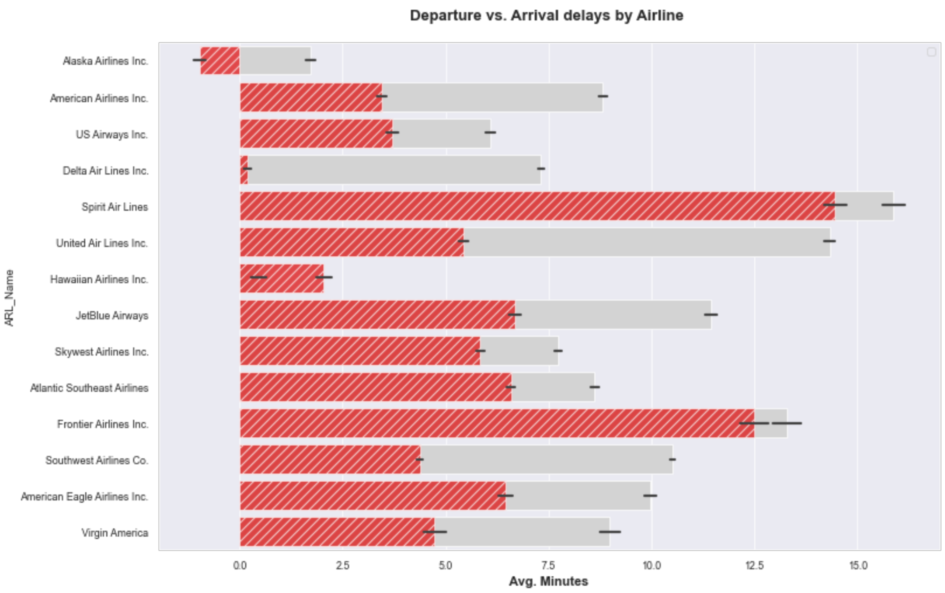
- **Delay reasons:**
  - E = Carrier caused
  - F = Weather
  - G = National Aviation System
  - H = Security
  - I = Late arriving aircraft
- **Reasons are not exclusive**

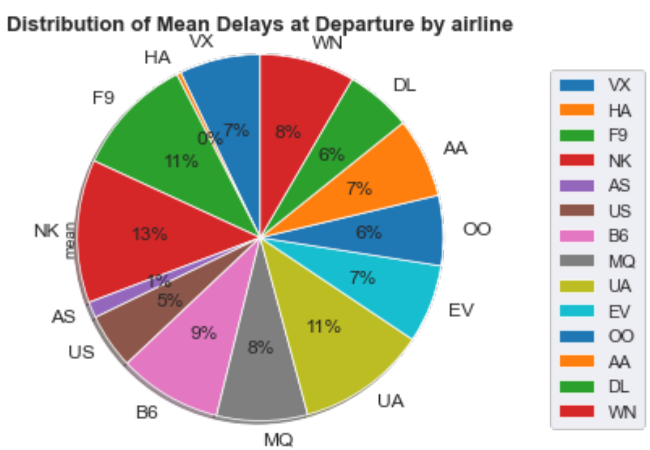


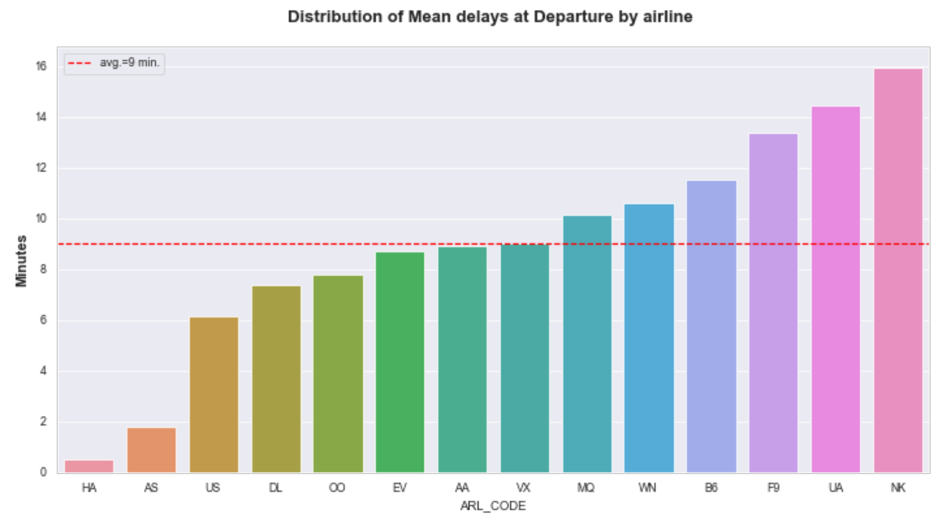


AIRLINES ADJUST FLIGHT SPEED TO REDUCE THE DELAY TIME FROM DEPARTURE TO ARRIVAL

- **The mean delay of 9' is significantly lowered by the large % of flights that are early and on-time**
- **Mean delays behave homogeneously among airlines except:**
  - Hawaiian and Alaska Airlines are top performers
  - Spirit & United airlines are lowest performers



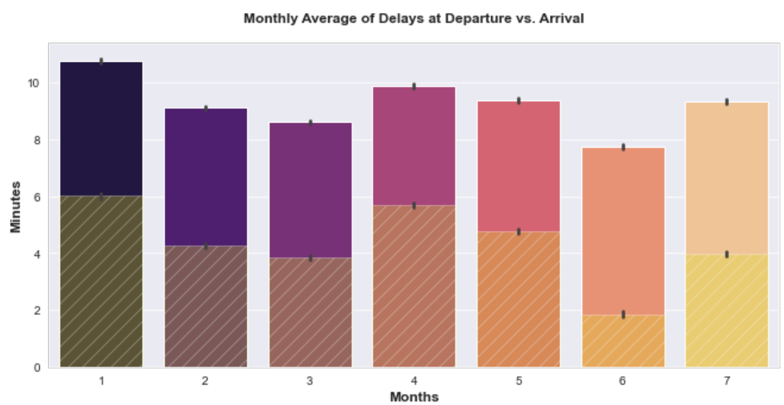


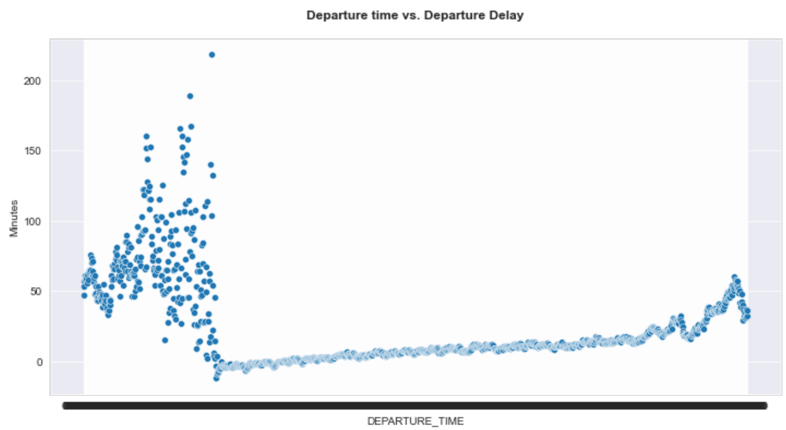


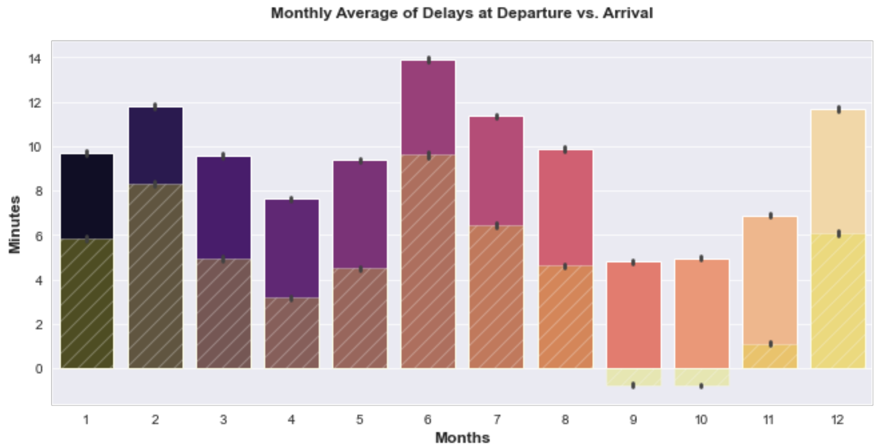
Performance metrics: Delays - Temporal analysis

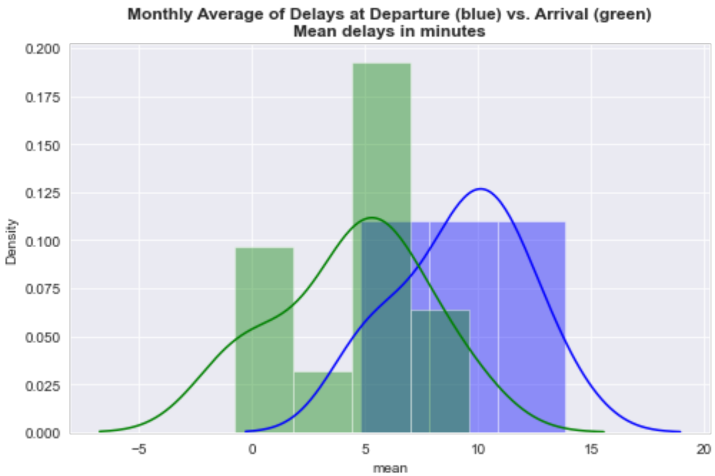
FALL , SPRING, & WEEKENDS ARE THE BEST TIMES TO MAKE UP FOR DELAYS

- ❑ **In average, airlines are able to reduce the delay time by half from departure to arrival**
- ❑ **Delays increase throughout the day**









OUTLINE

I. Introduction

II. Exploratory data analysis: Data examination & cleaning (APPENDIX)

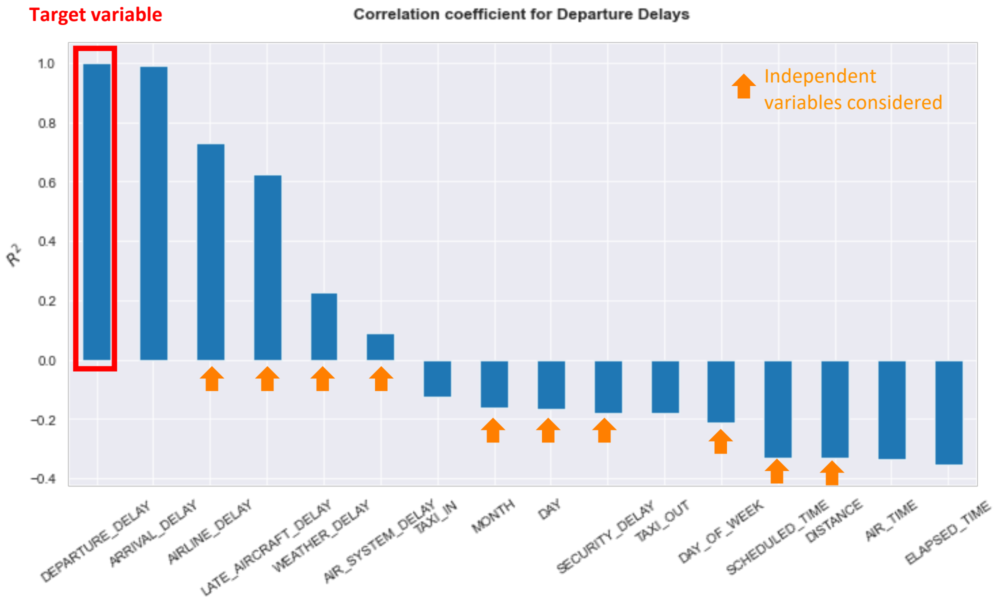
III. Exploratory data analysis: visualizations, statistics (descriptive & inferential), trends & relationships

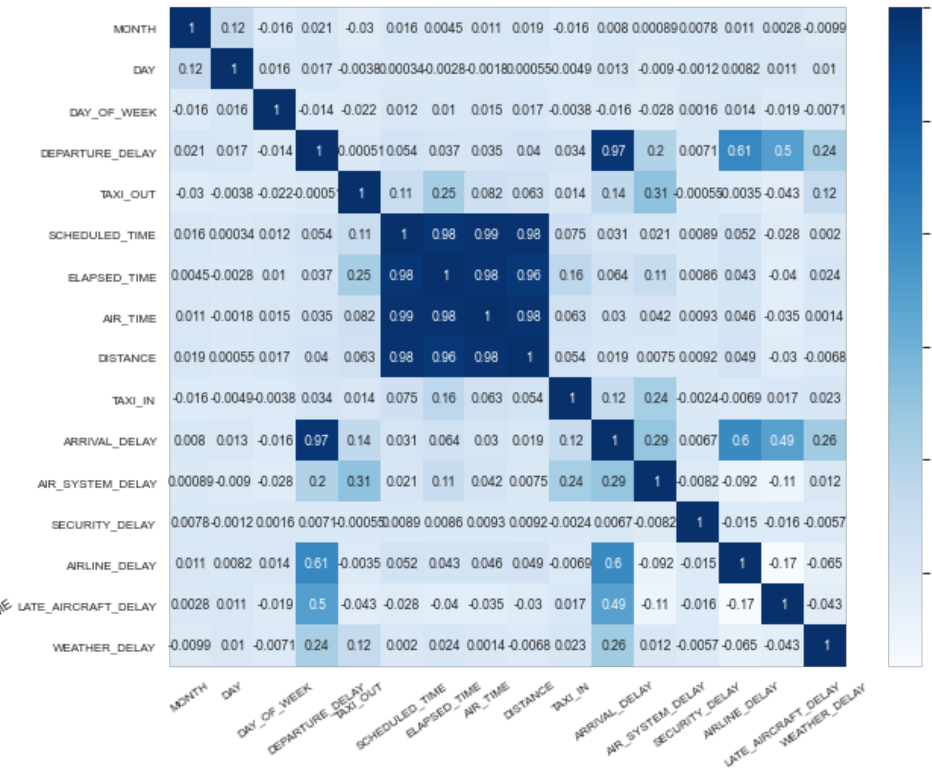
- Analysis of cancellation, delayed flights, and frequency of flights (MACRO)
- Analysis of flight delays: frequency, magnitude, reasons, temporal, spatial & carrier-based analysis (MICRO)

IV. Predictive Analytics: Modeling & Prediction

V. Conclusions & Key Insights

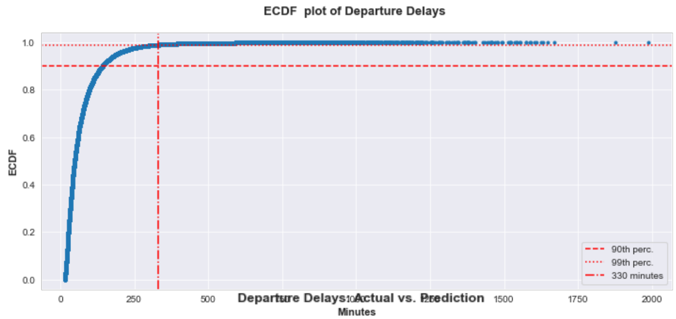
PREDICTIVE ANALYTICS: MODELING & PREDICTION

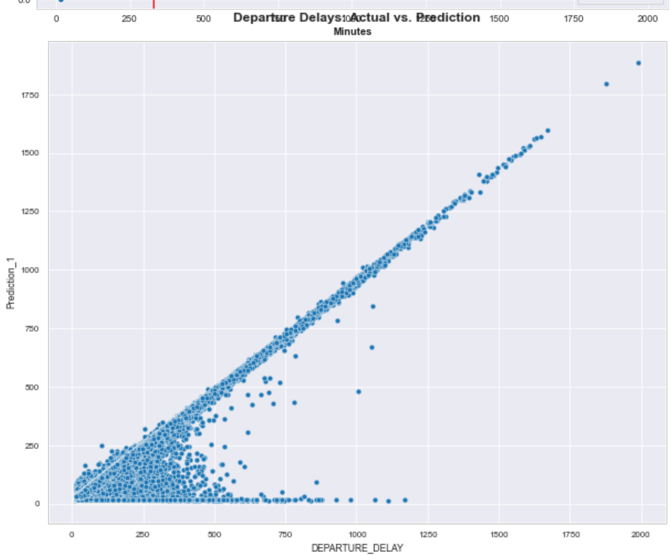




** A flight is considered delayed if it arrived at (or departed) the gate 15 minute or more after the scheduled arrival (departure) time*

MULTIVARIATE LINEAR REGRESSION MODEL_1: ALL DATA





OLS Regression Results

Dep. Variable:	DEPARTURE_DELAY	**R-squared:**	0.845
Model:	OLS	**Adj. R-squared:**	0.845
Method:	Least Squares	**F-statistic:**	8.917e+05
Date:	Sat, 02 Oct 2021	**Prob (F-statistic):**	0.00
Time:	12:45:35	**Log-Likelihood:**	-3.8736e+06
No. Observations:	819621	**AIC:**	7.747e+06
Df Residuals:	819615	**BIC:**	7.747e+06
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	15.8235	0.091	174.112	0.000	15.645	16.002
MONTH	0.2714	0.009	30.488	0.000	0.254	0.289
DAY_OF_WEEK	-0.3450	0.015	-22.750	0.000	-0.375	-0.315
AIRLINE_DELAY	0.9487	0.001	1664.755	0.000	0.948	0.950
LATE_AIRCRAFT_DELAY	0.9375	0.001	1444.347	0.000	0.936	0.939
WEATHER_DELAY	0.9346	0.001	715.833	0.000	0.932	0.937

Omnibus:	1053819.404	**Durbin-Watson:**	1.858
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	297557326.717
Skew:	7.040	**Prob(JB):**	0.00
Kurtosis:	95.276	**Cond. No.**	184.

Model accuracy metrics

Regression model	MAE	MAPE	RMSE	NRMSE*	Model acc. %
Model_1 (ALL data)	1.05E-13	0.13	27.31	0.01	87%

** Normalized RMSE using Max-Min*

OUTLINE

I. Introduction

II. Exploratory data analysis: Data examination & cleaning (APPENDIX)

III. Exploratory data analysis: visualizations, statistics (descriptive & inferential), trends & relationships

- Analysis of cancellation, delayed flights, and frequency of flights (MACRO)
- Analysis of flight delays: frequency, magnitude, reasons, temporal, spatial & carrier-based analysis (MICRO)

IV. Predictive Analytics: Modeling & Prediction

V. Conclusions & Key Insights

CONCLUSIONS & KEY INSIGHTS

Performance metrics: Cancellations & Diversions

- Weather is no. 1 cause for cancellations (54%) &, most frequent during Winter & Spring, least frequent during Fall
- Carrier is the no. 2 reason for cancellation (28%), most frequent during Summer when flights are most frequent
- Mondays & Tuesdays are most affected by cancellations; Fridays & Saturdays are the least affected
- Cancellations & Diversion frequency vary with the number of flights
- Top 5 airlines account for 75% of the cancellations

Performance metric: Flight frequency

- People travel most during Summer and least during Winter
- The top 3 busiest airlines contributed almost 50% to the total of US domestic flight volume in 2015
- The top 50% of the airlines provided almost 85% of the 2015 domestic flights

Performance metric: on-time vs. delays

- More flights are early than late (delayed)
- Although large delays are rare, they account for 55% of the delay time
- The most time was lost due to delay propagation
- The overall mean delays of 9’ is significantly lowered by the large % of flights that are early and on-time
- Hawaiian and Alaska Airlines are top performers; Spirit & United airlines are lowest performers in terms of delay time
- Fall , Spring, & weekends are the best times to make up for delays

Predictive Analytics: Modeling & Prediction

- An all-data encompassing multivariate linear regression model renders 87% accuracy prediction for delay magnitude

APPENDIX

OUTLINE

I. Introduction

II. Exploratory data analysis: Data examination & cleaning (APPENDIX)

III. Exploratory data analysis: visualizations, statistics (descriptive & inferential), trends & relationships
- Analysis of cancellation, delayed flights, and frequency of flights (MACRO)
- Analysis of flight delays: frequency, magnitude, reasons, temporal, spatial & carrier-based analysis (MICRO)

IV. Predictive Analytics: Modeling & Prediction

V. Conclusions & Key Insights

Exploratory data analysis: Data examination & cleaning - HIGHLIGHTS

- ❑ **Raw data: 4 files**
  - Main file very large: 5.8MM rows, 29 columns, 592 MB
- ❑ **Data preparation steps**
  - Optimize files for memory usage by changing datatype
  - Convert the time variables to time stamps
  - Fix the airport codes for the month of October
  - Merging data
- ❑ **Create a final clean file**
- ❑ **Due to file size, I broke the analysis into several Jupiter Notebooks**

Main challenges:

- **Dataset & file sizes**
- **Understanding the relationship among variables**

#	File Name	Description	No. of columns
1.	**flights_raw**	• the main data set to be used in analysis	29
2.	**airports_raw**	• reference file for the airports; • includes details for all airports such as name, city, state, & geographic coordinates; • to be merged with flights	7
3.	**airlines**	• reference file for the airlines; • includes the full names for all the airlines	2
4.	**October_2015_Flights**	• details of all flights for the month of October 2015 (see below why); • to be merged with flights	65

	Dtype	Non-null counts	Unique values	Missing Nulls	Missing (%)
YEAR	int64	5819079	1	0	0.0
MONTH	int64	5819079	12	0	0.0
DAY	int64	5819079	31	0	0.0
DAY_OF_WEEK	int64	5819079	7	0	0.0
AIRLINE	object	5819079	14	0	0.0
FLIGHT_NUMBER	int64	5819079	6952	0	0.0
TAIL_NUMBER	object	5804358	4897	14721	0.3
ORIGIN_AIRPORT	object	5819079	628	0	0.0
DESTINATION_AIRPORT	object	5819079	629	0	0.0
SCHEDULED_DEPARTURE	int64	5819079	1321	0	0.0
DEPARTURE_TIME	float64	5732926	1440	86153	1.5
DEPARTURE_DELAY	float64	5732926	1217	86153	1.5
TAXI_OUT	float64	5730032	184	89047	1.5
WHEELS_OFF	float64	5730032	1440	89047	1.5
SCHEDULED_TIME	float64	5819073	550	6	0.0
ELAPSED_TIME	float64	5714008	712	105071	1.8
AIR_TIME	float64	5714008	675	105071	1.8
DISTANCE	int64	5819079	1363	0	0.0
WHEELS_ON	float64	5726566	1440	92513	1.6
TAXI_IN	float64	5726566	185	92513	1.6
SCHEDULED_ARRIVAL	int64	5819079	1435	0	0.0
ARRIVAL_TIME	float64	5726566	1440	92513	1.6
ARRIVAL_DELAY	float64	5714008	1240	105071	1.8
DIVERTED	int64	5819079	2	0	0.0
CANCELLED	int64	5819079	2	0	0.0
CANCELLATION_REASON	object	89884	4	5729195	98.5
AIR_SYSTEM_DELAY	float64	1063439	570	4755640	81.7
SECURITY_DELAY	float64	1063439	154	4755640	81.7
AIRLINE_DELAY	float64	1063439	1067	4755640	81.7
LATE_AIRCRAFT_DELAY	float64	1063439	695	4755640	81.7
WEATHER_DELAY	float64	1063439	632	4755640	81.7

Predictive Analytics: Modeling & Prediction

- **Delays at Departure vs. Delays at arrival**
- **Compared model results: Scikit-Learn vs. Statsmodel**
- **Tested several variable combination iterations**
  - **Independent** variables considered**:** Month, Day, Day of week, Scheduled time, Distance, Delay causes (5 causes)
  - **Dependent** variables considered: Delays at departure, Delays at arrival
- **Created 3 linear multivariate regressions models for comparison**
  - Model_1: using all data for Delays at departure
  - Broke Model_1 into 2 models using p99 threshold

Predicting Delays Magnitude

Multivariate linear regression Model_2 & Model_3

OLS Regression Results

Dep. Variable:	DEPARTURE_DELAY	R-squared:	0.782
Model:	OLS	Adj. R-squared:	0.782
Method:	Least Squares	F-statistic:	5.808e+05
Date:	Sat, 02 Oct 2021	Prob (F-statistic):	0.00
Time:	12:45:47	Log-Likelihood:	-3.7479e+06
No. Observations:	811443	AIC:	7.496e+06
Df Residuals:	811437	BIC:	7.496e+06
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	17.1514	0.084	205.022	0.000	16.987	17.315
MONTH	0.2592	0.008	32.231	0.000	0.243	0.275
DAY_OF_WEEK	-0.3344	0.014	-24.419	0.000	-0.361	-0.308
AIRLINE_DELAY	0.9012	0.001	1222.768	0.000	0.900	0.903
LATE_AIRCRAFT_DELAY	0.9147	0.001	1385.869	0.000	0.913	0.916
WEATHER_DELAY	0.8928	0.002	590.484	0.000	0.890	0.896

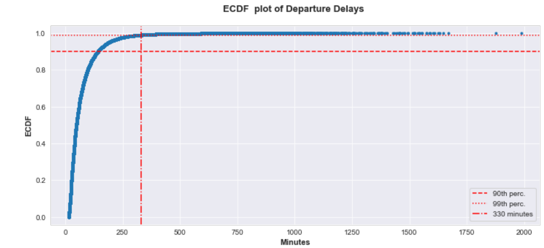
Omnibus:	813505.759	Durbin-Watson:	1.852
Prob(Omnibus):	0.000	Jarque-Bera (JB):	46269581.592
Skew:	5.023	Prob(JB):	0.00
Kurtosis:	38.603	Cond. No.	162.

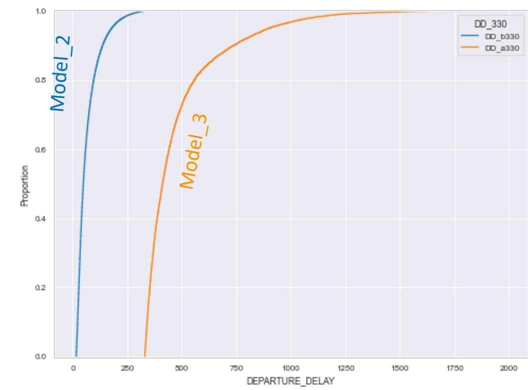
OLS Regression Results

Dep. Variable:	DEPARTURE_DELAY	R-squared:	0.746
Model:	OLS	Adj. R-squared:	0.746
Method:	Least Squares	F-statistic:	6006.
Date:	Sat, 02 Oct 2021	Prob (F-statistic):	0.00
Time:	12:45:51	Log-Likelihood:	-49013.
No. Observations:	8178	AIC:	9.804e+04
Df Residuals:	8173	BIC:	9.807e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	171.1200	3.154	54.255	0.000	164.937	177.303
MONTH	-0.3550	0.305	-1.163	0.245	-0.953	0.243
AIRLINE_DELAY	0.7376	0.005	153.703	0.000	0.728	0.747
LATE_AIRCRAFT_DELAY	0.6670	0.007	92.895	0.000	0.653	0.681
WEATHER_DELAY	0.7045	0.009	82.609	0.000	0.688	0.721

Omnibus:	5083.835	Durbin-Watson:	1.732
Prob(Omnibus):	0.000	Jarque-Bera (JB):	59606.547
Skew:	2.846	Prob(JB):	0.00
Kurtosis:	14.939	Cond. No.	1.17e+03





Model accuracy metrics – Comparison matrix

Regression model	MAE	MAPE	RMSE	NRMSE*	Model acc. %
Model_1 (ALL data)	1.05E-13	0.13	27.31	0.01	87%
Model_2 (<p99)	-8.13E-14	0.14	24.53	0.08	86%
Model_3 (>p99)	5.68E-13	0.03	96.97	0.06	97%

** Normalized RMSE using Max-Min*