EDA 과제
(Exploratory Data Analysis)

- FLIGHT DELAYS-

산업경영공학과 2017010146 이시현

INDEX

1. General Description

Background

- 항공교통 수요가 지속적으로 증가함에 따라 항공기 지연으로 인한 이용자들의 불편과 피해도 늘어나고 있다.
- 지연 발생 시 공항, 항공사, 항공교통이용자 모두에게 시간적·금전적 손해가 발생하며 공항 운영에 있어 상당한 피해를 입게 된다.
- 지연 발생을 예측할 수 있다면 공항 및 항공사 관계자가 사전에 공항운영에 대한 적절한 조치를 취할 수 있어 소비자가 입을 추가적인 피해에 대비할 수 있을 것이다.

1. General Description

Analytics Objectives

- 항공 일정, 날씨, 거리 등의 12개 변수 중 항공 지연을 발생시키는 주요 요인은 무엇인지 분석한다.
- 주어진 변수의 조합으로 항공이 제시간(ontime)에 도착하는지, 지연(delayed)되는지 파악한다.

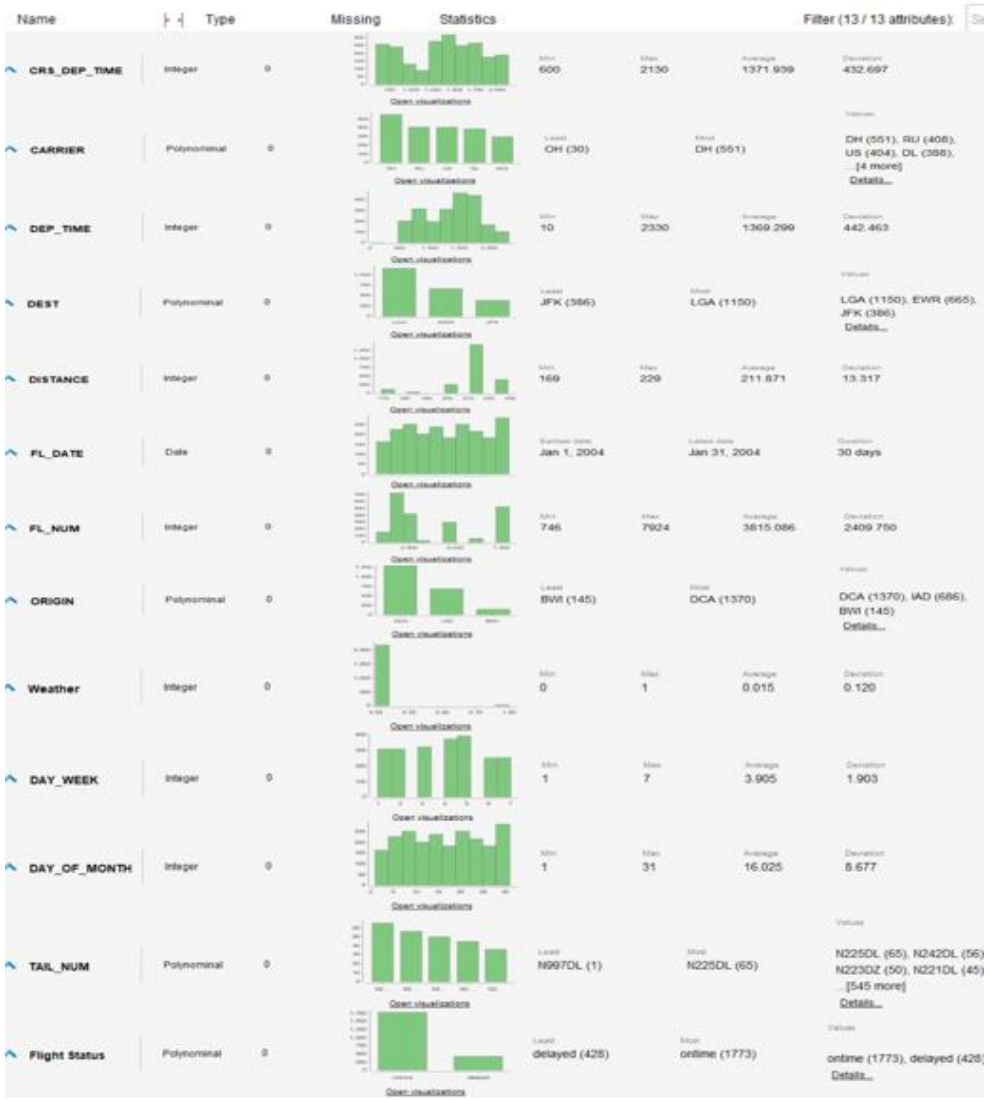
1. General Description

Data Set

- 총 2201개의 항공 데이터가 존재
- 13개의 속성이 존재
- 항공 지연 예측을 하기 위한 클래스(label)은 Flight Status 사용 = 속성 12개
- 결측치 없음

변수명	설명
CRS_DEP_TIME	Scheduled departure time
CARRIER	The airline
DEP_TIME	Actual departure time
DEST	Destination Airport in NY (LGA, EWR, JFK)
DISTANCE	Flight distance in miles
FL_DATE	Flight date (YEAR-MONTH-DATE)
FL_NUM	Flight number
ORIGIN	Departure Airport in Washington DC (DCA, IAD, BWI)
Weather	Whether the weather was inclement(1) or not(0)
DAY_WEEK	Day of week (1=MON ~ 7=SUN)
DAY_OF_MONTH	Day of month (1~31)
TAIL_NUM	This number is airplane specific
Flight Status	Whether the flight was delayed or on time(defined as arriving within 15 min of scheduled time)

2. Descriptive Statistics



변수명	속성	Min / Least	Max / Most	Avg / # of values	Deviation
CRS_DEP_TIME	수치형	600	21	1371.939	432.697
CARRIER	범주형	OH(30)	DH(551)	8	
DEP_TIME	수치형	10	2330	1369.299	442.463
DEST	범주형	JKF(386)	LGA(1150)	3	
DISTANCE	수치형	169	229	221.871	13.317
FL_DATE	수치형	Jan 1, 2004	Jan 31, 2004		
FL_NUM	범주형	746	7924	104	
ORIGIN	범주형	BWI(145)	DCA(1370)	3	
Weather	범주형	0	1	2	
DAY_WEEK	수치형	1	7		
DAY_OF_MONTH	수치형	1	31		
TAIL_NUM	범주형	N997DL(1)	N225DL(65)	549	
Flight Status	범주형	delayed(428)	ontime(1773)	2	

- Rapid Miner의 기술통계를 정리한 결과,
  수치형 변수는 6개, 범주형 변수는 7개로 나타났다.

2. Descriptive Statistics

- 비행은 6시에서 21시 사이에 진행되었으며, 총 8개의 항공사가 이용하였다. 이용한 비행기의 수는 549개이고, 항공편은 104개이다.

- 출발 공항, 도착 공항 모두 각각 3개이며, 비행 거리의 평균은 221.871마일, 분산은 13.317이다.

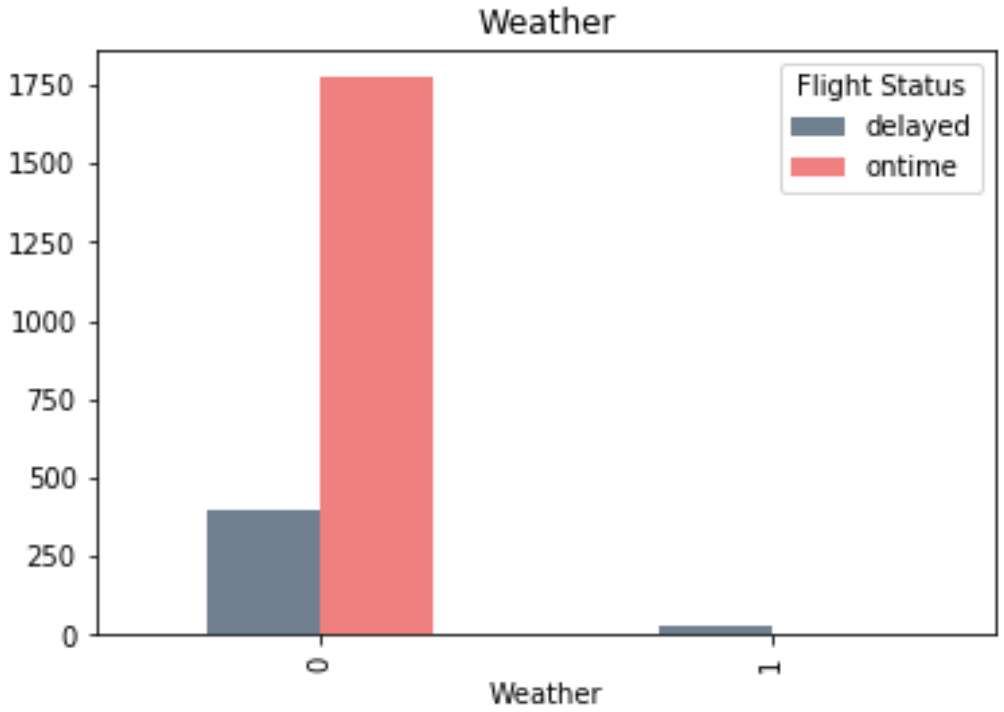
- 비행은 2004년 1월 한달간 진행되었으며, 요일을 1~7로 나타냈다.

- 항공 지연이 발생한 횟수는 428번, 제시간에 도착한 횟수는 1773번으로 항공 지연은 전체 데이터의 19.4%를 차지한다.

변수명	속성	Min / Least	Max / Most	Avg / # of values	Deviation
CRS_DEP_TIME	수치형	600	21	1371.939	432.697
CARRIER	범주형	OH(30)	DH(551)	8	
DEP_TIME	수치형	10	2330	1369.299	442.463
DEST	범주형	JKF(386)	LGA(1150)	3	
DISTANCE	수치형	169	229	221.871	13.317
FL_DATE	수치형	Jan 1, 2004	Jan 31, 2004		
FL_NUM	범주형	746	7924	104	
ORIGIN	범주형	BWI(145)	DCA(1370)	3	
Weather	범주형	0	1	2	
DAY_WEEK	수치형	1	7		
DAY_OF_MONTH	수치형	1	31		
TAIL_NUM	범주형	N997DL(1)	N225DL(65)	549	
Flight Status	범주형	delayed(428)	ontime(1773)	2	

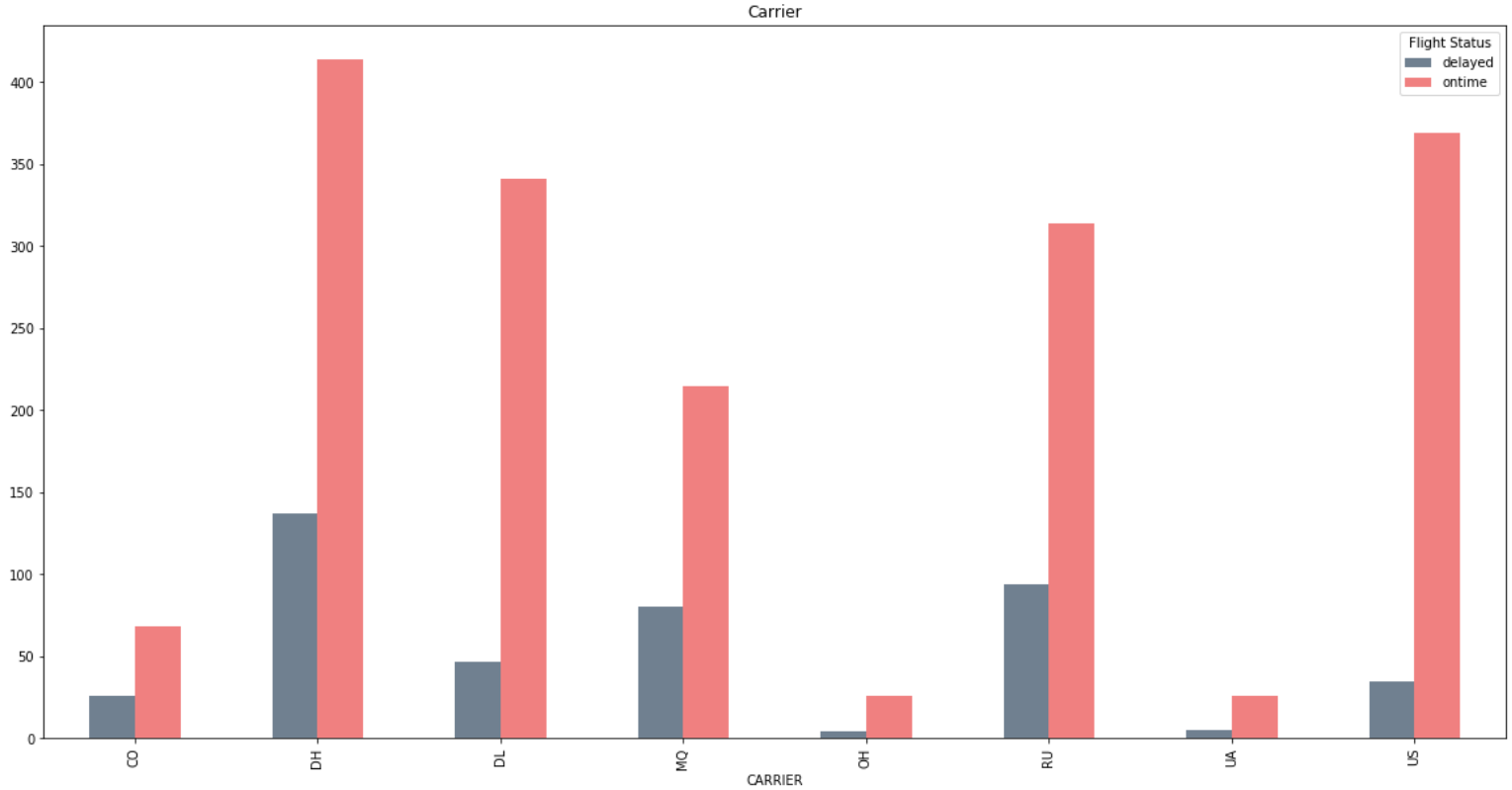
3. Data Visualization

- 날씨가 춥거나 비가 오는 등의 궂은 날씨(Weather)와 항공 지연과의 관계성을 보기 위해 막대 그래프를 사용하였다.
- 아래 그래프를 통해 알 수 있듯이, 궂은 날씨인 경우(Weather=1) 반드시 지연이 되었으며, 궂은 날씨가 아닌 날(Weather=0) 에도 항공 지연이 발생했다. 항공 지연은 날씨만의 문제가 아닌 것을 유추할 수 있다.



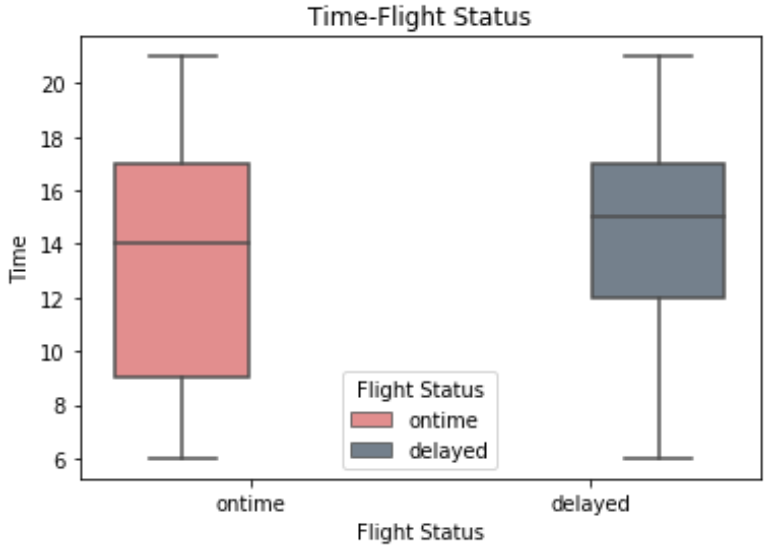
3. Data Visualization

- 항공사와 항공 지연의 관계성을 알아보기 위해 막대 그래프를 그려보았다.
- DL 항공사와 US 항공사가 비슷한 비행 횟수를 가지고 있지만 DL 항공사가 항공 지연 횟수가 더 많다.
- MQ 항공사는 네번째로 적은 비행 횟수를 가지고 있지만 지연 횟수가 세번째로 높다.

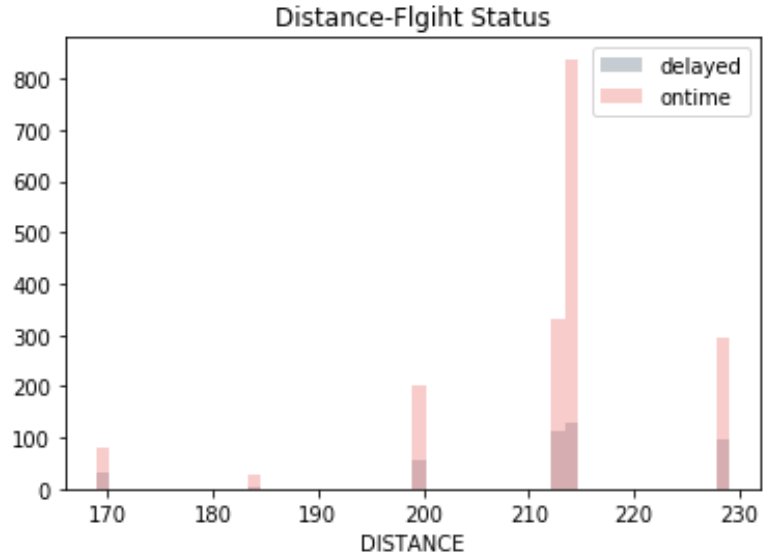


3. Data Visualization

- 예상 도착 시간과 항공 지연과의 관계성을 확인했다.
- 예상 도착 시간은 CRS_DEP_TIME에서 분은 제외시키고 시간만 추출한 변수(Time)를 사용했다.
- Time-Flight Status 박스 플랏을 통해 항공 지연이 될 때 시간들의 중앙값은 15시이며 12시에서 17시 사이에 전체 지연 중 50%가 발생함을 알 수 있다.



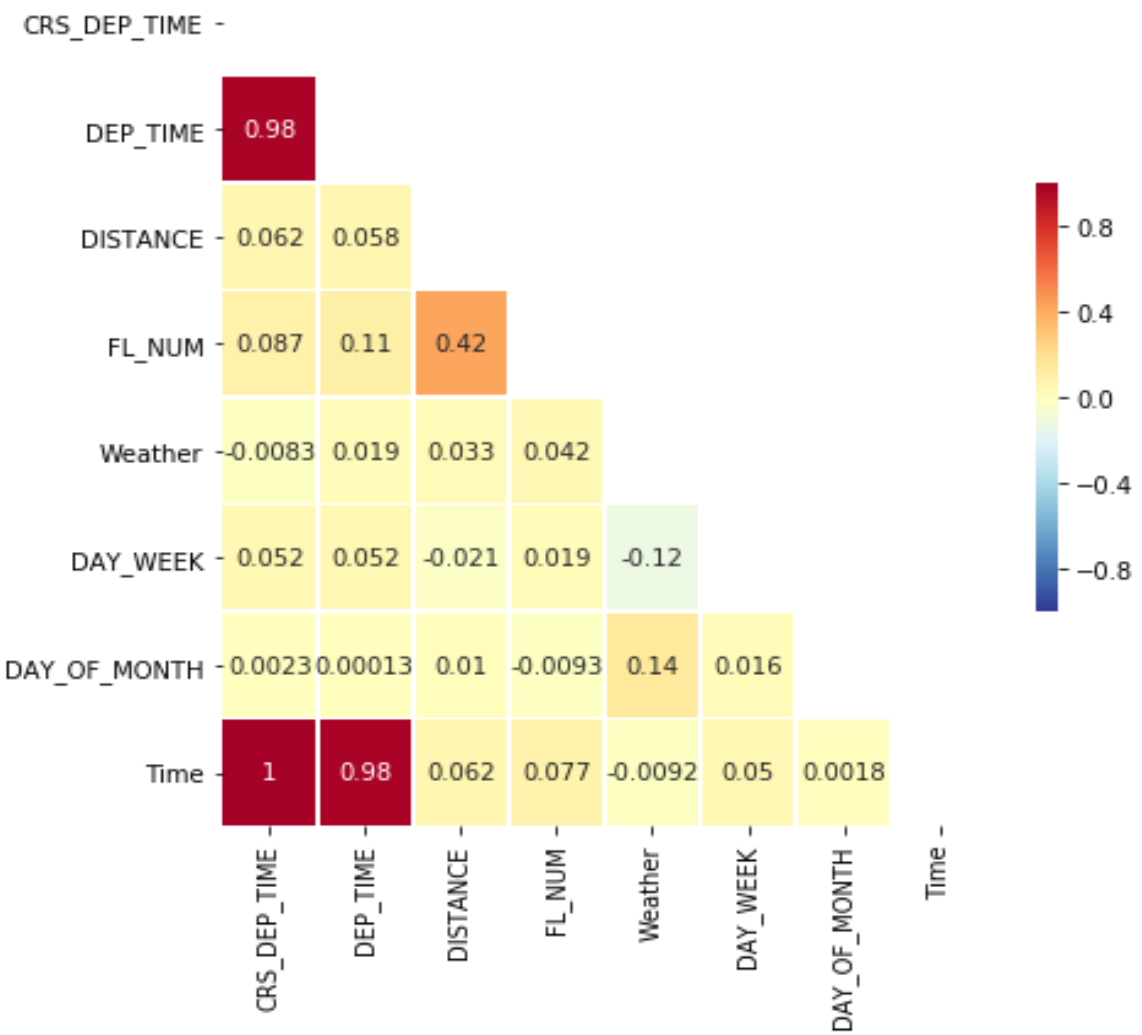
- 비행 거리와 항공 지연과의 관계성은 Distance-Flight Status 히스토그램을 통해 확인하였다.
- 출발공항과 도착공항 모두 각각 3개이기 때문에 거리의 분포가 다양하지 않음을 알 수 있다.
- 210~220 마일 사이의 비행 거리가 가장 많고 그만큼 지연 횟수도 높게 나타났다.



3. Data Visualization

- 항공 지연과 모든 변수들과의 관계성을 보기 위해 다변량 분석을 시행하였다.
- 예상 도착 시간과 실제 도착시간은 상관관계를 보이고, 예상 도착시간, 실제 도착시간, 비행 거리에서 항공 지연 분포의 차이가 발생하고 있다.

3. Data Visualization



- Time은 CRS_DEP_TIME 이용해 만들었기 때문에 Time과 CRS_DEP_TIME, DEP_TIME의 상관관계를 제외시킨 나머지의 상관관계를 분석하였다.
- 상관관계: 항공편 번호&거리 > 날씨&일
  (0.42) (0.14)
- 거리가 멀수록 항공편 번호는 더 높았으며, 일수가 높을수록 날씨가 1에 가까울 확률이 높아짐을 알 수 있다.

Result

EDA 결과, 항공 지연(Flight Status가 1인 데이터)들의 유형은 다음과 같다.

① Weather (궂은 날씨인지 아닌지)이 1인(궂은 날씨인 경우) 비행

② 지연율이 높은 항공사의 비행

③ 예상 도착 시간이 12시에서 17시 사이인 비행

④ 비행거리가 210마일 이상의 비행

위의 4가지 변수를 잘 고려하여 항공 지연 예측을 판단하여야 한다.

감사합니다