nycflight13 시각화

192STG11 우나영

CONTENTS

01

서론

데이터 설명

전처리

분석 목표

02

본론 I

항공사별 지표

03

본론 II

결항과 날씨

정시도착과 비행기성능

04

결론

요약 결론

한계

01

서론

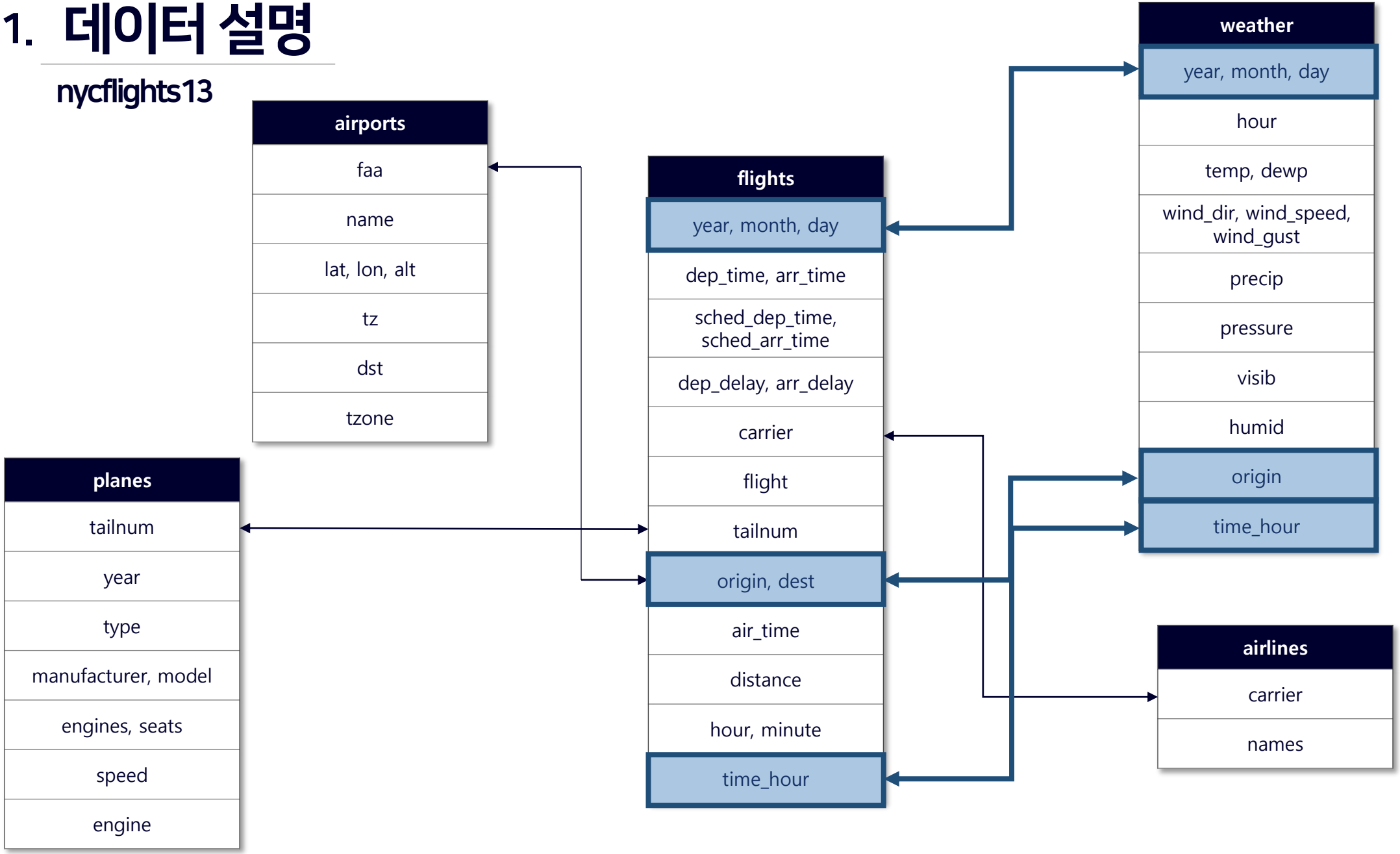
01. 데이터 설명

nycflights13

- R의 nycflight13 패키지 데이터
- NYC에서 출발한 비행편에 대한 실시간 데이터
- Relational Data : 총 5개의 data set으로 구성

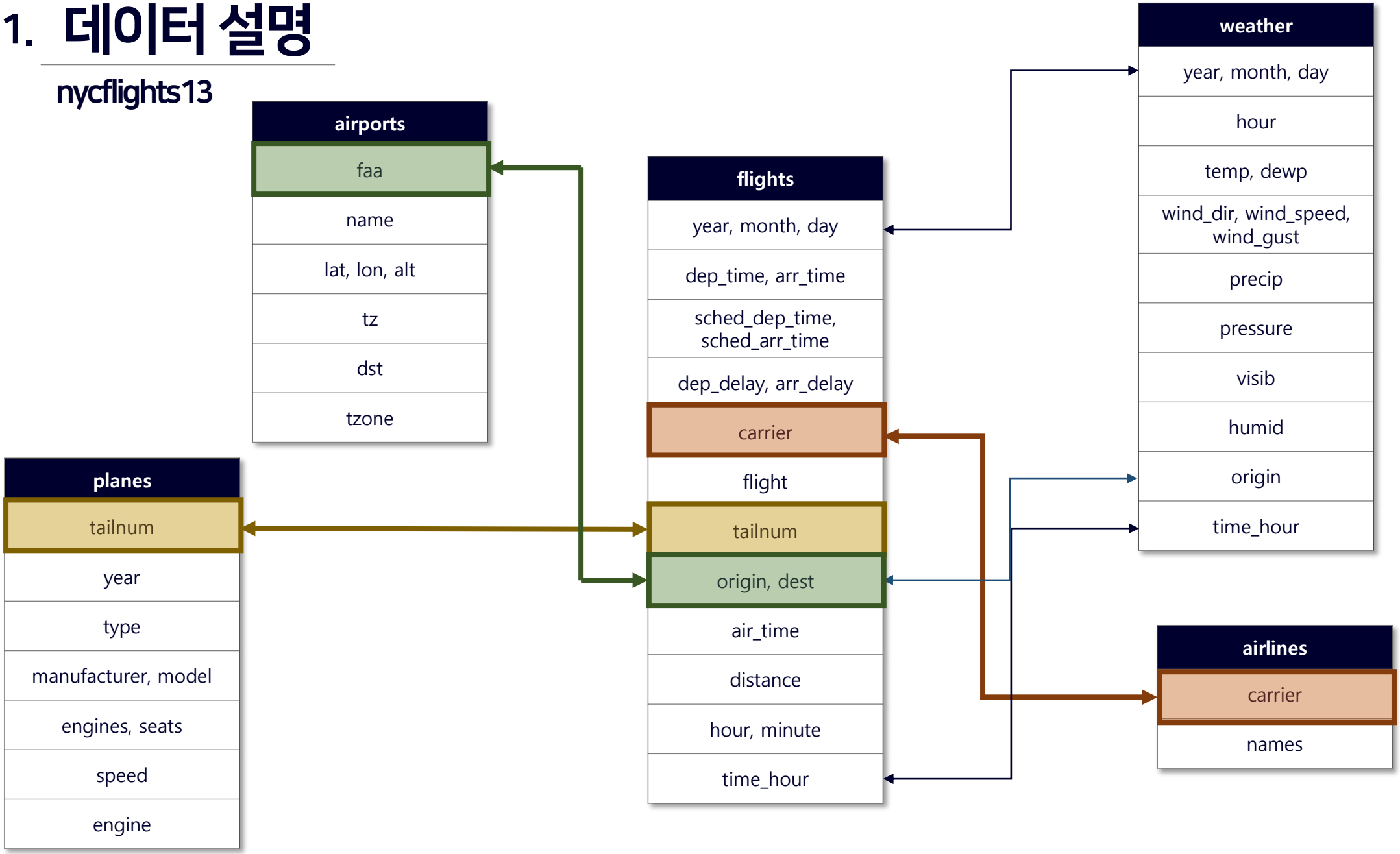
01. 데이터 설명

nycflights13



01. 데이터 설명

nycflights13



01. 전처리

```
> airports_df %>% filter(is.na(tzone))
# A tibble: 3 x 8
  faa   name                                  lat    lon   alt    tz dst   tzone
  <chr> <chr>                               <dbl>  <dbl> <dbl> <dbl> <chr> <chr>
1 EEN   Dillant Hopkins Airport              72.3   42.9   149    -5 A     NA
2 LRO   Mount Pleasant Regional-Faison Field 32.5  -79.5    12    -5 A     NA
3 YAK   Yakutat                             59.3 -139.     33    -9 A     NA
```

```
> airports_df %>% filter(faa %in% c("EEN", "LRO", "YAK"))
# A tibble: 3 x 8
  faa   name                                  lat    lon   alt    tz dst   tzone
  <chr> <chr>                               <dbl>  <dbl> <dbl> <dbl> <chr> <chr>
1 EEN   Dillant Hopkins Airport              72.3   42.9   149    -5 A     America/New_York
2 LRO   Mount Pleasant Regional-Faison Field 32.5  -79.5    12    -5 A     America/New_York
3 YAK   Yakutat                             59.3 -139.     33    -9 A     America/Yakutat
```

airports의 tzone **결측치**를 인터넷 조사를 바탕으로 대체

01. 분석 목표

What am I interested in?

1. 항공사 **실적** 지표 : 결항율 & 정시도착율

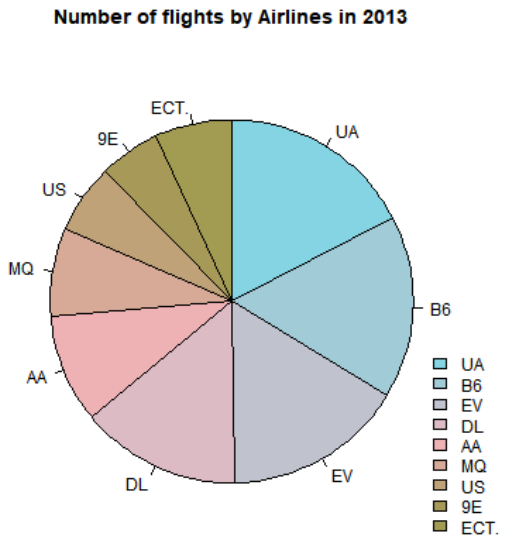
2. 결항율과 정시도착율에 미치는 **요인** : 날씨와 비행기 성능

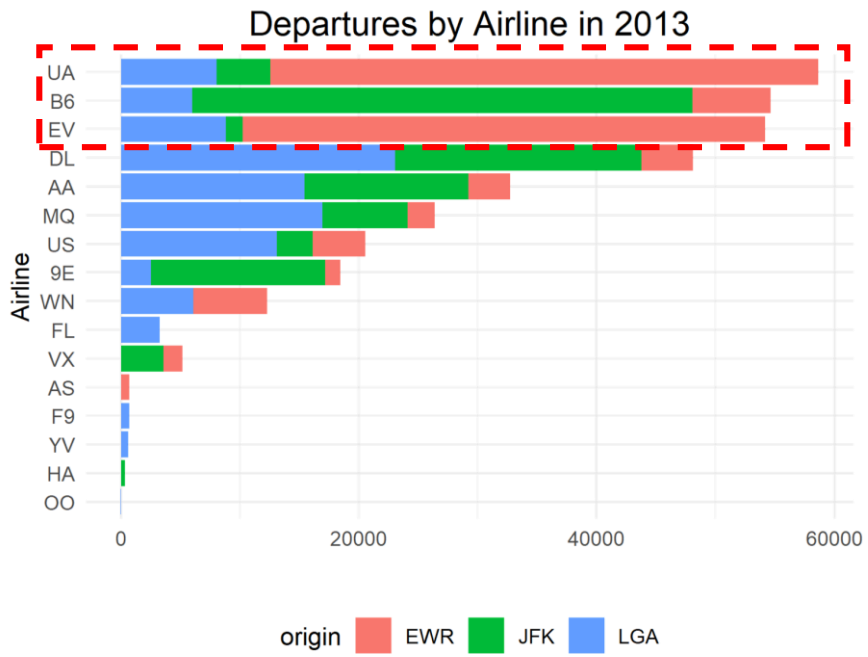
02

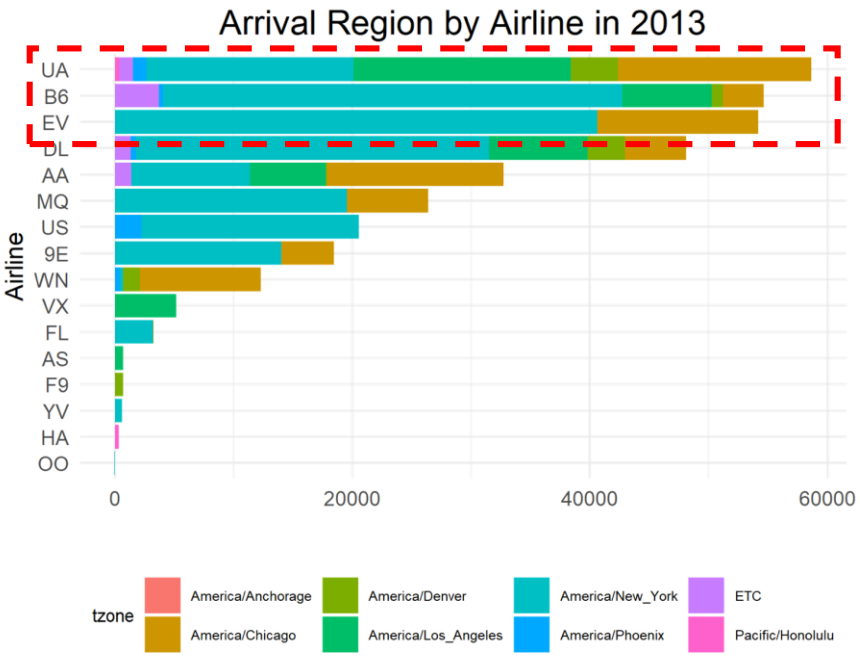
본문 I

02. 항공사별 지표

시장 점유율와 항공편 비율



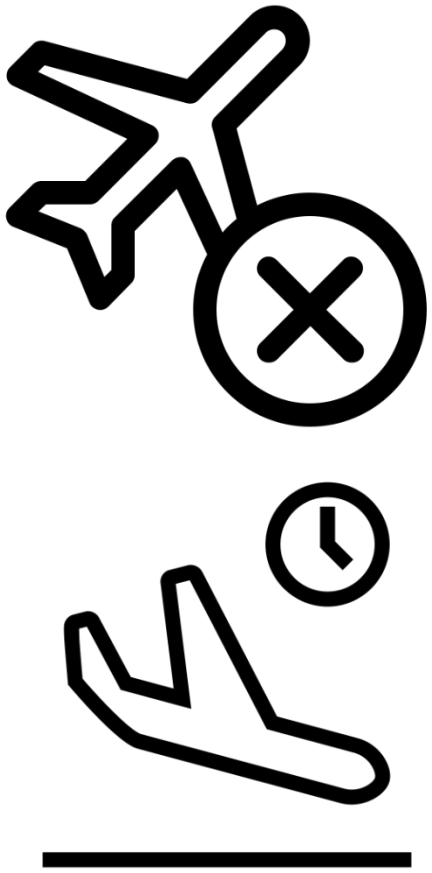




- 편의상 항공사 이름 대신 carrier code 사용
- 시장 점유율 : UA, B6, EV 등이 뉴욕발 비행편의 50% 차이
- UA와 EV는 대부분 EWR 공항 출발인 반면 B6는 대부분 JFK 공항 출발
- tzone을 기준으로 도착 공항 권역 구분
- B6와 EV의 도착 공항은 America/New_York 반면 UA는 다양한 편

02. 항공사별 지표

결항과 정시 도착 정의

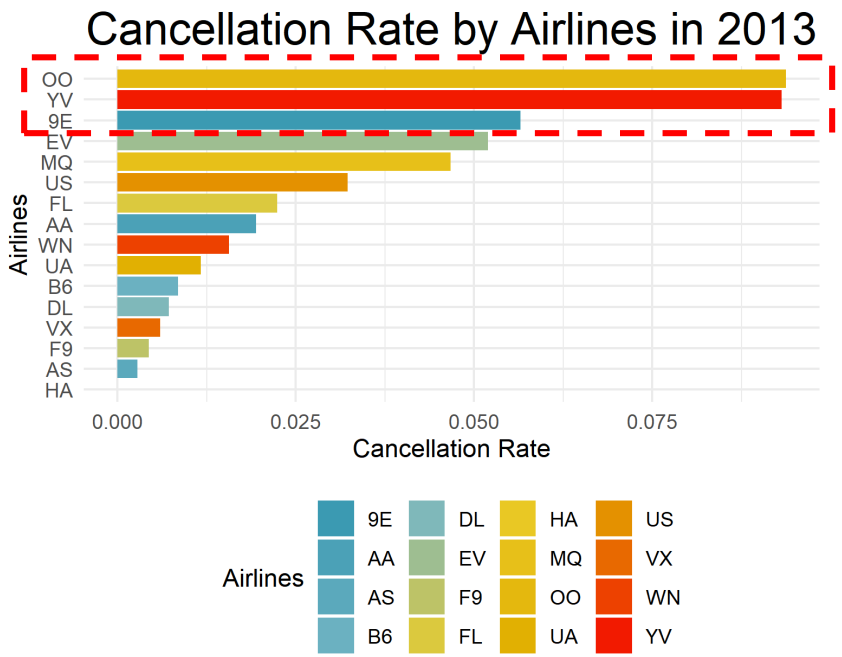


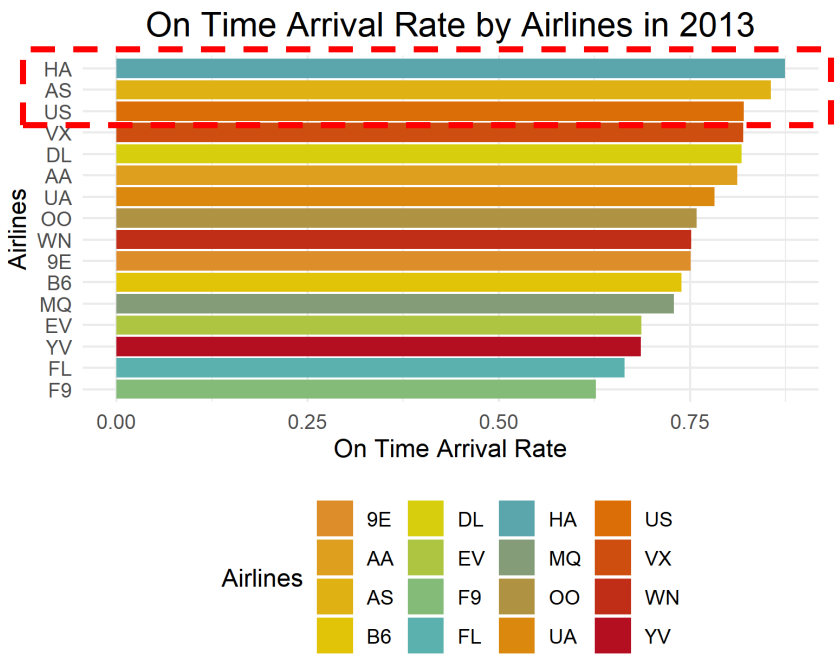
결항 : flights dep_time값이 NA(결측치)일 때,

정시 도착 : flights의 arr_delay 값이 15분 이하

02. 항공사별 지표

결항율과 정시도착율





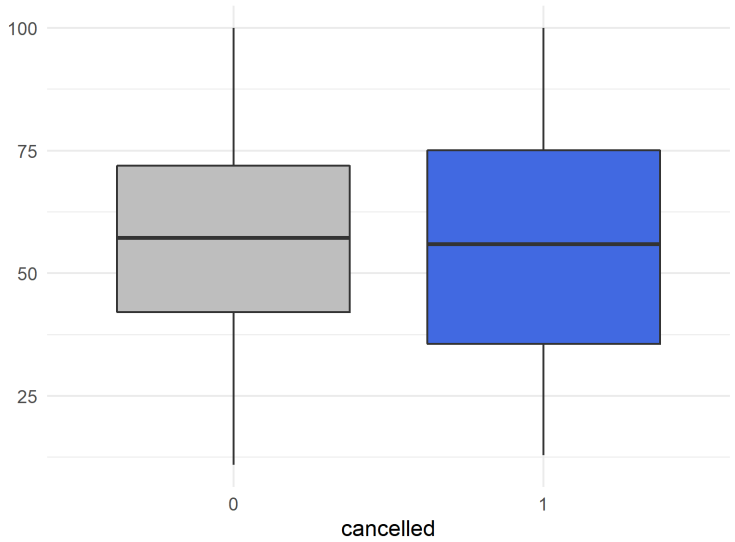
- 편의상 항공사 이름 대신 carrier code 사용
- 결항율 OO와 YV가 9%의 결항율로 선두 , 9E는 5%를 웃도는 수치로 3위
- 정시도착율 HA, AS 선두. US가 뒤를 따름.
- HA는 도착 공항이 HNL 뿐이며 AS는 도착공항이 SEA뿐이다.

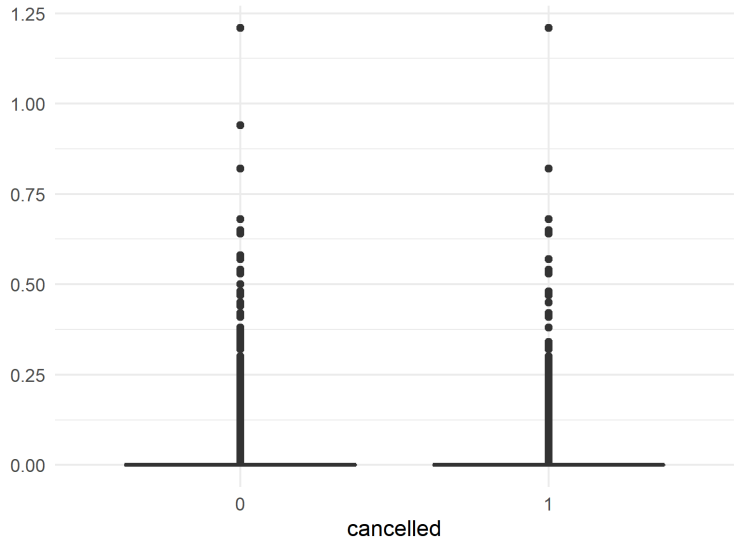
03

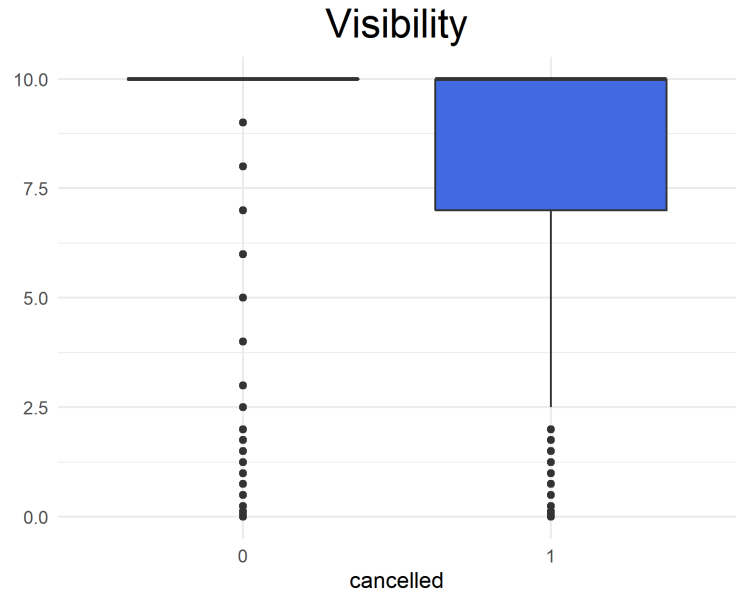
본문 II

03. 결항과 날씨

Boxplot – 차이 없음



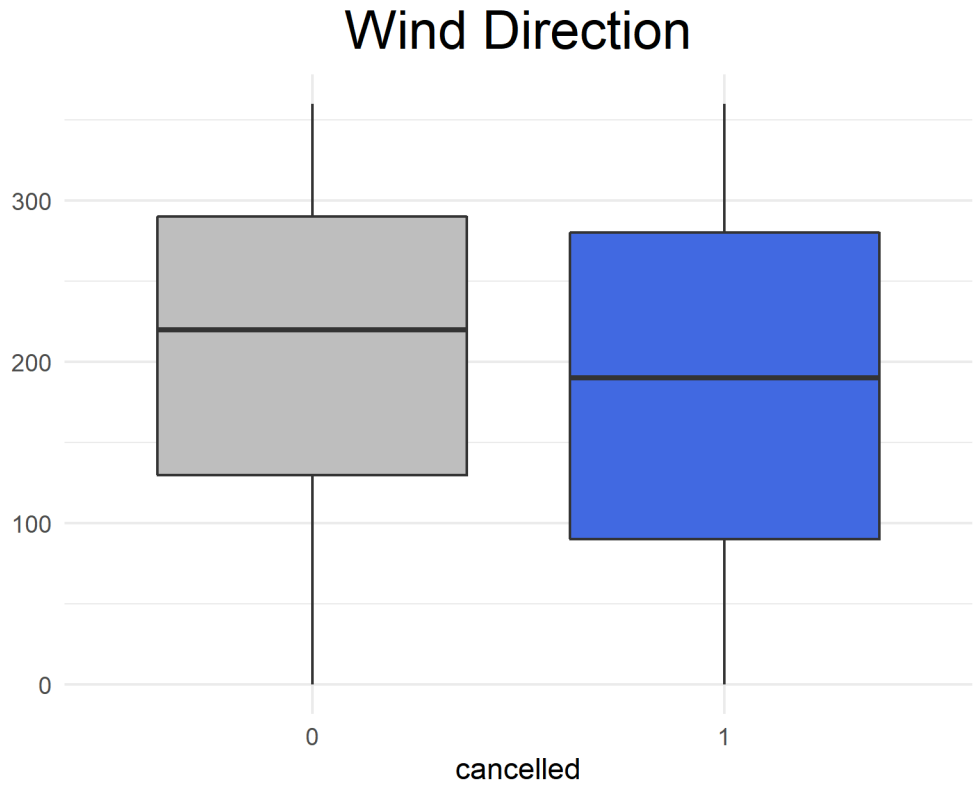


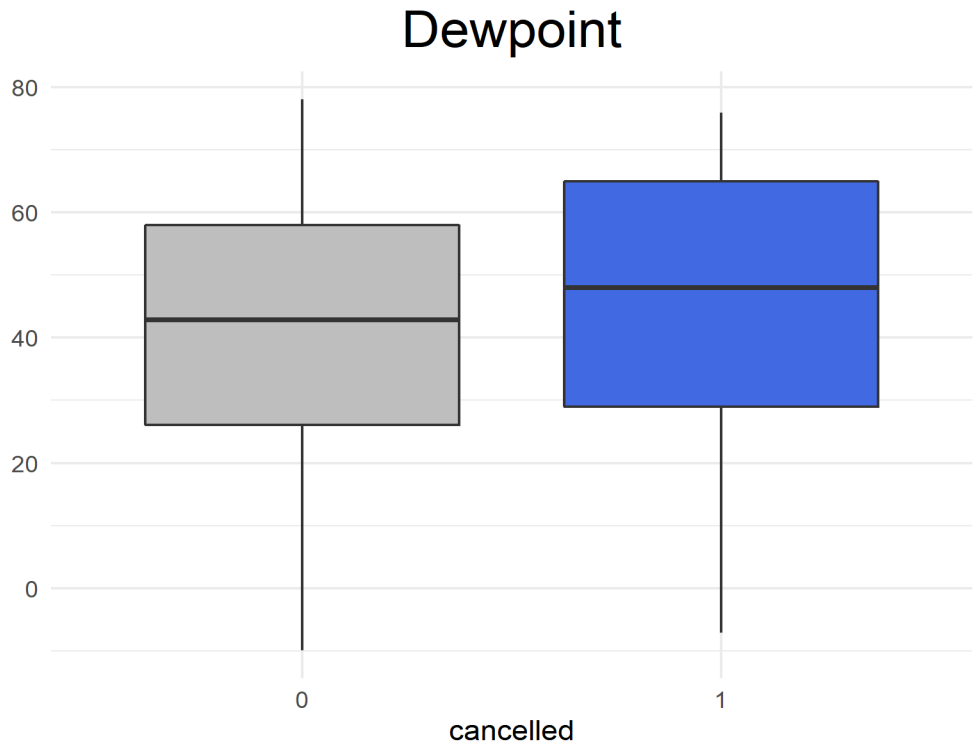


- 결항일때 1, 파란색 boxplot
- 결항 여부에 따라 temp, precipitation, visibility는 차이가 없어 보임.

03. 결항과 날씨

Boxplot – 차이 있음

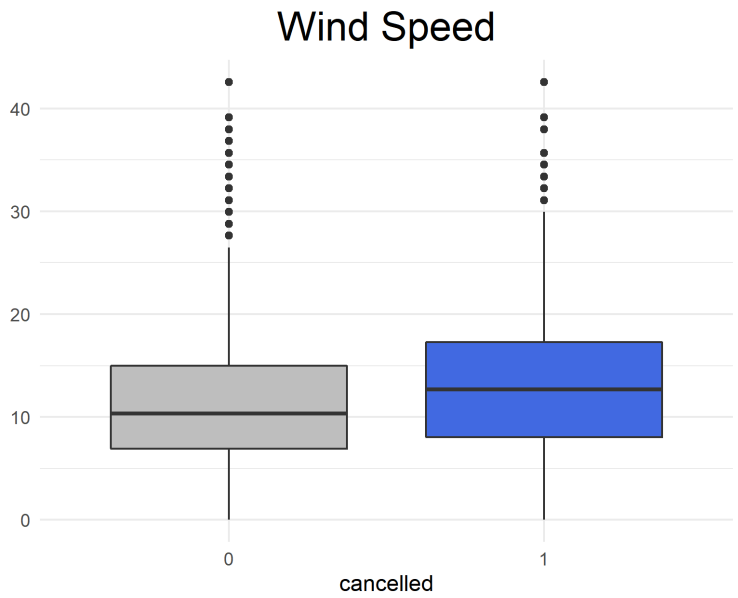


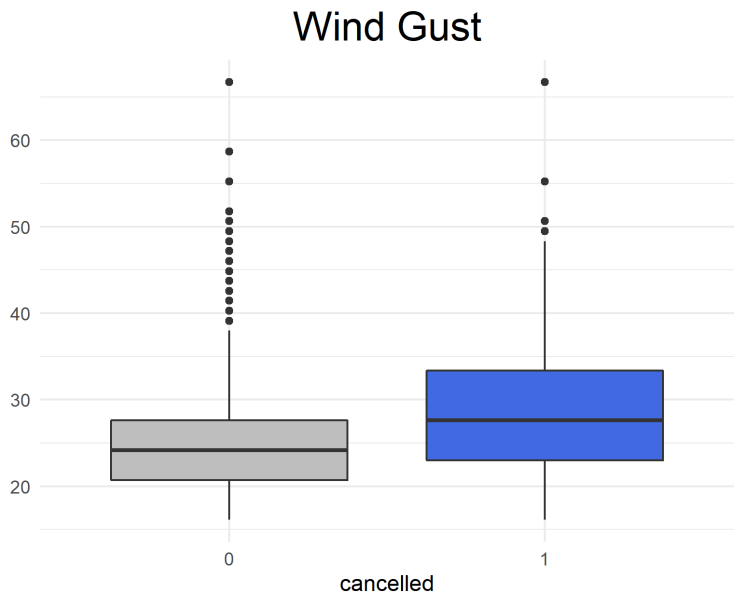


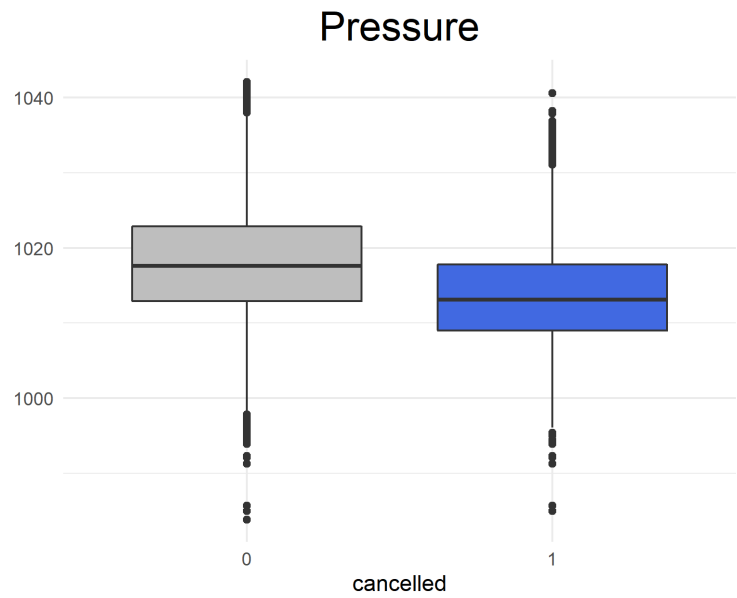
- 결항일때 1, 파란색 boxplot
- 결항 여부에 따라 wind direction, dewpoint 중앙값의 차이가 존재하는 것으로 보임.

03. 결항과 날씨

Boxplot – 차이 분명히 존재 그러나 outlier에 주목해야함



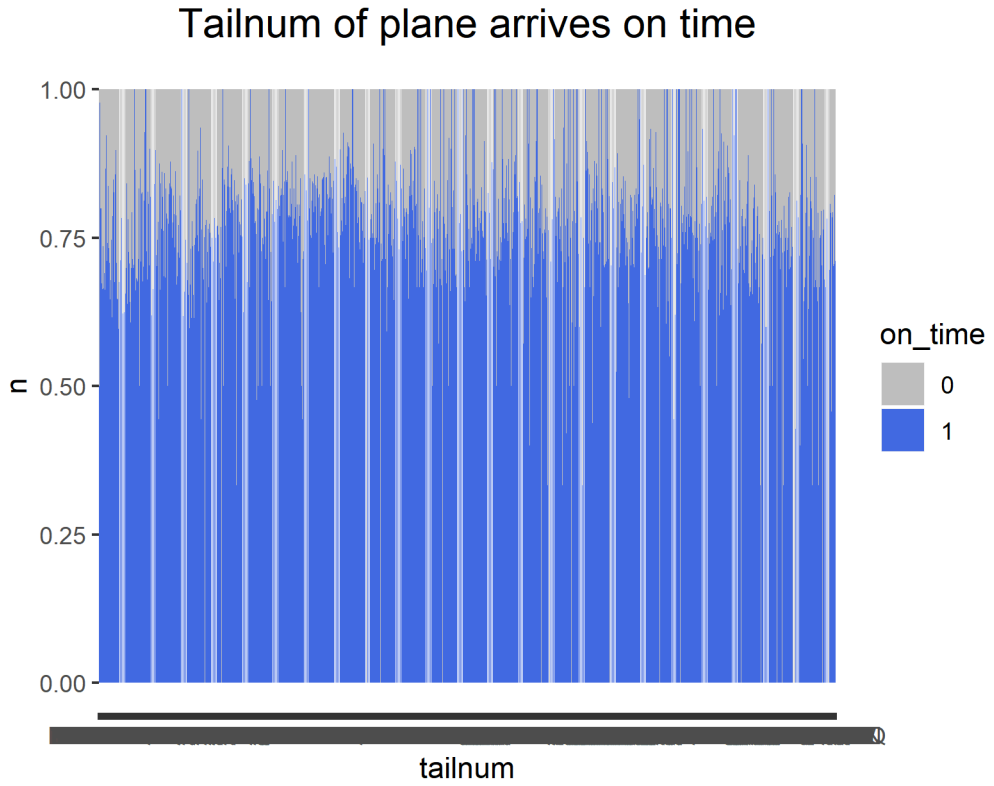




- 결항일때 1, 파란색 boxplot
- 결항 여부에 따라 wind speed, wind gust, pressure 중앙값의 차이가 존재하나 outlier에 주목하여 분석할 필요성 보임.

03. 정시도착과 비행기 성능

정시 도착율이 높은 비행기

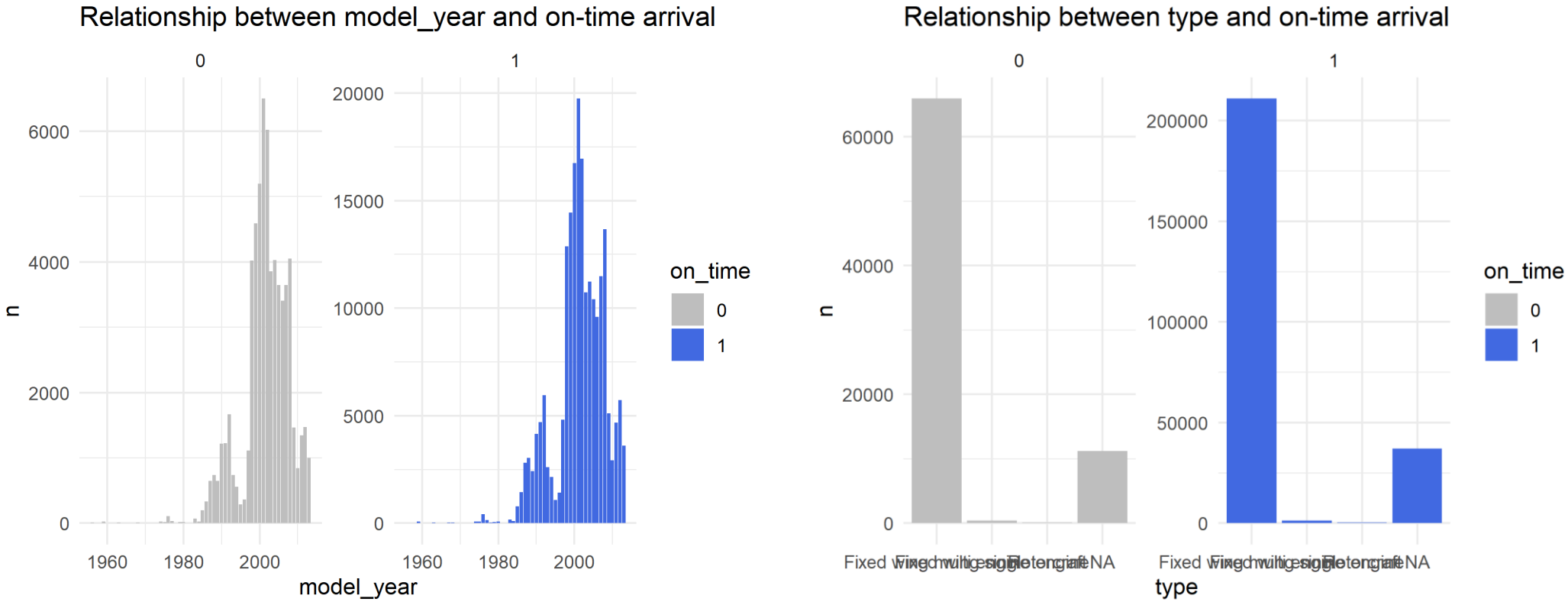


정시 도착	비율
0	23.71%
1	76.29%

- 정시 도착일때 1, 파란색
- 특정 비행기의 경우 표의 비율을 따르지 않는 다는 것을 알 수 있다.
- 따라서, 이를 통해 비행기에 따라 정시 도착율이 달라짐을 알 수 있다.

03. 정시도착과 비행기 성능

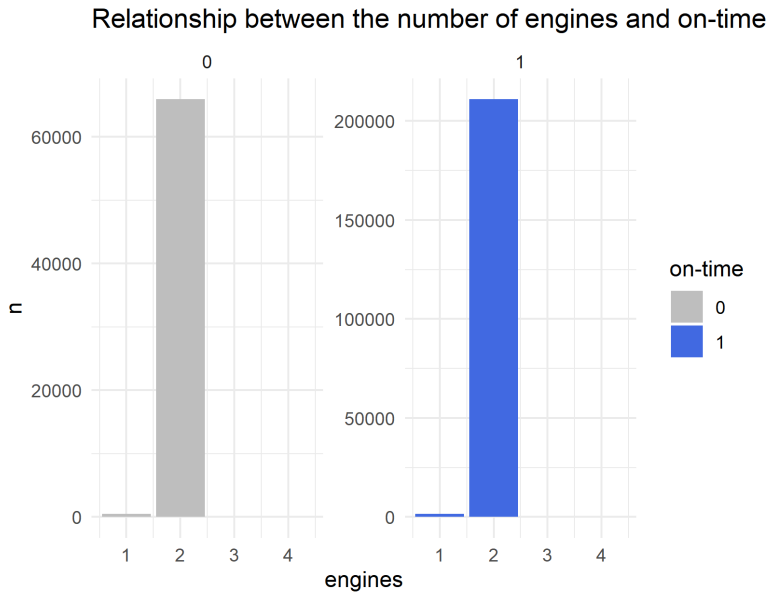
비행기 주요 성능이 정시도착에 끼치는 영향 – 영향 없어 보임

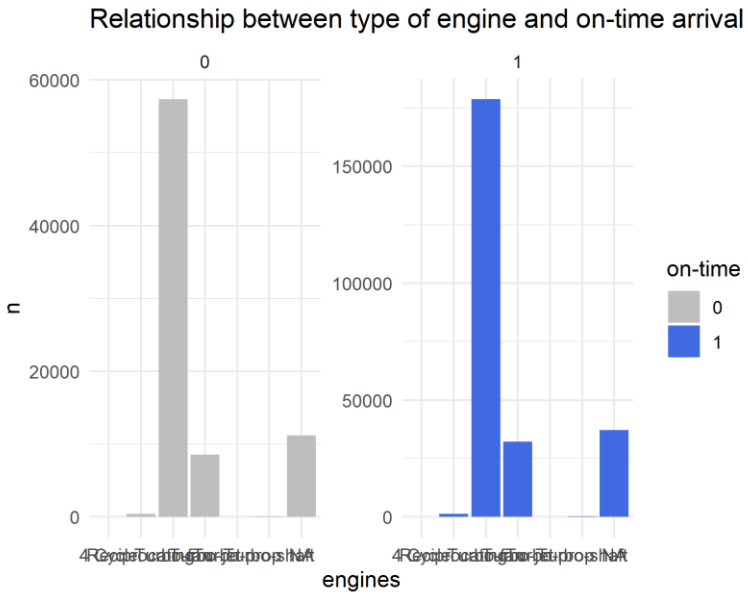


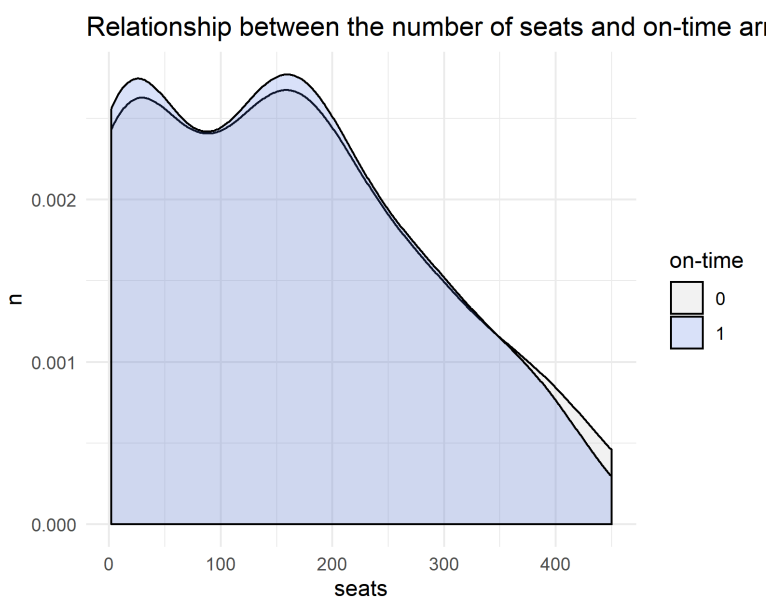
- 정시 도착일때 1, 파란색
- 비행기의 연식(model_year)과 비행기 종류(type)은 정시 도착과 관계 없어 보임.

03. 정시도착과 비행기 성능

비행기 주요 성능이 정시도착에 끼치는 영향 – 영향 없음



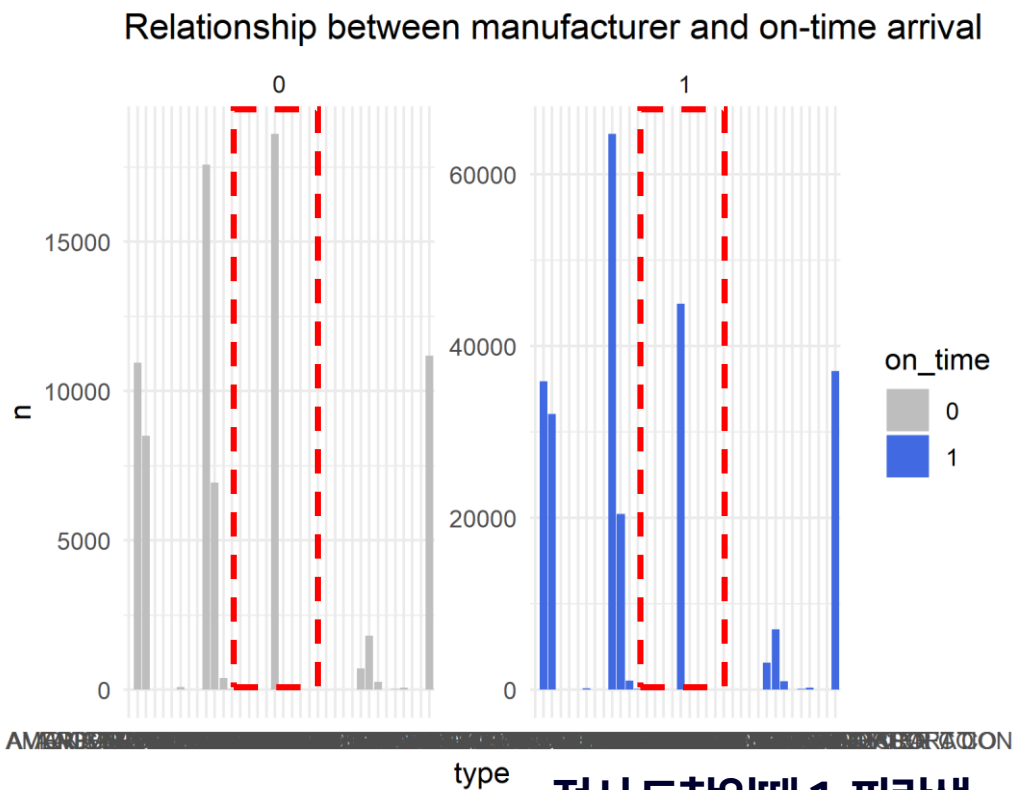


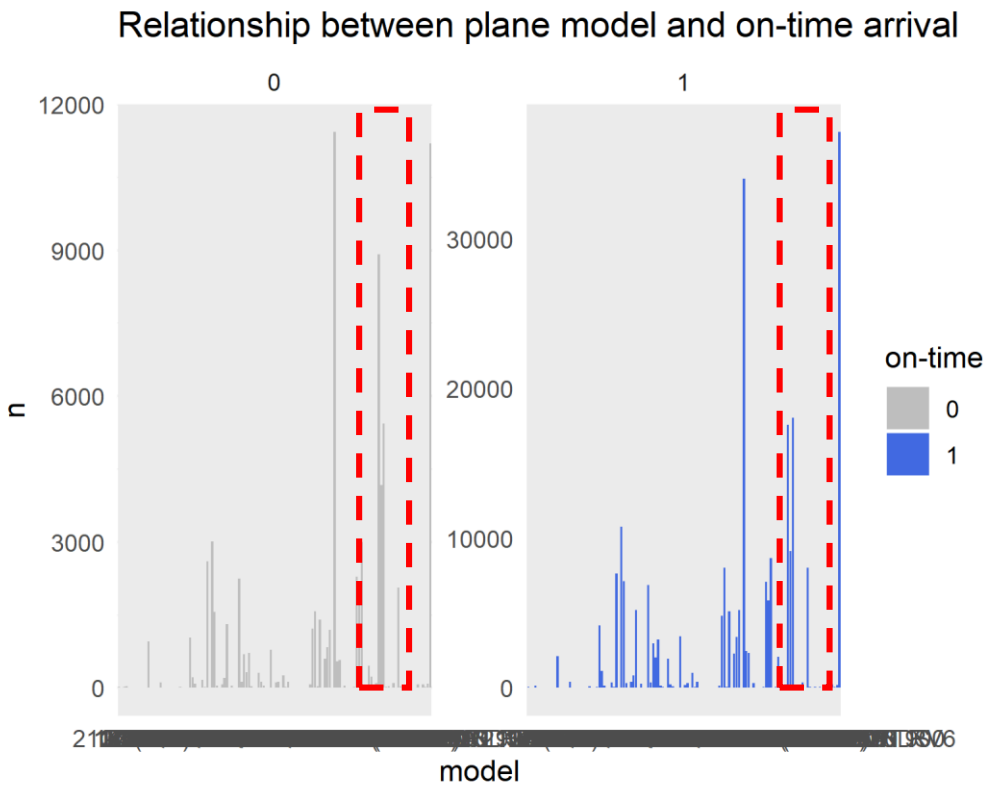


- 정시 도착일때 1, 파란색
- 비행기의 엔진 개수 및 엔진 종류, 좌석수는 정시 도착과 관계 없어 보임.

03. 정시도착과 비행기 성능

비행기 주요 성능이 정시도착에 끼치는 영향 – 영향 있음





- 정시 도착일때 1, 파란색
- 비행기의 생산회사(manufacturer)과 비행기 모델은 정시 도착과 관계 있어 보임.

04

결론

04. 요약 결론

Summary

- 항공사별 주요 실적 지표 : 결항율, 정시도착율
- 결항과 날씨는 밀접한 연관이 있을 것으로 보임.
- 정시 도착과 비행기 정보는 연관이 있는 것으로 파악됨.
  그러나, 비행 여건에 대한 분석이 추가적으로 필요함.

04. 요약 결론

What's next?

- 결항을 예측 모델링하기 위해 날씨 feature에 대한 이해
- 같은 조건에서 비행한 비행기들의 정보 비교(model, engine 등)
- 항공사가 결항율과 정시도착율 control 가능한 feature 찾기

04. 한계점

- **결측치** 고려 안함
- 결항과 날씨 부분의 boxplot 차이를 **통계적 검정**하지 않음
- 정시도착의 **이분법**적 정의 : 도착 지연 16분을 기준으로 정의
- 2개 이상의 **data merge** 시도 못 함
- 항공에 대한 **domain knowledge** 부족

Thank you