NYC flights analysis

Challenge questions

Flights departed from NYC airport

- Anomalies in the data?
- Flight distance influence on its delay?
- Best New York airport?

Data sets

What's in our data?

Flights.csv

→ 336 776 rows

→ year
month
day
departure
scheduled_departure
arrival
scheduled_arrival
carrier
flight_id
origin
destination
distance

Aiports.csv

→ 1 458 rows

→ FAA
name
latitude
longitude
altitude
UTC
DST
timezone

Airlines.csv

→ 16 rows

→ carrier
name

Weather.csv

→ 26 115 rows

→ origin
year
month
day
hour
temperature
dewpoint
humidity
wind_direction
wind_speed
wind_gust
precipitation
pressure
visibility

Anomalies in the data?

Issues in flights.csv

Missing values

- 8 255 null values in `departure`
- 8 713 null values in `arrival`

FAA Tags

- `EWR, LGA, JFK, ERW`
- No `ERW` in airports.csv
- `ERW` distribution analysis

Non-existing date

- 29th Feb. but no Feb. 28th 2013
- 2013 was not a leap year

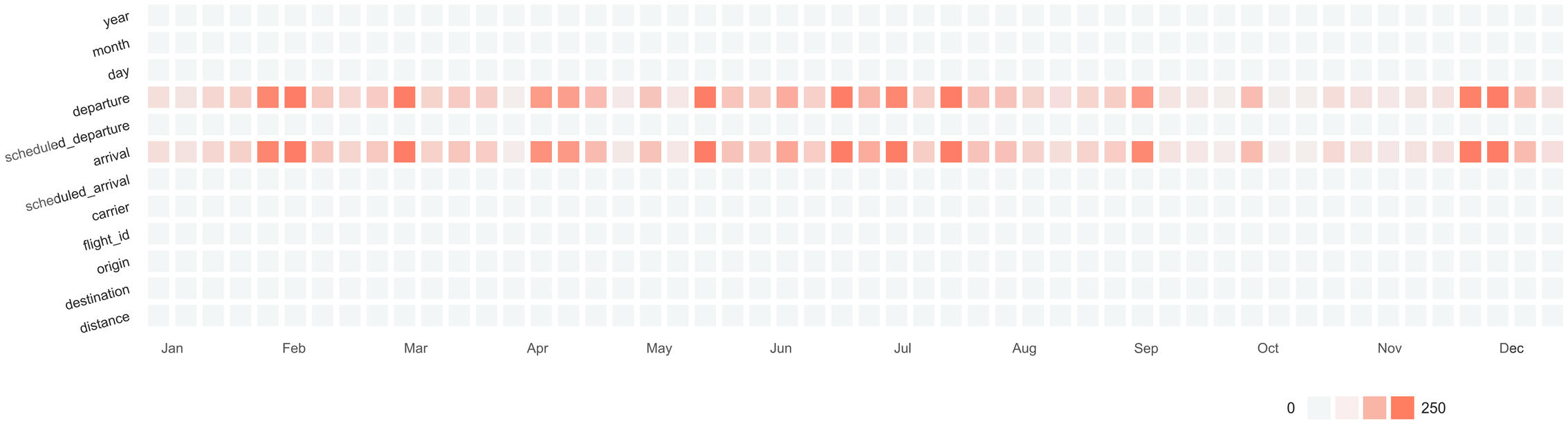
Hour transformation

- HHMM time format
- 24xx values, and 00xx
- MM anomalies

Anomalies in the data?

Inspecting missing values, by week

NA values count grouped by week, in flights.csv

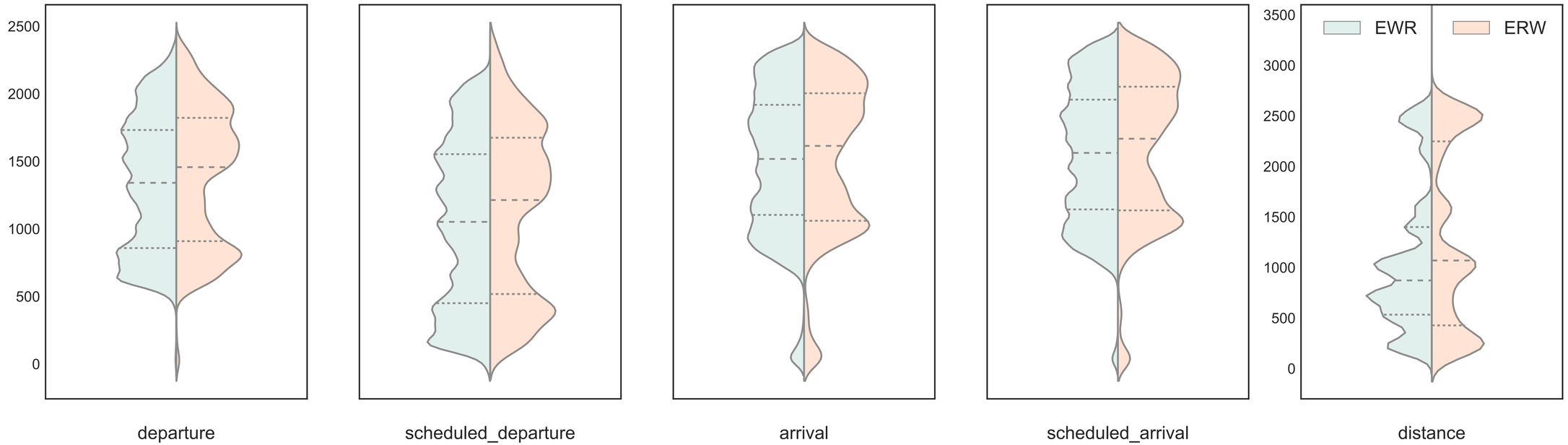


Note: NA only in `departure` and `arrival` variables, missing for both in same flights. Cancelled flights?

Anomalies in the data?

Comparing `ERW` and `EWR` distributions

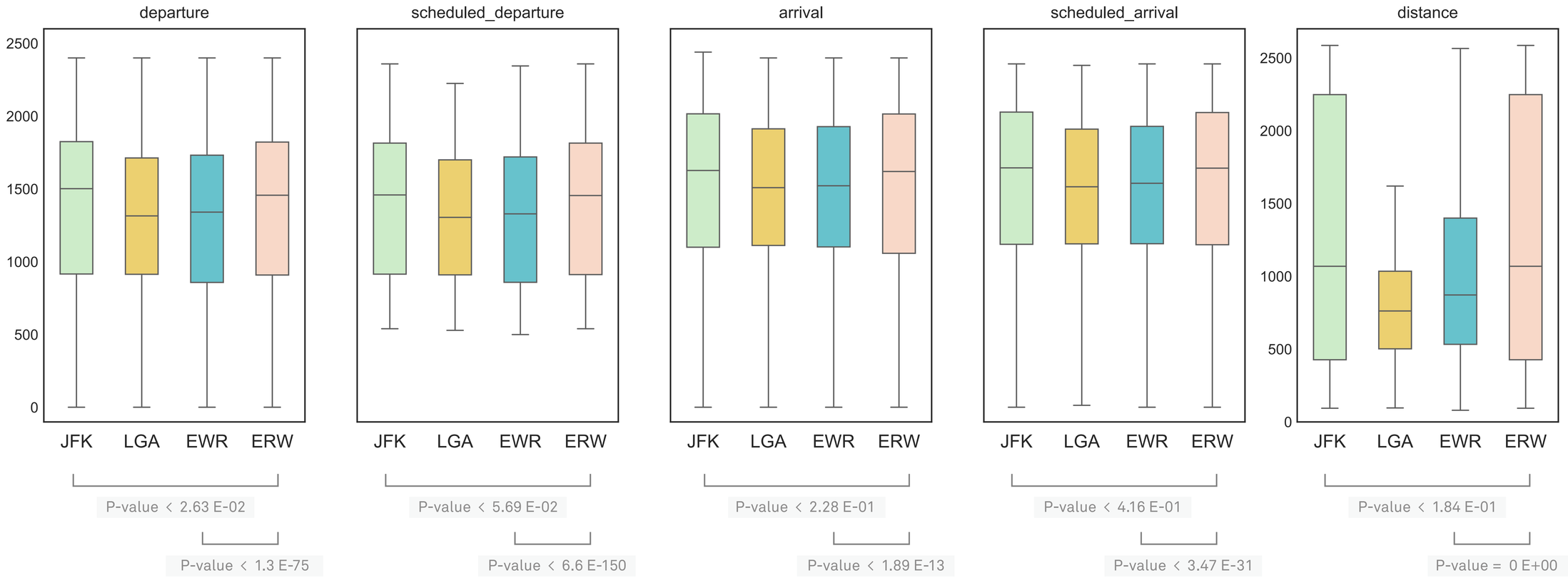
Violinplots comparison between `EWR` and `ERW` data



Note: `EWR` and `ERW` distributions differences, especially in flight distances.

Anomalies in the data?

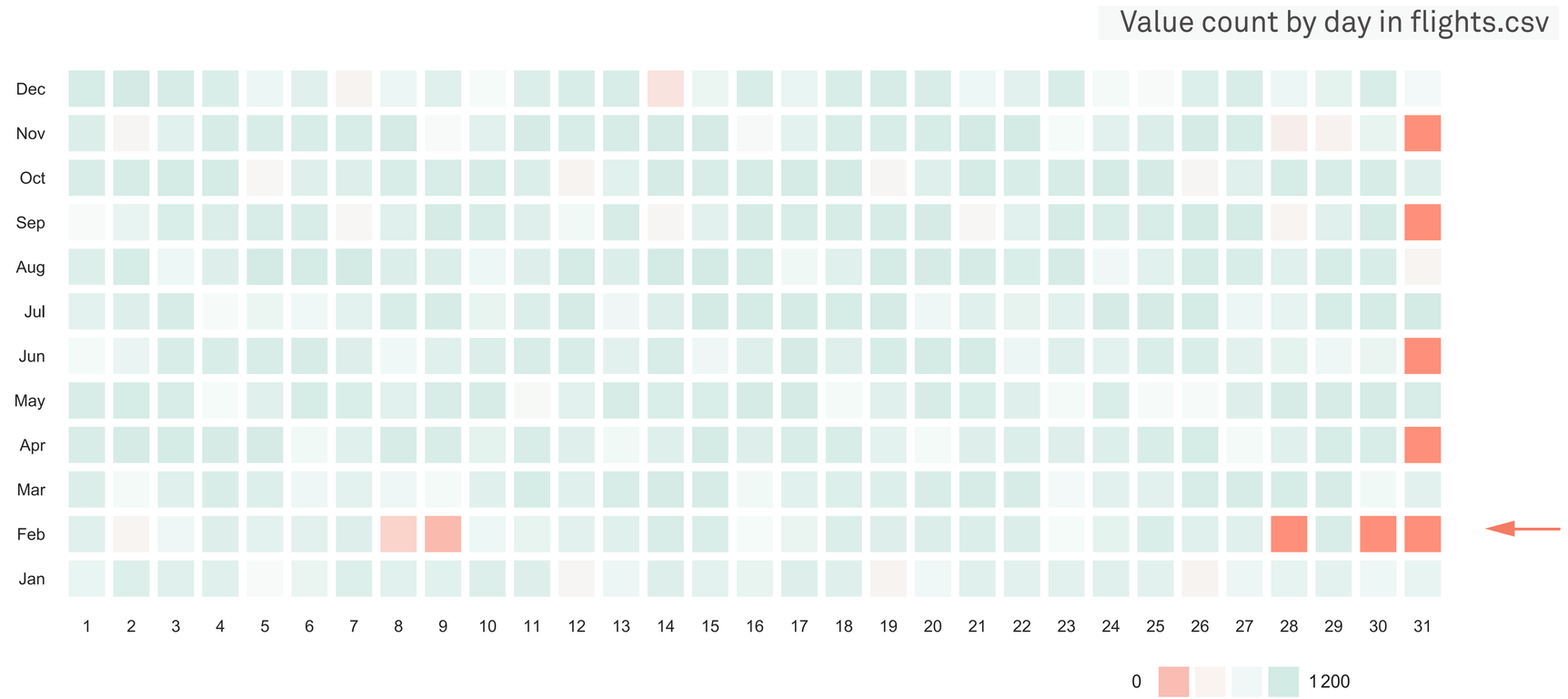
Distributions by origin



→ ERW have similar distributions than JFK

Anomalies in the data?

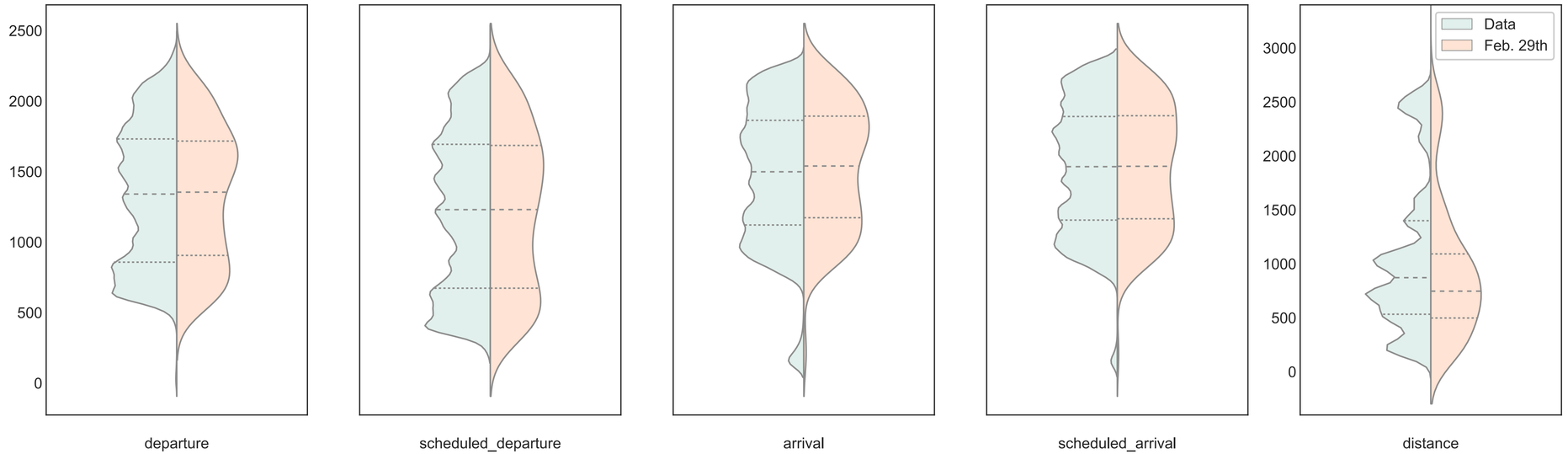
Inspecting dates



Note: No values for Feb. 28th, but for non-existing Feb. 29th.

Data

Violin plots of Feb. 29th time and distance variables



P-value < 7.9 E-01

P-value < 5.9 E-01

P-value < 5.8 E-04

P-value < 3.9 E-01

P-value = 4.6 E-02

Note: Feb. 29th distributions in same range as rest, except `arrival` and `distance`

Anomalies in the data?

Issues in flights.csv

Missing values

- 8 255 null values in `departure`
- 8 713 null values in `arrival`

→ **Drop NA**

FAA Tags

- `EWR, LGA, JFK, ERW`
- No `ERW` in airports.csv
- `ERW` distribution analysis

→ **Transform `ERW` into `JFK`**

Non-existing date

- 29th Feb. but no Feb. 28th 2013
- 2013 was not a leap year

→ **Transform Feb. 29th to 28th**

Hour transformation

- HHMM time format
- 24xx values, and 00xx
- MM anomalies

→ **Correct time format**

→ **Drop Feb. 29th (finally)**

Distance influence on delay

Calculate delays?

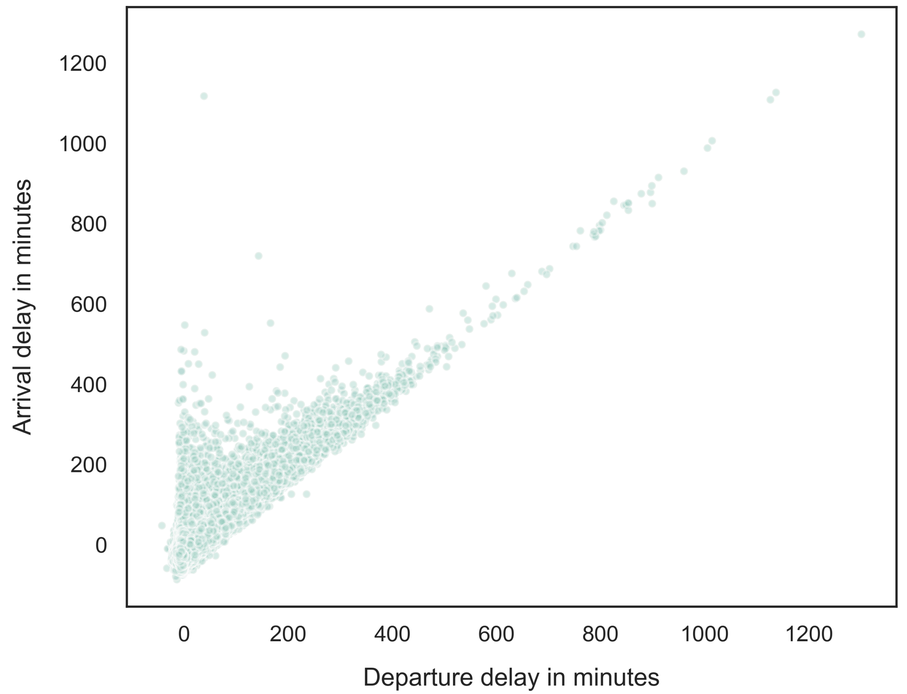
Delay anomalies

- Negative values
- Errors around midnight

→ **Correct calculations**

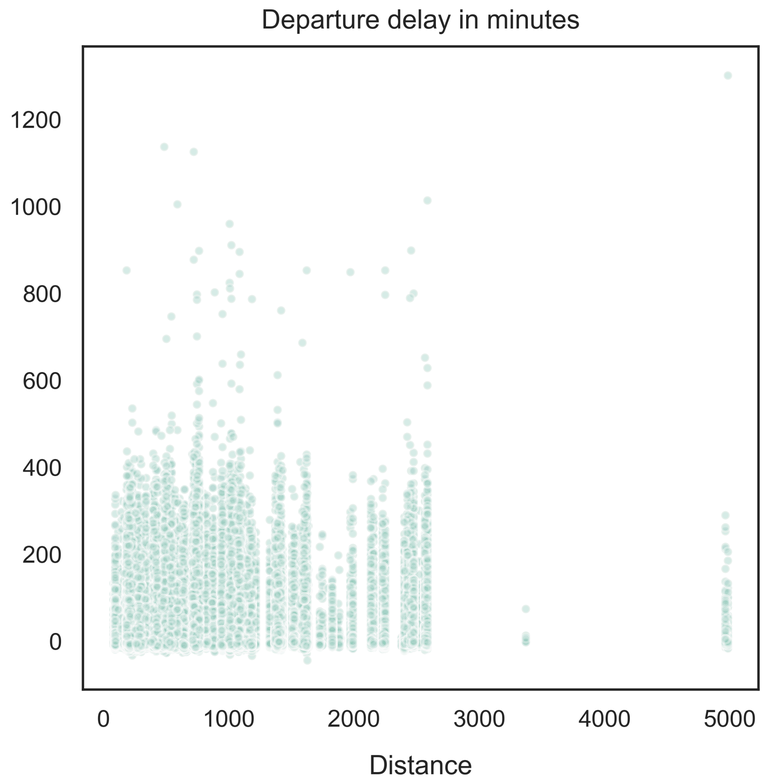
Distance

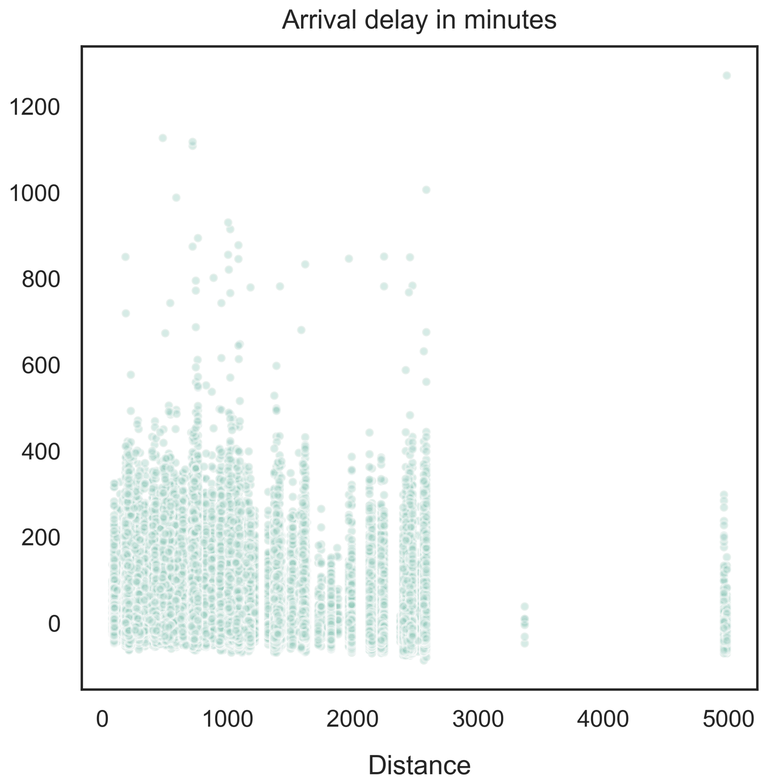
- Non-normal distribution
- `log` transform?



Distance influence on delay

`Departure` vs. `arrival` delay

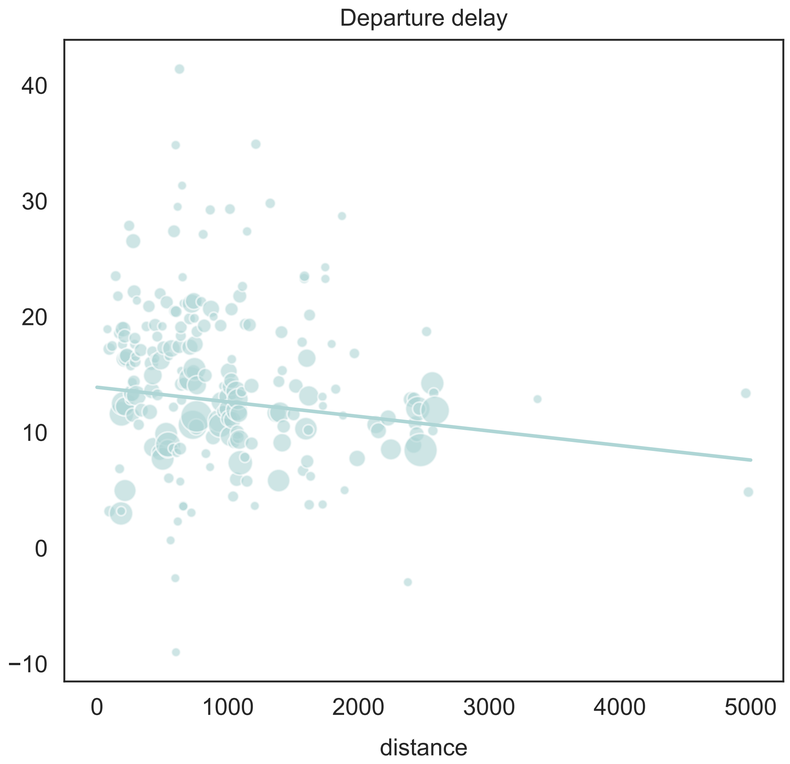


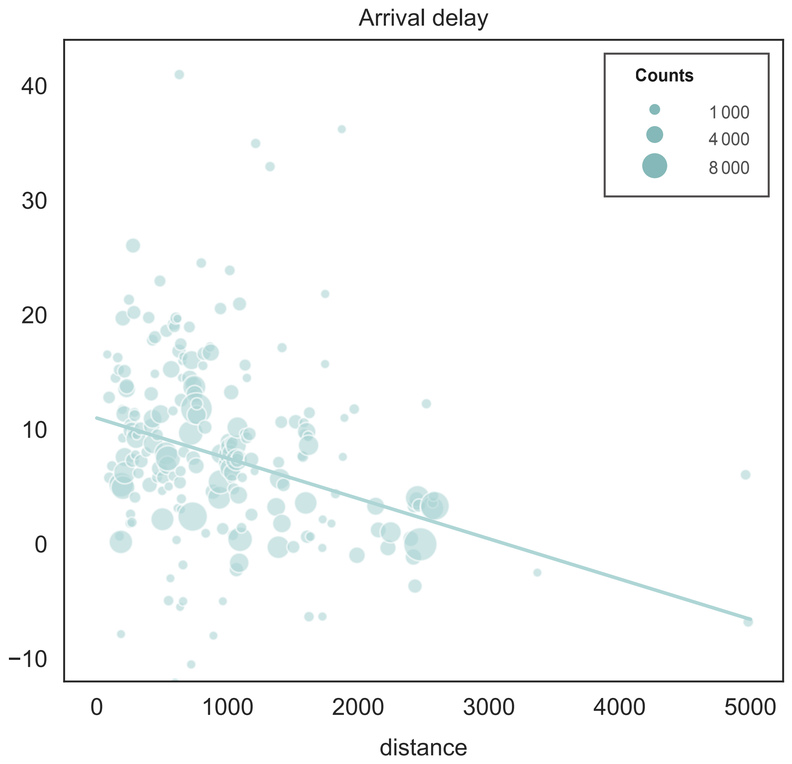


Note: Hard to observe a relationship, group by `distance`?

Distance influence on delay

Fitting a linear model





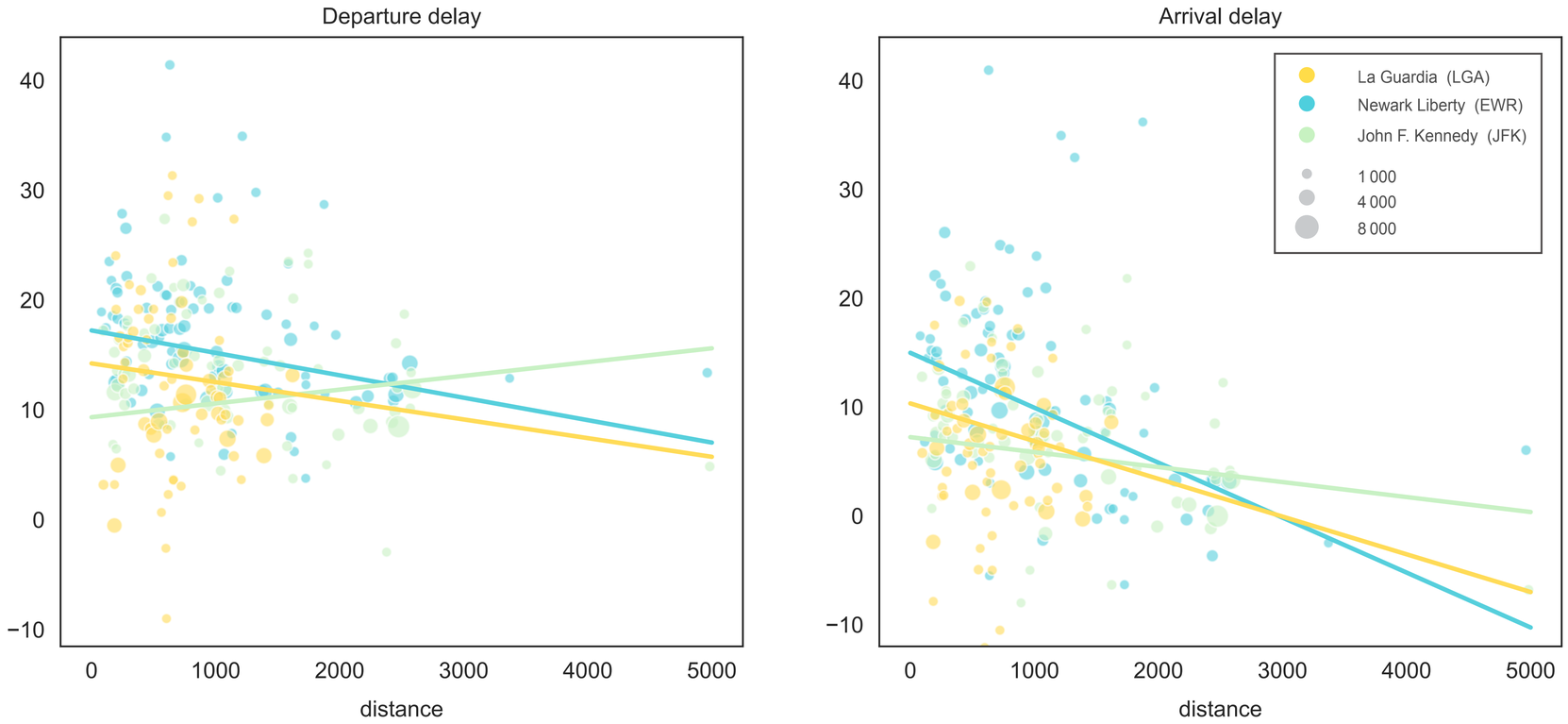
$$departure.delay = 13.81 - 0.0012 * distance$$

$$arrival.delay = 10.93 - 0.0035 * distance$$

→ opposite relationship between `distance` and `delay`

Distance influence on delay

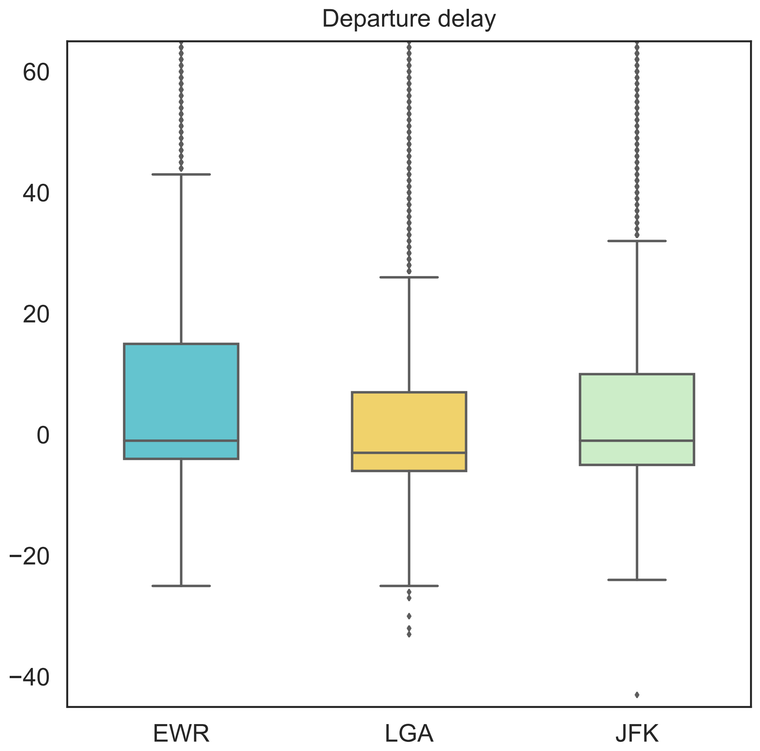
Airport origin comparison

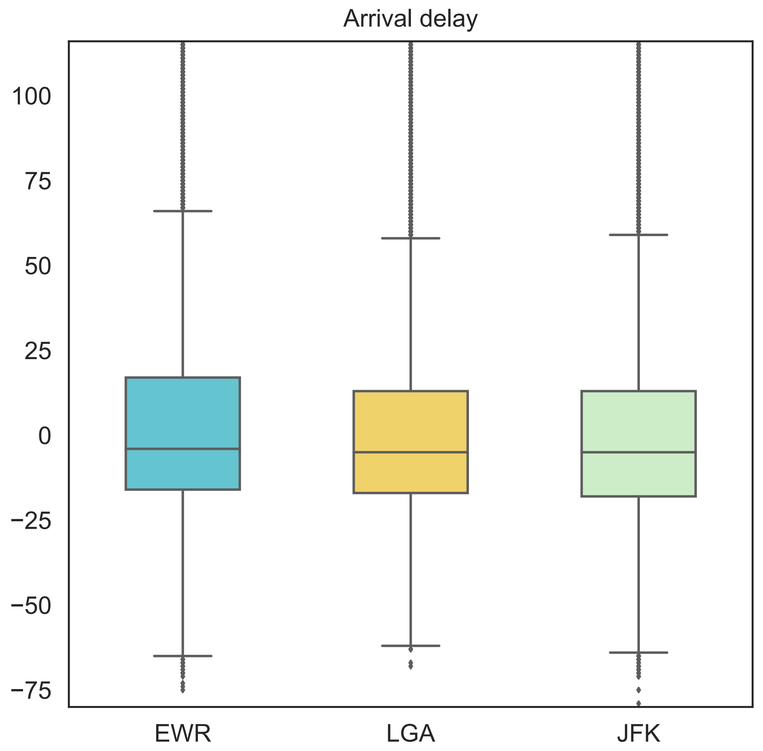


Note: JFK opposite relationship from LGA and EWR. JFK lower delay for small distances, LGA on average.

New York airports

Delay box plots by origin





$$departure.delay = 16.78 - 0.0016 * distance - 2.53 * JFK - 5.14 * LGA$$

$$arrival.delay = 13.79 - 0.0039 * distance - 2.96 * JFK - 4.66 * LGA$$

→ `LGA` is the best airport in term of delays

NYC Flights challenge

Closing and conclude

Many anomalies

- 16 968 NA values
- February 29th data
- HHMM time format, with MM > 60
- `ERW` encoding error
- Impossible negative delays

Best NYC airport?

- LaGuardia (LGA)
- Lower average delay, fixed distance
- Lower delayed count
- Lower delays distributions

Distance influence on delay?

- Moderate negative relationship
- Decrease in delay with flights distance
- More pronounced on arrival delay

Room for improvement

- Delay corrections (around midnight)
- Variables transformation?
- Improve model fitting. Add variables.
- Non-linear model?