Data Notes & Questions

* Missing Data for DEP\_DEL15 → from canceled flights (There are about 9000 flights in the full dataset that have DEP\_DEL15 as null but are not canceled flights. Did some spot checking. These are flights that departed on time. As part of data ingestion, need to set a valid value in this field if it is null.)
* Null values in TAIL\_NUM → will be a problem to deal with if we are looking up airplane features using TAIL\_NUM. Right now ignoring it. (Only canceled flights have null value for TAIL\_NUM)
* The arrival and departure times are strings and local times. Will need some work for us to correlate the weather data.
* FL\_DATE field has the date (local time based) of departure. Depending on the departure time, the arrival date could be different. For example, for a flight departing at say 23:30 having an arrival time of 630, we need to compute the date appropriately. Need to evaluate if it is safe to assume that the arrival date will never be more than departure date + 1 day.
* DEP\_TIME\_BLK and ARR\_TIME\_BLK might be useful. They identify the hour of the day. All flights arriving in the same hour have the same value in ARR\_TIME\_BLK. All flights departing in the same hour have the same value in DEP\_TIME\_BLK. These fields are in local time. Will be hard to translate to UTC. We may have two options - translate weather data to local TZ and use these fields or recompute these fields for UTC. These block fields have one block for 00:01 - 5:59 am and then hourly blocks. Values:
  + 0001-0559
  + 0600-0659
  + 0700-0759
  + 0800-0859
  + .
  + .
  + .
  + 2200-2259
  + 2300-2359
* There are 8 airports missing corresponding weather stations - PSE, PPG, OGS, SPN, SJU, TKI, GUM, XWA. Of these, SJU and GUM are large airports. Ignore the flights departing from these locations? We can recover weather data for SJU and XWA using geo proximity search. The other 6 can be dropped as they have low air traffic.
* Analysis of seasonality
  + Delays by hour of day
    - starter\_nb\_fp
  + Delays by day of week
    - [data-flights](https://docs.google.com/spreadsheets/d/1HQZzwVHnzkBs-iTTUnROYH5XR1QsdsgEmzVM9bkZdWQ/edit?usp=sharing)– initial eda (entire dataset)
  + Delays by month
    - [data-flights](https://docs.google.com/spreadsheets/d/1HQZzwVHnzkBs-iTTUnROYH5XR1QsdsgEmzVM9bkZdWQ/edit?usp=sharing)– initial eda (entire dataset)
  + Delays by quarter
    - [data-flights](https://docs.google.com/spreadsheets/d/1HQZzwVHnzkBs-iTTUnROYH5XR1QsdsgEmzVM9bkZdWQ/edit?usp=sharing)– initial eda (entire dataset)
  + Anything related to special occasions (holidays?)
* Do we understand if delays are more common for shorter vs. longer flights? E.g. less than 2 hours vs. more than 4? Or something like that.
* ~~Are carriers a factor in flight delay? Should we include OP\_CARRIER in our feature selection?~~
  + Yes this looks like a good variable to use, see initial analysis in [data-flights](https://docs.google.com/spreadsheets/d/1HQZzwVHnzkBs-iTTUnROYH5XR1QsdsgEmzVM9bkZdWQ/edit?usp=sharing)– initial eda (entire dataset)
* Outlier detection
* It looks like every single row in the full airlines dataset at /mnt/mids-w261/datasets\_final\_project/parquet\_airlines\_data/\* is duplicated. Total no of rows: 62513788, Total no of distinct rows: 31256894. Need to apply “distinct” as part of data cleanup.
* Canceled flights have a few issues:
  + Most data elements are null
  + Some of the canceled flights (ex: ORIGIN = 'ORD' AND TAIL\_NUM = 'N1EAMQ' AND FL\_DATE = '2015-02-02') have multiple rows with different destinations on the same day and same origin.