w261 Final Project

Team: FP_SectionK_Group11

8/10/2023

Steve Carr

Florencia Froebel

John Scott

Mike Varner

Berkeley
UNIVERSITY OF CALIFORNIA

Business Case - Predicting airline delays

Business Case

Delays in the domestic airline industry are extremely costly, to both airlines and passengers, with some estimates of net losses greater than $30B per year (Ball et al., 2010). Therefore, predicting when flights will be delayed and understanding the drivers will help to mitigate these costs. Using seven years of US flight and weather data, we aim to train four machine learning models to classify if a flight will be delayed two hours ahead of its scheduled departure by at least 15 minutes.

Metric Evaluation

Primarily concerned with Precision but also considered AUC and recall in our model evaluation and tuning process. Falsely predicting no-delay and upsetting passengers is likely more costly than falsely predicting a delay from the perspective of an airline.

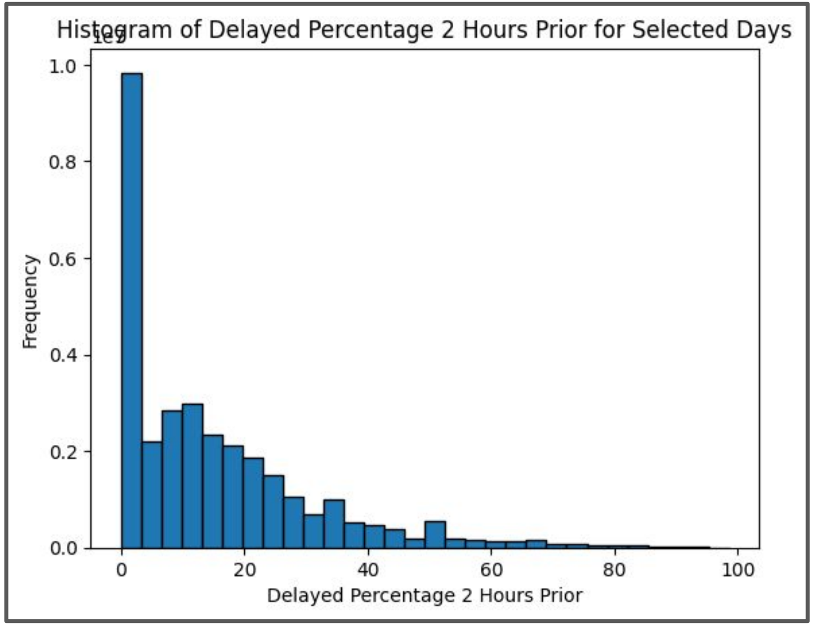
Feature Engineering Overview

- Kept 23 columns from original dataset
    - Converted columns like ‘sched_depart_date_time_UTC’ to date time datatype
    - Converted numeric columns to numbers
- Created the following:
    - **FL_DATE** to only date (no time)
    - **hour_of_day**: aggregate all flights within an hour without any minutes
    - **is_holiday**: binary indicator for all US holidays +- 1 day
    - **part_of_day**: groups flights into 3, 8-hour buckets (evening to morning; mid-morning to mid-afternoon; afternoon to evening)
    - **DELAYED_PERCENTAGE_2_HOURS_PRIOR**: % of delayed flights 2 hours prior by date, hour, and origin airport
    - **Pagerank** of airport
    - **Outdegree**/**Indegree** for the number of unique routes that depart/arrive from an airport

EDA of Percent of Flights Delayed 2 Hours Prior to Departure

Feature definition: *Assigns flights the percentage of flights delayed in that origin location 2 hours prior to departure datetime*

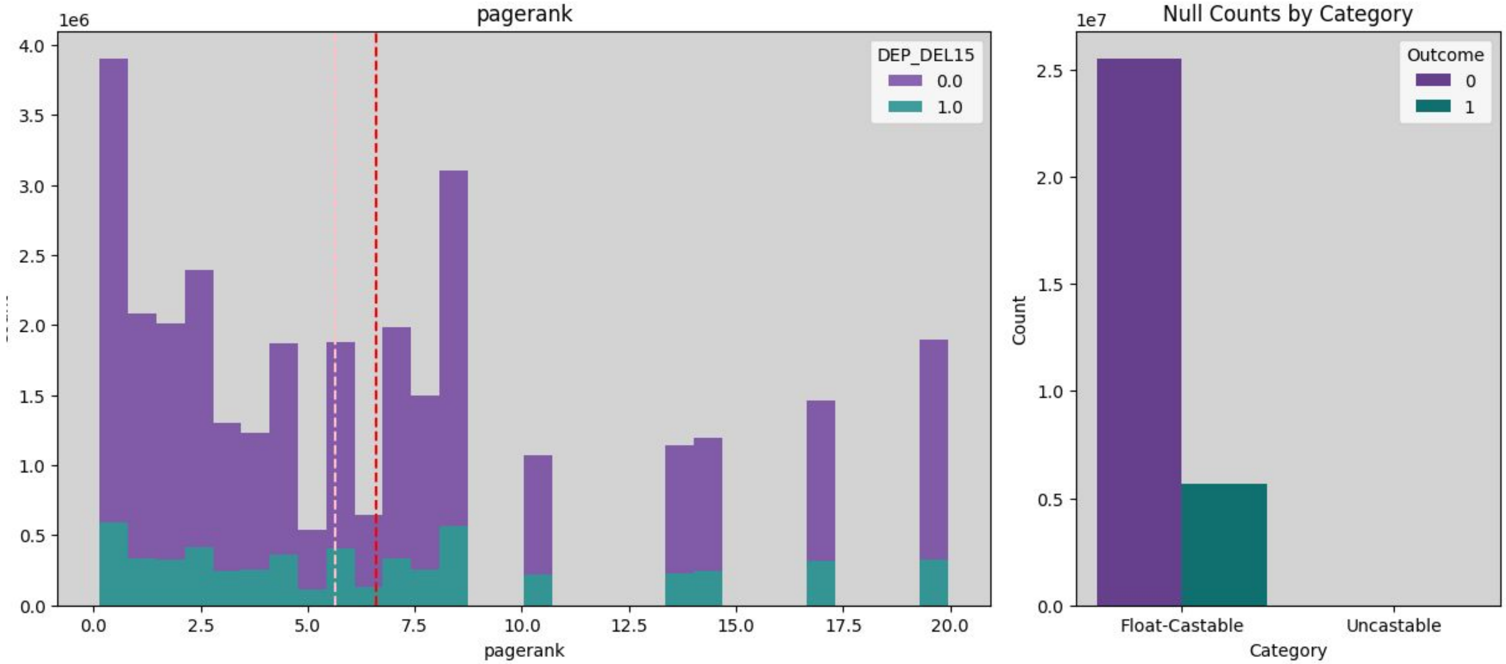
- Mean: 14.8%
- Standard Deviation: 16.3%
- Min: 0%
- Max: 98%
- Median: 10.95%



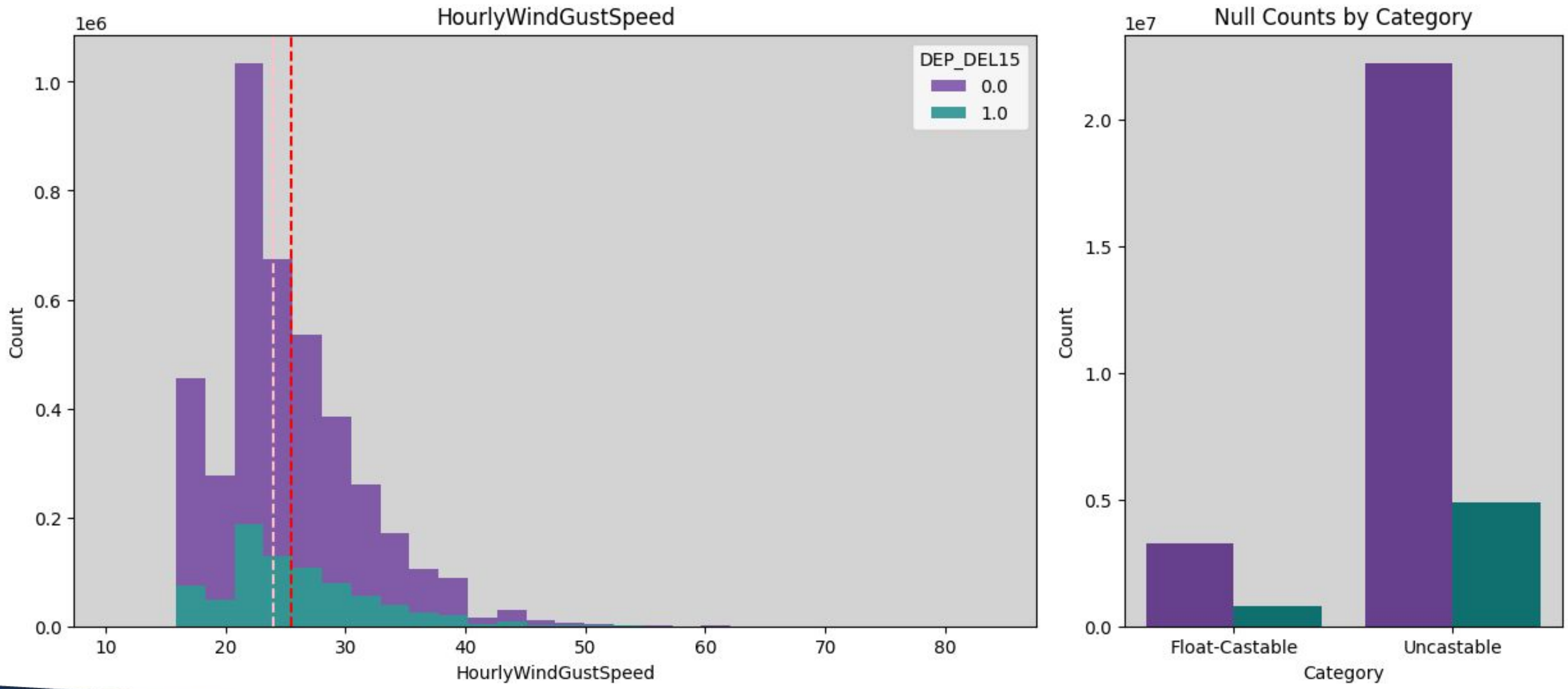
Berkeley
UNIVERSITY OF CALIFORNIA

EDA of PageRank feature

Feature definition: *The pagerank of the airport based on the graph built from the airports and the unique routes*



Rejected variable example



Berkeley
UNIVERSITY OF CALIFORNIA

Overview of Modeling Pipeline

Cleaning of Initial Dataset

- Feature selection
- Change data types
- Imputation of missing values (filled in values with 0 and dropped nulls – working on improvement)

Feature Engineering

- Create the following features:
    - is_holiday
    - part_of_day
    - Percent of delayed flights 2 hours prior (grouped by origin)
    - Pagerank (by origin airport)
    - In-degree (arrival unique routes #)
    - Out-degree (departure unique routes)

One-hot Encoding & Normalization

- Perform one-hot encoding to all categorical & non-numeric features
- Normalize numeric features

Cross Validation Split

- Initial split for test vs train data
- Test data included all of 2019
- Train data included 2015-2018
- Performed CV split on train data
- Created 5 CV folds, each training set spanning 250 days and validation set spanning 40 days
- Downsampling of train data

Modeling

- Train, validate and test data using the following models:
    - Logistic Regression
    - Logistic Regression with hyperparameter tuning
    - Random Forest
    - RF with hyperparameter tuning
    - Gradient Boosted Trees
    - Extreme Gradient Boosted Trees
    - MLP
- Generate performance metrics

Model Evaluation

- **Random Forests** vs **Others**

- Random Forest models had worse recall than others with no effects from hyperparameter tuning

- LR, LRreg, GBT, XGBT, and MLP had nearly identical performance

Experimental Results on Test Datasets by Model %					
Model	**AUC**	**Acc**	**Recall**	**Precision**	**Train Time (min)**
Logistic Regression (LR)	74.52%	82.94%	20.67%	63.18%	18
Logistic Regression - regularization (LRreg)	74.52%	82.94%	20.67%	63.18%	16.2
Random Forest (RF)	68.78%	81.99%	5.28%	75.23%	17.5
Random Forest w/ Hyperparameter tuning (RF+)	68.77%	81.99%	5.28%	75.23%	112.3
Gradient Boosted Trees (GBT)	75.52%	82.96%	20.14%	63.89%	17
Extreme Gradient Boosted Trees (XGBT)	75.55%	82.96%	20.14%	63.89%	18.4
Multilayer Perceptron: (14, 30, 2) (MLP)	75.20%	82.93%	20.97%	62.83%	32.2
Best % / Model	75.55% XGBT	82.96% GBT / XGBT	20.97% MLP	75.23% RF	-

Berkeley
UNIVERSITY OF CALIFORNIA

Analysis of Results - Methodology

***Objective of analysis**: see if the predicted flights that belong to false positive or false negative are different from the actual negative or positive flights*

Procedure:

- For each test results dataframe, create 2 new dataframes by filter for false positives and false negatives
- For each column in each of these dataframes, determine the average
- Compare that average to the average of either actual negative or positive flights:
    - False positives are compared to actual negatives
    - False negatives are compared to actual positives

Berkeley
UNIVERSITY OF CALIFORNIA

Analysis of Results - Results Overview

Category	Logistic Regression		XGBoost		Random Forest w/ Tuning	
	FN	FP	FN	FP	FN	FP
Hourly Sea Level Pressure	-0.025186	0.019044	-0.021664	-0.020305	-0.025999	-0.068346
Hourly Visibility	-0.024984	-0.279813	-0.062085	-0.414436	-0.10551	-1.666156
Hourly Precipitation	0.000617	0.002117	0.000933	0.003606	0.001274	0.020988
Distance	33.166226	5.280659	34.798681	-0.997749	27.187048	102.084401
Hourly Wind Speed	0.04969	3.708013	0.358179	0.450747	0.456404	0.17341
Pagerank	0.378874	-0.593636	0.277154	-0.44349	0.249119	1.286712
Hourly Relative Humidity	0.572311	1.665535	0.599328	7.107112	1.741551	23.976346
Hourly Wet Bulb Temperature	1.569604	-6.49105	0.853617	-1.361894	0.77721	1.403253
Hourly Dry Bulb Temperature	1.789644	-8.079636	0.968737	-3.430868	0.564374	-4.496227
Hourly Dew Point Temperature	1.852918	-6.702136	1.084382	0.145437	1.281119	6.719212
Hourly Wind Direction	3.641685	18.774804	5.752587	3.811973	7.108431	-10.839283
Out Degree	3.891144	-4.256577	3.043153	-3.261304	3.004349	12.738837
In Degree	3.946942	-4.263242	3.094946	-3.293972	3.052094	12.94189
% of Flights Delayed 2 Hours Prior by Origin	7.473148	10.631574	8.231332	27.410501	14.589649	34.546294

Berkeley
UNIVERSITY OF CALIFORNIA

Phase Conclusion & Next Steps

Conclusion

- Random Forest with and without hyperparameter tuning performs well in our primary metric of precision
  - Random forest without tuning would be the recommended model considering training time
- Model performance is surprisingly good considering we only used numeric data in the final model

Future Work

- Incorporate categorical variables into our final model pipeline
- Further feature engineering:
  - By origin determine peak flight periods
  - Deriving weather / geographic variables that can be applied to regions
- Utilizing SMOTE instead of downsampling
- Perform more model experiments:
  - Additional hyperparameter tuning for random forests
  - Larger and more complicated MLP
  - Consider using transformer architecture

Berkeley
UNIVERSITY OF CALIFORNIA

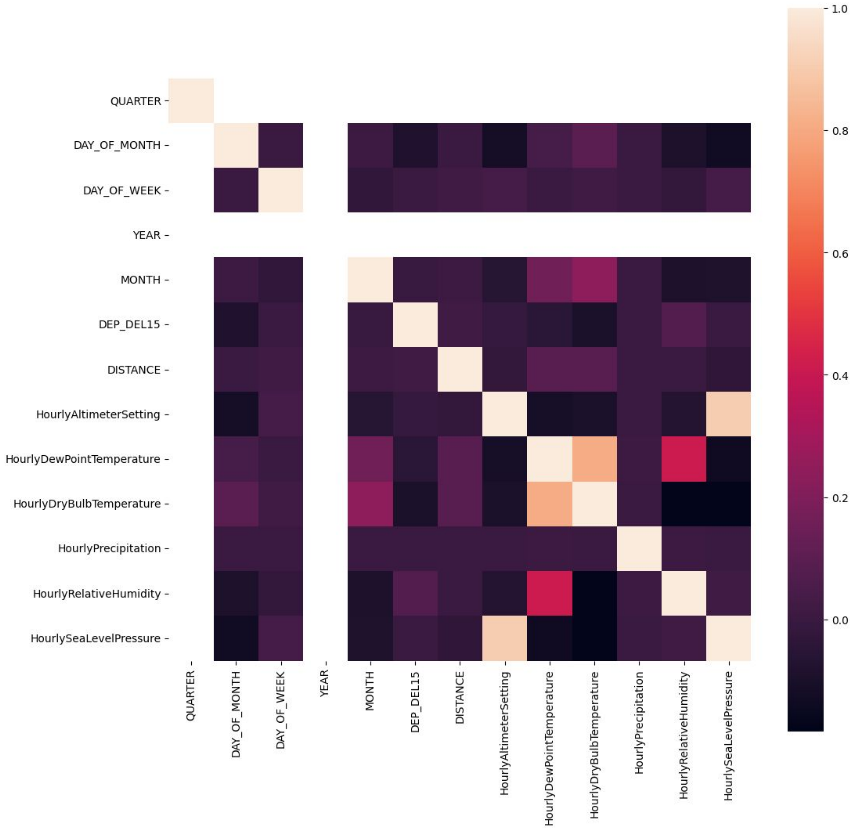
Questions & Answers

Appendix

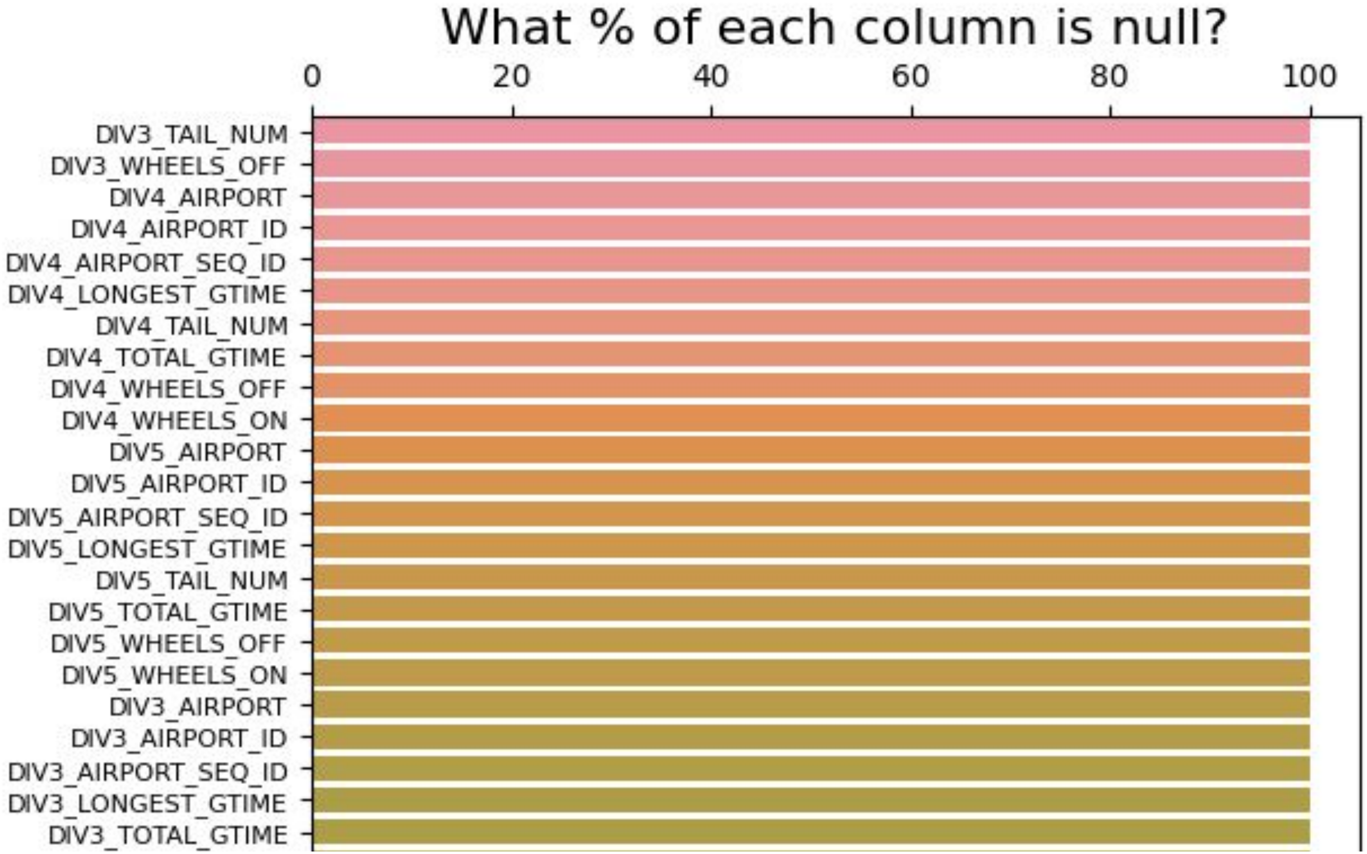
Berkeley
UNIVERSITY OF CALIFORNIA

Heatmap of key variables

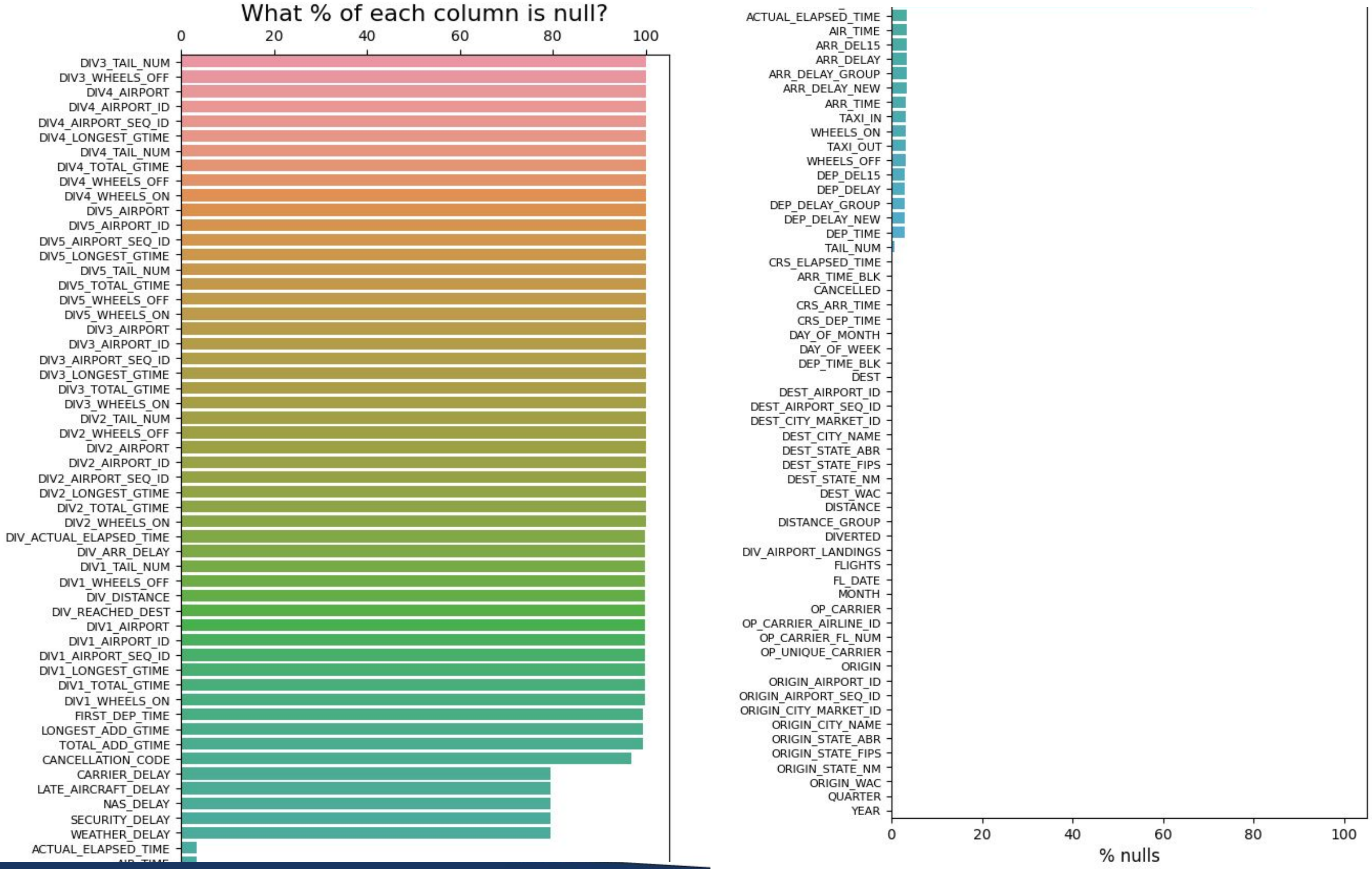
- 3 Month OTPW dataset
- Few variables of interest are highly correlat
- Hourly dew point temperature and relative to drop one of these variables



Berkeley
UNIVERSITY OF CALIFORNIA

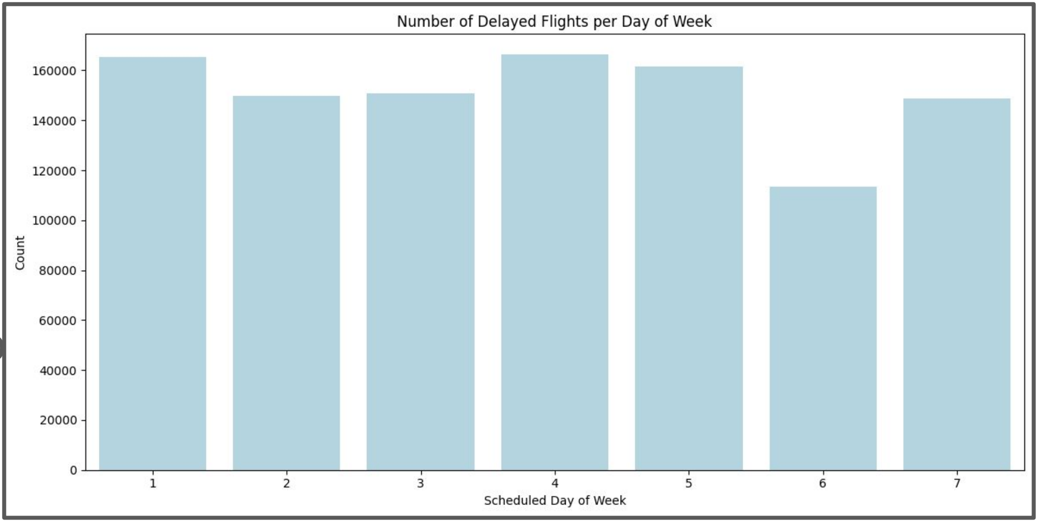


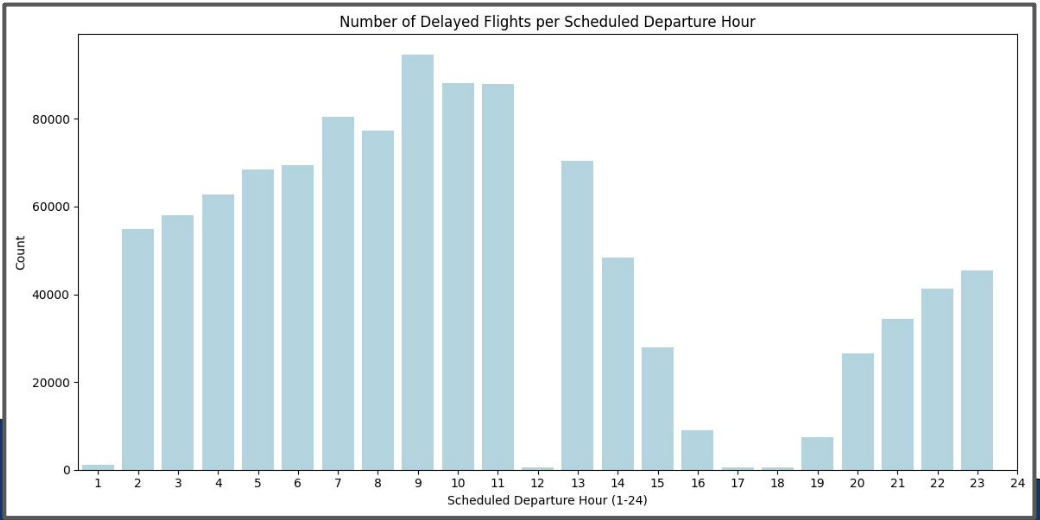
Berkeley
UNIVERSITY OF CALIFORNIA


What % of each column is null?
0
20
40
60
80
100
DIV3_TAIL_NUM
DIV3_WHEELS_OFF
DIV4_AIRPORT
DIV4_AIRPORT_ID
DIV4_AIRPORT_SEQ_ID
DIV4_LONGEST_GTIME
DIV4_TAIL_NUM
DIV4_TOTAL_GTIME
DIV4_WHEELS_OFF
DIV4_WHEELS_ON
DIV5_AIRPORT
DIV5_AIRPORT_ID
DIV5_AIRPORT_SEQ_ID
DIV5_LONGEST_GTIME
DIV5_TAIL_NUM
DIV5_TOTAL_GTIME
DIV5_WHEELS_OFF
DIV5_WHEELS_ON
DIV3_AIRPORT
DIV3_AIRPORT_ID
DIV3_AIRPORT_SEQ_ID
DIV3_LONGEST_GTIME
DIV3_TOTAL_GTIME
DIV3_WHEELS_ON
DIV2_TAIL_NUM
DIV2_WHEELS_OFF
DIV2_AIRPORT
DIV2_AIRPORT_ID
DIV2_AIRPORT_SEQ_ID
DIV2_LONGEST_GTIME
DIV2_TOTAL_GTIME
DIV2_WHEELS_ON
DIV_ACTUAL_ELAPSED_TIME
DIV_ARR_DELAY
DIV1_TAIL_NUM
DIV1_WHEELS_OFF
DIV_DISTANCE
DIV_REACHED_DEST
DIV1_AIRPORT
DIV1_AIRPORT_ID
DIV1_AIRPORT_SEQ_ID
DIV1_LONGEST_GTIME
DIV1_TOTAL_GTIME
DIV1_WHEELS_ON
FIRST_DEP_TIME
LONGEST_ADD_GTIME
TOTAL_ADD_GTIME
CANCELLATION_CODE
CARRIER_DELAY
LATE_AIRCRAFT_DELAY
NAS_DELAY
SECURITY_DELAY
WEATHER_DELAY
ACTUAL_ELAPSED_TIME
ACTUAL_ELAPSED_TIME
AIR_TIME
ARR_DEL15
ARR_DELAY
ARR_DELAY_GROUP
ARR_DELAY_NEW
ARR_TIME
TAXI_IN
WHEELS_ON
TAXI_OUT
WHEELS_OFF
DEP_DEL15
DEP_DELAY
DEP_DELAY_GROUP
DEP_DELAY_NEW
DEP_TIME
TAIL_NUM
CRS_ELAPSED_TIME
ARR_TIME_BLK
CANCELLED
CRS_ARR_TIME
CRS_DEP_TIME
DAY_OF_MONTH
DAY_OF_WEEK
DEP_TIME_BLK
DEST
DEST_AIRPORT_ID
DEST_AIRPORT_SEQ_ID
DEST_CITY_MARKET_ID
DEST_CITY_NAME
DEST_STATE_ABR
DEST_STATE_FIPS
DEST_STATE_NM
DEST_WAC
DISTANCE
DISTANCE_GROUP
DIVERTED
DIV_AIRPORT_LANDINGS
FLIGHTS
FL_DATE
MONTH
OP_CARRIER
OP_CARRIER_AIRLINE_ID
OP_CARRIER_FL_NUM
OP_UNIQUE_CARRIER
ORIGIN
ORIGIN_AIRPORT_ID
ORIGIN_AIRPORT_SEQ_ID
ORIGIN_CITY_MARKET_ID
ORIGIN_CITY_NAME
ORIGIN_STATE_ABR
ORIGIN_STATE_FIPS
ORIGIN_STATE_NM
ORIGIN_WAC
QUARTER
YEAR
0
20
40
60
80
100
% nulls

Flight Delays

- Data from 12m OTPW dataset
- Day of Week does not seem to
  except for Saturday
- Delays by hour is more like to

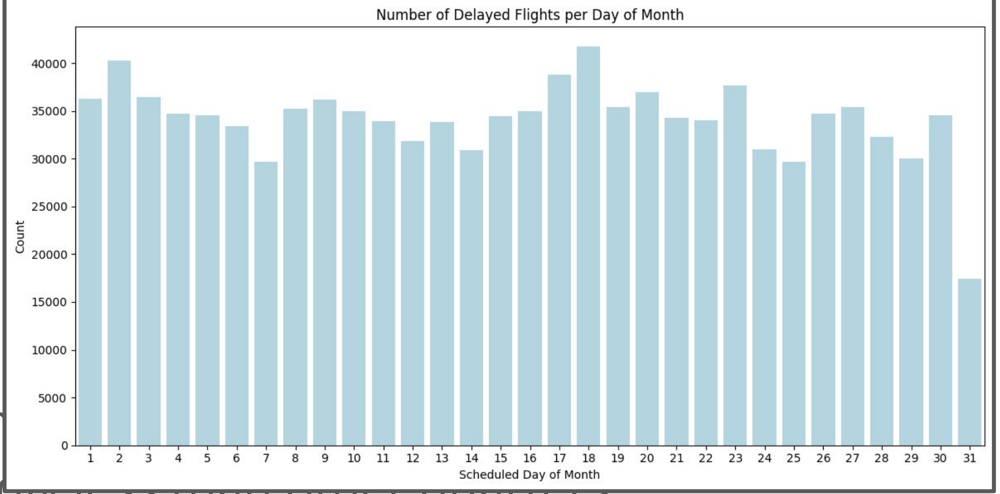


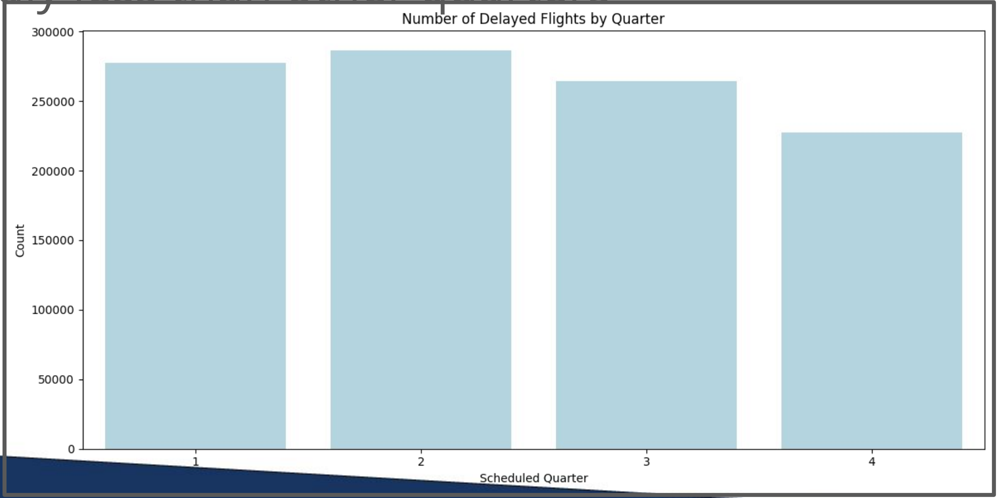


Berkeley UNIVERSITY OF CALIFORNIA

Flight Delays cont.

- Data from 12m OTPW dataset
- Day of Month has some variatio
  deflated due to not all months h
- Quarter 4 delays are unexpectedly less than other quarters

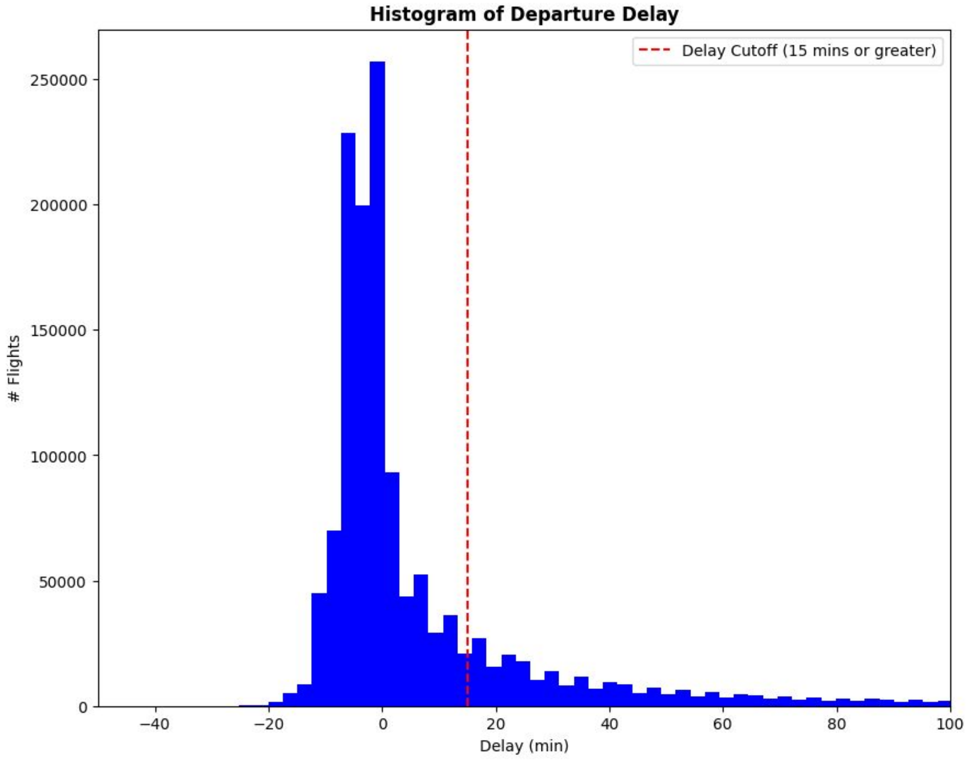




Berkeley
UNIVERSITY OF CALIFORNIA

Flight Departure Delays

- Data from 3 month OTPW datase
- Long right tail as expected
- Flights delayed more than 15 mir



UNIVERSITY OF CALIFORNIA