# Dataset

This report utilizes the NTSB Census of US Civil Aviation Accidents. This dataset contains extensive information on the vast majority of aviation accidents and incidents since 2008. This includes everything from the date and location of the accident to the flight hours of each crew member and much, much more. This report uses data from April 23rd, 2024.

### Data Collection

Data was downloaded from http://data.ntsb.gov/avdata. The NTSB distributes the data in an MDB file that must be loaded into Microsoft Access. I then exported the tables that I would be using as Excel files, which I finally converted into CSV files that were loaded into a MySQL server. Once in the MySQL server, I dropped columns of entirely null values and columns that consisted of unimportant data for this report such as the NTSB employee that input the data.

# Background Information

This project is inspired by the work of Xiaoge Zhang, Prabhakar Srinivasan, and Sankaran Mahadevan in their paper "Sequential deep learning from NTSB reports for aviation safety prognosis" (https://doi.org/10.1016/j.ssci.2021.105390). During my reading of their report, I found their problem statement had the order of operations in reverse. In their paper, they began with the textual "Sequence of Events" and predicted the damage done to the aircraft, if fatalities occurred, and whether the event was categorized as an accident or incident. I believe that is backwards, as in an investigation, the sequence of events would be written last after all facts are known, and the predicted information would be some of the first information to be known.

### Problem Statement

Accordingly, their research inspired two main questions that will be addressed in this report:

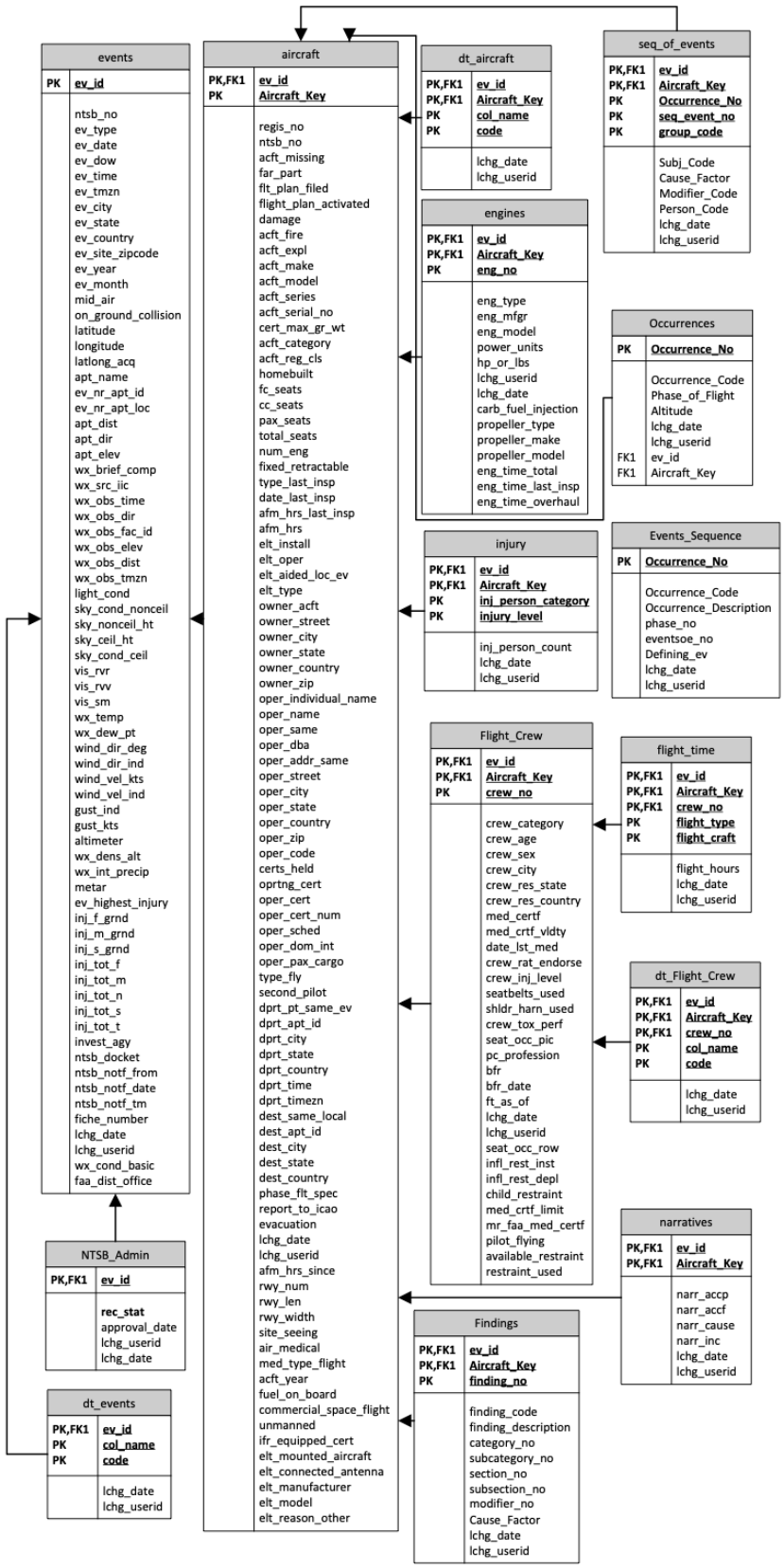
1. Can surface-level features (eg. data that can be immediately gathered at the beginning of an investigation) be used to predict the cause of the accident?
2. What does this tell us about the causes of accidents?

I hope that answering these two questions will add to the wealth of information surrounding the causality of aviation accidents and will help to reduce the amount of avoidable aviation accidents.

# SQL Interactions

After my cursory cleaning in MySQL, I then imported the data into this Jupyter Notebook for further preparation and analysis.

-Data Schema


events
PK ev_id
ntsb_no
ev_type
ev_date
ev_dow
ev_time
ev_tmzn
ev_city
ev_state
ev_country
ev_site_zipcode
ev_year
ev_month
mid_air
on_ground_collision
latitude
longitude
latlong_acq
apt_name
ev_nr_apt_id
ev_nr_apt_loc
apt_dist
apt_dir
apt_elev
wx_brief_comp
wx_src_iic
wx_obs_time
wx_obs_dir
wx_obs_fac_id
wx_obs_elev
wx_obs_dist
wx_obs_tmzn
light_cond
sky_cond_nonceil
sky_nonceil_ht
sky_ceil_ht
sky_cond_ceil
vis_rvr
vis_rvv
vis_sm
wx_temp
wx_dew_pt
wind_dir_deg
wind_dir_ind
wind_vel_kts
wind_vel_ind
gust_ind
gust_kts
altimeter
wx_dens_alt
wx_int_precip
metar
ev_highest_injury
inj_f_grnd
inj_m_grnd
inj_s_grnd
inj_tot_f
inj_tot_m
inj_tot_n
inj_tot_s
inj_tot_t
invest_agy
ntsb_docket
ntsb_notf_from
ntsb_notf_date
ntsb_notf_tm
fiche_number
lchg_date
lchg_userid
wx_cond_basic
faa_dist_office
aircraft
PK,FK1 ev_id
PK Aircraft_Key
regis_no
ntsb_no
acft_missing
far_part
flt_plan_filed
flight_plan_activated
damage
acft_fire
acft_expl
acft_make
acft_model
acft_series
acft_serial_no
cert_max_gr_wt
acft_category
acft_reg_cls
homebuilt
fc_seats
cc_seats
pax_seats
total_seats
num_eng
fixed_retractable
type_last_insp
date_last_insp
afm_hrs_last_insp
afm_hrs
elt_install
elt_oper
elt_aided_loc_ev
elt_type
owner_acft
owner_street
owner_city
owner_state
owner_country
owner_zip
oper_individual_name
oper_name
oper_same
oper_dba
oper_addr_same
oper_street
oper_city
oper_state
oper_country
oper_zip
oper_code
certs_held
oprtng_cert
oper_cert
oper_cert_num
oper_sched
oper_dom_int
oper_pax_cargo
type_fly
second_pilot
dprt_pt_same_ev
dprt_apt_id
dprt_city
dprt_state
dprt_country
dprt_time
dprt_timezn
dest_same_local
dest_apt_id
dest_city
dest_state
dest_country
phase_flt_spec
report_to_icao
evacuation
lchg_date
lchg_userid
afm_hrs_since
rwy_num
rwy_len
rwy_width
site_seeing
air_medical
med_type_flight
acft_year
fuel_on_board
commercial_space_flight
unmanned
ifr_equipped_cert
elt_mounted_aircraft
elt_connected_antenna
elt_manufacturer
elt_model
elt_reason_other
dt_aircraft
PK,FK1 ev_id
PK,FK1 Aircraft_Key
PK col_name
PK code
lchg_date
lchg_userid
seq_of_events
PK,FK1 ev_id
PK,FK1 Aircraft_Key
PK Occurrence_No
PK seq_event_no
PK group_code
Subj_Code
Cause_Factor
Modifier_Code
Person_Code
lchg_date
lchg_userid
engines
PK,FK1 ev_id
PK,FK1 Aircraft_Key
PK eng_no
eng_type
eng_mfgr
eng_model
power_units
hp_or_lbs
lchg_userid
lchg_date
carb_fuel_injection
propeller_type
propeller_make
propeller_model
eng_time_total
eng_time_last_insp
eng_time_overhaul
Occurrences
PK Occurrence_No
Occurrence_Code
Phase_of_Flight
Altitude
lchg_date
lchg_userid
FK1 ev_id
FK1 Aircraft_Key
injury
PK,FK1 ev_id
PK,FK1 Aircraft_Key
PK inj_person_category
PK injury_level
inj_person_count
lchg_date
lchg_userid
Events_Sequence
PK Occurrence_No
Occurrence_Code
Occurrence_Description
phase_no
eventsoe_no
Defining_ev
lchg_date
lchg_userid
Flight_Crew
PK,FK1 ev_id
PK,FK1 Aircraft_Key
PK crew_no
crew_category
crew_age
crew_sex
crew_city
crew_res_state
crew_res_country
med_certf
med_crtf_vldty
date_lst_med
crew_rat_endorse
crew_inj_level
seatbelts_used
shldr_harn_used
crew_tox_perf
seat_occ_pic
pc_profession
bfr
bfr_date
ft_as_of
lchg_date
lchg_userid
seat_occ_row
infl_rest_inst
infl_rest_depl
child_restraint
med_crtf_limit
mr_faa_med_certf
pilot_flying
available_restraint
restraint_used
flight_time
PK,FK1 ev_id
PK,FK1 Aircraft_Key
PK,FK1 crew_no
PK flight_type
PK flight_craft
flight_hours
lchg_date
lchg_userid
dt_Flight_Crew
PK,FK1 ev_id
PK,FK1 Aircraft_Key
PK,FK1 crew_no
PK col_name
PK code
lchg_date
lchg_userid
narratives
PK,FK1 ev_id
PK,FK1 Aircraft_Key
narr_accp
narr_accf
narr_cause
narr_inc
lchg_date
lchg_userid
Findings
PK,FK1 ev_id
PK,FK1 Aircraft_Key
PK finding_no
finding_code
finding_description
category_no
subcategory_no
section_no
subsection_no
modifier_no
Cause_Factor
lchg_date
lchg_userid
NTSB_Admin
PK,FK1 ev_id
rec_stat
approval_date
lchg_userid
lchg_date
dt_events
PK,FK1 ev_id
PK col_name
PK code
lchg_date
lchg_userid

# Cleaning

In this cleaning stage, I started with general filters on the events table, then cleaned each column of the table. I then merged the child tables onto the cleaned events table to filter away data on the events that were dropped from the events table. I then went through each column of the child tables cleaning each. Finally, I began early EDA by showing details about each column. Exact steps are shown in code.

## Findings

As part of the cleaning process for the findings table, I dropped any of the findings beyond the 5th finding. This was to reduce the repetition in the training data to a manageable amount while preserving the quality of multiple findings.

# EDA

*Beyond the analysis shown during the cleaning stage

For my EDA, I merged all the tables into one. While the above visuals show the distibution for each column, I needed to know the correlation and association of the variables. To resolve this, I performed a Chi^2 Test for Association between all the categorical variables, a Correlation Test between all of the continuous variables, and a One-Way-ANOVA Test between the categorical and continuous variables.

# Encoding Categoricals

In my research, I found conflicting and confusing information on the many different strategies for encoding high cardinality nominal categorical features. Three, however, were regularly recommended: One-Hot-Encoding, Binary Encoding, and Feature Hash Encoding. I chose to test all three and pick the strategy that gave the best results. For the features that only had two options (Y/N | 1/0 | T/F | etc.), I binary encoded them in all three instances.

Main resources:

https://www.kdnuggets.com/2021/05/deal-with-categorical-data-machine-learning.html

https://kantschants.com/complete-guide-to-encoding-categorical-features#heading-nominal-categorical-features

# Modelling (using Random Forest Classifier)

For modelling, I chose an 85%, 15% train-test-split to maintain a large training set, while still having plenty of records for validation. With the 85% test set, I then performed 10-fold cross validation to train a sklearn Random Forest Classifier for 17-class classification. Finally, the top performing model -the Binary Encoded model- predicted the top 4 causes for the accident.

# Results

The results of the model validation against the test set are below.

Accuracy: 82.94%

Recall: 71.05%

Precision (Low-Bound): 46.43%

Precision (High-Bound): 71.05%

### Category Performance

The model was most accurate at predicting classes 12, 16, 24, 32, 33.

The model had it's highest recall predicting classes 16, 24, 26, 32, 33.

Finally, the model was most precise on classes 16, 24, 26, 32, 33.

- 12 - Aircraft-Aircraft Systems
- 16 - Aircraft-Aircraft Oper/Perf/Capability
- 24 - Personnel-Action/Decision
- 26 - Personnel-Task Performance
- 32 - Environmental-Physical Environment
- 33 - Environmental-Conditions/Weather/Phenomena

# Interpretation

To address my two main questions:

**1. Can surface-level features be used to predict the cause of the accident?**

In general, yes. The accuracy above 80% shows that the surface level features do hold information that indicates a potential cause for an accident. In addition, the continuous features seem to show more about the particular causes, so the model may be more accurate with features that I had to drop due to data cleanliness, like air pressure.

**2. What does this tell us about the causes of accidents?**

There are interactions between temperature, age of crew, wind velocity, total non-injuries, number of passenger seats, number of flight crew seats, day of week of the flight, and month of the flight that influence the cause of the accident. More research must be done on the particulars of the accident, but it is not unreasonable to assume that a summer month, weekend day of week, high age, and low number of flight crew will result in a higher chance of an accident caused by human error. Conversely, high temperatures reduce air density and lower overall aircraft performance, and high wind velocities can overspeed the aircraft. Therefore, those may point to environmental factors as the cause of the accident. Adding strength to these hypotheses, the model has the highest scores in predicting the causes that would be most closely associated with each.

# Model Issues

The model was best at predicting the most common causes of the aviation accidents. While, the model still may be useful, this suggests it may have trouble identifying the more obsure causes for accidents. Possible future solutions could include upsampling the underrepresented causes, downsampling the most common causes, or using a model that allows for class weighting.

# Acknowledgements

Finally, I would like to thank Dr. Shashi Jha (College of Charleston) for his guidance on this report; Jess Thomas (NTSB) for sharing the causes corresponding to the finding codes; and Xiaoge Zhang, Prabhakar Srinivasan, and Sankaran Mahadevan for inspiring this report.