Ben Gruher

CPSC 5830 Lab 2

January 20th, 2021

1. **Number of observations by Aircraft.** Our next step will be to look at a sequence of observations for a single aircraft and see if we can segment it into flights. It's worth seeing whether the number of observations per aircraft in a two-day window meets rough expectations. That is, if flights are between one hour and eight hours, and an observation is taken every 15 seconds, then we would expect somewhere between 240 and 1920 observations per flight, assuming it is being observed by only one receiver at a time. If many/most aircraft do not fall within those bounds, it is at least worth further exploration. For the two-day period Oct 15 and Oct16 generate a histogram showing number of observations and count of aircraft having that many observations. Does the result give you confidence that we can build flight segments from this data set?

For this analysis, I joined the aircraft dataframe with the observation dataframe on 'Icao' code. I then grouped the observations by 'Icao' code to get the observations for each plane. For the two days included in this analysis (Oct 15-16), there were observations of 61,838 different planes. For the most part, there were many observations for each plane within those two days. Figure 1 shows the 20 planes with the most observations within that time span. As you can see, the planes with the most observations had around 15,864-22,150 observations. Using between 240 and 1920 observations per flight, the plane with the most observations would have gone on between 11 and 92 flights during those days. 92 flights in two days does not sound reasonable, but 11 could be possible. However, a more likely explanation could be that multiple receivers picked up that plane's signal during its flight(s). This would be somewhat concerning if our project depended on only one receiver picking up the plane's signal.

```
+------+-----+
|  Icao|count|
+------+-----+
|A11062|22150|
|A16927|22150|
|A1FFB5|22022|
|A1F0D9|20606|
|A26F6D|18982|
|A0B8FB|18794|
|ACCA19|18308|
|A3965C|17696|
|A7DA54|17682|
|ABB6B5|17528|
|A3650B|17418|
|A37A02|17406|
|A88CCF|16802|
|A67883|16568|
|A033FF|16560|
|ADDA24|16470|
|A89086|16146|
|A0DDFB|16102|
|AC10D8|15958|
|A10BBC|15864|
+------+-----+
only showing top 20 rows
```

*Figure 1 – Number of observations per plane in the two-day time interval. Results are ranked in descending order.*

Now that we have explored the upper-end for number of observations per aircraft, we will examine the lower-end. Out of 61,838 planes, there were 16,293 planes with fewer than 240 observations, our approximation for the number of observations for a single flight. While this may seem concerning, I hypothesize that many of these planes had one flight scheduled during/near that two day window and that their departure time and arrival time straddles one of the cutoffs for our two day interval.

Figure 2 shows the distribution of number of observations per plane during our two-day window. Edit: unfortunately, I was unable to create a histogram using MatPlotLib because my EMR cluster had an old version of Numpy installed and I was unable to update the package. Figure 2 shows the bins and counts for the histogram. The figure shows that most observations fall in the first bin (35,075 of the 61,838 planes have between one observation and 2216 observations). We see over 4,000 planes have between 2216 and 4431 observations in those two days. This suggests that the majority of planes did not have more observations than we expected. We can assume that the few planes with a very high number of observations are outliers on the upper end. However, we should explore the planes that had very few observations.

```
Bins:  [1.0, 2215.9, 4430.8, 6645.700000000001, 8860.6, 11075.5, 13290.400000000001, 15505.300000000001, 17720.2, 19935.1000000000
02, 22150]
Counts:  [35075, 4150, 2265, 1505, 461, 169, 60, 17, 3, 4]
```

*Figure 2 – Bins and Counts for histogram showing the distribution of number of observations per plane during two-day interval.*

2. **From and To fields.** It will be important to identify where a flight is leaving from and going to. The From and To fields in the observation are supposed to provide that information, but the documentation says the fields is often wrong, do don't depend on it. Focus only on a few major American commercial airline companies--Alaska, Southwest, JetBlue, American, United --and examine all observations on aircraft operated by those carriers. Are the From and To fields complete (i.e. is the fill rate high on those observations)? Are they consistent (i.e. on all observations for the same aircraft "close in time," are the From and To fields the same)?

For this analysis question, I first joined the observation data with the operator data. This allowed me to restrict the observations to only those that include a plane operated by Alaska, Southwest, JetBlue, American, or United (see Figure 3).

| Op | count |
|---|---|
| United Airlines | 2331135 |
| American Airlines | 4322836 |
| JetBlue | 6938 |
| Alaska Airlines | 1193859 |
| Southwest Airlines | 3432734 |

*Figure 3 – Number of observations for each of the major airlines during the two-day window.*

From this new dataframe, I was able to find the fill rate for the 'From' and 'To' attributes. For just these major airlines, the fill rate for both the 'From' and 'To' attributes was around 87.7%, making these columns relatively complete. In order to evaluate the consistency of these fields, I created dataframes that have the unique values for the fields based on a grouping by 'Icao' code as well as a 10-minute time interval. This allowed me to see if observations from around the same time and for the same plane had different 'From' and 'To' values. As seen in Figure 4, only about 0.17% of planes had multiple observations from within 10 minutes of each other with different values in the 'From' column. Likewise, we can see that in Figure 5, only about 0.15% of planes had multiple observations from within 10 minutes of each other with different values in the 'To' column. For both these columns, I see my results as a very positive sign for the reliability of the 'To' and 'From' column. The small number of planes with multiple 'To' and 'From' values within 10 minutes could have very easily been making a quick stop at an airport and reset its destination and departure airports. My implementation also allows me to adjust the time interval if needed (see notebook).

From column:

| Icao | time_interval | collect_set(From) | UniqueFrom |
|---|---|---|---|
| AAE582 | 2020-10-16 23:00:00 | [CLT Charlotte Do... | 2 |
| A0BCB2 | 2020-10-16 21:00:00 | [DFW Dallas Fort ... | 2 |
| AA7DAD | 2020-10-15 15:50:00 | [RDU Raleigh Durh... | 2 |
| AA7DAD | 2020-10-16 14:30:00 | [COS City of Colo... | 2 |
| A78A52 | 2020-10-15 23:50:00 | [PHL Philadelphia... | 2 |
| AA64CA | 2020-10-15 00:00:00 | [DFW Dallas Fort ... | 2 |
| ADC17B | 2020-10-15 13:10:00 | [DFW Dallas Fort ... | 2 |
| AA1BE0 | 2020-10-16 22:40:00 | [DFW Dallas Fort ... | 2 |
| AB2729 | 2020-10-15 14:40:00 | [CLT Charlotte Do... | 2 |
| AC90A7 | 2020-10-16 23:00:00 | [DFW Dallas Fort ... | 2 |
| AC7216 | 2020-10-15 16:40:00 | [FAT Fresno Yosem... | 2 |
| ADC8E7 | 2020-10-15 14:50:00 | [DFW Dallas Fort ... | 2 |
| A0B199 | 2020-10-15 14:00:00 | [MDT Harrisburg, ... | 2 |
| A1804C | 2020-10-15 23:50:00 | [PHX Phoenix Sky ... | 2 |
| A9034C | 2020-10-16 17:20:00 | [CLT Charlotte Do... | 2 |
| AB4EA8 | 2020-10-16 15:30:00 | [PHL Philadelphia... | 2 |
| A4F59C | 2020-10-15 14:30:00 | [DFW Dallas Fort ... | 2 |
| AB34A7 | 2020-10-16 14:50:00 | [PIT Pittsburgh, ... | 2 |
| AD9BC8 | 2020-10-16 21:40:00 | [CLT Charlotte Do... | 2 |
| A4ACAE | 2020-10-16 16:40:00 | [DFW Dallas Fort ... | 2 |

only showing top 20 rows

Percentage of observations with multiple 'From' values: 0.0016707295320768179

*Figure 4 – Statistics for the 'From' field: the unique values for each plane and the percentage of planes with multiple values for the field (bottom). Table is sorted so that the planes with the most values for the 'Species' attribute are at the top.*

To column:

| Icao | time_interval | collect_set(To) | UniqueTo |
|---|---|---|---|
| ADC17B | 2020-10-15 13:10:00 | [LAX Los Angeles,... | 2 |
| AA7DAD | 2020-10-15 15:50:00 | [RDU Raleigh Durh... | 2 |
| AA64CA | 2020-10-15 00:00:00 | [CLT Charlotte Do... | 2 |
| A4ACAE | 2020-10-16 16:40:00 | [CLT Charlotte Do... | 2 |
| A1804C | 2020-10-15 23:50:00 | [PHX Phoenix Sky ... | 2 |
| AA5237 | 2020-10-15 23:40:00 | [RIC Richmond, Un... | 2 |
| A0B199 | 2020-10-15 14:00:00 | [DFW Dallas Fort ... | 2 |
| AC90A7 | 2020-10-16 23:00:00 | [GYE Simon Boliva... | 2 |
| AA1BE0 | 2020-10-16 22:40:00 | [RDU Raleigh Durh... | 2 |
| AD9BC8 | 2020-10-16 21:40:00 | [PHL Philadelphia... | 2 |
| A9034C | 2020-10-16 17:20:00 | [SMF Sacramento, ... | 2 |
| AAE582 | 2020-10-16 23:00:00 | [TPA Tampa, Unite... | 2 |
| AA0E62 | 2020-10-16 22:10:00 | [PHX Phoenix Sky ... | 2 |
| A4BF41 | 2020-10-16 16:20:00 | [HNL Honolulu, Un... | 2 |
| ADDA24 | 2020-10-16 22:50:00 | [MCO Orlando, Uni... | 2 |
| A78A52 | 2020-10-15 23:50:00 | [PHX Phoenix Sky ... | 2 |
| AC7216 | 2020-10-15 16:40:00 | [DFW Dallas Fort ... | 2 |
| AA35E1 | 2020-10-15 15:20:00 | [ORD Chicago O'Ha... | 2 |
| AB2729 | 2020-10-15 14:40:00 | [DFW Dallas Fort ... | 2 |
| AB4EA8 | 2020-10-16 15:30:00 | [NAS Lynden Pindl... | 2 |

only showing top 20 rows

Percentage of observations with multiple 'To' values: 0.0015458707414233903

*Figure 5 – Statistics for the 'To' field: the unique values for each plane and the percentage of planes with multiple values for the field (bottom). Table is sorted so that the planes with the most values for the 'Species' attribute are at the top.*

3. **Aircraft attributes.** The attributes collected for the aircraft are put in the observation payload by database lookup. Are they complete –i.e. does every aircraft have these attributes? Are they consistent –i.e. are the attributes the same for a single aircraft?

The attributes in question here are: Species, Mil, Cou, and Type. For each attribute, I found the fill rate by filtering observations from the aircraft dataset where Species=0, Mil=null, Cou=null, and Type=null, respectively. As seen in Figure 6, all four columns are complete. 'Mil' and 'Cou' have a 100% fill rate, while 'Species' and 'Type' have a fill rate of over 99%.

I then created dataframes for each attribute showing the different values seen in observations for each plane. Ideally, we would see a single unique value for each plane because a plane can only be of one species, military status, country, and type. If we see multiple values for these attributes, we should be concerned. However, Figures 7-10 show that all 4 attributes in question have only one unique value for each plane. This suggests that these columns are consistent.

```
Species fill rate:
0.9925651663329473
Mil fill rate:
1.0
Cou fill rate:
1.0
Type fill rate:
0.996097283534554
```

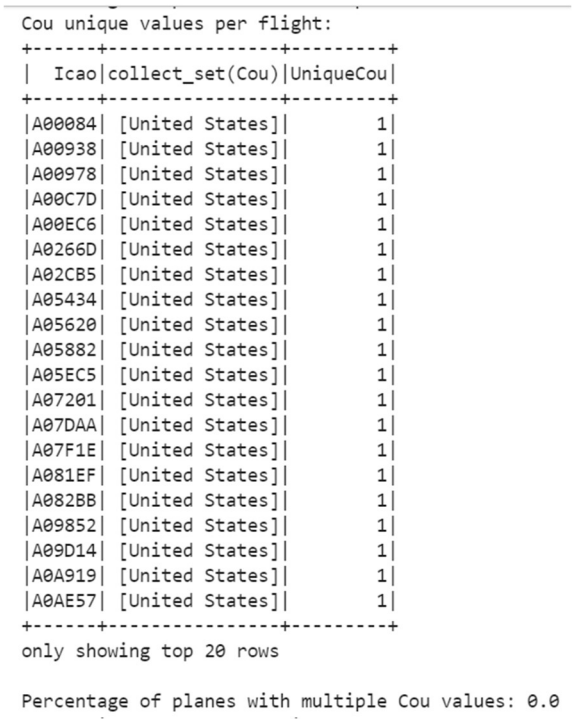
*Figure 6 – Fill rates for 'Species', 'Mil', 'Cou', and 'Type' columns.*

```
Species unique values per flight:
+------+--------------------+-------------+
|  Icao|collect_set(Species)|UniqueSpecies|
+------+--------------------+-------------+
|A0026E|                 [1]|            1|
|A01EE6|                 [4]|            1|
|A02FFF|                 [1]|            1|
|A03F75|                 [1]|            1|
|A03F84|                 [1]|            1|
|A0492E|                 [1]|            1|
|A05D05|                 [1]|            1|
|A05EDF|                 [4]|            1|
|A062CC|                 [1]|            1|
|A07ECC|                 [1]|            1|
|A095B4|                 [1]|            1|
|A09977|                 [1]|            1|
|A0A21A|                 [1]|            1|
|A0C963|                 [1]|            1|
|A0CCAB|                 [1]|            1|
|A0CF0E|                 [1]|            1|
|A0EFDF|                 [1]|            1|
|A0FCD6|                 [4]|            1|
|A101C3|                 [1]|            1|
|A11AD4|                 [1]|            1|
+------+--------------------+-------------+
only showing top 20 rows

Percentage of planes with multiple Species values: 0.0
```

*Figure 7 – Statistics for the 'Species' field: the unique values for each plane and the percentage of planes with multiple values for the field (bottom). Table is sorted so that the planes with the most values for the 'Species' attribute are at the top (there are no planes with more than 1 unique value).*

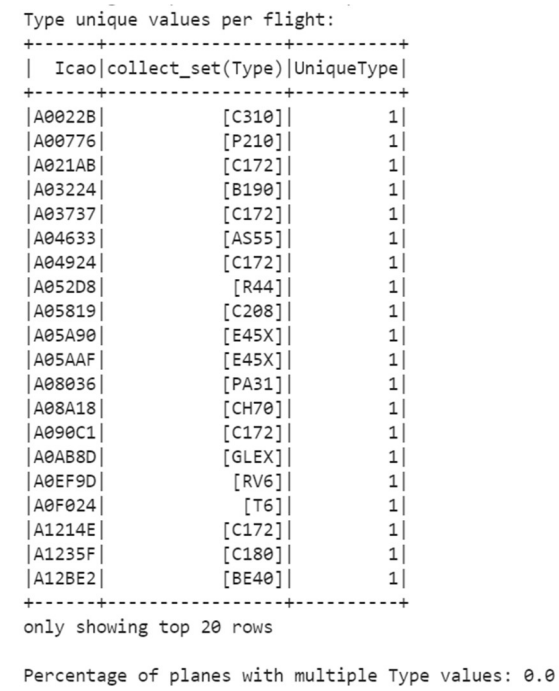
```
Mil unique values per flight:
+------+---------------+---------+
|  Icao|collect_set(Mil)|UniqueMil|
+------+---------------+---------+
|A0022B|         [false]|        1|
|A00776|         [false]|        1|
|A021AB|         [false]|        1|
|A03224|         [false]|        1|
|A03737|         [false]|        1|
|A04633|         [false]|        1|
|A04924|         [false]|        1|
|A052D8|         [false]|        1|
|A05819|         [false]|        1|
|A05A90|         [false]|        1|
|A05AAF|         [false]|        1|
|A08036|         [false]|        1|
|A08A18|         [false]|        1|
|A090C1|         [false]|        1|
|A0AB8D|         [false]|        1|
|A0EF9D|         [false]|        1|
|A0F024|         [false]|        1|
|A1214E|         [false]|        1|
|A1235F|         [false]|        1|
|A12BE2|         [false]|        1|
+------+---------------+---------+
only showing top 20 rows

Percentage of planes with multiple Mil values: 0.0
```

*Figure 8 – Statistics for the 'Mil' field: the unique values for each plane and the percentage of planes with multiple values for the field (bottom). Table is sorted so that the planes with the most values for the 'Mil' attribute are at the top (there are no planes with more than 1 unique value).*



```
Cou unique values per flight:
+------+---------------+---------+
|  Icao|collect_set(Cou)|UniqueCou|
+------+---------------+---------+
|A00084| [United States]|        1|
|A00938| [United States]|        1|
|A00978| [United States]|        1|
|A00C7D| [United States]|        1|
|A00EC6| [United States]|        1|
|A0266D| [United States]|        1|
|A02CB5| [United States]|        1|
|A05434| [United States]|        1|
|A05620| [United States]|        1|
|A05882| [United States]|        1|
|A05EC5| [United States]|        1|
|A07201| [United States]|        1|
|A07DAA| [United States]|        1|
|A07F1E| [United States]|        1|
|A081EF| [United States]|        1|
|A082BB| [United States]|        1|
|A09852| [United States]|        1|
|A09D14| [United States]|        1|
|A0A919| [United States]|        1|
|A0AE57| [United States]|        1|
+------+---------------+---------+
only showing top 20 rows

Percentage of planes with multiple Cou values: 0.0
```

*Figure 9 – Statistics for the 'Cou' field: the unique values for each plane and the percentage of planes with multiple values for the field (bottom). Table is sorted so that the planes with the most values for the 'Cou' attribute are at the top (there are no planes with more than 1 unique value).*



```
Type unique values per flight:
+------+----------------+----------+
|  Icao|collect_set(Type)|UniqueType|
+------+----------------+----------+
|A0022B|          [C310]|         1|
|A00776|          [P210]|         1|
|A021AB|          [C172]|         1|
|A03224|          [B190]|         1|
|A03737|          [C172]|         1|
|A04633|          [AS55]|         1|
|A04924|          [C172]|         1|
|A052D8|           [R44]|         1|
|A05819|          [C208]|         1|
|A05A90|          [E45X]|         1|
|A05AAF|          [E45X]|         1|
|A08036|          [PA31]|         1|
|A08A18|          [CH70]|         1|
|A090C1|          [C172]|         1|
|A0AB8D|          [GLEX]|         1|
|A0EF9D|           [RV6]|         1|
|A0F024|            [T6]|         1|
|A1214E|          [C172]|         1|
|A1235F|          [C180]|         1|
|A12BE2|          [BE40]|         1|
+------+----------------+----------+
only showing top 20 rows

Percentage of planes with multiple Type values: 0.0
```

*Figure 10 – Statistics for the 'Type' field: the unique values for each plane and the percentage of planes with multiple values for the field (bottom). Table is sorted so that the planes with the most values for the 'Type' attribute are at the top (there are no planes with more than 1 unique value).*

4. **Aircraft to operator.** The relationship of aircraft to operator Icao and name is also marked as "only as good as the database." Are they complete –i.e. does every aircraft have these attributes? Are they consistent –i.e. are the attributes the same for a single aircraft?

For this analysis question, I joined the aircraft dataframe with the operator dataframe and looked specifically at the 'Op' and 'OpIcao' fields. As seen in Figure 11, the 'Op' attribute is complete for 99.995% of observations. In addition, there are no planes that have multiple 'Op' values (only one unique value). Therefore, this attribute is both complete and consistent. Figure 12 shows that the 'OpIcao' attribute is complete for only 27.74% of observations. As with 'Op', there are no planes with multiple 'OpIcao' values. Therefore, this attribute is consistent but not very complete.

```
Attribute:  Op
Fill rate:  0.9999482112153792
+------+--------------------+--------+
|  Icao|     collect_set(Op)|UniqueOp|
+------+--------------------+--------+
|A00021|[RUSSELL KELLUM L...|       1|
|A001ED|[INTERNATIONAL PR...|       1|
|A01406|[LMC ANGELS LLC  ...|       1|
|A051EE|[KAI TAK LLC     ...|       1|
|A05534|[ISR PLATFORMS LL...|       1|
|A05B7B|           [Private]|       1|
|A062AE|[BCC LEASING LLC ...|       1|
|A067D7|[BAY VENTURE MANA...|       1|
|A08B59|[GT AVIATION LLC ...|       1|
|A08F76|           [Private]|       1|
|A091FB|[DEPARTMENT OF AG...|       1|
|A0A7AE|[MCCOY JOSEPH W  ...|       1|
|A0B860|[AVIATION CAPITAL...|       1|
|A0CA6E|[KING AIR INVESTM...|       1|
|A155AC|[SANTA BARBARA CO...|       1|
|A161B3|[1887 LEASING LLC...|       1|
|A168D3|[RINCON GABRIEL R...|       1|
|A1885A|[SUNBELT LLC     ...|       1|
|A1AFEA|    [Boeing Company]|       1|
|A1B7E8|[AIM AIRCRAFT LEA...|       1|
+------+--------------------+--------+
only showing top 20 rows

Percentage of planes with multiple Op values: 0.0
```

*Figure 11 – Statistics for the 'Op' field: fill rate (top), the unique values for each plane, and the percentage of planes with multiple values for the field (bottom). Table is sorted so that the planes with the most values for the 'Op' attribute are at the top (there are no planes with more than 1 unique value).*

```
Attribute:  OpIcao
Fill rate:  0.27742216145671494
+------+------------------+------------+
|  Icao|collect_set(OpIcao)|UniqueOpIcao|
+------+------------------+------------+
|A288E4|              [UAL]|           1|
|AC61B1|              [EDV]|           1|
|A2D2F2|              [AAY]|           1|
|A6751E|              [ASH]|           1|
|ABADC0|              [SWA]|           1|
|A742C2|              [AAL]|           1|
|A05A90|              [UCA]|           1|
|A7FDAB|              [ASA]|           1|
|A0AB8D|              [EJA]|           1|
|A8352B|              [NKS]|           1|
|A3918F|              [UPS]|           1|
|A8847A|              [SIS]|           1|
|A4CE28|              [ATN]|           1|
|A8FB06|              [PMS]|           1|
|A5BD0B|              [SWA]|           1|
|AA70F7|              [FDX]|           1|
|A03224|              [AMF]|           1|
|AA80C0|              [EJA]|           1|
|A318EA|              [SWA]|           1|
|AACFE9|              [SKW]|           1|
+------+------------------+------------+
only showing top 20 rows

Percentage of planes with multiple OpIcao values: 0.0
```

*Figure 12 – Statistics for the 'OpIcao' field: fill rate (top), the unique values for each plane, and the percentage of planes with multiple values for the field (bottom). Table is sorted so that the planes with the most values for the 'OpIcao' attribute are at the top (there are no planes with more than 1 unique value).*