Budapesti Mûszaki és Gazdaságtudományi Egyetem
Visual Analysis of Measurement Data (BMEVIMIAV16)

Homework specification
Air traffic and noise complaints in San Francisco

Rippl Balázs Róbert(FLHM6J)

October 22, 2022

1 Dataset

For this homework I found two different, but highly correlated datasets. Air Traffic Landings Statistics[1] and Aircraft Noise Complaint Data[2] are both publicly available datasets published by San Francisco's government.

1.1 Licensing

The datasets are licensed under Open Data's PDDL[3].

1.2 Metadata

The information below is considered before any cleaning or tidying of the data.

Dataset	**Landing**	**Noise complaint**
Column count	14	5
Row count	∼28200	4749
File format	csv	csv
File size	3129 KB	111 KB
Measurement start	2005 07	2005 01
Measurement end	2022 06	2019 12

After filtering both datasets, to only contain the same measurement timeframe, I can use the dates as binding data.

2 Basic profiling of the dataset

For this part I decided to analyse the two sets independently.

2.1 Landing data profile

Since there are 14 columns, I only specify the more important ones I think I will definitely use.

2.1.1 Activity period

This variable is a six digit number. The first four digits represent the year, and the last two digits represent a month within that year. This column seems perfect, as there are no missing values, and cardinality is also what should be: there are 204 distinct values. That is exactly the 17 years the minimum 200507 and maximum 202206 set. It also shows high correlation to aircraft version, which seems reasonable.

2.1.2 Landing aircraft type

A low cardinality variable having only three different values: passenger, freighter, combi. There are no missing values.

2.1.3 Aircraft body type

Also a low cardinality variable holding only four distinct values. It's about the planes' size, so high correlation to departure region and manufacturer, and 0% missing rate indicate that it's correct and usable data.

2.1.4 Aircraft model

The main data I am most interested in. Holds 103 different values with 0% missing rate.

2.1.5 Aircraft version

Also a categorical variable. Has a high missing rate, about 46%. But that is totally alright, as this column is just an extension for the model. Some models simply don't have different versions.

2.1.6 Landing count

One of the actually measured data, holding whole numbers. Contains no zero or missing values. It ranges from 1 to 2245, averaging at around 108.5. It's Q1 is 13, Q2 is 30, Q3 is 78. Has a standard deviation of 240.

2.1.7 Total landed weight

The other actually measured data, holding real numbers. It is probably in pounds, since that is the standard weigh for US flight data. Contains no zero or missing values. It ranges from 6850 all the way up to 275840010, averaging at around 18440374. It's Q1 is 2944000, Q2 is 9324000, Q3 is 19363150. Has a standard deviation of 29486272.

2.2 Noise complaint data profile

2.2.1 Year

Year of measurement. Ranges from 2005 to 2019, no missing values. Q1 is at 2011, Q2 at 2015, Q3 at 2017. Averages at 2014.

2.2.2 Month

Month of measurement. I'm going to concatenate it to the year, so it matches the format in the other dataset.

2.2.3 Community

Categorical data with 128 distinct values. Specifies different areas. Has no missing values.

2.2.4 Total complaints

The number of monthly complaints associated by community. Has only a single missing value. Q1 is 3, Q2 is 35, Q3 is 602. It ranges from 1 up to 93498, averaging at 2276 with a deviation of 7721.

2.2.5 Total number of callers

The number of Complainants associated by community. The number of people who did all the complaints in the previous column. Ranges from 1 to 40859, averaging at 50. It's Q1 is 1, Q2 is 3, Q3 is 11. Has a deviation of 770.

3 Questions raised

I would like to find, whether any variable from the landing data causes a higher number of noise complaints, above the trivial "more landings = more complaints". I'd also like to see, if different areas of San Francisco react to changes in the traffic differently. Lastly, it would be interesting, how the COVID pandemic affected the individual datasets. How did traffic change, and did people in home office complain more?

4 Planned technology usage

I want to create a report in the form of a jupyter notebook, using pandas, bamboolib and seaborn.

References

[1] `https://data.sfgov.org/Transportation/Air-Traffic-Landings-Statistics/fpux-q53t`

[2] `https://data.sfgov.org/Transportation/Aircraft-Noise-Complaint-Data/q3xd-hfi8`

[3] `https://opendatacommons.org/licenses/pddl/summary/`