Carolyn Morikawa (ckm65)
May 2019
INFO 2950: Data Science Final Project

DataSheet

Performance_Master_Aircraft

A merged dataset of three unique large datasets

Motivation:

1. **For what purpose was the dataset created?**

The purpose of this dataset was created to analyze delayed airplane flights arrival and departures. The dataset was analyzed on three main variables including: model, location, and day of the week in order to find significance in specific delayed flights. The dataset was analyzed to support my main hypothesis that there is a positive correlation between delayed departures and delayed arrivals.

- The purpose of the performance dataset was to map the scheduled and actual departure and arrival times reported by certified US air carriers. It was created to measure the performance of specific airplanes.
- The purpose of the master dataset was to dataset contains the records of all U.S Civil Aircrafts maintained by the FAA, Civil Aviation Registry, and Aircraft Registration branch.
- The purpose of the aircraft reference dataset contains registry information about various aircraft models. It was created to analyze which aircrafts and being used and for what.

2. **Who created this data set and behalf of which entity?**

The performance_master_aircraft data was created and cleaned my Carolyn Morikawa.

- The performance dataset was created by the Office of Airline Information and the Bureau of Transportation Statistics.
- The master dataset was created by the U.S Department of Transportation and the Federal Aviation Administration.
- The aircraft reference dataset was created by the U.S Department of Transportation and the Federal Aviation Administration.
-

3. **Who funded the creation of the dataset?**

David Mimno, and several Teaching Assistants of the class INFO 2950: Introduction to Data Science funded the creation of this dataset.

- The performance dataset was funded by the Bureau of Transportation statistics and the U.S Department of transportation.
- The master dataset was funded by the Bureau of Transportation, U.S Department of transportation and the Federal Aviation Registry.
- The aircraft reference dataset was funded by the Bureau of Transportation, U.S Department of transportation and the Federal Aviation Registry.

Composition:

1. **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**

The instances comprised in the dataset represent aircraft represented airports, aircraft type,time, and flight information. This instances are representative and records of past and current flight history.

- The instances in the performance dataset represent airports, time and flight information
- The instances in the master dataset represent aircraft type and flight information
- The instances in the aircraft reference dataset reprepresent aircraft type.

2. **How many instances are there in total?**

There are 4 total instances: airports, aircraft type,time, and flight information

- There are 3 instances in the performance dataset airports, time and flight information
- There are 2 instances in the master dataset:aircraft type and flight information
- There is 1 instance in the aircraft reference dataset: aircraft type.

3. **Does the dataset contain all possible instances or is it a sample of instances from a larger set?**

The dataset contains a sample of instances from the larger dataset. The dataset is comprised of three very large different dataset. In order to clean my dataset I only extracted a few instances from each of the datasets to compile into my results

4. **What data does each instance consist of?**

The data that each instance consists of:

Airport : Destination Airport, City Name, Origin Airport, City Name

Time: Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.Difference in minutes between scheduled and actual departed time. Early arrivals set to 0. Day of Week, year/

Aircraft Type: aircraft Tail Number, A code assigned to the aircraft manufacturer, model and series, name of the aircraft manufacturer,name of the aircraft model and series,A code assigned to the engine manufacturer and model.

Flight Information: this is a combination of all of the information listed above including the average air time.

5. Is there a label or target associated with each instance?

There is no target for specific airports. It includes all Airports within the US. The target for Time specifically focuses on the Day of the Week. The target for Aircraft focuses on the model number and manufacturer of the plane. The flight information is a compilation of all remaining and necessary data.

6. Is any information missing from individual instances? N/A

Are relationships between individual instances made explicit? Yes, the words are categorized by their use in respective chapters.

7. Are there recommended data splits?

No there are no recommended data splits. I spit the data based on the amount of digits in the N_Number and the Tail_Number. I then split the datasets based on the corresponding code of the aircraft.

8. Are there any errors, sources of noise, or redundancies in the dataset?

There could be potential errors in the dataset relating towards the overlapping of the variables. There are redundancies in the dataset. For example, a tail number occurs multiple times because it is tracked when it enters or leaves a destination. There are also redundancies in the type of aircraft used, which makes sense because there are only so many different types of aircraft.

9. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self contained, it does not rely on external resources for specific analysis purposes. However, in order to create the dataset, 3 datasets were linked.

10. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?

No, this dataset does not contain data that might be considered confidential.

11. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No, the dataset does not contain information that might be offensive, insulting, threatening or might otherwise cause anxiety.

12. Does the dataset relate to people?

N/A.The data does not relate to people.

13. Does the dataset identify any subpopulations (e.g., by age, gender)?

The subpopulations of this dataset represent the day, month, year and day of the week for flight departures and arrivals. However, each individual dataset has many subpopulations. For example location it has origin, origin city, originstate, origin airport.

14.Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

No

15.Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opin- ions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? No

16.Any Other Comments? No

Collection Process

1. **How was the data associated with each instance acquired?**

The data was associated with each instance acquired as it was directly observable based on departing and arriving airplanes.Each instances was recorded based on when it was registered in an online system tracking the specific information based on when it arrived and departed from specific locations. All the information was tracked my specific airports and online mechanism along with the tracker located on aircrafts.

2. **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?**

None, the data was taken collected three specific API's from the U.S Department of Transportation and the Federal Aviation Administration.

3. **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deter- ministic, probabilistic with specific sampling probabilities)?**

My dataset is a combination of three very large datasets. In strategy sampling for specific variables was looking at which variable most closely related to my key objectives.

- Does one type of aircraft typically arrive and depart late to specific destinations? If so, which aircraft model?
- Does a specific location constantly have a delayed departure and receive delayed arriving flights? If so, which location and why do you think that is?
- Does a specific day in the week consistently have more delayed departing and arriving flights.

4. **Who was involved in the data collection process (e.g., students, crowdworkers, contrac- tors) and how were they compensated (e.g., how much were crowdworkers paid)?**

There were no individuals involved in the dataset. However the dataset included the specific instances mentioned in the above section. The airport cities were involved in this dataset along with specific models and types of aircraft.

5. **Were any ethical review processes conducted (e.g., by an institutional review board)?**

No, there were not any ethical review processes conducted.

6. **Does the dataset relate to people?**

The dataset does not relate to people.

7. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

N/A. The dataset does not relate to people.

8. **Were the individuals in question notified about the data collection?**

N/A. There were no individuals involved in this dataset.

9. **Did the individuals in question consent to the collection and use of their data?**

N/A.There were no individuals involved in this dataset.

10. **If consent was obtained, were consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**

The merged dataset does not provide personal information about individuals or their specific flight history. Therefore, there is no concerns regarding consenting individuals.

11. **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted?**

No, it does not provide information relating to how many people were on the flights and who was on it. Therefore the potential impact of the dataset does not breach any individual privacy concerns.

12. **Any other comments?** No.

Preprocessing/cleaning/labeling:

1. **Was any preprocessing/cleaning/labeling of the data done (e.g.,discretizationorbucket- ing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**

Overall the dataset as a whole was preprocessed by extracting specific variables from the performance, master and aircraft reference/registry datasta in order to provide evidence to support my hypothesis. The variables extracted from each individual dataset are mentioned in the previous questions.

When merging the performance and the master dataset I had to clean the dataset concatenating an N to make sure that the Tail_Number and N_Number matched up to find the numbers that were overlapping. A column was added to the master dataset copying the values of the N_Number and setting the column name to Tail_Number. I merged the performance and master dataset on overlapping/interaction of the Tail_Numbers.

When merging the performance_master dataset(created above) , with the aircraft registration dataset I added a column to the performance_master_dataset named CODE. I then merged the specific MFR MDL CODE variable from the performance_master dataset with the CODE variable in the registry dataset.

Furthermore,the datasets were cleaned based on a filter to remove any blank entries of “Nan” values. The packaged utilized to do this was isnull().

The final dataset contains information about the Tail_Number, DayOfWeek, Year, AirTime, ArrDelayMinutes, DepDelayMinutes, CODE, OriginCityName, DestCityName, MFR, MODEL

2. **Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**

The raw data was not saved in addition to the preprocessed/cleaned/labeled data. Since the three individuals datasets were very large I only extracted the variables to further my analysis. I did not include all the raw data because it would become messy and not precise for the reader.

3. **Is the software used to preprocess/clean/label the instances available?**

No software was used. Only python pandas, numpy, pyplot, scikit-learn packages were utilized in jupyter notebook to complete these extractions.

4. **Any other comments?** No.

Uses

1. **Has the dataset has been used for any tasks already?**

The dataset has not been used for any tasks already because it is a unique and new merged and cleaned dataset.

- The performance dataset has been used to track specific performances calculating the percent on time for specific airports
- The master dataset has been used to compare registration it across the years.
- The aircraft registration dataset is mainly used as a subset of the master dataset for tasks.

2. **Is there a repository that links to any or all papers or systems that use the dataset?**

The U.S Department of Transportation and the Federal Aviation Administration are systems that could utilize this data set in the future.

3. **What (other) tasks could the dataset be used for?**

This dataset could be used to analyze specific days in aircraft model to track arrival and delayed departures. It can also be used as matching the Tail Number to specific model codes and then tracking the average airtime for specific aircrafts. The individual datasets are comprised of a lot of information but in order to support my hypothesis and key objectives I specifically extracted the recorded instances.

4. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**

The dataset was preprocess/cleaned/labeled to support and analyze my specific hypothesis. The variables extracted were specifically for my findings but there could be multiple other extractions to complete the same information. I chose these variables because it was very clear how to analyze the dataset. I had to filter out some of the ‘blank entries’ which could become a problem if an individual wants to look specifically at data from all information including ‘NaN entries.

5. **Are there tasks for which the dataset should not be used?**

The dataset should not be used solely for reading. Datasets are created to be analyzed and identify common trends.

6. **Any Other comments?** N/A