README

1. Zachary Chun (chunz2), Ryan Feng (rfeng), Yuma Tou (yumatou)

2. washington_flights

3. This dataset contains info about airplane flights departing from Washington state airports in January 2017. It includes information about destination and origin, amount of time of any delayed departures, date, carrier. These features are useful to answer the question: "For a given flight, how many minutes late (or early) will the flight arrive?" which is question that many people wonder about when they travel. We think that these features (destination, origin, carrier, etc) should lead to a correlation with predicting arrival delays.

   To ensure quality: we ensured the following:

   - no military times were greater than or equal to 2400 or less than 0
   - no negative or 0 distances traveled
   - all the destinations airports exist (has a valid ID and name)
   - all the departure airports are from Washington state
   - no date outside January (31 days)
   - no day of the week outside of Monday to Sunday (1 to 7)
   - every sample has all 10 features included

4. Features in order:

   (a) Day of Month (1-31)
   (b) Day of Week (1-7, 1 = Monday)
   (c) Carrier
   (d) Origin Airport ID
   (e) Destination Airport ID
   (f) Departure Time (Military Time without leading 0s)
   (g) Departure Delay (Minutes, negative if early departure)
   (h) Arrival Time (Military Time without leading 0s)
   (i) Distance (Miles)

   For any given input file, the features listed above will appear in a comma separated list (in a csv file). The corresponding output file will have Arrival Delay (Minutes, negative if early arrival). Note that if anyone looking at this is curious as to what airports the ids correspond to, we included airport_ids.csv.

5. We decided to split our data into 90 : 10 (training : test) by partitioning randomly. This should give us a fair distribution of input and output values in both sets. Choosing the 90 : 10 split was based on the convention that we want lots of data to train on, and don't need as much to test on.

6. We believe our dataset is ethical because this is public data that doesn't infringe on any individual's personal information. Since our evaluation is on the airline level, individual people are not targeted in anyway. In addition, the people have a right to know about how public services rank versus each other. This data is compiled from observable instances anyways.

Our dataset comes from: https://www.transtats.bts.gov/Tables.asp?DB_ID=120