EDA with R
airline on-time performance data

Veena Mendiratta
Adjunct Professor
MLDS Program
October 2023

https://www.transtats.bts.gov/

Data

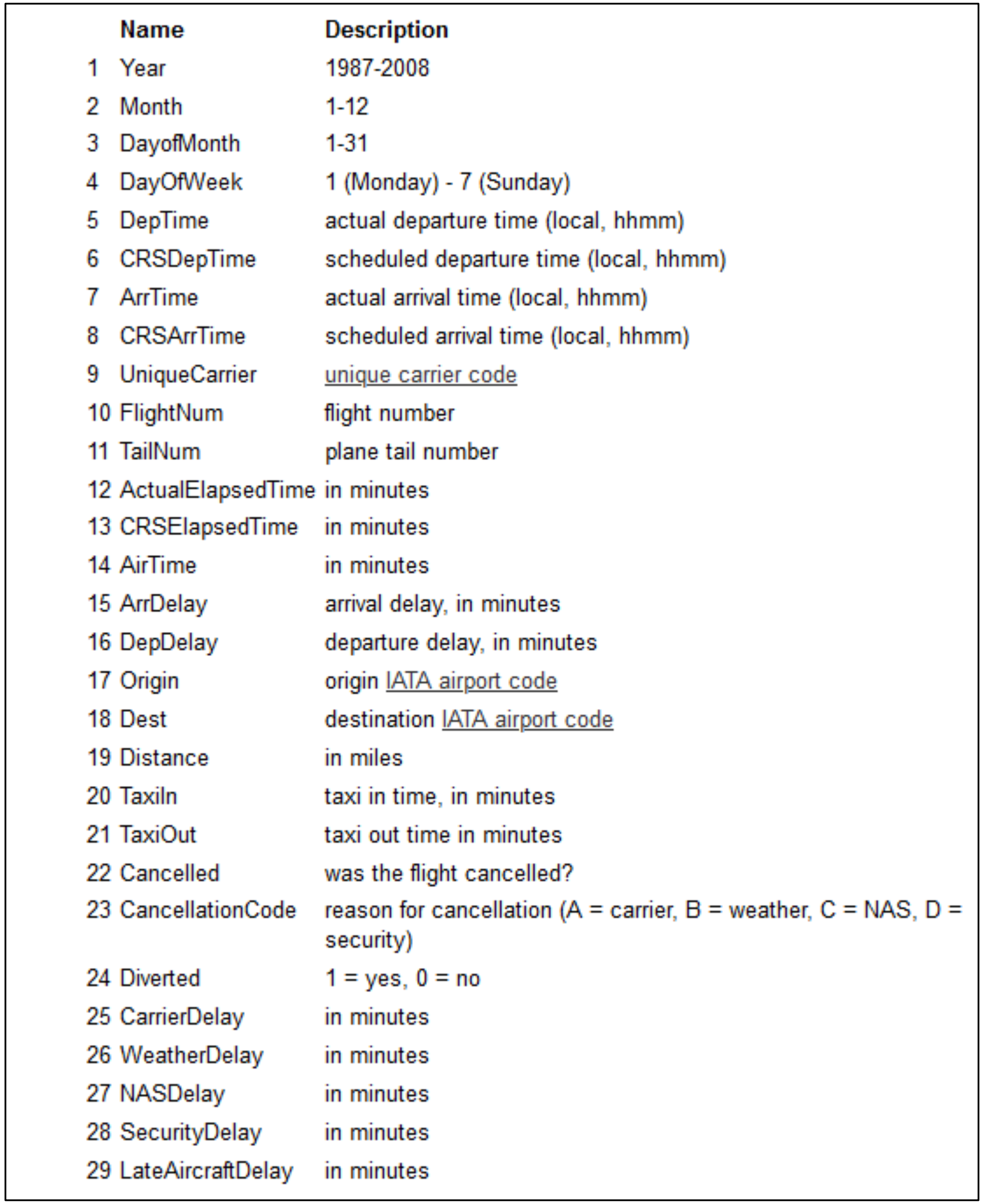
The data consists of:

— flight arrival and departure details,

— for all commercial flights within the USA,

— from October 1987 to present.

Variables of Interest



	Name	Description
1	Year	1987-2008
2	Month	1-12
3	DayofMonth	1-31
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	unique carrier code
10	FlightNum	flight number
11	TailNum	plane tail number
12	ActualElapsedTime	in minutes
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes
16	DepDelay	departure delay, in minutes
17	Origin	origin IATA airport code
18	Dest	destination IATA airport code
19	Distance	in miles
20	TaxiIn	taxi in time, in minutes
21	TaxiOut	taxi out time in minutes
22	Cancelled	was the flight cancelled?
23	CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24	Diverted	1 = yes, 0 = no
25	CarrierDelay	in minutes
26	WeatherDelay	in minutes
27	NASDelay	in minutes
28	SecurityDelay	in minutes
29	LateAircraftDelay	in minutes

Typical steps for EDA

- State the problem you are trying to solve.
- Clean and prepare the data (most time-consuming task).
- Explore the data.
  - Get to know its properties and quirks.
  - Check for missing values and perform imputation.
  - Check summaries of numerical variables and tables of the categorical data.
  - Plot univariate and multivariate representations of the variables;
    ➔ can get an overview of the data quality and find outliers.
  - Check if variables need to be transformed;
    ➔ often a logarithmic transformation of skewed measurements.
- Choose a model and train it on the data.

Problem

With this data, it is possible to answer many interesting questions. Examples include:

- Do planes with a delayed departure fly with a faster average speed to make up for the delay?
- How does the delay of arriving flights vary during the day? Are planes more delayed on weekends?
- Are there differences in delays between January and July?
- How has the market share of different airlines shifted over the years?
- Are there specific planes that tend to have longer delays? What characterizes them? Maybe the age, or the manufacturer?

dlookr R package

Collection of tools that support data diagnosis, exploration, and transformation

- Data diagnostics provides information and visualization of missing values and outliers and unique and negative values to help you understand the distribution and quality of your data.
- Data exploration provides information and visualization of the descriptive statistics of univariate variables, normality tests and outliers, correlation of two variables, and relationship between target variable and predictor.
- Data transformation supports binning for categorizing continuous variables, imputes missing values and outliers, resolving skewness.
- Creates automated reports that support these three tasks.

Reference manual:	dlookr.pdf
Vignettes:	Exploratory Data Analysis Data quality diagnosis Introduce dlookr Data Transformation

Other R Packages for EDA

All packages work on full datasets
All packages contain functions for summarizing datasets

- arsenal - comparison of two data frames that can detect shared variables (compare function)
- autoEDA - GitHub-based tool for univariate and bivariate visualizations
- DataExplorer – simple data transformations
- dataMaid - two main functions: checks of data consistency and validity, and summarize each column
- dlookr - collection of tools that support data diagnosis, exploration, and transformation
- ExPanDaR package – designed for panel data exploration
- explore package – full dataset summaries, uni- and bivariate visualizations, and modeling based on decision trees or logistic regression.
- exploreR package - analysis is based on linear regression; univariate regression model for each independent variable, plotting target variable against each independent variable, feature standardization by scaling.
- funModeling - rich set of tools for EDA connected to the book by Casas (2018).
- inspectdf package – summarize dataset with number of missing values, etc.; univariate analysis; bivariate relationships described by Pearson correlation coefficient.