PRAC2: Neteja i anàlisi de les dades

Marc Ferrer Margarit (mferrermargarit@uoc.edu) i Marc Ramos Bruach (mramosbru@uoc.edu)

5/15/2021

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre? (0.5) (MF)

El dataset que farem servir per a aquesta pràctica conté els retards i cancel · lacions dels vols 2015 als EUA. Normalment els motius principals dels retards de vol són relacionats amb el temps, però en alguns casos també hi ha retards de vols relacionats amb les companyies aèries o aeroports. Aquest document examina i mostra les causes de retard i cancel · lació en diversos aspectes. Així doncs aquest dataset és important per saber quines són les principals causes dels retards que s'hi han produït durant el 2015.

La pregunta o problema que es vol respondre és quines són les causes més comunes que han provocat els retards o les cancel · lacions dels vols i també veure quins són els llocs o el conjunt de destinacions que és més comú que es produeixi una cancel · lació o quines són les aerolínies més propenses a tenir cancel · lacions.

Aquest dataset ha sigut obtingut a partir d'una pràctica anterior realitzada durant el màster de Data Sciene de la UOC. També es pot obtenir el dataset complet, amb totes les dades dels vols (aprox. 600 MB) al següent enllaç: https://www.kaggle.com/usdot/flight-delays?select=flights.csv

2. Integració i selecció de les dades d'interès a analitzar. (0.5) (MF)

Per veure les dades que conté el dataset el carregarem i mostrarem les columnes que conté i la mida del dataset:

```
fligths <- read.csv("../data/flights.csv")
dim(fligths)
```

```
## [1] 145952     28
```

```
colnames(fligths)
```

```
##  [1] "YEAR"                "MONTH"               "DAY"
##  [4] "DAY_OF_WEEK"         "AIRLINE"             "FLIGHT_NUMBER"
##  [7] "ORIGIN_AIRPORT"      "DESTINATION_AIRPORT" "SCHEDULED_DEPARTURE"
## [10] "DEPARTURE_TIME"      "DEPARTURE_DELAY"     "TAXI_OUT"
## [13] "WHEELS_OFF"          "SCHEDULED_TIME"      "ELAPSED_TIME"
## [16] "AIR_TIME"            "DISTANCE"            "WHEELS_ON"
## [19] "TAXI_IN"             "SCHEDULED_ARRIVAL"   "ARRIVAL_TIME"
## [22] "ARRIVAL_DELAY"       "DIVERTED"            "CANCELLED"
## [25] "CANCELLATION_REASON" "AIR_SYSTEM_DELAY"    "LATE_AIRCRAFT_DELAY"
## [28] "WEATHER_DELAY"
```

Com podem veure tenim 28 columnes i un total de 145.592 dades en el dataset obtingut. També podem veure quines són les dades que conté el nostre dataset. Com que volem veure els retards o problemes que poden

haver-hi, les causes i en quines aerolínies només cal que seleccionem aquelles columnes que ens proporcionin aquesta informació. En aquest cas serien:

- YEAR
- MONTH
- DAY
- AIRLINE
- FLIGHT_NUMBER
- ORIGIN_AIRPORT
- DESTINATION_AIRPORT
- SCHEDULED_DEPARTURE
- DEPARTURE_TIME
- DEPARTURE_DELAY
- SCHEDULED_TIME
- SCHEDULED_ARRIVAL
- ARRIVAL_TIME
- ARRIVAL_DELAY
- DIVERTED
- CANCELLED
- CANCELLATION_REASON
- AIR_SYSTEM_DELAY
- LATE_AIRCRAFT_DELAY
- WEATHER_DELAY

Amb aquestes dades ja podem fer un anàlisi complet per tal de donar resposta a les preguntes proposades.

3. Neteja de les dades. (2) (MR)

3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos? (1)

3.2. Identificació i tractament de valors extrems. (1)s

#4. Anàlisi de les dades. (2.5) (MR) ## 4.1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar). (0.5) ## 4.2. Comprovació de la normalitat i homogeneïtat de la variància. (1) ## 4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents. (1)

5. Representació dels resultats a partir de taules i gràfiques. (2) (MF)

6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema? (0.5) (MR)

7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python. (2) (MF)