**Components del grup: Núria Palahí**

**1.- Descripció del dataset. Perquè és important o quina pregunta/problema pretén respondre?**

He escollit el dataset <https://www.kaggle.com/c/bike-sharing-demand/>

M'ha semblat interesant poder estudiar aquestes dades veure la predicció de lloguer des de diferents punts de vista, temps, temporada, moment del dia.

**2.- Neteja de les dades.**

**2.1.- Selecció de les ades d'interès a analitzar. Quins són els camps més rellevants per tal de respondre al problema**?

Les dades a utilitzar son els arxius train.csv i test.csv. Cadascun d'aquests arxius té la següent informació:

* datetime: dia i hora
* season: camp numèric que indica hivern, primavera, tardor o estiu
* holiday: camp binari que indica si és festiu
* workingday: camp binari que indica si és laborable
* weather: camp numeric que indica com ha estat el dia (1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog)
* temp: temperatura en C
* atemp: temperatura de sensació
* humidity: humitat
* windspeed: velocitat del vent
* count: total del lloguer

Utilitzarem totes les dades, però potser podríem classificar com menys rellevants la temperatura de sensació i el de l'estació de l'any. És clar que els lloguers, temperatura, temps i data serien els més rellevants, ja ens deixarien fer una bona aproximació a la solució.

**2.2. Les dades contenen zeros o elements buits? Valors extrems? Com gestionaries cadascun d'aquests casos?**

En el model que he escollit no es contemplen.

Si trobéssim elements buit podríem tractar-los en r amb una sentència tipus b[is.na(b)]<-"".

Donat que en aquest exemple estem contemplant dades numèriques, el 0 l’hauríem de donar com a vàlid.

Els valors extrems els tractarem sobretot en les sentencies gplot, delimitant els límits dels valors x i y.

**3.- Anàlisi de les dades**

**3.1.- Selecció dels grups de dades que es volen analitzar/comparar.**

Compararem les dades dels arxius train i test, aquests dos arxius contenen. Train les dades amb lloguers, train les mateixes dades però sense els lloguers per poder fer-ne la predicció de l'estudi. Comprarem totes les dades i més tard les anirem agrupant segons com vulguem visionar els resultats, estació de l'any, hores segons el tipus de dia, etc...

**3.2.- Comprovació de la normalitat i homogeneïtat de la variància. Si és necessari (i possible), aplicar transformació que normalitzin les dades.**

No he tingut temps d’explicar aquest punt.

**4.- Representació dels resultats a partir de taules i gràfiques.**

**Imagen que contiene mapa, texto

Descripción generada con confianza muy altaImagen que contiene texto, mapa

Descripción generada con confianza muy altaImagen que contiene texto, mapa

Descripción generada con confianza muy altaImagen que contiene texto, mapa

Descripción generada con confianza muy altaImagen que contiene texto, mapa

Descripción generada con confianza muy altaImagen que contiene texto, mapa

Descripción generada con confianza muy alta**

**5.- Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?**

Si, l’estudi es concloent i es podria utilitzar en un pla de negoci per obrir una botiga de lloguer.

1. Gent lloga més bicis a la tardor que a la primavera
2. Hi han més lloguer quan el temps es bo
3. Es lloguen més bicis als matins/vespres en dies laborables, i per dies sencers als caps de setmana
4. La possibilitat de temps normal es major a la primavera
5. La possibilitat de mal temps és major a l'estiu i hivern
6. La possibilitat de bon temps és sempre alta

**6.- Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i presentació de les dades.**