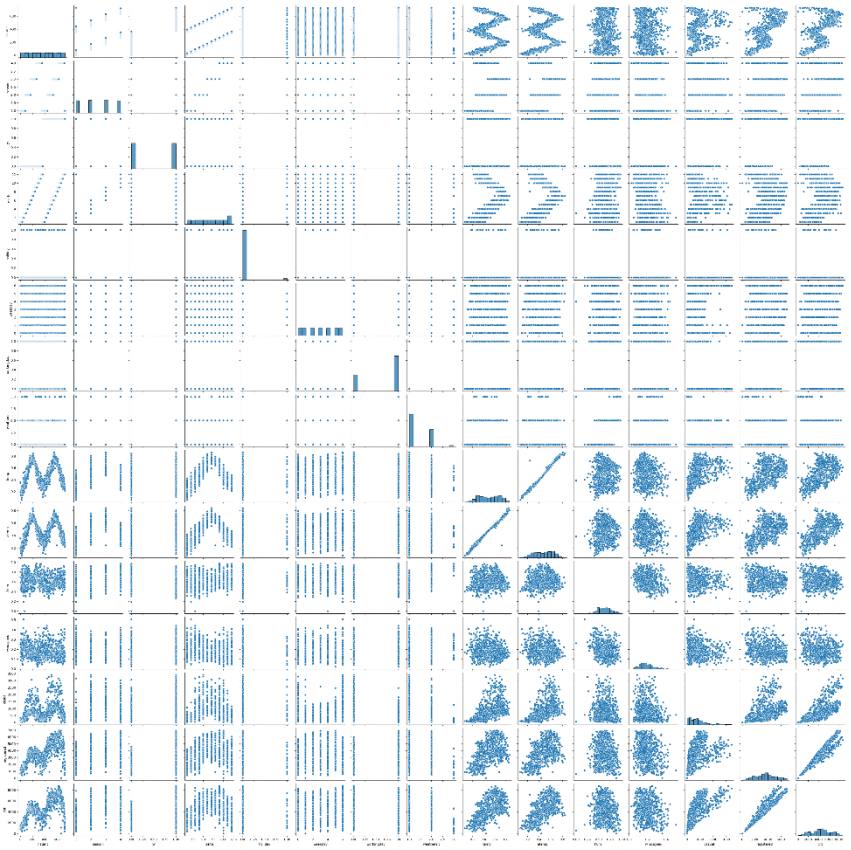
PRÀCTICA 1 – ANÀLISI I REGRESSIÓ DE DADES



Grup: GPA202-0930

Daniel Alcover Neto - 1562889

Alexandre Galvany Pardo - 1566586

Jordi González González - 1526878

Introducció a la Pràctica

En aquesta pràctica aplicarem els diferents coneixement que hem obtingut amb l'anàlisi i la regressió de dades usant una base de dades assignada a través de Kaggle.

En el nostre cas tenim una base de dades relacionada amb el Bike Sharing el qual consisteix en el lloguer de bicicletes dins de la ciutat de Washington D.C. per tal de reduir el tràfic i la contaminació que produeixen els cotxes així com el de millorar la salut dels diferents usuaris.

Referent a la base de dades en si es poden extreure tot tipus de dades a part de les del temps que s'utilitzen o les distàncies que recorren i és aquí on el nostre grup aplicarà aquest anàlisis.

Aquestes són els atributs que trobem més interessants dins de la pròpia:

- **casual**: Nombre d'usuaris casuals que utilitzen el servei sense estar registrats.
- **registered**: Nombre d'usuaris registrats que utilitzen el servei.
- **cnt**: Nombre total de lloguer de bicis d'usuaris tant casuals com registrats.

L'atribut que estudiarem en profunditat serà el cnt. Aquest atribut ens ajudarà a predir respecte els altres com poden ser el temps, l'estació de l'any, la temperatura, si el dia és festiu o el dia de la setmana i quina és la demanda de bicicletes que podem tenir per aquell dia amb aquells atributs en concret.

Creiem que el fet de predir quin es el nombre de bicicletes que s'utilitzaran cada dia permet, per exemple, augmentar la oferta de bicicletes els dies on haguem predit que se'n farà un ús major i els dies en els que aquesta predicció mostri que l'ús es reduït es podrà aprofitar pel manteniment d'algunes estacions de bicicletes o de les pròpies bicicletes en si.

Podeu trobar tots els arxius en el nostre repositori: https://github.com/Jordigg2000/APC_LAB

Llibreries Utilitzades

- **Numpy**: És una llibreria que està especialitzada en el càlcul numèric i anàlisis de dades per a un volum normalment elevat de dades com és en el nostre cas.
- **Scikit-learn**: És una llibreria molt útil dins del apartat de ML ja que ens proporciona algorismes d'aprenentatge tant supervisats com no supervisats. Aquests algorismes ens permetran fer un estudi més profund en les nostres dades.
- **Matplotlib**: Llibreria que ens servirà per poder tenir una bona visualització de les dades ja que ofereix gràfics que faran que aquestes siguin comprensibles. Alguns exemples serien diagrames de dispersió, histogrames, gràfics de barres...
- **Scipy**: És una llibreria ideal per l'anàlisi de dades que volem fer ja que ens permet manipular les dades i visualitzar-la a través de una ample llista de comandes d'alt nivell de Python.
- **Pandas**: Aquesta llibreria ens serà molt útil ja que ens permetrà netejar i analitzar les dades. Esta especialitzada en dades que estan desordenades del món real com és en el nostre cas.

Apartat (C): Analitzant Dades

Anàlisi Simple Atributs Base de Dades

La nostre base de dades consta d'un total de 16 atributs que són els següents:

- **instant**: Record index.
- **dteday**: Data.
- **season**: Estacions del any (1:primavera, 2:estiu, 3:tardo, 4:hivern).
- **yr**: Any (0: 2011, 1:2012).
- **mnth**: Mes (1 a 12).
- **hr**: Hores (0 a 23).
- **holiday**: Si el dia és festiu o no.
- **weekday**: Dies de la setmana.
- **workingday**: En el cas de que el dia sigui festiu o cap de setmana és un 1, la resta un 0.
- **temp**: Dades de la temperatura normalitzada en Celsius. La temperatura està posada en tmin, t_max, calculada per hores.
- **atemp**: Sensació tèrmica normalitzada en Celsius. La temperatura està posada en tmin, t_max, calculada per hores.
- **hum**: Humitat normalitzada. Els valors estaven dividits entre 100 (max)
- **windspeed**: Velocitat del vent normalitzada. Els valors estaven dividits entre 67 (max)
- **casual**: Nombre d'usuaris casuals que utilitzen el servei sense estar registrats.
- **registered**: Nombre d'usuaris registrats que utilitzen el servei.
- **cnt**: Nombre total de lloguer de bicis d'usuaris tant casuals com registrats.

Per saber els tipus dels atributs de la nostre BD el que farem serà utilitzar una funció de la llibreria **pandas** que es diu **dtypes**:

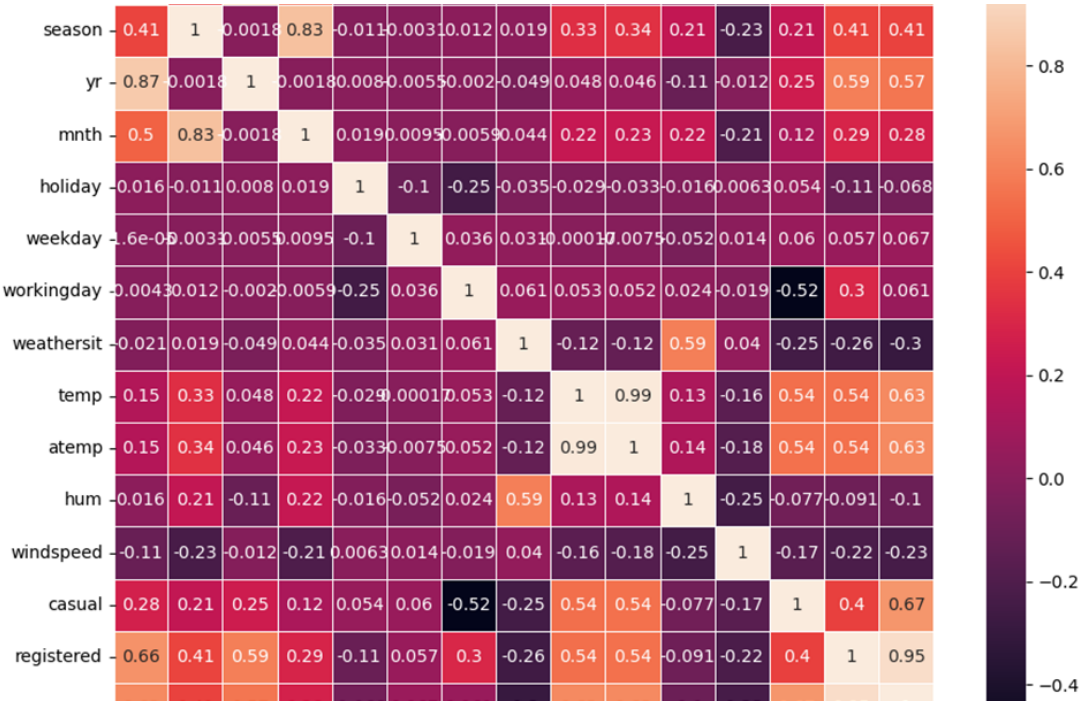
```
In [19]: print("Visualitzem el tipus de les dades de la nostre base de dades:")
         dataset.dtypes

         Visualitzem el tipus de les dades de la nostre base de dades:

Out[19]: instant         int64
         dteday         object
         season          int64
         yr              int64
         mnth            int64
         holiday         int64
         weekday         int64
         workingday      int64
         weathersit      int64
         temp          float64
         atemp         float64
         hum           float64
         windspeed     float64
         casual          int64
         registered      int64
         cnt             int64
         dtype: object
```

Correlació dels Atributs

El atribut amb una major correlació respecte els altres és el de cnt per això podem arribar a la conclusió que la nostre aposta inicial per aquest atribut és correcte ja que és el que està més interrelacionat amb els altres.



En aquest mapa de calor es pot observar les correlacions que tenen tots els atributs entre ells. Fa que sigui més fàcil de diferenciar gràcies a la gamma de colors.

Els atributs amb els que està més relacionat son els següents:

- Instant
- A_temp
- Season
- Casual
- Registered
- Year

En el cas de Instant el descartarem ja que aquest valor només representa un identificador de cadascuna de les mostres de la base de dades i no aporta valor en el nostre estudi.

També per a la selecció que farem servir seran els atributs de Casual, Registered i el de Atemp (fa referència a la sensació tèrmica) ja que pensem que seràn els que, a part de tenir una correlació més important, ens ajudaran a predir millor el cnt.

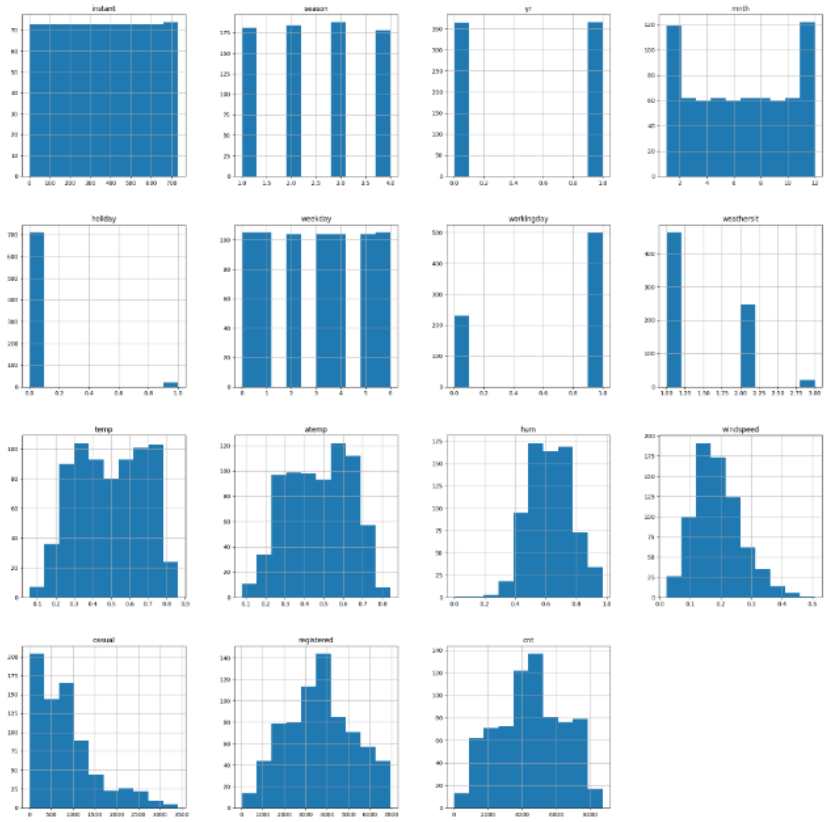
Distribució Normal de les Dades

Per seguir entenent millor el comportament de les dades hem generat un histograma de cada atribut de la base de dades. Per fer això hem usat una funció de la llibreria **pandas** que es diu **hist** que internament usa la llibreria **Matplotlib**. Amb aquesta informació podem observar l'evolució que tenen i el tipus de distribució.

Per exemple, una distribució coneguda i de les més habituals és la distribució normal o Gaussiana. En el nostre cas les dades que tenen aquesta distribució son la temperatura, la humitat, la velocitat del vent, el registered i el "count" que representa el nombre total de lloguers de bicicletes.

Observant l'atribut "cnt", com a exemple, podem adonar-nos que les condicions climatològiques (temperatura, humitat, velocitat del vent…) estan molt presents, influencien molt el nombre de lloguers fets.

```
In [19]: dataset.hist(figsize=(25,25))
         <IPython.core.display.Javascript object>
```



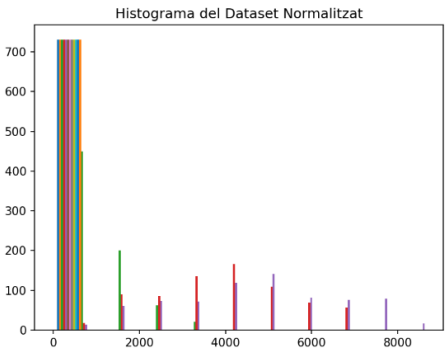
Apartat (B): Primeres regressions

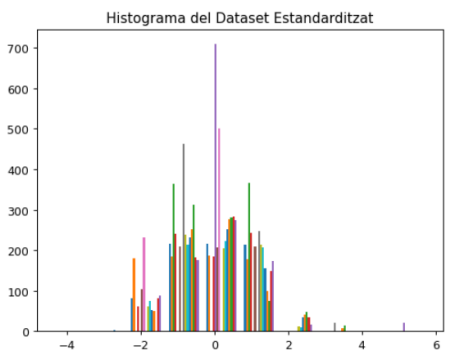
Si fem un estudi de la correlació entre el nostre Atribut objectiu “cnt” y la resta d’atributs podem extreure una conclusió y valorar quins atributs son els mes importants per fer una bona predicció.

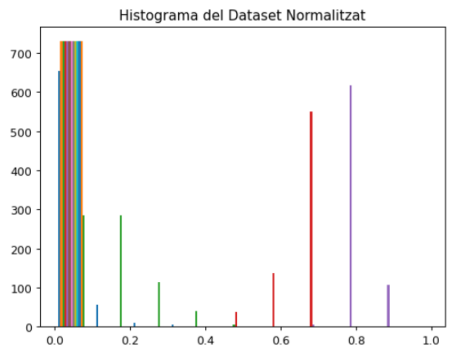
```
cnt           1.000
registered    0.946
casual        0.673
atemp         0.631
instant       0.629
temp          0.627
yr            0.567
season        0.406
mnth          0.280
weekday       0.067
workingday    0.061
holiday      -0.068
hum          -0.101
windspeed    -0.235
weathersit   -0.297
Name: cnt, dtype: float64
```

Podem observar que els atributs Registered, Casual i atemp son els més representatius per tant seràn els que més utilitzarem durant l’estudi en aquesta primera pràctica.

Ara valorarem que és més eficient a l’hora de de normalitzar o estandarditzar les nostres dades.



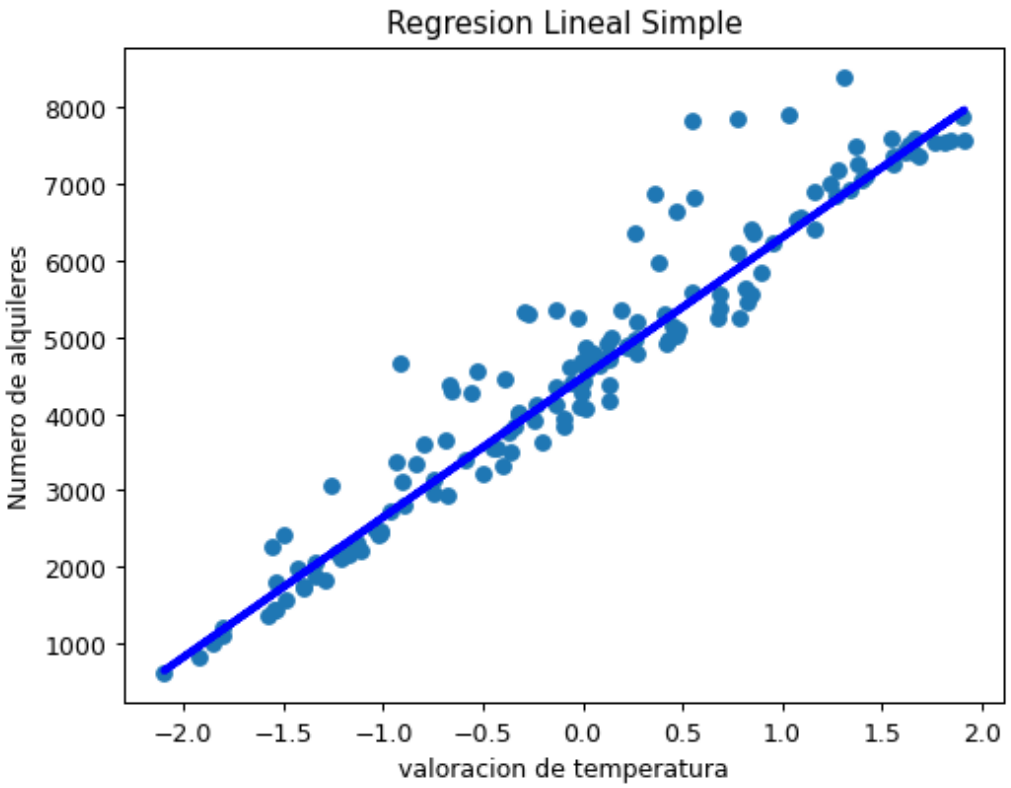




A l’hora de tractar amb les dades busquem l’histograma que s’apropi més a la distribució Gaussiana per a que sigui més còmode pel model per manipular les dades.

Per tant utilitzarem sempre el dataset Estandaritzat per fer l'anàlisi de la base de dades.

Fent la Regressió lineal simple de l'atribut "registered" estandaritzat observem aquesta línea.



i a més calculem el seu MSE…

```
Mean squeared error:  316785.7044683083
R2 Score:  0.9141379297420034
```

Veiem que efectivament tenim un "Score" molt alt, és a dir, una **correlació entre atributs** molt alta amb un error quadràtic menor.

A continuació farem una Regressió lineal **múltiple** per relacionar tots els atributs:

```
Mean squeared error:  11552.633052037552
R2 Score:  0.9969614907942353
```

Com hem utilitzat tots els atributs tenim un **Score gairebé perfecte**, però ara veurem que si treiem els atributs menys rellevants i ens quedem amb els més significatius…

```
Mean squared error:  3447.549862588747
R2 Score:  0.9991125620380082
```

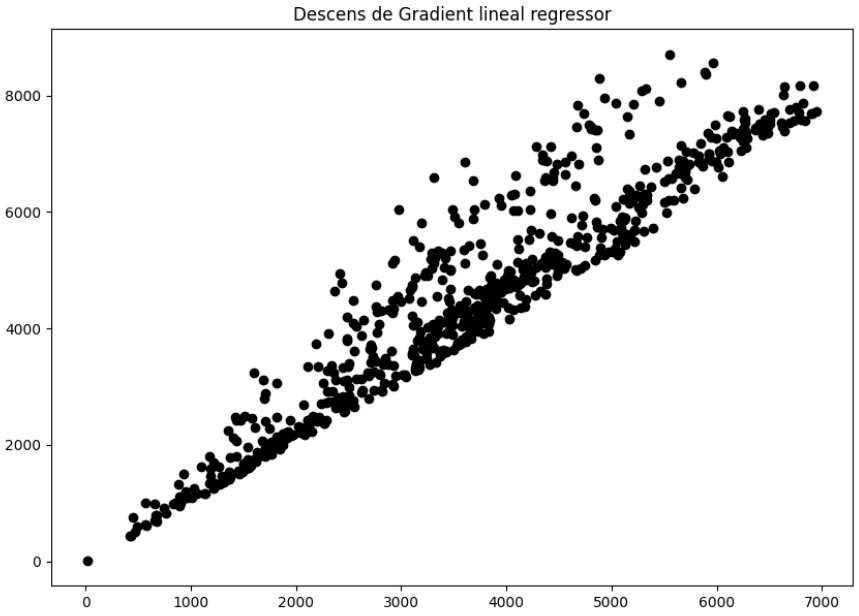
…tenim inclús un **Score més alt!**

Per tant **podem concloure** que el més important per a fer una predicció és tenir clar quins son els atributs amb informació més rellevant, és a dir, els **atributs amb major correlació amb l'atribut objectiu.**

Apartat (A): El descens del gradient

Amb el càlcul del descens del gradient obtenim els valors òptims de x i y que millor s'ajusten a la nostre recta del núvol de punts i una llista amb l'històric de la funció de cost, per poder estudiar com ha anat decreixent l'error.

A continuació, executarem la funció del Descens del Gradient sobre l'atribut "registered", qué com ja hem comentat, és l'atribut amb més correlació amb l'atribut objectiu.



i ara també mostrem un gràfic 3D utilitzant a més l'atribut "casual" que es el segon atribut més rellevant.