Biking Sharing in Washington D.C.

Aprenentatge Computacional - MatCAD

Sofia Di Capua - NIU:1603685, Marc Bosom Saperas - NIU: 1606776

10 d'octubre de 2022

1 Introducció

El dataset que ens ha tocat l'hem descarregat aquí. Aquest dataset recull les dades del nombre de bicicletes llogades pels nous sistemes de *bike sharing* a Washington D.C. durant els anys 2011 i 2012. També s'hi recullen altres tipus de dades (com per exemple si el dia era festa o laboral, climàtiques, ...) que podrien arribar a afectar a l'hora de decidir si s'agafa o no la bicicleta.

Els sistemes *bike sharing* són un nou procés de lloguer de bicicletes on el procés de fer-se soci, agafar la bicicleta i tornar-la es fa quasi automàtic. A través d'aquests sistemes qualsevol és capaç de fer servir una bicicleta fàcilment des de diferents localitzacions i tornar-la en alguna altra. En el 2011 ja hi havia al voltant de 500 programes de *bike sharing* al voltant de món, amb milers de bicicletes a la disposició de tothom. A més, avui en dia aquests tipus de programes són de gran importancia pel medi ambient i la reducció de l'ús de vehicles de combustió.

L'anàlisi del dataset s'ha fet amb el llenguatge de programació `Python` des de `Jupyter Notebook`. La llibreria que s'ha fet servir majoritàriament és `SKLearn` i `Pandas`. Aquestes dues llibreries et permeten fer les construccions dels datasets i t'aporten els mètodes necessaris per a fer els estudis de regressió corresponents.

2 Estructura del dataset

El dataset forma part del grup de datasets de Kaggle. En l'enllaç que ens va tocar podem trobar tres documents diferents. Primer trobem el *Readme.txt*, d'on podrem llegir la informació detallada de la informació i de com estan organitzades les dades. Després trobem dos fitxers diferents que contenen informació: *day.csv* i *hour.csv*. La diferencia és la manera en la que es registren les dades, les quals són diariament i per hores, respectivament. En el nostre cas hem decidit descarregar-nos i estudiar el dataset *day.csv*, ja que ens va semblar més interessant fer l'estudi diariament.

El fitxer anterior té un total de 17 columnes diferents, cadascuna d'elles amb 731 files (que corresponen amb el nombre de dies que hi ha en dos anys). Aquestes columnes corresponen als atributs del dataset, i són els següents:

- INSTANT: un valor int per l'índex de la fila.
- DTEDAY: un string amb la data del dia.
- SEASON: un valor int que fa referencia a cada estació de l'any. (1: Primavera, 2: Estiu, 3: Tardor, 4: Hivern).
- YR: un valor int que fa referencia a l'any. (0: 2011, 1: 2012)

- MNTH: un valor int que va des del 0 fins l'11 que fa referencia al mes de l'any corresponent.
- HOLIDAY: un valor boleà per indicar si el aquell dia era festiu sense incloure caps de setmana (0 no ho era, 1 si que ho era).
- WEEKDAY: un valor int entre el 0 i el 6 per indicar el dia de la setmana.
- WORKINGDAY: un valor boleà per indicar si el dia era laborable (0 no ho era, 1 si que ho era).
- WEATHERSIT: un valor int entre l'1 i el 4 per indicar diferents tipus de clima. (1: cel clar, 2: cel ennuvolat, 3: pluja lleu, 4: tempesta)
- TEMP: un valor float que indica la temperatura.
- ATEMP: un valor float que indica la sensació de la temperatura.
- HUM: un valor float que indica el grau d'humitat.
- WINDSPEED: un valor float que indica la velocitat del vent.
- CASUAL: un valor int que compta el nombre de bicicletes llogades de manera casual aquell dia.
- REGISTERED: un valor int que compta el nombre de bicicletes llogades per persones registrades al sistema aquell dia.
- CNT: un valor int amb el total de bicicletes llogades aquell dia (casual + registered) .

Totes aquestes variables construeixen una taula on es recull tota la informació.

	instant	dteday	season	yr	mnth	holiday	weekday	workingday
0	1	2011-01-01	1	0	1	0	6	0
1	2	2011-01-02	1	0	1	0	0	0
2	3	2011-01-03	1	0	1	0	1	1
3	4	2011-01-04	1	0	1	0	2	1
4	5	2011-01-05	1	0	1	0	3	1
..	...	...	...	..	...	...	...	...
726	727	2012-12-27	1	1	12	0	4	1
727	728	2012-12-28	1	1	12	0	5	1
728	729	2012-12-29	1	1	12	0	6	0
729	730	2012-12-30	1	1	12	0	0	0
730	731	2012-12-31	1	1	12	0	1	1

	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	2	0.344	0.364	0.806	0.160	331	654	985
1	2	0.363	0.354	0.696	0.249	131	670	801
2	1	0.196	0.189	0.437	0.248	120	1229	1349
3	1	0.200	0.212	0.590	0.160	108	1454	1562
4	1	0.227	0.229	0.437	0.187	82	1518	1600
..	...	...	...	...	...	...	...	...
726	2	0.254	0.227	0.653	0.350	247	1867	2114
727	2	0.253	0.255	0.590	0.155	644	2451	3095
728	2	0.253	0.242	0.753	0.124	159	1182	1341
729	1	0.256	0.232	0.483	0.351	364	1432	1796
730	2	0.216	0.223	0.578	0.155	439	2290	2729

[731 rows x 16 columns]

Figura 1: Taula del dataset inicial

3 Objectiu i adaptació del dataset

El nostre objectiu és estudiar la relació entre els dies laborables i el nombre de bicicletes llogades. Nosaltres creiem que durant els dies de festa la CNT serà més gran que els dies que no. També ens agradaria analitzar el valor de la CNT en funció del clima.

Un cop definit l'objectiu i les nostres hipòtesis, trobem necessari fer una reducció del nostre dataset perquè tenim moltes columnes amb informació irrellevant per la nostra investigació. A continuació justificarem perquè hem eliminat les columnes següents:

- DTEDAY: és una variable string, la qual cosa ens dificulta la nostra investigació. Nosaltres només necessitem saber si el dia és laborable o si no ho és, per tant, ens és irrellevant.
- SEASON: l'informació que ens podria aportar aquesta variable son les condicions meteorològiques que volem estudiar. Però altres variables ja ens donen aquesta informació de forma concreta, per tant, no ens és útil.
- YR, MNTH, WEEKDAY: com que només volem estudiar si el dia és laborable o no ho és, no necessitem saber ni l'any, ni el mes, ni el dia de la setmana en el que ens trobem.
- HOLIDAY: de totes les dades que hi ha en aquesta columna, només el 2.87% és true. Per tant, per falta d'informació, hem decidit prescindir d'aquesta variable i centrar-nos únicament en workingday per determinar si el dia és laborable o no ho és.
- CASUAL, REGISTERED: ens hem centrat en el nombre total de bicicletes directament.

Un cop arribats a aquest punt, tenim la informació que necessitem per a dur a terme els nostres anàlisis de les dades.

Si ens fixem en les dades de la columna WEATHERSIT, es tracten d'uns valors arbitraris entre l'1 i el 4 que indiquen diferents tipus de clima. Aquesta assignació de números és aleatòria, així que no té cap significat més enllà de ser una etiqueta. Això fa que WEATHERSIT sigui un atribut categòric nominal. Cercar alguna relació amb aquesta variable seria assumir que la seva ordenació existeix, i això seria considerar una informació falsa. És per això que hem decidit formar 4 variables boleanes diferents a partir d'aquesta. Aquest procès s'anomena "one-hot encoding", i aquestes noves variables són: CLEAR, MISTCLOUDY, LIGHTRAIN i ICEPALLETS.

Així ens queden variables boleanes, les quals estan formades per valors binaris (0 i 1). Per tant em de mirar si els valors estan ben distribuïts, ja que si la majoria de valors són iguals aquesta columna no ens aportaria cap informació rellevant pel nostre anàlisis. Hem calculat el percentatge de valors *true* i hem obtingut els següents percentatges:

Variable	Percentatge (%)
WORKINGDAY	68.39945280437757
CLEAR	63.33789329685362
MIST&CLOUDY	33.78932968536252
LIGHTRAIN	2.8727770177838577
ICEPALLETS	0.0

Les dues últimes variables tenen un percentatge nul o quasi nul de valors *true*. És per això que em decidit eliminar les variables LIGHTRAIN i ICEPALLETS.

Aplicant els canvis que hem justificat abans, el dataset que tenim queda de la següent manera:

	workingday	temp	atemp	hum	windspeed	Clear	Mist&Cloudy	cnt2
0	0	0.344	0.364	0.806	0.160	0	1	985
1	0	0.363	0.354	0.696	0.249	0	1	801
2	1	0.196	0.189	0.437	0.248	1	0	1349
3	1	0.200	0.212	0.590	0.160	1	0	1562
4	1	0.227	0.229	0.437	0.187	1	0	1600
..	...	...	...	...	...	...	...	...
726	1	0.254	0.227	0.653	0.350	0	1	2114
727	1	0.253	0.255	0.590	0.155	0	1	3095
728	0	0.253	0.242	0.753	0.124	0	1	1341
729	0	0.256	0.232	0.483	0.351	1	0	1796
730	1	0.216	0.223	0.578	0.155	0	1	2729

[731 rows x 8 columns]

Figura 2: Taula del dataset després de ser modificat

Cal recalcar que el nom de la variable CNT s'ha canviat per CNT2. Això ho hem fet per tenir la columna corresponent al càlcul total de bicicletes llogades a l'última columna de l'esquerra i així facilitar-nos la feina a l'hora d'accedir-hi. Tot i així, al llarg de la memoria ens referirem a aquesta variable com a CNT.

4 Anàlisi de les dades

És molt útil saber com es distribueixen les dades per moltes raons. Mentre analitzem el dataset, saber la distribució de les dades ens pot ajudar a trobar valors extrems (outliers), dades que falten o valors que s'han introduit incorrectament (33.3 en comptes de 333, per exemple). També és beneficios per poder formar-nos hipòtesis sobre la correlació de les dades.

Abans de calcular la correlació que hi ha entre les variables del nostre dataset, farem un anàlisi de les dades per a tenir una idea de com es distribueixen i, així, entendre-les millor.

Fent servir la comanda `describe()` podem veure la següent taula:

	workingday	temp	atemp	hum	windspeed	Clear	Mist&Cloudy	cnt2
count	731.000	731.000	731.000	731.000	731.000	731.000	731.000	731.000
mean	0.684	0.495	0.474	0.628	0.190	0.633	0.338	4504.349
std	0.465	0.183	0.163	0.142	0.077	0.482	0.473	1937.211
min	0.000	0.059	0.079	0.000	0.022	0.000	0.000	22.000
25%	0.000	0.337	0.338	0.520	0.135	0.000	0.000	3152.000
50%	1.000	0.498	0.487	0.627	0.181	1.000	0.000	4548.000
75%	1.000	0.655	0.609	0.730	0.233	1.000	1.000	5956.000
max	1.000	0.862	0.841	0.973	0.507	1.000	1.000	8714.000

Figura 3: Resultat després d'executar la comanda describe()

Aquesta taula ens permet veure quina és la mitjana de cada valor, el seu error estàndard (std) i quin és el seu valor màxim i mínim.

Després d'això hem contruït diferents histogrames per veure com de distribuïdes estan les dades del dataset actual. Els histogrames són els següents:

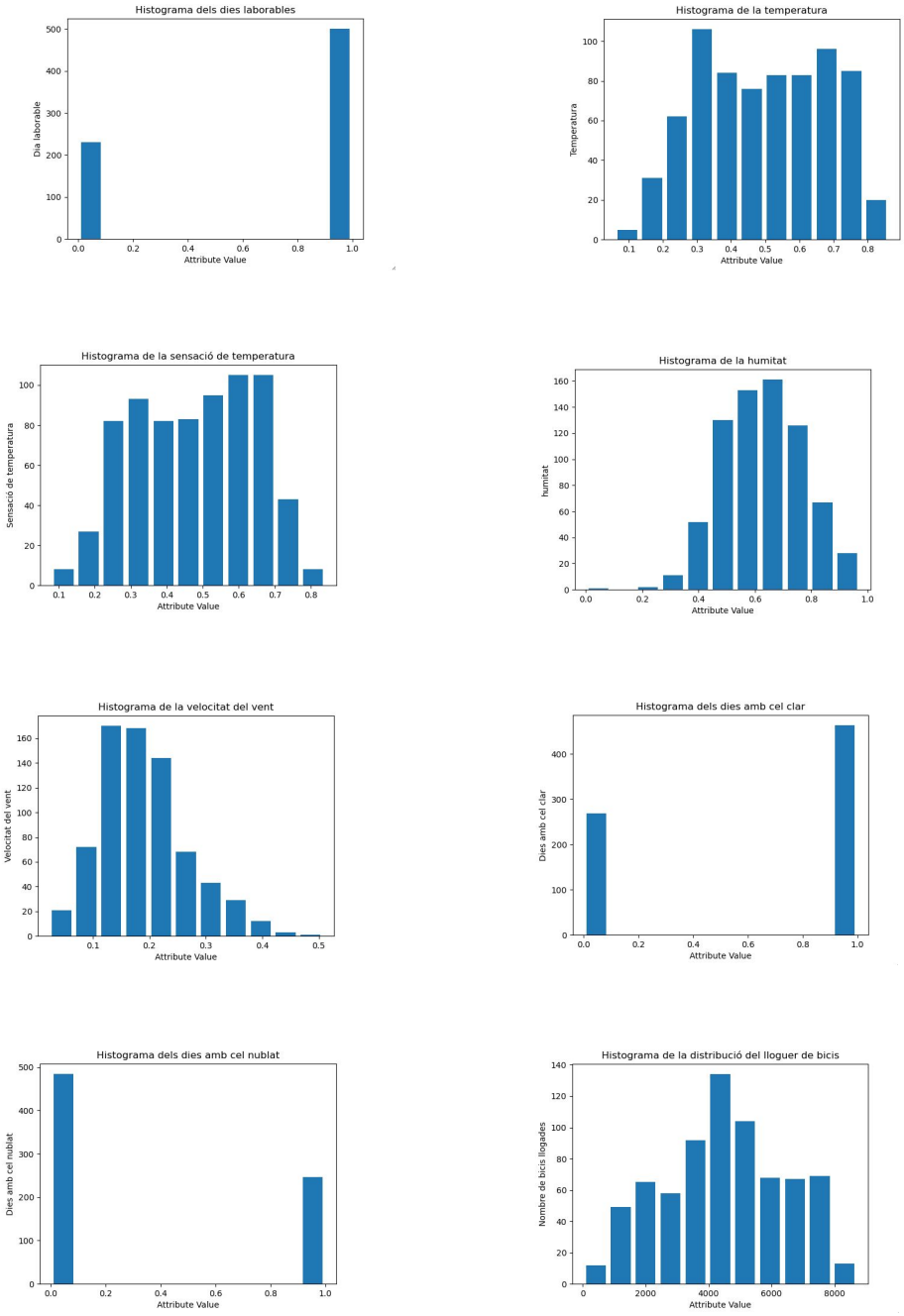


Figura 7: Histogrames del conjunt de dades del nostre dataset

Aquests histogrames ens han ajudat a veure on estan els valors màxims i mínims de les columnes que formen el nostre dataset i les formes que tenen. Queda clar que en el dataset hi ha més dies laborables que no laborables, i que TEMP i ATEMP tenen dades molt similars. Això és d'esperar, doncs la sensació de temperatura que fa ha de ser similar a la temperatura de l'entern. Dels gràfics que indiquen els dies que hi havia cel clar i dels dies que hi havia núvols podríem arribar a considerar que el clima durant aquests dos anys era més aviat bo que no pas dolent. I dels tres gràfics que queden que podrien tractar-se d'una distribució normal. Per poder assegurar-nos-en d'això, caldrà realitzar alguns testos per estar-ne segurs.

Per mirar quines de les dades no boleanes tenen una distribució normal hem realitzat dos testos diferents: test de normalitat de Shapiro-Wilk i el test de normalitat d'Agostino.

En estadística, la prova de Shapiro-wilk s'utilitza per a contrastar un conjunt de dades. Es planteja com a hipòtesis nul·la (H_0) que una mostra $x_1, ..., x_n$ provingui d'una població normalment distribuïda. L'estadístic de la prova és:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

on:

- $x_{(i)}$ és el nombre que ocupa la i-éssima posició de la mostra.
- $\overline{x}$ és la mitjana mostral.

La hipòtesis nul·la es rebutjarà si W és propera a 0, doncs serà un valor que oscil·larà entre 0 i 1.

En canvi el test d'Agostino la seva manera per derminar si la distribució és normal o no és la següent:

$$D_{exp} = \frac{\sum_{i=1}^{n} i x_i - \frac{(n+1)\sum_{i=1}^{n} x_i}{2}}{\sqrt[n]{n\left(\sum_{i=1}^{n} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)}}$$

Els resultats els recollim en les taules següents:

Dades	Resultats	Conclusió
temp	stat=0.966, p=0.000	No Gausiana
atemp	stat=0.974, p=0.000	No Gausiana
hum	stat=0.993, p=0.002	No Gausiana
windspeed	stat=0.971, p=0.000	No Gausiana
cnt	stat=0.980, p=0.000	No Gausiana

Taula 1: Resultats després d'aplicar el test de normalitat Shapiro-Wilk

Dades	Resultats	Conclusió
temp	stat=294.297, p=0.000	No Gausiana
atemp	stat=144.352, p=0.000	No Gausiana
hum	stat=0.683, p=0.711	Gausiana
windspeed	stat=51.425, p=0.000	No Gausiana
cnt	stat=62.708, p=0.000	No Gausiana

Taula 2: Resultats després d'aplicar el test de normalitat d'Agostino

El valor de *stat* és el resultat de l'hipotesis nul·la (H_0), mentre que p és el p-valor. Els resultats dels testos només indiquen quin és el resultat més segur, per tant són una orientació. Tot i així, en el primer test no hi ha cap variable que sigui gaussiana, mentre que en el segon l'única que manté una distribució normal és l'humitat.

5 Estudi de la correlació entre les nostres dades

Un cop estudiades les dades individualment, hem continuat estudiant la relació que aquestes dades tenen les unes amb les altres. Seguint el nostre objectiu principal, vam començar fent diferents gràfics de dispersió comparant el nombre total de bicicletes llogades (la nostra variable objectiu) amb la resta de valors que tenim. Els resultats són els següents:

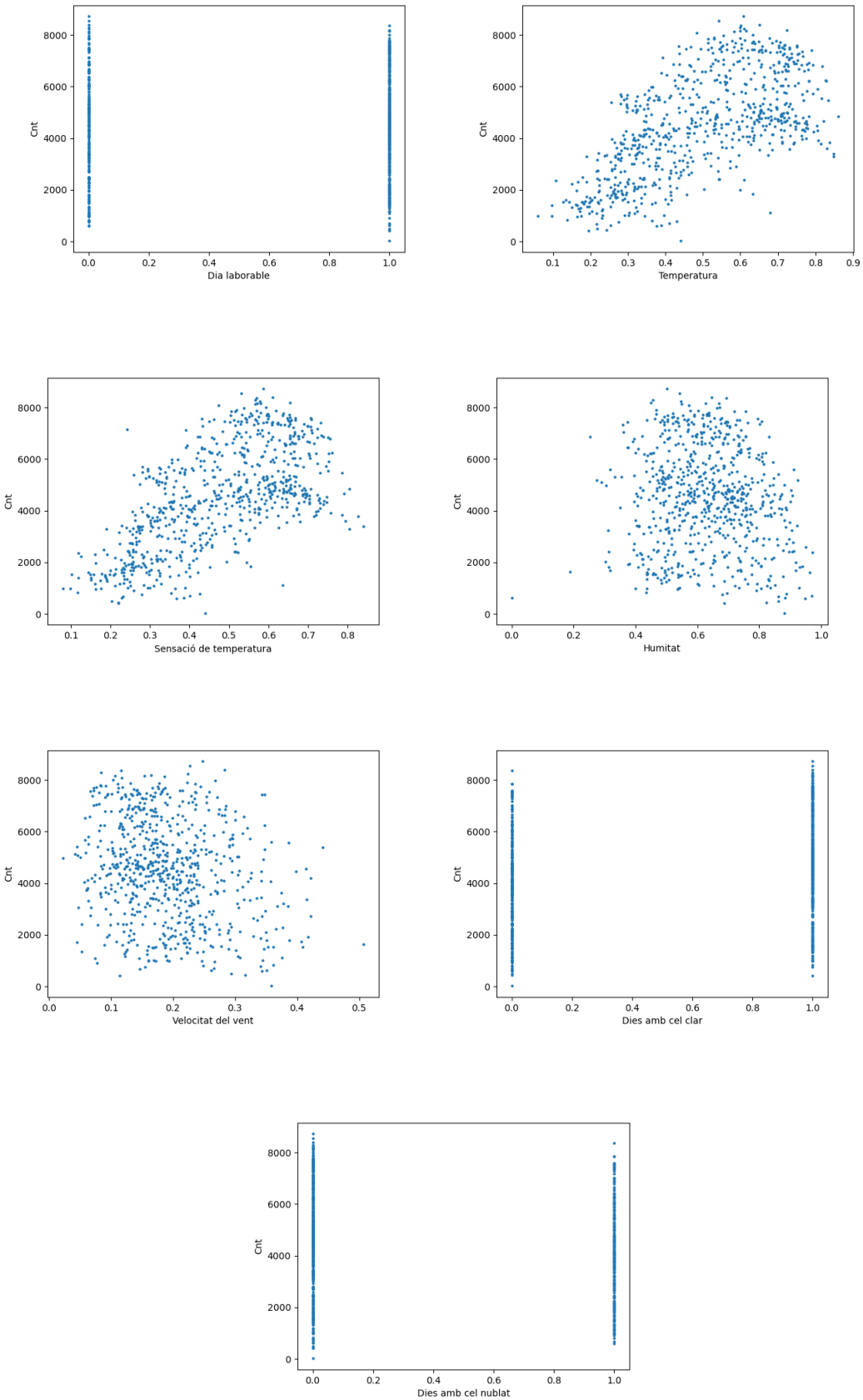


Figura 11: Gràfics de dispersió del total de dades del nostre dataset

Els gràfics amb valors boleans s'identifiquen fàcilment perquè es tracta només de dues columnes sobre el 0 i l'1. Dels altres gràfics de dispersió podem veure si els punts segueixen una

tendència o no. Si ens fixem en la sensació de temperatura i la temperatura podem veure com els punts tenen tendència a augmentar juntament amb el valor de la CNT. En canvi l'humitat i la velocitat del vent tenen tots els punts concentrats a un únic punt i no mostren cap variació.

Per veure com es relacionen les dades entre si (i comprobar si es relacionen linealment) fem servir la comanda `pairplot()` de la llibreria `seaborn`. Aquesta comanda mostra la relació per (n,2) combinacions de variables en forma d'una matriu de gràfics. Els gràfics de la diagonal univariats, on es comparen les variables amb elles mateixes. És d'allà d'on podem veure si segueixen una distribució normal. El resultat del gràfic és el següent:

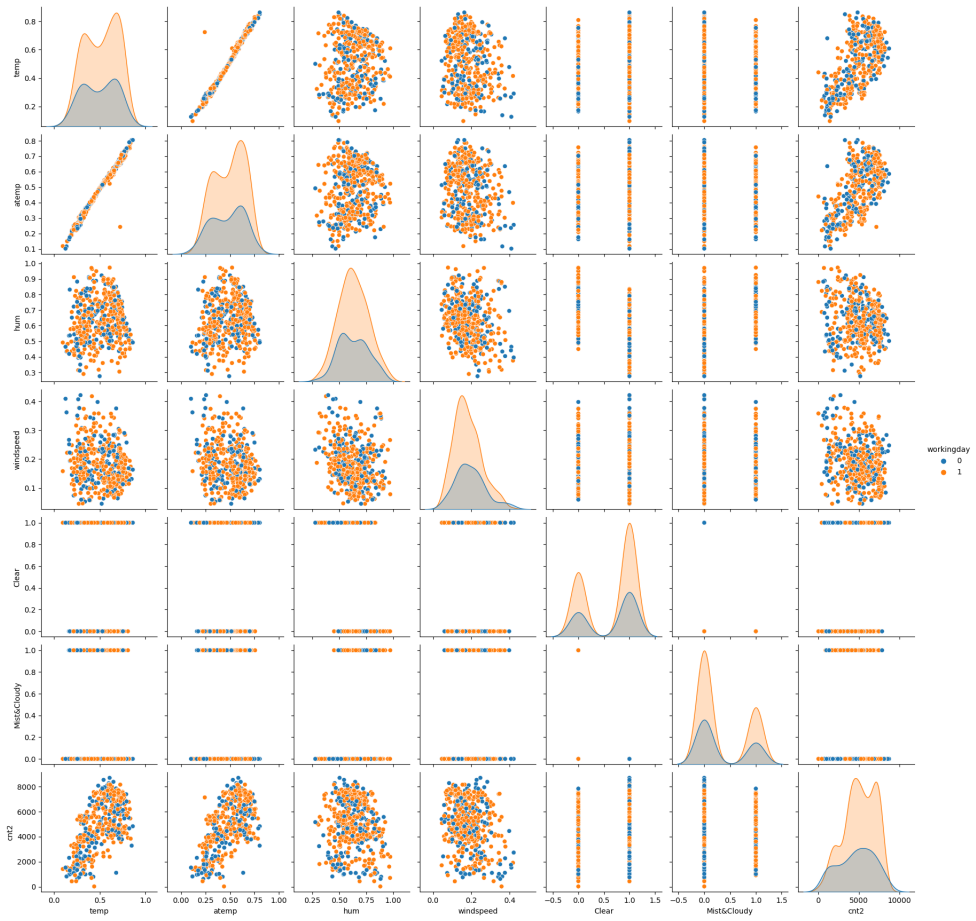


Figura 12: Resultat d'executar la comanda `pairplot()`

Després d'executar aquest gràfic tenim gaire clara la relació entre les variables. Podem veure com ATEMP i temp tenen una relació lineal molt clara, i també podem descartar les variables binàries perquè no ens poden aportar cap tipus d'informació.

De la resta de gràfiques no es pot treure informació concreta ja que són simplement núvols de punts. Hem pensat que aquest problema pot ser degut a la diferencia de dies laborables (500) amb dies no laborables (231). Per tant, hem creat un dataset puntual on el nombre de dies laborables i no laborables siguin el mateix amb l'intenció de veure amb més claredat aquesta matriu de gràfiques. Els dies laborables que excloem estan seleccionats aleatòriament. Després de fer aquesta reducció hem obtingut el següent:

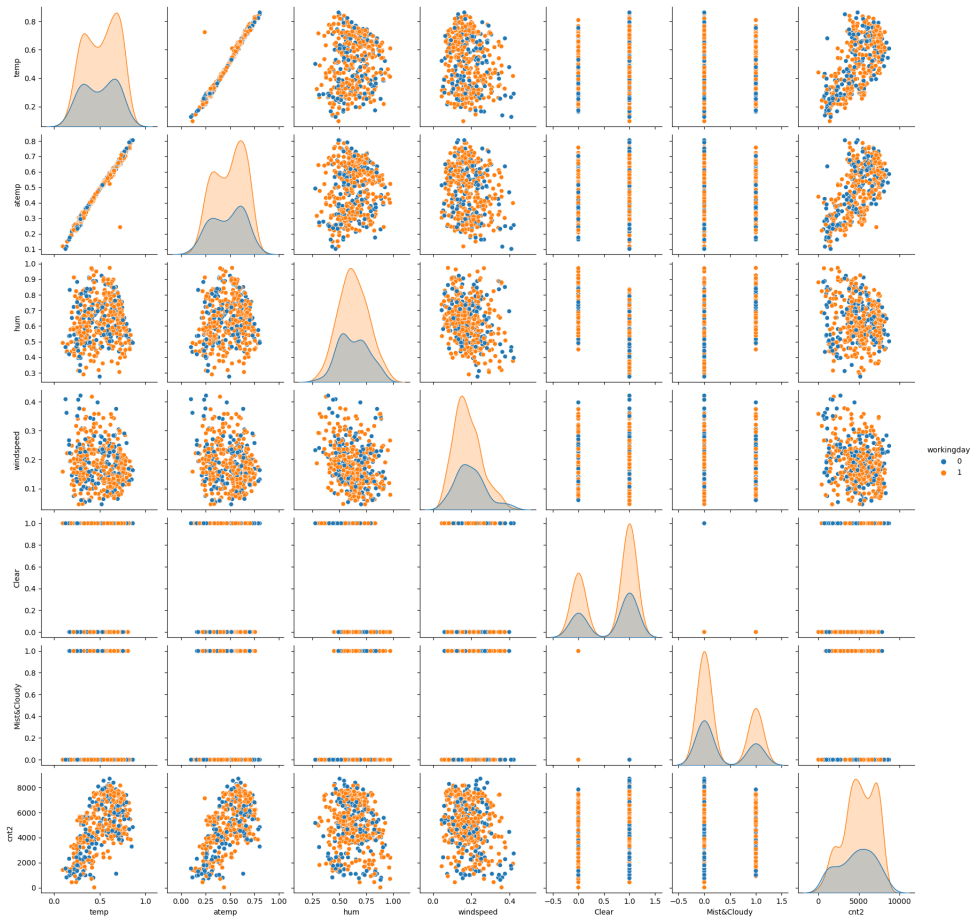


Figura 13: Resultat d'executar la comanda `pairplot()` reduïnt el nombre de dies laborables

Tot i haver reduït el nombre de dies laborables, tampoc som capaços de veure-hi la relació entre les dades, així que necessitem buscar una altra estratègia per poder estudiar si les dades estan o no relacionades.

Calcular la correlació és una manera de veure com de relacionades estan dues columnes de forma numerica. El resultat d'aquest analísis és un valor entre -1 i 1. El signe inidica el tipus de relació; si és positiu indica que existeix una relació positiva (quan un creix l'altre també), si és negatiu indica el contrari (si un creix l'altre decreix). Si dues variables són independents la seva correlació serà zero. La força de la relació lineal incrementa a mesura que la correlació s'aproxima a -1 o +1.

Per calcular la correlació entre les dades hem fet servir la comanda `heatmap()` de la llibreria `seaborn`. El resultat és un mapa de calor feta amb la correlació de les dades :

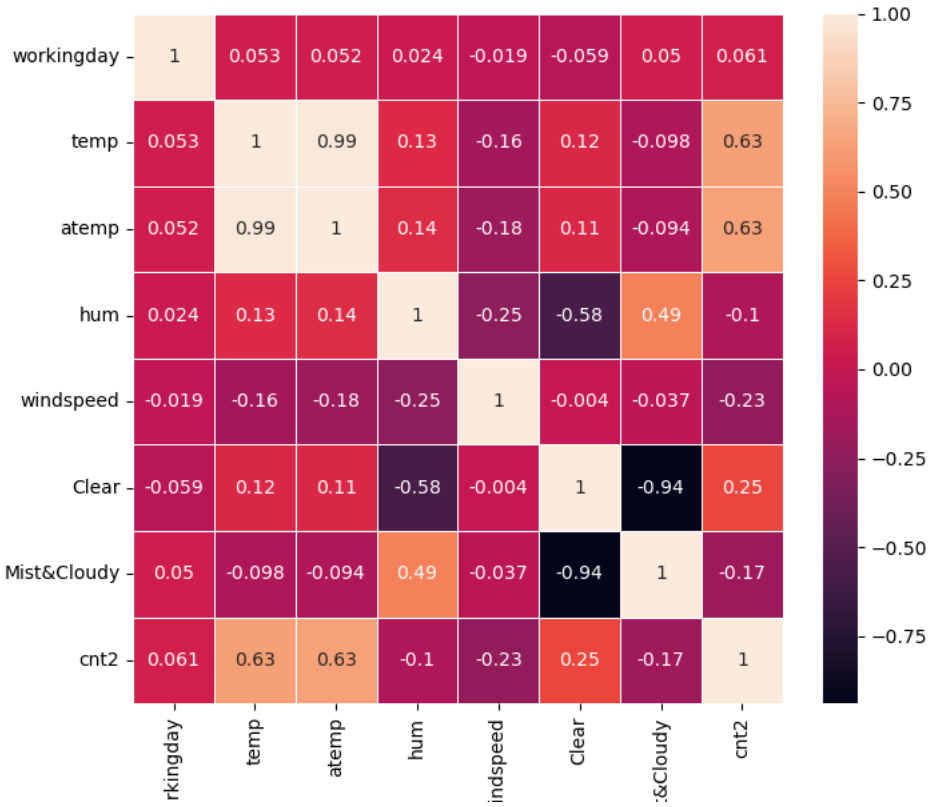


Figura 14: Mapa de calor de les correlacions

A partir d'aquesta taula podem estudiar quins son els atributs que tenen més correlació entre ells. Podem destacar les següents relacions:

- ATEMP i TEMP estan altament relacionades entre si de forma positiva. La qual cosa té sentit, perquè com més elevada és la temperatura, major és la sensació de calor.
- CLEAR i MIST&CLOUDY tenen una correlació molt elevada i negativa. La qual cosa té sentit, doncs no pot ser que el cel estigui clar i nublat alhora. Com més clar estigui el cel, menys núvols i boira hi haurà.
- CNT amb TEMP i ATEMP. Tenen una correlació proxima a 1 i seria interessant estudiar-la amb més detall.
- HUM amb MISTCLOUDY i CLEAR. Amb la primera variable hi ha una correlació elevada i positiva mentres que amb l'altre s'aproxima a -1. Pensant-ho amb claredat sembla lògic, si hi ha boira l'humitat aumentarà però si hi ha cels clars disminuirà.

A partir d'aquest mapa de calor també podem veure com les variables WORKINGDAY i WINDSPEED tenen una correlació propera a cero així que podem concloure que no són condicions que determinin la decisió d'agafar o no la bicicleta.

Hem realitzat també un test de correlació de Pearson en les nostres dades per a comprovar si obtenim els mateixos resultats. Aquest test ens indicarà si les diferents dades que tenim estan o no relacionades amb la nostra CNT. El test de correlació de Pearson és una mesura de dependència lineal entre dos variables aleatòries quantitatives, i és independent de l'escala de la mesura de les variables. Donades dues variables aleatories (X, Y), el coeficient de correlació

de Pearson es defineix com:

$$\rho_{(X,Y)} = \frac{\sigma_{(X,Y)}}{\sigma_X \sigma_Y} = \frac{Cov(X,Y)}{\sqrt{Var(x)Var(Y)}}$$

On

- $\sigma_{(X,Y)}$ és la covariança de (X,Y).
- σ_X és la desviació estàndard de la variable X.
- σ_Y és la desviació estàndard de la variable Y.

El coeficient de la correlació de Pearson és un valor entre -1 i 1. Depent del valor extreiem una conclusió o una altra.

- Si r=1, aleshores vol dir que existeix una correlació positiva perfecta.
- Si 0 < r < 1, aleshores existeix una correlació positiva.
- Si r = 0, aleshores no existeix cap tipus de correlació entre les dades.
- Si -1 < r < 0, aleshores existeix una correlació negativa.
- Si r = -1, aleshores existeix una correlació negativa perfecta.

Recollim els resultats en la taula següent:

Dades	Resultats	Conclusió
TEMP	stat=0.627, p=0.000	Dependent
ATEMP	stat=0.631, p=0.000	Dependent
HUM	stat=-0.101, p=0.006	Dependent
WINDSPEED	stat=-0.235, p=0.000	Dependent

Taula 3: Resultats després d'aplicar el test de correlació de Pearson

A partir d'aquest test hem vist que totes les dades que són no binaries del nostre dataset tenen alguna dependencia amb la CNT. Després de comparar les correlacions que hem calculat amb dels del HEATMAP hem descobert que la llibreria `SKLearn` fa servir la correlació de Pearson per a calcular la correlació. Per tant hem aprofitat per calcular la correlació a partir de la llibreria i manualment.

És doncs a partir d'aquest moment on hem replantejat el nostre objectiu principal. Tot i que en el test de correlació de Pearson totes les variables siguin dependents, el valor de la correlació entre WORKINGDAY i la CNT és de 0.061, el qual és molt proper a 0. En canvi la temperatura i la sensació de temperatura són variables que tenen una correlació de 0.63 amb CNT. Per tant, el nostre objectiu ara mateix és estudiar la relació entre la temperatura exterior i el nombre de bicicletes que es lloguen, i per això construirem un model de regressió lineal.

6 Regressions del nostre model

Per predir una variable correctament cal entendre com es comporta la nostre variable objectiu, és a dir, com es distribueix. Distribucions hi ha moltes i molt diferents, per aquest motiu hem analizat tant les nostres dades en els anteriors apartats per poder fer una bona regressió.

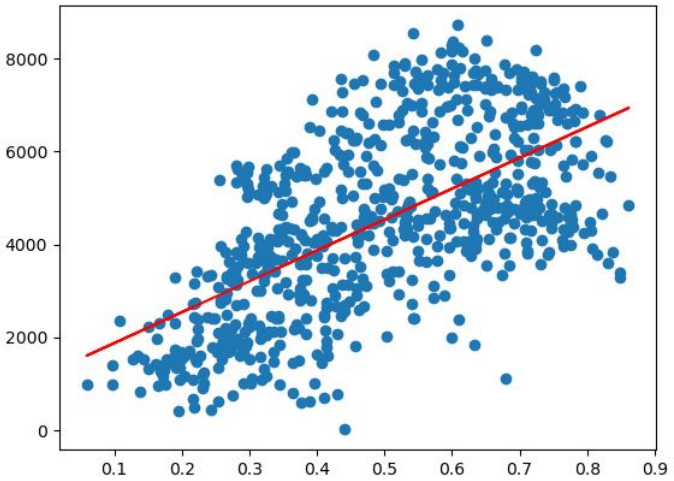
Quan fem una regressió lineal estem buscant els valors β_0 i β_1 que compleixin l'equació:

$$Y = \beta_0 + \beta_1 X$$

Amb l'objetiu de predir el valor de la variable Y a partir de totes les observacions X. Tot i així, mai s'espera que les Y predites siguin exactament iguals a les Y donades. No perquè el model sigui incorrecte, sinó perquè no es té en compte l'error. Degut a aquests errors els valors de Y donats uns valors de X fluctuaran al voltant del valor mitjà de la mostra. Aquests errors s'anomenen errors residuals, són la diferència entre el valor observat i el valor estimat en una mateixa mostra, i idealment valdrien zero.

En un model lineal s'assumeix que els errors residuals segueixin una distribució normal, ja que, si ho són, significa que pots confiar en els resultats. Mirant-ho amb més detall, si els errors residuals segueixen una normal: la probabilitat de trobar outliers és més baixa, es distribuiran els valors de les mostres de forma més o menys equitativa al voltant del valor mitjà i també allunyats d'aquest punt. Això provoca no tenir totes les dades concentrades en el centre i equival a més informació per a la pendent de la regressió (que s'extreu dels valors més extrems, que si els errors segueixen una normal, tenen menys error residual i, per tant, són més fiables). Aquest fet fa que busquem que els errors reisduals de la nostra regressió segueixin una distribució normal, ja que és més fàcil fer una predicció i amb més precisió.

Basant-nos en els valors de la correlació que hem obtingut en el `heatmap`, hem realitzat una regressió lineal entre la sensació de temperatura i la CNT, aquesta última com la variable que volem predir. La regressió té aquesta forma:

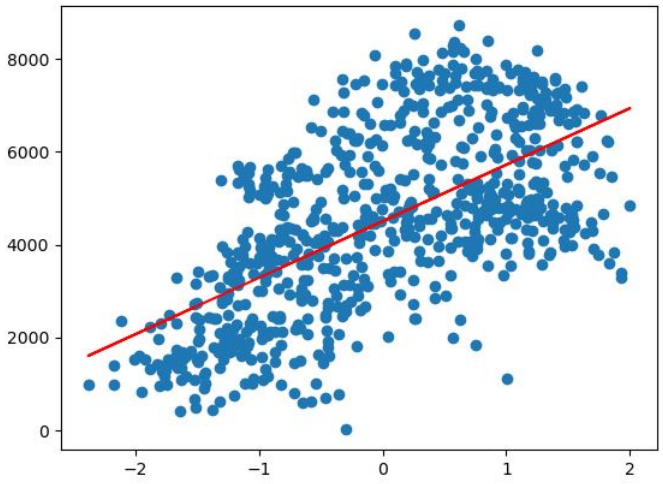



```
Mean squeared error:  2272020.2556456523
R2 score:  0.3937487313729242
```

Figura 15: Regresió Lineal amb el dataset original

L'R^2 és un valor estadístic que indica la proporció de la variança d'una variable depenen que s'explica per una o diverses variables independents d'un model de regressió. És un valor que sempre es troba entre 0 i 1, com més proper sigui a 1 millor estarà construit el model. Si ens fixem amb el valor de l'MSE podem veure que és del valor de $2 \cdot 10^6$. això s'explica observant l'eix de les y, ja que es divideix en intervals de 2000 unitats. Hi ha punts doncs que es troben a més de 2000 unitats de distancia respecte la recta de regressió vermella. Aquest valor elevat al quadrat ens dona nombres del valor de 10^6, i fent el sumatori de les diferencies, amb la regressió que tenim és normal obtenir un MSE tan gran.

Un cop obtenida aquesta regressió hem provat a calcular-ne una altra estandaritzant el valor de la sensació de la temperatura, i la regressió que hem obtingut es mostra en el gràfic següent:



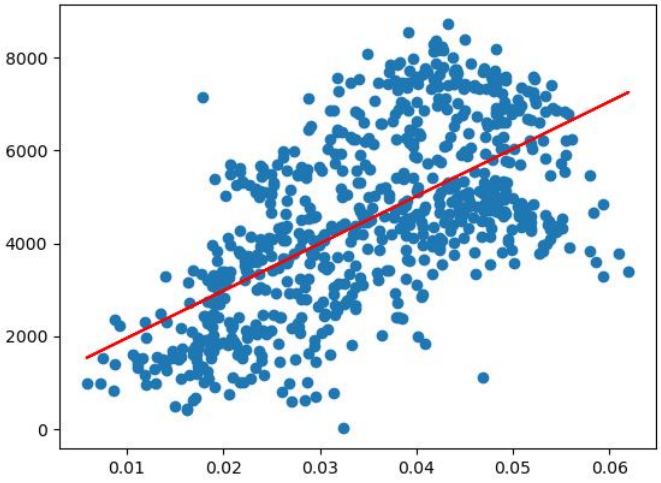

```
Mean squeared error:  2272020.2556456523
R2 score:  0.3937487313729242
```

Figura 16: Regresió Lineal amb la X estandaritzada

Estandaritzar les dades consisteix en escalar les dades a un rang més petit amb l'objectiu de reduïr el valor de l'MSE. Però les dades estandaritzades són les del valor de la x, ja que el valors de la y no s'han de modificar. Això últim ho expliquem més endavant.

Finalment hem volgut normalitzar les dades de l'eix de les x. D'aquesta manera reduïm els valors de la sensació de temperatura a un rang entre 0 i 1. La normalització és un procés reversible, així que sempre es poden recuperar les dades originals invertint la transformació, i és molt útil tant per a la visualització de les dades com per l'entrenament d'algorismes i en les regressions lineals. Però és molt important tenir en compte que no s'ha de normalitzar mai les variables que es volen predir, ja que alteraria totalment la solució del problema que volem estudiar i, per tant, no té cap sentit normalitzar la y.

El resultat de la regressió amb les dades normalitzades és el següent:




```
Mean squeared error:  2255173.8513220353
R2 score:  0.39824391752613697
```

Figura 17: Regresió Lineal amb la X normalitzada

En totes tres regressions podem veure com, tot i que la distribució de les dades sí que ha variat en totes tres (només l'eix de les x), el valor de R^2 i l'MSE és el mateix i no varia. Això és degut al que s'ha explicat anteriorment dels valors de l'eix de les y.

Arribats a aquest punt ens preguntem què passaria si calculem els valors de R^2 i de l'MSE a partir d'una regressió polinòmica de diferents graus. Els resultats que hem obtingut són els següents:

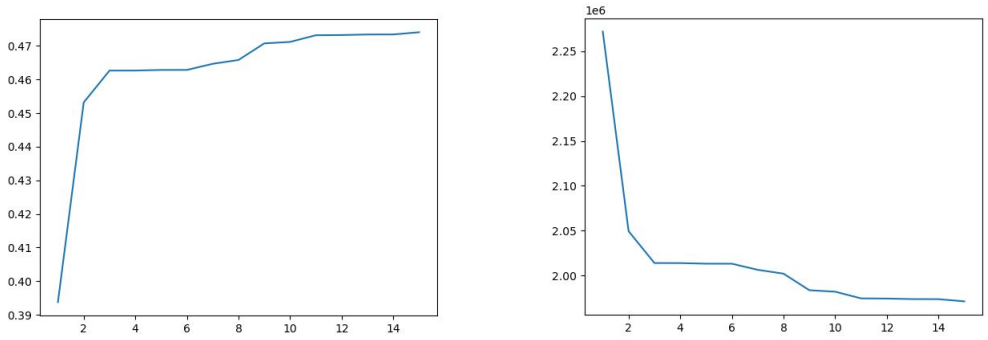


Figura 18: Evolució de l'R^2 i de l'MSE, respectivament, en funció del grau del polinomi

En aquest gràfic veiem com, augmentant el valor del grau del polinomi amb el que fem la regressió, el valor de l'R^2 augmenta mentre que l'MSE disminueix. Això s'explica dient que un polinomi de grau més alt permet ajustar-se millor a la distribució de les dades. És important fixar-se que els valors de l'eix de les y del gràfic de l'MSE són del valor de 10^6.

Per comprovar que la nostra regressió és correcte, hem comprovat la distribució dels nostres errors i de les nostres dades predites.

Després d'aplicar els testos de Shapiro-Wilk i Agostino, podem veure que probablement les nostres dades no segueixen cap distribució gaussiana. Això pot ser un indicador de que la nostra regressió no és fiable, però com tenim uns valors de l'R^2 i de l'MSE a favor, continuem considerant que el nostre model és correcte.

Després de fer les nostres investigacions ens hem adonat que la variable CNT no té cap mena de relació amb els dies laborables, la qual cosa ens ha sorprès i ha provocat que haguem de canviar el nostre objectiu principal. Però si que té relació amb TEMP. Després de realitzar les nostres regresions amb la CNT i TEMP ens hem adonat que ens veiem en la necessitat de conèixer més sobre l'anàlisi de les dades per poder fer un anàlisi més fiable.

7 Preguntes

7.1 Quin és el tipus de cada atribut ?

La resposta a questa pregunta es troba en l'apartat Estructura del dataset. per veure el tipus de cada atribut, cliqueu aquí.

7.2 Quins atributs tenen una distribució Guassiana?

Durant el nostres anàlisis hem realitzat dos testos per determinar si les nostres dades són o no gausianes. Segons el test de Shapiro-Wilk les dades de les que disposem són probablement no Gaussianes, però pel test d'Agostino l'única dada que probablement té una distribució normal és el grau d'humitat. Podem veure els histogrames en la figura 7.

7.3 Quin és l'atribut objectiu? Per què?

El nostre atribut objectiu és la cnt. Aquesta variable és la que recull el recompte total de bicicletes llogades al llarg dels anys 2011 i 2012, que és el que volem predir.

7.4 Quin són els atributs més importants per fer una bona predicció?

Els valors més importants a l'hora de fer la regressió són aquells que tenen una dependencia i una correlació elevada amb la variable que volem predir. Si ens fixem amb el nostre cas, després de fer el test de Pearson totes les variables són dependents, i després del `heatmap` podem veure com la sensació de temperatura i la temperatura són les variables amb la correlació més elevada, i per això són les més bones per a fer una predicció.

7.5 Amb quin atribut s'assoleix un MSE menor?

Després de fer la regressió amb les diferents varaibles del nostre dataset podem omplir la taula següent:

Variable	R^2	**MSE**
WORKINGDAY	0.0037400640490624637	3733637.9674631804
TEMP	0.3937487313729242	2272020.2556456523
ATEMP	0.39824391752613697	2255173.8513220353
HUM	0.010132146131519582	3709682.652697436
WINDSPEED	0.05501135581553118	3541490.883482077
CLEAR	0.06394301215055309	3508018.1220054883
WINDSPEED	0.02988726636435568	3635647.2887412277

A partir d'aquesta taula doncs podem veure com els atributs que assoleixen un MSE més petit són la temperatura i la sensació de temperatura, que són precisament els que hem fet servir per a fer la regressió.

7.6 Quina correlació hi ha entre els atributs de la vostra base de dades?

Les correlacions que hem calculat estan agrupades en la figura 14, que correpon amb el `heatmap` de l'apartat 5.

7.7 Com influeix la normalització en la regressió?

Després de normalitzar les nostres dades no hem pogut veure cap millora ni en l'MSE ni en l'R^2. Si ens fixem amb els gràfics 15, 16 i 17 podem veure com la regressió no canvia, i que els valors de l'MSE i de l'R^2 és mantenen pràcticament consntants

7.8 Com millora la regressió quan es filtren aquells atributs de les mostres que no contenen informació?

Per calcular la regressió nosaltres hem fet servir directament les variables que ens eren útils per a fer-la. Per determinar-les el que hem fet és dur a terme un procés d'anàlisi de les dades per determinar quina d'elles estava més relacionada amb la CNT, que era la nostra variable objectiu. Un cop vam determinar que la variable amb més correlació era la temperatura, vam realitzar totes les regressions que vam trobar necessàries per a concloure la nostra investigació.

7.9 Si s'aplica un PCA, a quants components es redueix l'espai? Per què?

Un anàlisi de components principals o ACP (PCA en àngles) cerca la projecció segons la qual les dades quedin millor representades en termes de mínims quadrats. Converteix un conjunt d'observacions de variables possiblement correlacionades en un conjunt de valors de variables sense correlació lineal anomenades components principals. Per aquest motiu no ens fixarem en la variable CNT2, ja que volem relacionar-la amb la resta de variables.
Aquest anàlisis consisteix en fer una transformació lineal que escull un nou sistema de coordanades pel conjunt original de les dades en el cual la variança més gran de les dades es captura en el primer eix, la segona més gran en el segon eix, i així sucesivament. La transformació lineal descrita redueix la dimensionalitat de les dades.

Fent un PCA obtenim la següent taula amb les components resultants:

	workingday	temp	atemp	hum	windspeed	Clear	Mist&Cloudy
PC1	0.046	-0.142	-0.138	0.455	-0.062	-0.620	0.603
PC2	-0.076	-0.652	-0.657	-0.256	0.262	0.035	-0.042
PC3	-0.880	-0.069	-0.056	0.163	-0.428	0.068	-0.058
PC4	-0.463	0.214	0.192	-0.125	0.801	-0.157	0.145
PC5	-0.053	0.082	0.080	-0.819	-0.315	-0.210	0.411
PC6	-0.010	0.032	-0.004	-0.120	-0.056	-0.735	-0.664
PC7	-0.000	0.704	-0.709	0.010	-0.016	0.020	0.016

Figura 19: Resultats després de fer el PCA

Cada component principal (PCi) s'obté per combinació lineal de les variables originals. Es poden entendre com a noves variables obtingudes en combinar d'una determinada forma les variables originals. La primera component principal del nostre grup de variables (WORKINGDAY, TEMP, ..., MISTCLOUDY) és la combinació lineal normalitzada d'aquestes variables que té major variància:

$$PC1 = \phi_{11}\text{WORKINGDAY} + *\phi_{21}\text{TEMP} + ... + \phi_{p1}\text{MISTCLOUDY}$$

Els termes $\phi_{11}, \ldots, \phi_{p1}$ reben en el nom de loadings i són els que defineixen a la component. ϕ_{11} és el loading de la variable WORKINGDAY de la primera component principal. Els loadings poden interpretar-se com el pes/importància que té cada variable en cada component i, per

tant, ajuden a conèixer que tipus d'informació recull cadascuna de les components.

Una vegada calculada la primera component (PC1) es calcula la segona (PC2) repetint el mateix procés, però afegint la condició que la combinació lineal no pot estar correlacionada amb la primera component. Això equival a dir que PC1 i PC2 han de ser perpendiculars. EL Procés es repeteix de manera iterativa fins a calcular totes les possibles components o fins que es decideixi detenir el procés.

Analitzar amb detall el vector que forma cada component (és a dir, cada fila) pot ajudar a interpretar quin tipus d'informació recull cadascuna de les variables. La primera component és el resultat de la següent combinació lineal de les variables originals:

$$PC1 = -0.024 * \text{WORKINGDAY} - 0.498 * \text{TEMP} - 0.499\text{ATEMP} + 0.144 * \text{HUM} + 0.135 * \text{WINDSPEED}$$

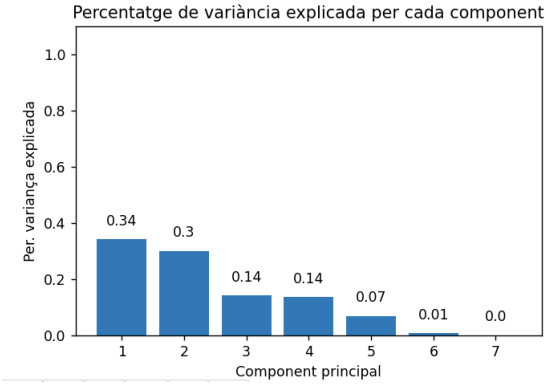
$$-0.36 * \text{CLEAR} + 0.333 * \text{MISTCLOUDY} - 0.472 * \text{CNT2}$$

Per interpretar cada component principal, s'ha d'examinar la magnitud i la direcció dels coeficients de les variables originals. Com més gran és el valor absolut del coeficient, més important és la variable corresponent en el càlcul de la component. El gran que ha de ser el valor absolut d'un coeficient per a considerar-lo important és subjectiu. Per aquesta raó cal emprear el nostre coneixement especialitzat en la base de dades per a determinar a quin nivell és important el valor de correlació.

En la figura 19, la primera component principal té una gran associació negativa amb CLEAR i grans associacions positives amb HUM i MISTCLOUDY, així que aquest component mesura principalment les diferències en el clima tenint en compte l'humitat. La segona component té grans associacions negatives amb TEMP i ATEMP, de manera que aquesta component mesura principalment la temperatura. La tercera component té una alta associació negativa amb WORKINGDAY, fent que els altres coeficients baixin dràsticament excepte WINDSPEED que també té una relació negativa, per la qual cosa aquesta component mesura la velocitat del vent en els dies no laborables. La PC4 té una molt alta associació positiva amb WINDSEPEED fent que la resta disminueixi excepte WORKINGDAY que té una associació negativa menor, en valor absolut, que WINDSPEED; aquesta component pot estar mesurant el mateix que PC3 pero de forma inversa, prioritzant WINDSPEED en comptes de WORKINGDAY.

Una de les preguntes més freqüents que sorgeix després de realitzar un PCA és: Quanta informació present en el set de dades original es perd en projectar les observacions en un espai de menor dimensió? O cosa que és el mateix: Quanta informació és capaç de capturar cadascuna de les components principals obtingudes? Per a contestar a aquestes preguntes es recorre a la proporció de variància explicada per cada component principal.
La proporció de variància explicada és un valor de gran utilitat a l'hora de decidir el nombre de components principals a utilitzar en les anàlisis posteriors. Si es calculen totes les components principals d'un set de dades, llavors, encara que transformada, s'està emmagatzemant tota la informació present en les dades originals.
Una vegada calculades les components principals, es pot conèixer la variància explicada per cadascuna d'elles, la proporció respecte al total i la proporció de variància acumulada.

Fent un plot de la proporció de la variància per cada component obtenim:



Nota: l'ordre de les components és el mateix que en la taula anterior.

Com es pot observar, la primera component explica el 34% de la variança observada en les dades i la segona el 30%. Les dues següents components expliquen un 14% cadascuna. Mentre que la resta no superen per separat l'1% de variància explicada, la qual cosa ens indica que no són significatives.

La component que correspon a PC1 compensa al voltant del 34% la variança explicada. El que sifgnifica que una part important (100% –34% = 66%) de les observacions és distribueix en més d'una dimensió. També per aquest motiu no ens hem de fixar en la variable CNT2, perquè no forma part de les observacions sino que volem relacionar-la amb aquestes.

Una altra manera d'abordar el resultat és preguntar: Quantes components es requereixen per a cobrir més del X% de la variància? Per reduir la dimensionalitat de les dades i conservar almenys el 90 % de variació de les dades originals. Llavors s'ha d'incloure 4 components principals (les 4 primeres en el gràfic) per aconseguir el 92% de la variància explicada en aquest cas. Amb un total de 8 variables en el conjunt de dades original (tenint en compte CNT2), l'abast per a reduir la dimensionalitat és limitat però no impossible.

En conclusió, el gràfic ens indica que, per mantenir un 92% de la variància expicada de les dades originals, hem descollir les 4 primeres components principals. I la taula ens indica que, d'aquestes 4 primeres components, les variables que tenen els coeficients en valor absolut més elevats (i, per tant, son més importants per calcular la component) són totes les variables del nostre dataset. Per aquesta raó no s'hauria d'eliminar cap variable, és a dir, no seria aconsellable reduir més la dimesionalitat de les nostres dades.