MEMÒRIA
PRÀCTICA 1 - APC
PRIMERS PASSOS

Universitat Autònoma de Barcelona

Jhoe Gabriel Chirinos Sullcany, 1601291
Oriol Camps Isús, 1558778

ANÁLISI DE BASE DE DADES

La nostra base de dades és *Human Resources Data Set,* on es veu la informació sobre la situació laboral i personal dels treballadors d'una empresa.
En aquesta base de dades hi ha un total de 36 atributs, aquests son:

Atribut	**Descripció**	**DataType**
Employee Name	Nom del treballador	Text
EmpID	ID única per a cada treballador	Text
MarriedID	Indica si la persona está casada(1=SI, 0=NO)	Binary
MaritalStatusID	ID que mostra l'estat de MaritalDesc com un int	Integer
EmpStatusID	Codi que mostra l'estat de EmploymentStatus com un int	Integer
DeptID	Codi del departament on treballa el traballador	Integer
PerfScoreID	Codi de la puntuació de rendiment que coincideix amb la puntuació de rendiment més recent del treballador	Integer
FromDiversityJobFairID	Difrencia si l'empleat prové de la feria de empleo Diversity (1=SI, 0=NO)	Binary
Salary	Salari d'un any del treballador en dòlars	Float
Termd	Si el treballador ha sigut acomiadat (1=SI, 0=NO)	Binary
PositionID	Un int que diferencia la posició del treballador a l'empresa	Integer
Position	El nom/títol de la posició que ocupa el treballador	Text
State	Estat on viu la persona.	Text
Zip	Codi postal on viu treballador	Text
DOB	Aniversari del treballador	Date
Sex	Sexe del treballador-(M=home, F=dona)	Text
MaritalDesc	Estat civil de la persona (divorciat, solter, vidu, separat,etc.)	Text
CitizenDesc	Etiqueta per diferenciar si la persona es un ciutadà comunitari o no.	Text

HispanicLatino	Diferencia si l'empleat es hispano/llatí(SI/NO)	Text
RaceDesc	Raça amb la que s'identifica el treballador	Text
DateofHire	Data de la contractació de la persona.	Date
DateofTermination	Data on el treballador va ser acomiadat, només en el cas on Termd = 1.	Date
TermReason	Una descripció del motiu pel qual el treballador va ser acomiadat.	Text
EmploymentStatus	Descripció categòrica del estat del treballador, Active, Voluntarily Terminated, Terminated for cause…	Text
Department	Nom del departament on treballa el treballador.	Text
ManagerName	Nom de la persona encarregada del treballador.	Text
ManagerID	Identificador del encarregat.	Integer
RecruitmentSource	El nom de la font de contractació d'on es va contractar el treballador.	Text
PerformanceScore	Font de rendiment, Fully Meets, Partially Meets, etc.	Text
EngagementSurvey	Resultat de l'última enquesta de participació feta per un soci extern.	Float
EmpSatisfaction	La puntuació de satisfacció del treballador de l'última enquesta, valors de l'1 al 5.	Integer
SpecialProjectsCount	Numero de projectes especials fets per el treballador durant els últims 6 mesos.	Integer
LastPerformanceReviewDate	La data més recent de l'última revisió del rendiment de la persona.	Date
DaysLateLast30	El número de vegades que el treballador ha arribat tard a la feina	Integer
Absences	El número de vegades que el treballador ha faltat a la feina	Integer

A continuació vem mirar una descripció de les variables numèriques més detalladament per a comprovar els valors, mitjanes, quartils, desviació estàndard, etc. de cada variable i així fer-nos una idea de com estaven distribuïdes les dades.

	EmpID	MarriedID	MaritalStatusID	GenderID	EmpStatusID	DeptID
count	311.000	311.000	311.000	311.000	311.000	311.000
mean	10156.000	0.399	0.810	0.434	2.392	4.611
std	89.922	0.490	0.943	0.496	1.794	1.083
min	10001.000	0.000	0.000	0.000	1.000	1.000
25%	10078.500	0.000	0.000	0.000	1.000	5.000
50%	10156.000	0.000	1.000	0.000	1.000	5.000
75%	10233.500	1.000	1.000	1.000	5.000	5.000
max	10311.000	1.000	4.000	1.000	5.000	6.000

PerfScoreID	FromDiversityJobFairID	Salary	Termd	PositionID	Zip
311.000	311.000	311.000	311.000	311.000	311.000
2.977	0.093	69020.685	0.334	16.846	6555.482
0.587	0.291	25156.637	0.473	6.223	16908.397
1.000	0.000	45046.000	0.000	1.000	1013.000
3.000	0.000	55501.500	0.000	18.000	1901.500
3.000	0.000	62810.000	0.000	19.000	2132.000
3.000	0.000	72036.000	1.000	20.000	2355.000
4.000	1.000	250000.000	1.000	30.000	98052.000

ManagerID	EngagementSurvey	EmpSatisfaction	SpecialProjectsCount	DaysLateLast30	Absences
303.000	311.000	311.000	311.000	311.000	311.000
14.571	4.110	3.891	1.219	0.415	10.238
8.078	0.790	0.909	2.349	1.295	5.853
1.000	1.120	1.000	0.000	0.000	1.000
10.000	3.690	3.000	0.000	0.000	5.000
15.000	4.280	4.000	0.000	0.000	10.000
19.000	4.700	5.000	0.000	0.000	15.000
39.000	5.000	5.000	8.000	6.000	20.000

1. Descripció de les variables númeriques.

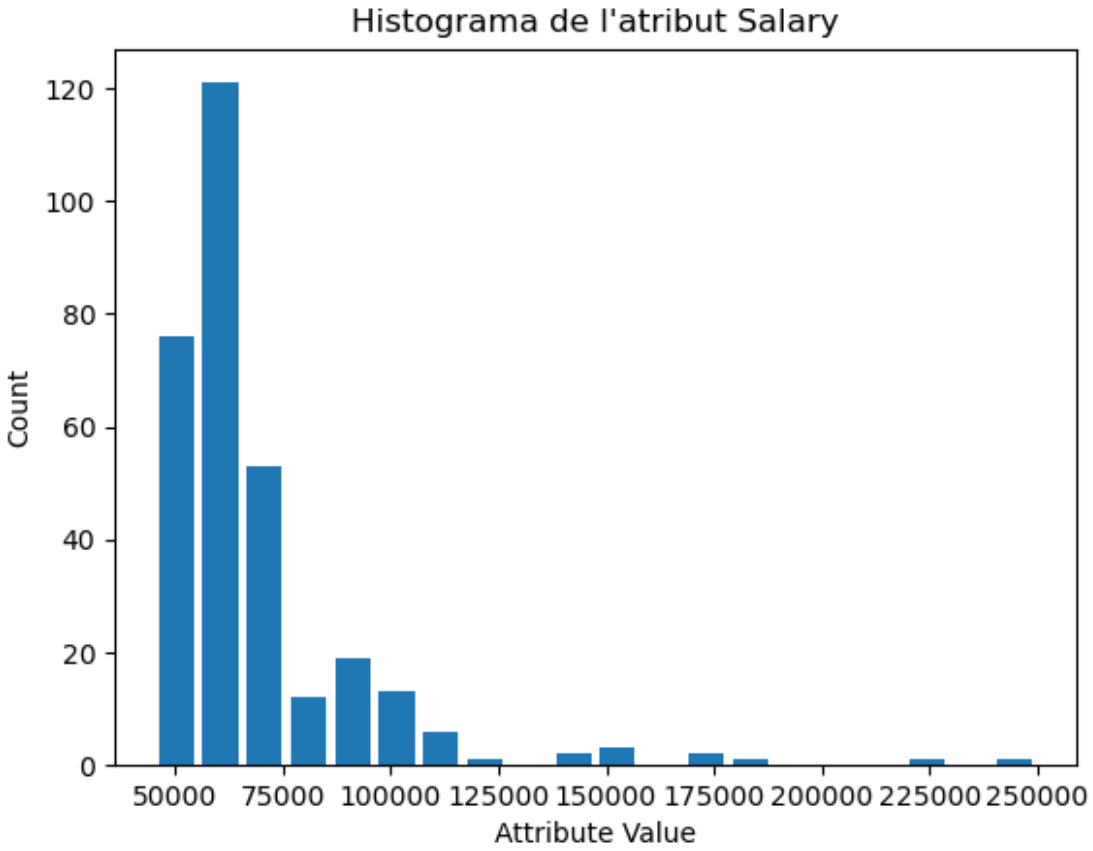
Arribats a aquest punt, vam decidir netejar la base de dades ja que hi havia molts atributs els quals no eren útils, el primer cas eren tots els identificadors ja que no ens donaven informació rellevant del treballador, només eren un simple distintiu, per això els vem el eliminar del conjunt de dades, el segon cas eren totes aquelles dades categòriques, ja que a l'hora de comparar i utilitzar els regressors lineals no ens serien d'utilitat i no serien una bona representació del resultat, per últim els atributs binaris ja que, al igual que en el cas anterior, no son atributs viables per a la regressió.

Al final ens ha quedat una base de dades amb aquests 7atributs.

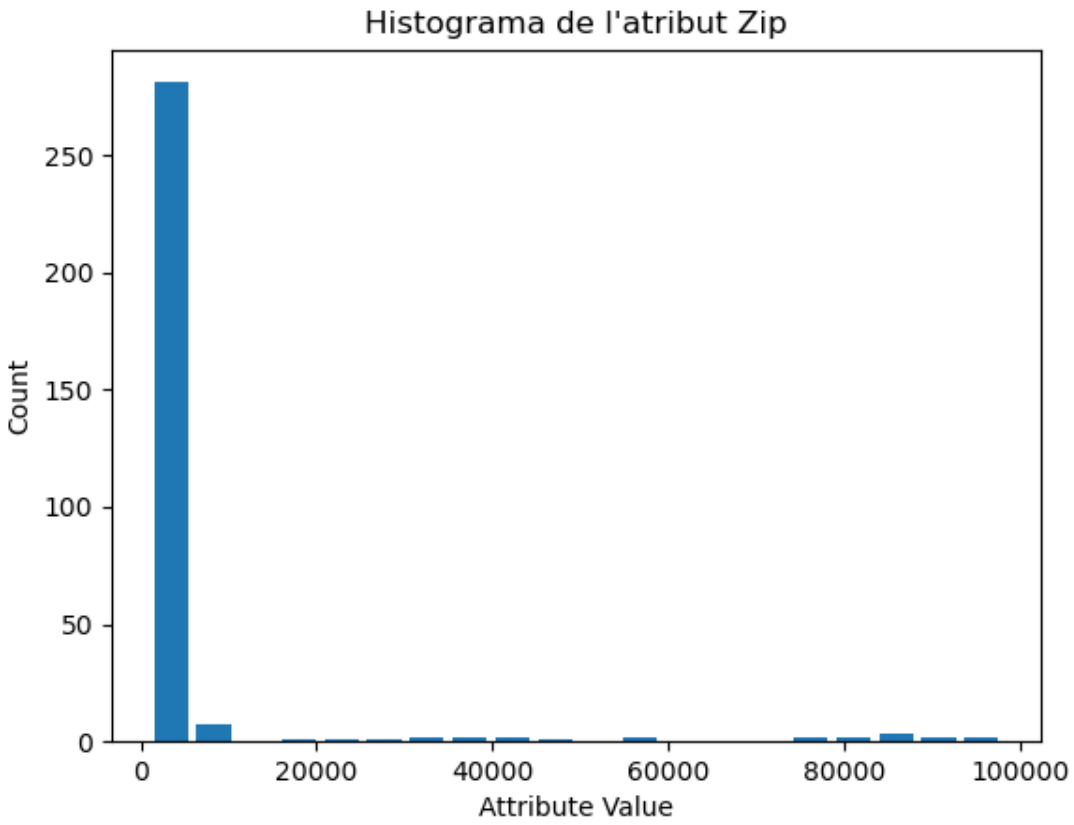
```
Salary                   int64
Zip                      int64
EngagementSurvey       float64
EmpSatisfaction          int64
SpecialProjectsCount     int64
DaysLateLast30           int64
Absences                 int64
```

2. Descripció de les variables númeriques.

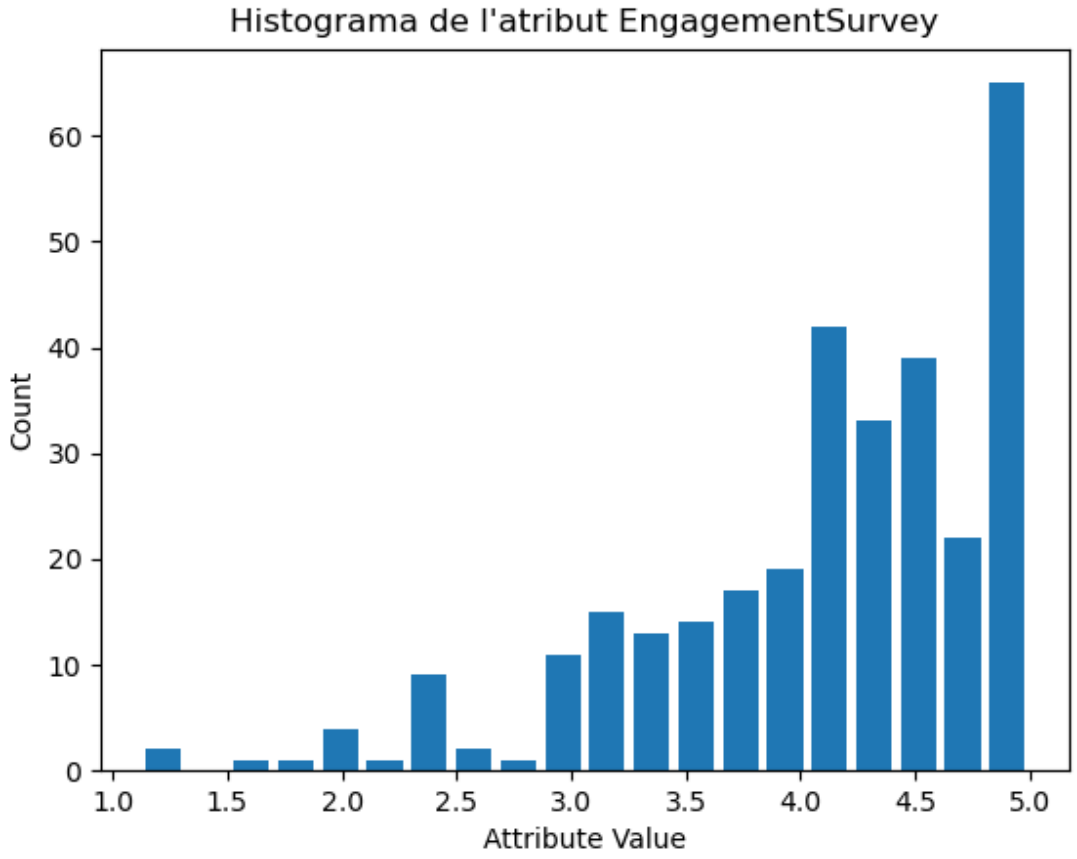
Un cop ben netejada la base de dades, vem buscar una manera més gràfica per a representar-la, el primer que vem fer va ser un histograma de cada una de les variables per saber si alguna segueix una distribució normal.



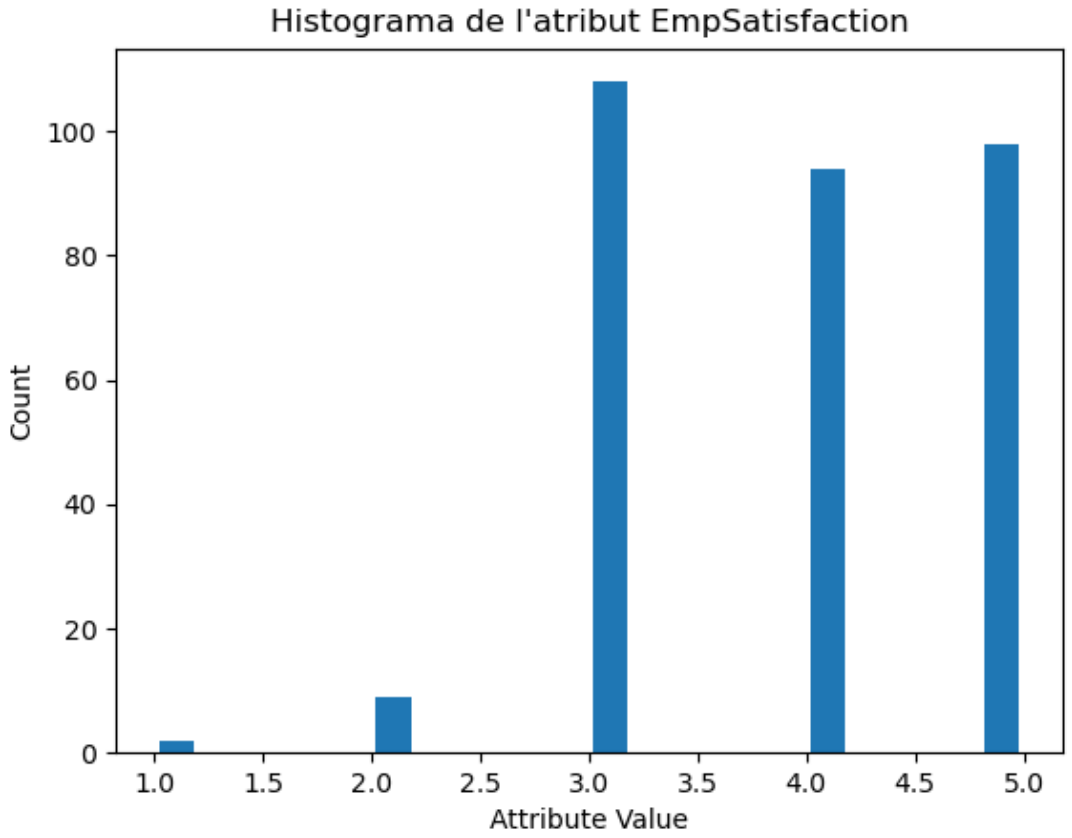
3. Histograma de l'atribut salary



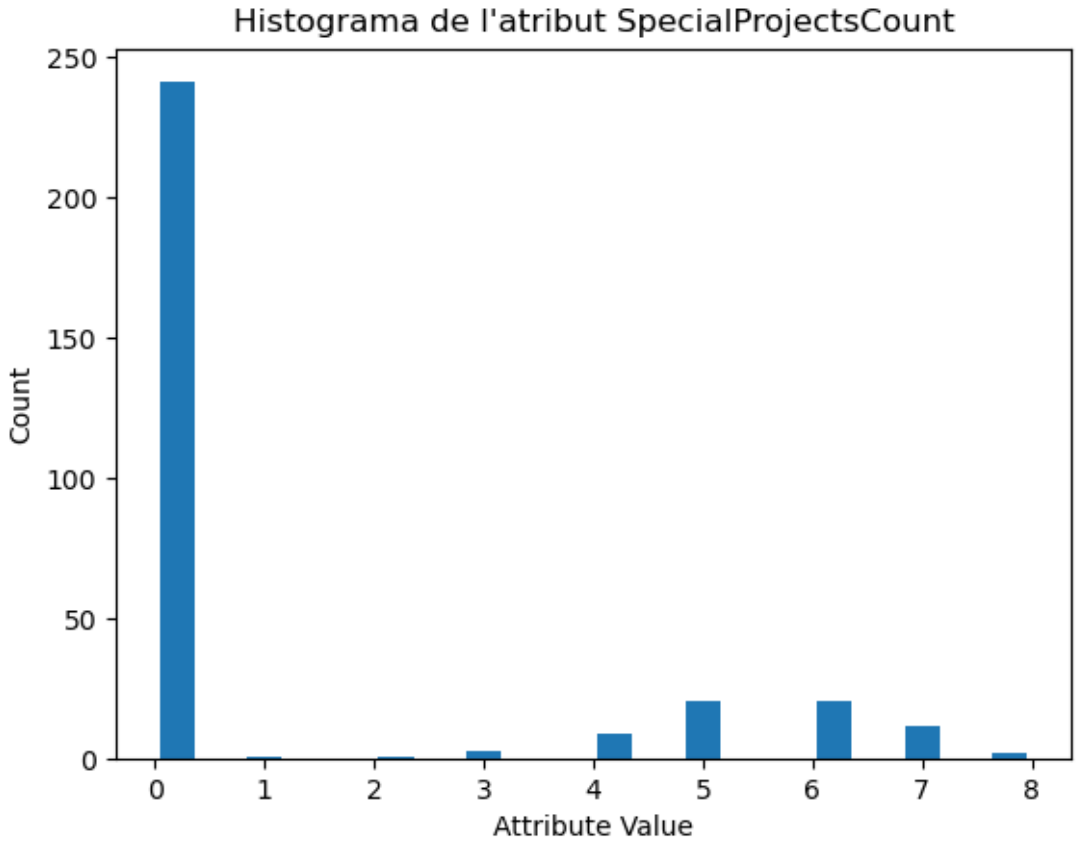
4. Histograma de l’atribut zip



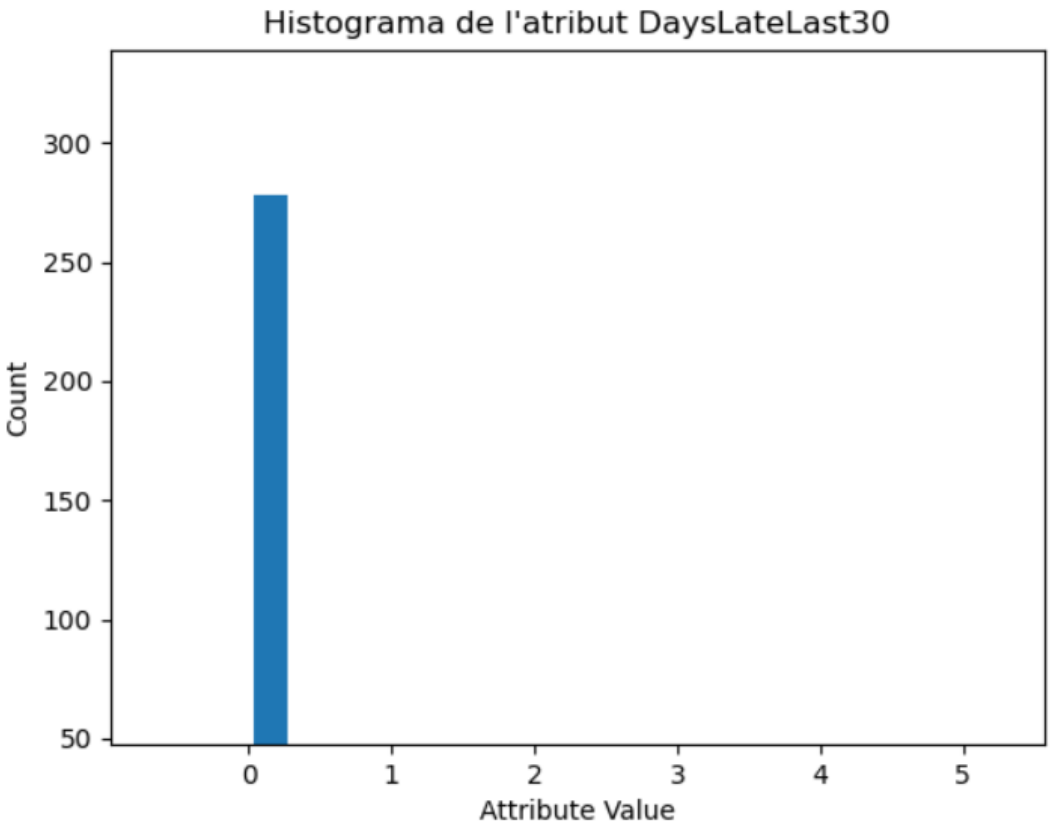
5. Histograma de l’atribut EngagementSurvey



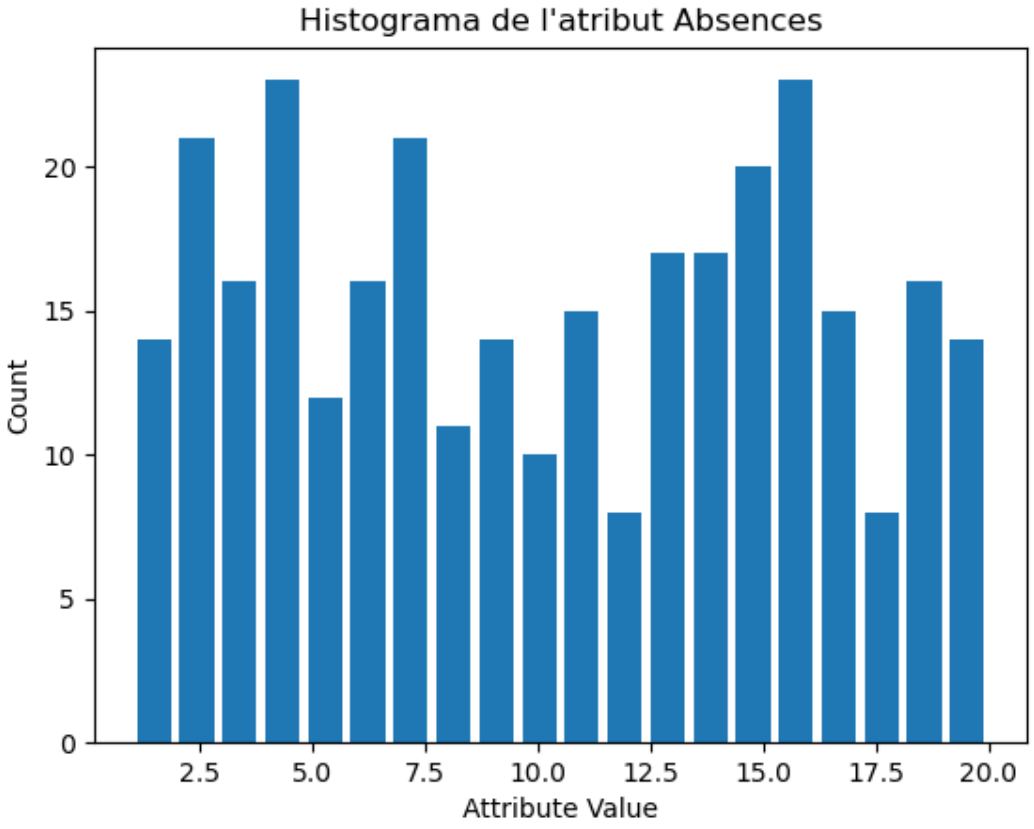
6. Histograma de l'atribut EmpSatisfaction



7.Histograma de l'atribut SpecialProjectsCount



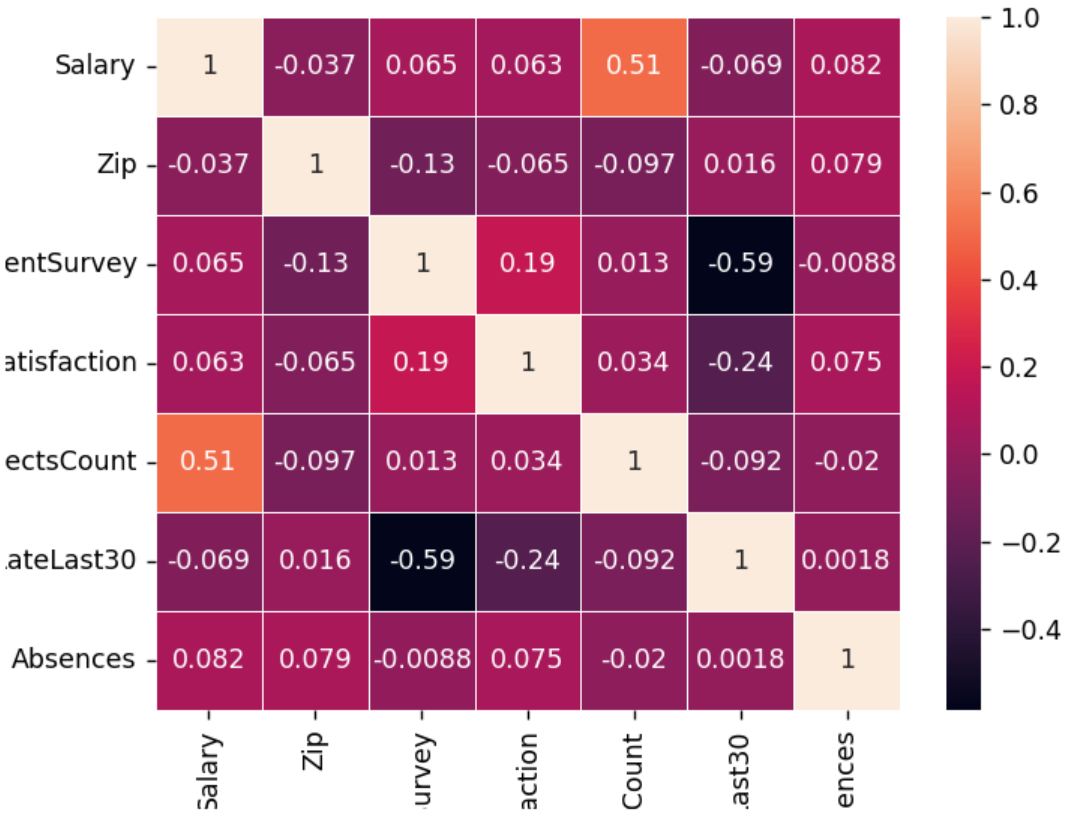
8. Histograma de l’atribut DayLateLast30



9. Histograma de l’atribut Absences

Com es pot comprovar ninguna de les variables segueix una distribució normal, per lo tant ninguna és un candidat clar per a ser un atribut objectiu.

A continuació vem mirar la correlació que hi ha entre les variables per tal de veure si alguna era bona opció.



10. HeatMap dels atributs més rellevants

Com es pot veure, no hi ha molta correlació entre les dades ja que la majoria son variables independents, hi ha alguns casos en que una parella de variables té una correlació major però no és significativament més alta ja que no existeix cap manera de correlacionar aquestes variables.

Donat aquest punt, vem decidir buscar un altre manera de representar dades. En aquest punt vem fer un pairplot entre totes les variables per aixi intentar veure com es comportaven els valors de les diferents variables respecte altres.

11. PairPlot dels atributs mes rellevants

Veient aquestes gràfiques vem notar que el salari tenia una distribució més adequada respecte les altres variables, ja que, al ser una variable continua millora la relació amb altres atributs.

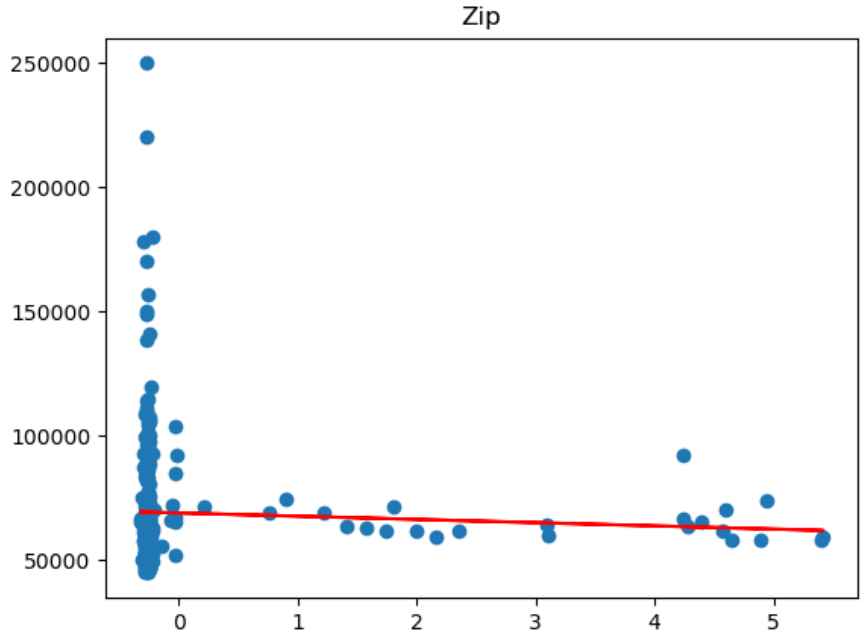
Arribats a aquest punt, vem decidir que la nostre variable objectiu seria Salary ja que és la que més interes pot tenir a l'hora de ser predita per la regressió i també és la que té millor relació amb la resta d'atributs.

Primeres regressions

Una vegada hem escollit el atribut objectiu ens tocava fer les prediccions per a aquest atribut. Per a fer valorar les regressions lineals hem calculat tant el error quadratic mitja(mse) com el R2 score.

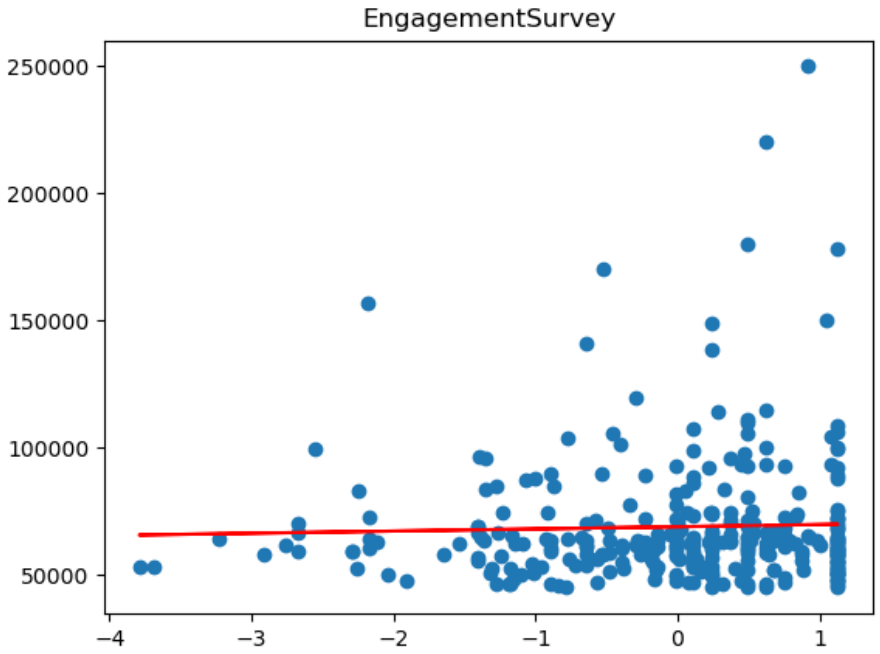
En aquest cas no hem considerat la estandarditzación de les dades ja que els resultats que ens donen les regressions un cop estandarditzades, no reflecteixen la realitat de cada atribut.

Els resultat obtinguts en cada regressió han sigut els següents



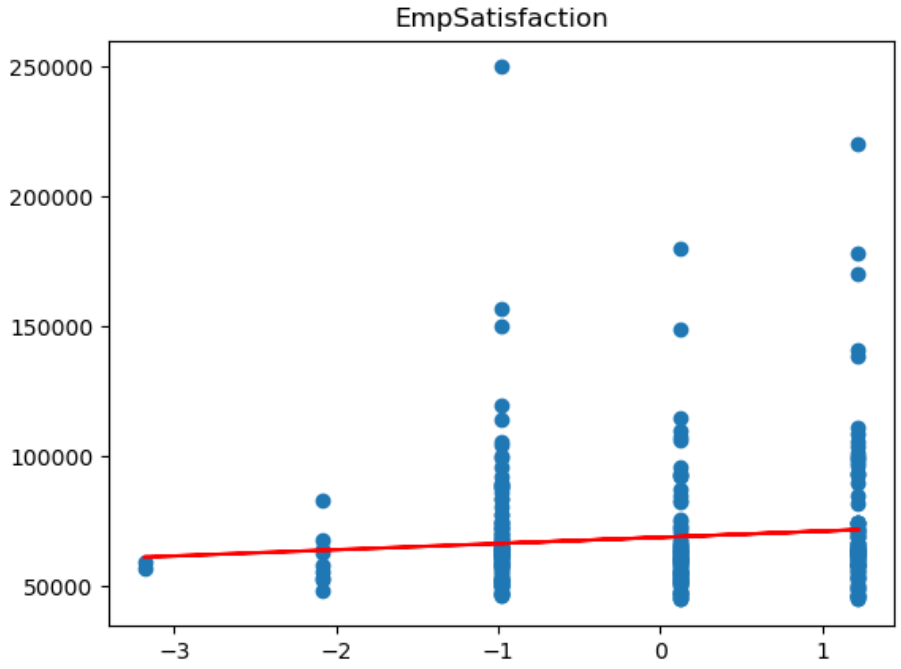

```
Zip
Mean squeared error:  523501620.44452566
R2 score:  0.003434097586788698
```

12. Regressió lineal de l'atribut ZIP



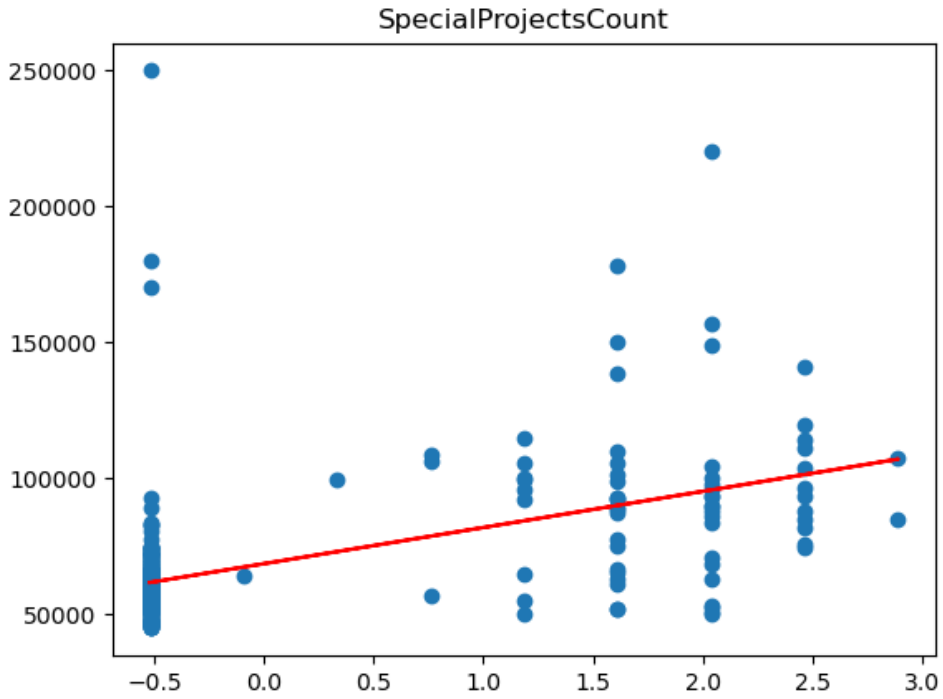

```
EngagementSurvey
Mean squeared error:  524530750.38026625
R2 score:  0.0014749903690481903
```

13. Regressió lineal de l'atribut EngagementSurvey



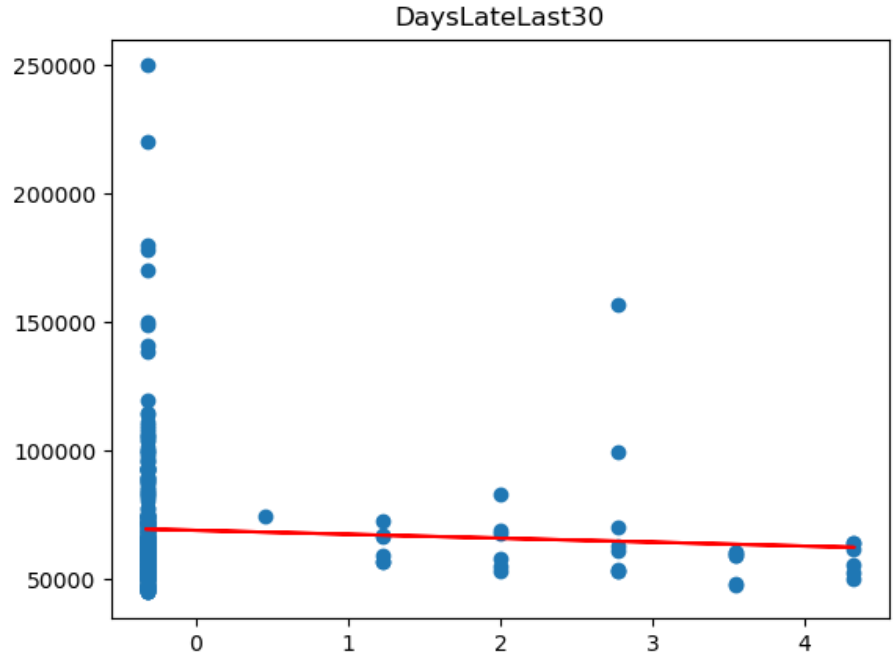

```
EmpSatisfaction
Mean squeared error:  519450963.4926485
R2 score:  0.011145146498511505
```

14. Regressió lineal de l'atribut EmpSatisfaction



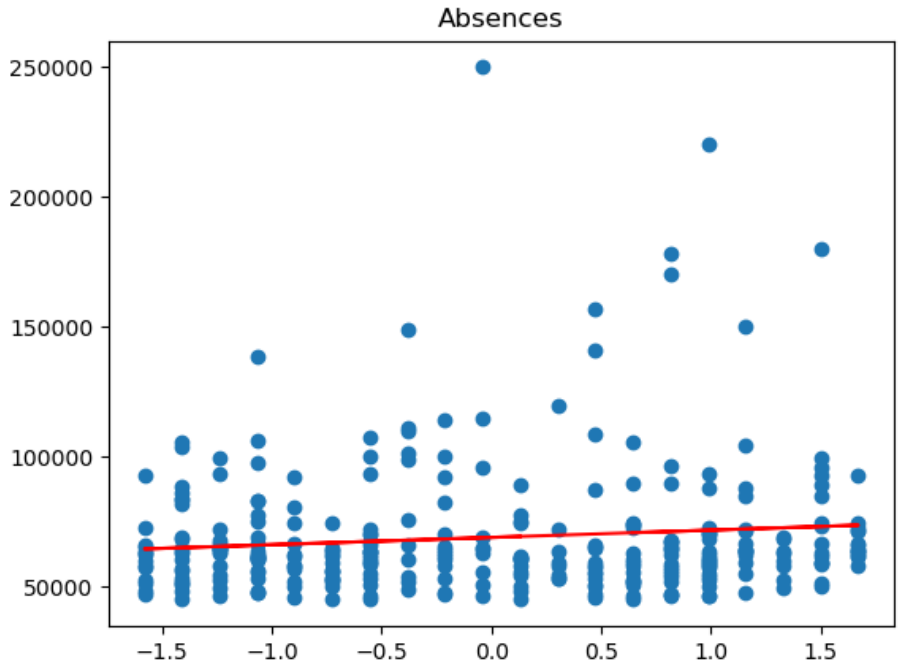

```
SpecialProjectsCount
Mean squeared error:  343442967.7709206
R2 score:  0.3462034543207917
```

15. Regressió lineal de l'atribut SpecialProjectsCount




```
DaysLateLast30
Mean squeared error:  711538536.6055325
R2 score:  0.0033940904884876
```

16. Regressió lineal de l'atribut DayLateLast30




```
Absences
Mean squeared error:  706029598.6394683
R2 score:  0.011110102833094548
```

17. Regressió lineal de l'atribut absences

Salary vs Zip
Mean squeared error:
691918120.9533377
R2 score:
0.0020263096286673443

Salary vs EmpSatisfaction
Mean squeared error:
691360034.4934449
R2 score:
0.002831254010176032

Salary vs DaysLateLast30
Mean squeared error:
692521818.1481009
R2 score:
0.0011555795537062652

Salary vs EngagementSurvey
Mean squeared error:
689327658.9212464
R2 score:
0.00576260858608002

Salary vs SpecialProjectsCount
Mean squeared error:
540223630.4221958
R2 score:
0.22081969852817884

Salary vs Absences
Mean squeared error:
685704998.6732619
R2 score:
0.010987677141390373

Com es pot observar, en la llista valors de les regressions, els valors del MSE son molt grans i els valors de R2 Score molt petits, això es deu a que les variables no tenen molta correlació amb la variable objectiu, la que millor podria funcionar seria el atribut *SpecialProjectsCount,* ja que es el que te un R2 Score més alt i el MSE més baix, tot i que no és recomanable utilitzar aquesta regressió per a predir el salari òptim del treballador.

GITHUB

Al nostre repositori trobarem tots els documents que hem utilitzat i modificat per poder treballar la nostre pràctica, aquests serien:

- **HRDataset_v14.csv** : Base de dades de la nostre pràctica.
- **Practica1-Regressio2022-GEI.ipynb** : Enunciat de la pràctica que no vam modificar per poder seguir-lo com a pauta.
- **Practica1.ipynb**: Notebook del document utilitzat per fer la pràctica.
- **Practica1.py** : Codificació utilitzada a la pràctica
- **PLAB 1 -APC.pdf**: Memòria final de la pràctica.
- **PRESENTACIÓ PRÀCTICA 1.pdf** : Presentación de la memòria en PDF.

Com es podrà veure a la rama main no hi ha molts commits, això és degut a que nosaltres com a grup vam poder treballar presencialment i per tant els únics commits es poden trobar son les pujades finals dels documents definitius de la pràctica.

➢ URL del nostre github

CONCLUSIONS

Després de fer totes les neteges, proves, regressions, etc. hem arribat a la conclusió que aquesta base de dades no és la indicada per a fer un estudi d'aquest tipus ja que la naturalesa de les dades no dóna peu a que es puguin fer regressions adequades a les necessitats.
Les causes d'aquest problema son molt variades, en primer lloc, ninguna variable té una distribució gaussiana, lo qual és un indicador de que les regressions no donaran resultats precisos, la següent causa pot ser la falta d'atributs no categòrics ja que de 36 atributs totals hem tingut que reduir el número d'atributs a 6, lo qual ens dona una base per a predir el objectiu molt reduïda.

Tenint en conte els resultats de les regressions, es podria arribar a concluir que per a fer una bona predicció del salari d'un treballador farien falta altres atributs, com per exemple les hores treballades, les hores extres fetes pel treballador, els dies de vacances demanats, etc. ja que en aquesta empresa, les dades personals no son importants a l'hora de decidir el salari dels seus treballadors.