Universitat Autònoma de Barcelona
Facultat de Ciències

Pràctica 1

Autors:

Andrea González & Gerard Lahuerta & Ona Sánchez

1603921 — 1601350 — 1601181

11 d'Octubre del 2022

Índex

1 Introducció

L'objectiu d'aquesta pràctica és, mitjançant la interfície proporcionada per Jupyter Notebook, estudiar i predir un valor en funció d'un conjunt de paràmetres que es calcularan mitjançant un conjunt de dades.
Les dades han sigut proporcionades per la web de Kaggle, concretament, la base de dades de RRHH.

L'objectiu d'aquesta primera pràctica és trobar models que descriguin les dades i permetin generar noves conclusions. Així doncs, després d'un estudi de les dades s'ha decidit intentar predir el *Salary* en funció de les altres variables.

El dataset que s'utilitza es pot trobar al següent enllaç:
https://www.kaggle.com/datasets/rhuebner/human-resources-data-set?resource=download.

Aquest dataset va estar creat per la Dra. Carla Patalano i el Dr. Rich.
Si bé el dataset està pensat per a predir si un treballador acabarà el seu contracte o no, s'ha pogut utilitzar també (no sense problemes i dificultats) per al nostre objectiu.

Les dades de recursos humans poden ser difícils d'aconseguir, analitzar i visualitzar, per la qual cosa s'ha treballat per tal d'obtenir la millor pressició possible.

2 Presentació de les funcions

2.1 Llibreries i importacions

Per tal de poder dur a terme la nostra tasca és imprescindible tenir intal·lades les següents llibreries, ja que s'utilitzen les funcions següents (d'entre altres).

Llibreria	**Funció utilitzada**
sklearn.datasets	make_regression
numpy (as np)	shuffle isnan min max floor reshape array
pandas (as pd)	read_csv DataFrame
matplotlib pyplot (as plt)	figure subplots plot hist
seaborn (as sns)	heatmap pairplot
sklearn.linear_model	LinearRegression Lasso BayesianRidge
math	*operacions aritmètiques vàries*
sklearn.metrics	r2_score
ipywidgets	interact
mpl_toolkits.mplot3d	axes3d
intertools	combinations
plotly.express (as px)	scatter_matrix
sklearn.decomposition	PCA

2.2 Funcions programades

2.2.1 trobar_moda

- Entrada:
  - **DataFrame** dt
  - **string** col_e
  - **string** col_a
  - **string** val
- Sortida: **float** pos
- Funcionament: Agafa les files que tenen el mateix valor *val* en l'atribut *col_a*, troba quin és el valor de *col_e* més comú i el retorna.
- Informació rellevant: Funció utilitzada exclusivament per a tractar els nulls de l'atribut ManagerID.

2.2.2 ordenar_salary

- Entrada:
  - **DataFrame** dt
  - **string** col
- Sortida: **dict** lista
- Funcionament: Crea una llista amb els diversos tipus de l'atribut *col* i els ordena en funció del salari al diccionari que retorna.
- Informació rellevant: Funció utilitzada exclusivament per a tractar el canvi de tipus de variable (d'objecte a int).

2.2.3 standarize

- Entrada:
  - **np.array** X
- Sortida: **np.array** x
- Funcionament: Per cada atribut, calcula la mitjana i la desviació estàndar, posteriorment normalitza cada dada restant la mitjana i dividint per la desviació estàndar.
- Informació rellevant: Funció utilitzada exclusivament com a pas intermedi per a estudiar el dataset i millorar la seva predicció.

2.2.4 mse

- Entrada:
  - **np.array** y1
  - **np.array** y2
- Sortida: **float** mse
- Funcionament: Es comprova que la mida de y1 i y2 sigui igual i es calcula

$$\frac{1}{n}\sum_{i=1}^{n}(y_{1i}-y_{2i})^2$$

- Informació rellevant: La funció és utilitzada per tots els regressors per estudiar l'error que cometen.

2.2.5 regression

- Entrada:
  - **np.array** x
  - **np.array** y
- Sortida: Regressor lineal (*LinearRegressor()*) ja entrenat amb les dades x i y.
- Funcionament: Es crea un objecte de regressió amb *sklearn* i es retorna el model entrenat amb el mètode *fit*.
- Informació rellevant: Funció utilitzada tant per a fer el regressor lineal simple com per al regressor multilineal simple, amb la tranformació logarítmica i sense.

2.2.6 split_data

- Entrada:
  - **np.array** x
  - **np.array** y
  - **float** train_ratio
- Sortida: **np.array** x_train, **np.array** y_train, **np.array** x_val i **np.array** y_val.
- Funcionament: mitjançant els mètodes *shuffle* i *floor* de la llibreria numpy creem les 4 llistes que retornem amb els components de les x i y aleatòriament ordenats i dividits.
- Informació rellevant: Funció utilitzada tant per validar el comportament dels regressors lineals com els multilineals i evitar el overfitting. Aquesta funció en alguns casos ha sigut modificada afegint condicions per evitar tenir en compte outliners. En aquests casos la funció ha sigut reescrita i assenyalada explícitament. En cas de no introduir el train_ratio, aquest tindrà el valor per defecte 0.8.

2.2.7 combination

- Entrada:
  - **list** A
  - **int** n_conj
- Sortida: **list** aux
- Funcionament: Es crea una llista amb totes les combinacions de *n_ conj* elements de la llista *A* usant la funció *combinations* de la llibreria *intertools*.
- Informació rellevant: Funció utilitzada únicament per a crear totes les combinacions possibles d'atributs rellevants en la regressió multilineal.

2.2.8 lasso

- Entrada:
  - **np.array** x
  - **np.array** y
  - **float** a
- Sortida: Regressor lineal (*Lasso()*) ja entrenat amb les dades x i y.
- Funcionament: Es crea un objecte de regressió *Lasso* amb *sklearn* i es retorna el model entrenat amb el mètode *fit*.
- Informació rellevant: Funció utilitzada tant per a fer el regressor lineal com per al regressor multilineal amb el mètode de la llibreria sklearn *Lasso*, amb la tranformació logarítmica i sense. En cas de no introduir la a, aquesta tindrà el valor per defecte 0.1.

2.2.9 Bayes

- Entrada:
  - **np.array** x
  - **np.array** y
  - **float** t
- Sortida: Regressor lineal (*BayessianRidge()*) ja entrenat amb les dades x i y.
- Funcionament: Es crea un objecte de regressió BayessianBrigde amb *sklearn* i es retorna el model entrenat amb el mètode *fit*.
- Informació rellevant: Funció utilitzada tant per a fer el regressor lineal com per al regressor multilineal amb el mètode de la llibreria sklearn *BayessianBrigde*, amb la tranformació logarítmica i sense. En cas de no introduir la t, aquesta tindrà el valor per defecte 10^{-6}.

3 Gestió del dataset

3.1 Explicació del Dataset

El dataset tracta sobre una empresa fictícia de 311 treballadors (que estan o han estat actius) i 36 característiques que s'han recollit sobre ells.
Per tant, el nostre dataset és de mida 311x36 (files x columnes).
Els 36 atributs recollits dels treballadors són els següents:

Atribut	**Explicació**	**Tipus de dada**
Employee_Name	Nom del treballador	string
EmpID	Identificador empleat	naturals
MarriedID	Casat / no casat	binari
MaritalStatusID	Estat Civil	naturals [0-4]
EmpStatusID	Estat del treballador	naturals [1-5]
GenderID	Sexe	binari
DeptID	Identificador del departament	naturals [1-6]
PerfScoreID	Qualitat del treball	naturals [1-4]
FromDiveristyJobFairID	Participació en fires	binari
Salary	Salari anual	float [45.0k-250k]
Termd	Tipus de jornada	binari
PositionID	Identificador de la posició	naturals [1-30]
Position	Càrrec	string
State	Estat del treballador	string
Zip	Codi Postal	int
DOB	Data de naixement	string
Sex	Sexe	binari
MaritalDesc	Estat Civil	string
CitizenDesc	Residència / no resident	string
HispanicLatino	Procedència llatina	string
RaceDesc	Ètnia	string
DateofHire	Data de contractació	string
DateofTermination	Data de finalització	string
TermReason	Tipus de contracte	string
EmploymentStatus	Situació Laboral	string
Department	Departament	string
ManagerName	Nom del mànager	string
ManagerID	Identificador manager	naturals [1-39]
RecruitmentSource	Font de captació	string
PerformanceScore	Puntuació de rendiment	string
EngagementSurvey	Enquesta de participació	float [1.12-5]
EmpSatisfaction	Satisfacció del treballador	naturals [1-5]
SpecialProjectsCounts	Nombre de projectes	naturals [0-8]
LastPerformanceReview_Date	Última revisió	string
DaysLateLast30	Dies de retard	naturals [0-6]
Absences	Nombre d'abscències	naturals [0-20]

Cal comentar que en informar-se sobre el dataset es troben bastants incongruències com:

- Els coordinadors dels treballadors no estàn a la base de dades
- Alguns valors que haurien d'estar correlacionats no ho estan, com EmploymentStatus, EmpID i EmpStatusID

S'ha tractat amb aquestes incongruencies asumint que:

- Els coordinadors no pertanyen a l'empresa (són gent que subcontracta els serveis de la mateixa)
- Tots els atributs són independents i no estàn correlacionats entre ells fins que es demostri mitjançant càlculs el contrari i es pugui trobar una relació directa.

3.2 Gestió dels valors nulls

Per motius diversos hi ha atributs que no contenen totes les dades. Aquests atributs són (amb el nombre de dades que hi falten):

Atribut	**#Nulls**
DateofTermination	207
ManagerID	8

Observant aquests valors obtenim la següent conclusió:

- La variable DateofTermination només conté valors si el treballador ha acabat el seu contracte (ja no es troba actiu).
- La variable ManagerID està corrupta, ja que tota la informació perduda es correspon als càrrecs Production Technician I i Production Technician II.

S'han gestionat els valors nulls de la següent manera:

- En tenir més del 50% dels valors nulls, es procedeix a eliminar la columna DateOfTermination en no ser útil per a predir el *Salary*.
- Enl tenir menys del 50% dels valors nulls a la variable ManagerID, aquests se substituiran per una estimació, la moda del ManagerID dels treballadors que tenen el mateix càrreg. Se suposa que un mànager organitza un conjunt de treballadors similars, pel que un conjunt de treballadors amb la mateixa posició ha de ser gestionat pel mateix conjunt de mànagers. Del conjunt s'escull el més probable (la moda) per a susbtituir el valor null.

Important: La gestió de les dades així com l'anàlisi de les mateixes no inclouen l'atribut DateOfTermination.

3.3 Gestió del tipus de dades

Per la correcta gestió de les dades és necessari que totes tinguin valors numèrics (*int64* i *float64*), pel que s'ha decidit passar tots els valors de tipus *object* a *int64* i *float64*.
El resultat d'aquesta transformació es mostra a continuació:

Atribut	**Tipus de dada**	
	Original	Transformada
Employee_Name	Object *(string)*	int64
EmpID	int64	int64
MarriedID	int64	int64
MaritalStatusID	int64	int64
EmpStatusID	int64	int64
GenderID	int64	int64
DeptID	int64	int64
PerfScoreID	int64	int64
FromDiveristyJobFairID	int64	int64
Salary	int64	int64
Termd	int64	int64
PositionID	int64	int64
Position	Object *(string)*	int64
State	Object *(string)*	int64
Zip	int64	int64
DOB	Object *(string)*	int64
Sex	Object *(string)*	int64
MaritalDesc	Object *(string)*	int64
CitizenDesc	Object *(string)*	int64
HispanicLatino	Object *(string)*	int64
RaceDesc	Object *(string)*	int64
DateofHire	Object *(string)*	int64
TermReason	Object *(string)*	int64
EmploymentStatus	Object *(string)*	int64
Department	Object *(string)*	int64
ManagerName	Object *(string)*	int64
ManagerID	float64	float64
RecruitmentSource	Object *(string)*	int64
PerformanceScore	Object *(string)*	int64
EngagementSurvey	float64	float64
EmpSatisfaction	int64	int64
SpecialProjectsCounts	int64	int64
LastPerformanceReview_Date	Object *(string)*	int64
DaysLateLast30	int64	int64
Absences	int64	int64

Es fa la transformació de les dades (tot i que existeixen atributs numèrics que representen alguns dels objectes, com GenderID i Sex) perquè hi ha alguns atributs que haurien d'estar correlacionats de manera directa però no ho estan. Pel que (mitjançant la decisió explicada a 3.1) es decideix transformar totes les dades.

4 Estudi del dataset

4.1 Decisions preses abans d'estudiar el dataset

Per tal d'estudiar el dataset, es va decidir comparar els valors obtinguts per a la regressió lineal simple estandaritzant les dades i sense fer-ho. Una vegada vist si hi existeix millora o no amb l'estandaritzat escollim amb quina de les dues formes d'analitzar-lo ens quedem per a fer la resta d'anàlisis.

Per evitar treballar amb moltes dades, s'ha intentat crivar les variables per només usar les més significatives i, així, asseguar que utilitzem atributs que tenen una correlació directa (o prou alta) amb l'atribut objectiu (*Salary*).

Per tal de veure si existeixen relacions entre les dades (així com si les dades segueixen alguna distriució), part de l'anàlisi s'ha centrat a observar histogrames de les dades i gràfics de relació dels atributs amb l'atribut objectiu. D'aquesta manera podem crivar dades que tenen distribucions que sabem que no ajuden a predir ni millorrar la predicció (com atributs que segueixen distribucions uniformes, etc...).

Per altra banda, hem categoritzat 2 valors de criva diferents a l'hora de discutir si un atribut és rellevant o no mitjançant el *R2 score*:

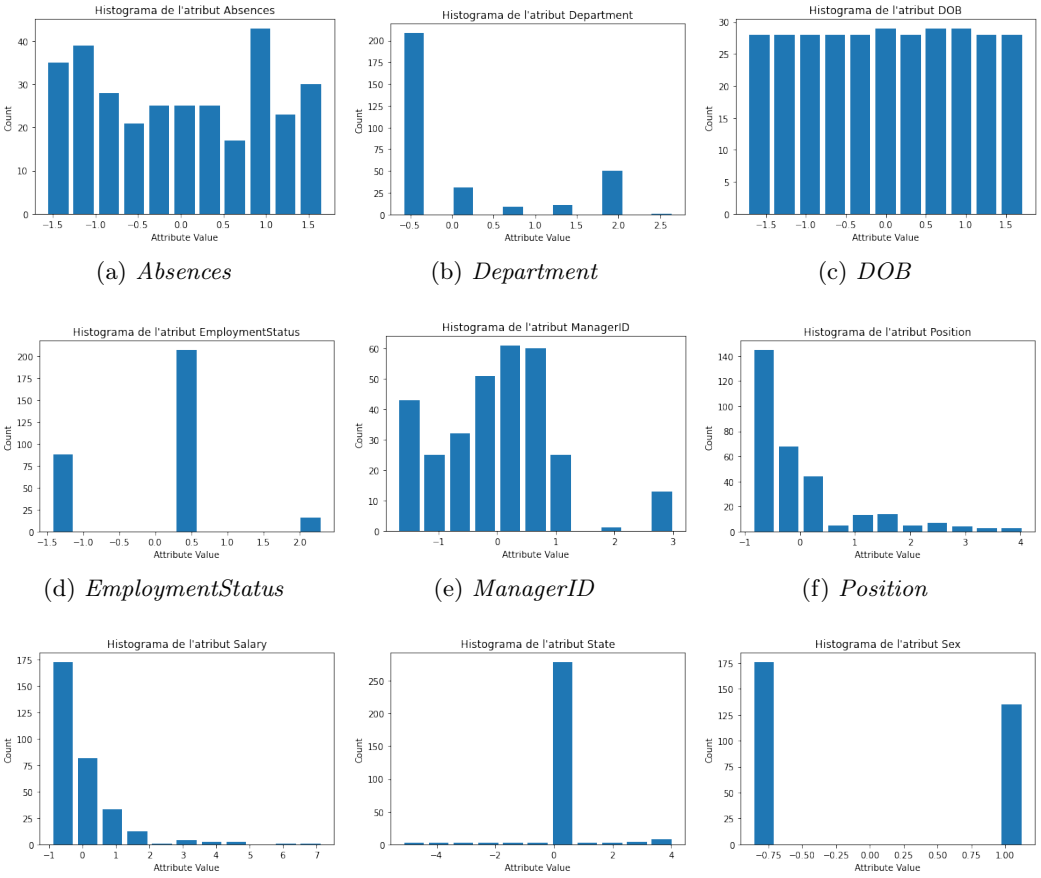
- Si aquest és superior a 0.5 (en valor absolut) diem que és rellevant per a predir el *Salary*.
- Si aquest és superior a 0.3 (en valor absolut) diem que és afí per a predir el *Salary*.

Fem servir aquests dos valors per assegurar que es treballa amb:

1. Valors significatius per a predir/millorar la predicció.
2. Valors fortament correlacionats amb el conjunt de les dades.

4.2 Distribució de les dades

Per tal de poder utilitzar propietats de distribucions (com la normal, etc...) en les dades, representem els histogrames de les variables significatives.



(a) *Absences* (b) *Department* (c) *DOB*

(d) *EmploymentStatus* (e) *ManagerID* (f) *Position*

(g) *Salary* (h) *State* (i) *Sex*

Figura 1: Histogrames dels atributs

S'observa dels histogrames[1] que cap atribut segueix una distribució normal.
La majoria d'atributs segueixen distribucions discretes o uniformes (pel que el seu impacte en la predicció serà poc rellevant o nul).

Per altra banda, existeixen alguns atributs (com el que volem predir, el *Salary*) que sembla que segueixin una distribució exponencial. S'intentarà utilitzar aquesta informació més endavant per a millorar la nostra predicció.

IMPORTANT: Els histogrames han sigut fets amb les dades estandaritzades per així poder visualitzar de millor manera com estan distribuïdes les dades, ja que en alguns casos els valors eren molt dispars i es dificultava la interpretació del gràfic.

[1] Es mostren part dels histogrames obtinguts, la resta es troben a 9.1

4.3 Correlació de les variables

Mostrem ara l'anàlisi de les correlacions dels atributs de la base de dades.
Mencionar que els càlculs de les correlacions han sigut amb el dataset sense estandaritzar.

Atribut	**# afins**
Employee_Name	10
EmpID	4
MarriedID	0
MaritalStatusID	1
EmpStatusID	4
GenderID	1
DeptID	9
PerfScoreID	5
FromDiveristyJobFairID	1
Salary	10
Termd	4
PositionID	2
Position	10
State	0
Zip	1
DOB	10
Sex	1
MaritalDesc	1
CitizenDesc	0
HispanicLatino	0
RaceDesc	1
DateofHire	10
TermReason	5
EmploymentStatus	4
Department	10
ManagerName	11
ManagerID	11
RecruitmentSource	0
PerformanceScore	2
EngagementSurvey	3
EmpSatisfaction	1
SpecialProjectsCounts	10
LastPerformanceReview_Date	13
DaysLateLast30	3
Absences	0

Taula 1: Taula d'atributs afins

Atribut	**Corr.**
Employee_Name	0.756
Position	0.917
DOB	0.751
DateofHire	0.586
Department	0.622
ManagerName	0.654
SpecialProjectsCount	0.508

Taula 2: Taula d'atributs rellevants

S'observa a la figura 2 mostrada que alguns dels atributs amb més relacions són *Employee_Name*, *Salary*, *Position* i *DOB*, entre d'altres. S'escull d'entre els atributs amb més relacions el *Salary* com a objectiu per fer-ne l'estudi, ja que és dels pocs que són continus, numèrics i afins amb la resta.

S'observa, a més, que existeix un conjunt nombrós d'atributs amb el mateix nombre de correlacions que *Salary*, i per això s'intueix que estan relacionats entre ells, afirmant així que existeix una forta relació entre l'atribut *Salary* i la resta d'atributs de la base de dades.

Per altra banda, s'observa a la figura 3 la correlació entre l'atribut *Salary* i la resta d'atributs del dataset. Cal recalcar que només es mostren els atributs que presenten una correlació major al 0.5 (els atributs rellevants).

Una vegada reduït el conjunt de dades a analitzar de 36 a 7 atributs, s'estudia si tenen alguna distribució (respecte a la resta) que indiqui alguna duplicitat de les dades o alguna distribució que útil per predir el valor del *Salary*.

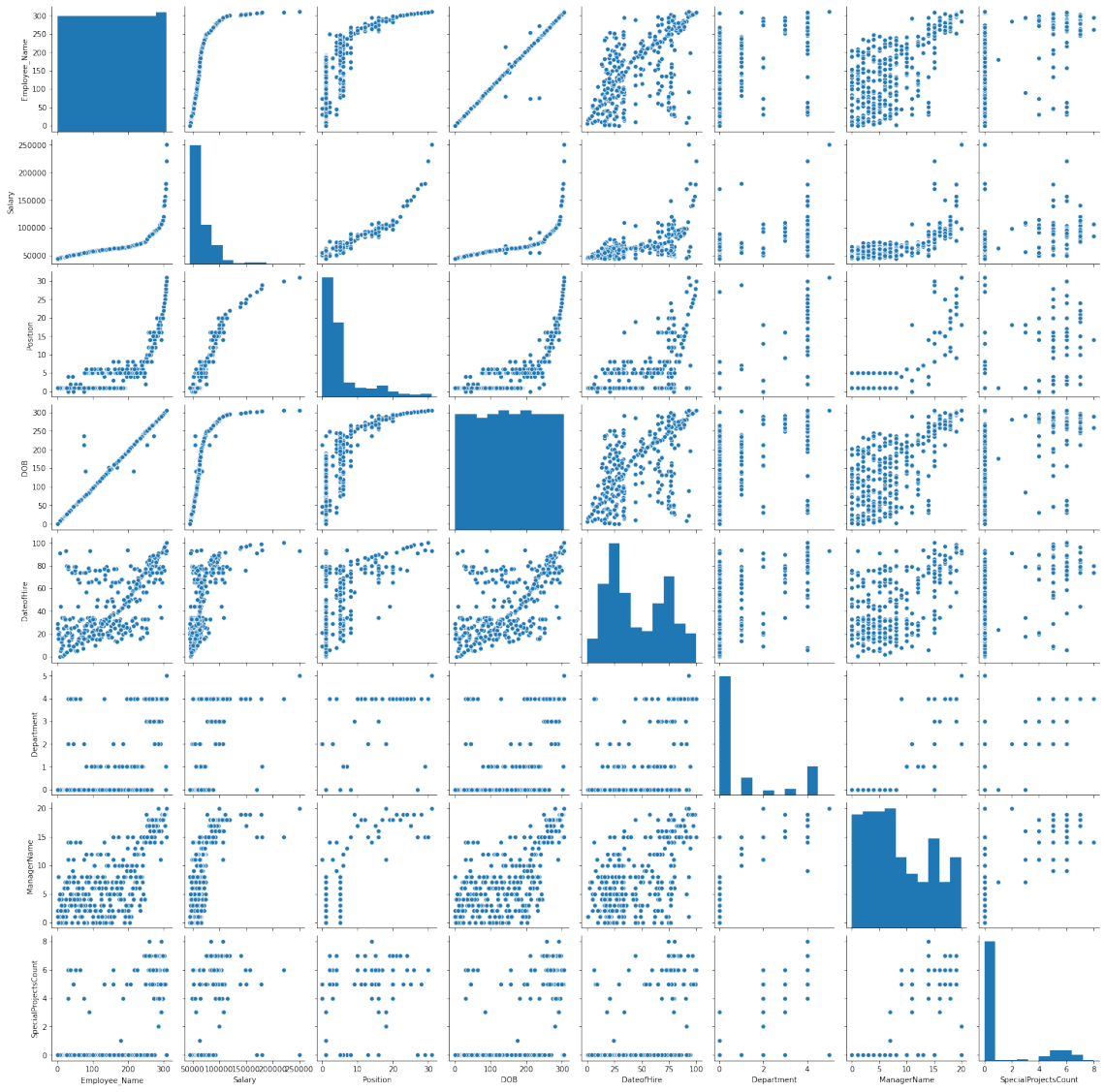


Figura 2: Pairplot dels atributs rellevants i l'objectiu

Els gràfics estan ordenats de la següent forma: Employee_Name, Salary, Position, DOB, DateofHire, Department, ManagerName i SpecialProyectsCount.
S'observa com els atributs Employee_Name, DOB, Position i ManagerName semblen seguir distribució exponencial respecte al Salary.
Per altra banda, hi ha atributs com el DateofHire que tot i que sembla tenir alguna relació, té molts valors dispersos.
També hi ha atributs com SpecialProyectCount que no es veu a simple vista quina relació hi ha.

S'observa com les variables DOB i Employee_name tenen distribucions uniformes i una relació entre elles massa lineal. Estudiem mitjançant un mapa de calor quina relació hi ha entre elles, entre la resta d'atributs rellevants i entre el Salary.

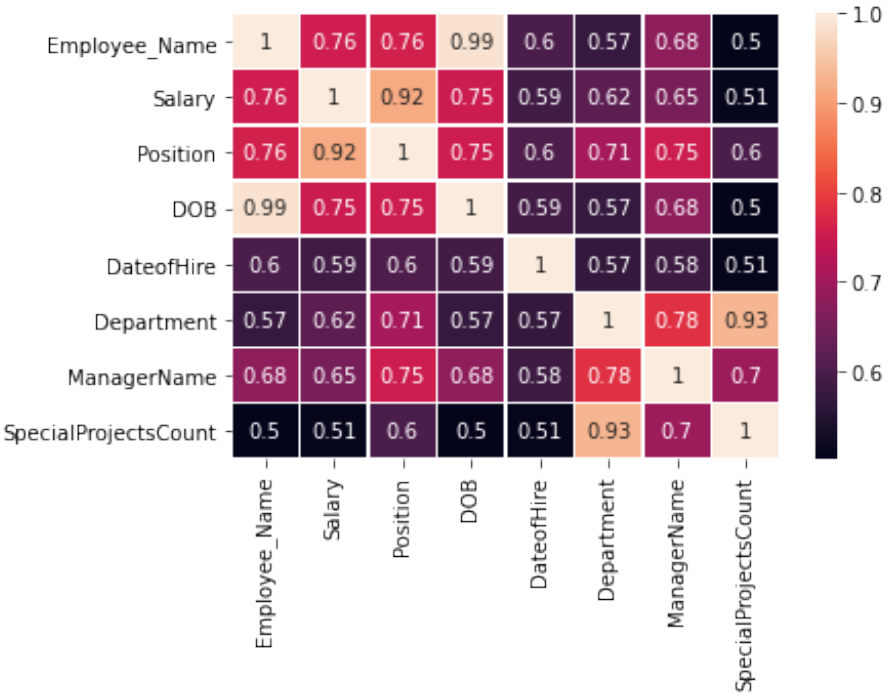


Figura 3: Heatmap dels atributs rellevants i l'objectiu

S'observa del mapa de calor que existeix una forta correlació entre totes les dades rellevants.
A més, s'observa com el millor atribut (pel que fa a que té la millor correlació amb l'objectiu) és el *Position*.

Per altra banda, destaca el fet que la correlació entre els atributs *Employee_ Name* i *DOB* és quasi de 1, encara que no hauria de tenir tanta correlació, ja que en una empresa real amb tants treballadors, el normal és que entre diversos treballadors hi hagués alguns amb el mateix nom i/o la mateixa data de naixement. Es dedueix el següent:

1. Tant el *DOB* com l'*Employee_ Name* segueixen distribucions uniformes pel que no es repeteixen ni dates de naixement ni noms en la majoria dels casos.

2. L'empresa no té suficients treballadors com perquè succeeixi una repetició significativa de les dades de *DOB* i *Name*.

Per aquest fenomen, deduïm que no es pot utilitzar la *DOB* i el *Employee_ Name* per predir el *Salary*.

Mencionar que el fet que l'empresa sigui fictícia perjudica les prediccions, perquè no permet tenir en compte efectes socials per a predir el *Salary*, com racisme o sexisme, ja que no compleix cap desigualtat que podria patir una empresa real. Per altra banda, representen les dades que hauria de tenir una empresa sense cap "brecha" salarial a causa de discriminacions, i per això ens dediquem a estudiar les autèntiques bases de tenir un sou més elevat o no: els càrrecs de l'empresa, el treball extra, el coordinador, etc...

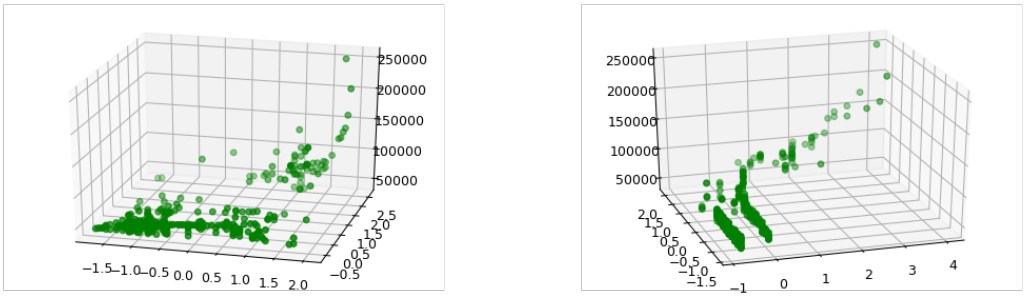
Concloem l'estudi de les correlacions crivant les dades a únicament *Position*, *DateofHire*, *Department*, *ManagerName* i *SpecialProjectsCount*.

4.4 Representació de les dades

Decidim representar les dades del Salary respecte dels atributs rellevant en R3 per veure si hi existeix alguna relació visual amb el valor del Salary.

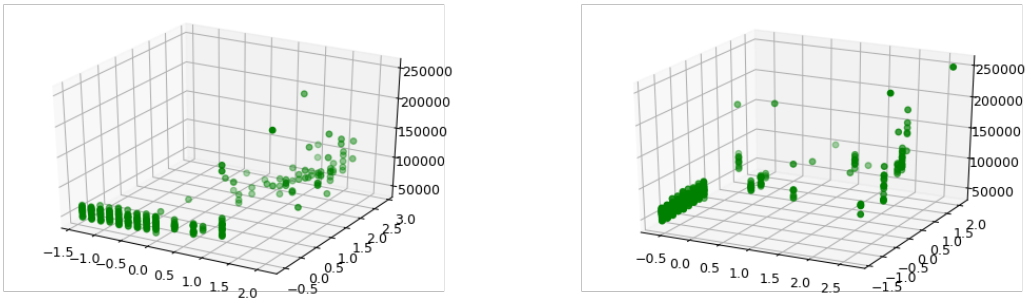
4.4.1 Representació 2 a 2 en R3

A l'hora de representar a R3 el Salary respecte de 2 atributs ens adonem que hi ha un conjunt de gràfics on es veu fàcilment una relació i un altre conjunt on no es veu cap relació aparent. Mostrem ara una mostra dels gràfics estudiats[2].
En aquest cas, els gràfics s'han fet amb els atributs estandaritzats i el Salary amb els valors originals per a poder enfatitzar el creixement exponencial d'alguns dels gràfics.



(a) DateofHire & Department (b) ManagerName & Position

Figura 4: Gràfics amb relacions aparents



(a) ManagerName & SpecialProyectCount (b) Department & ManagerName

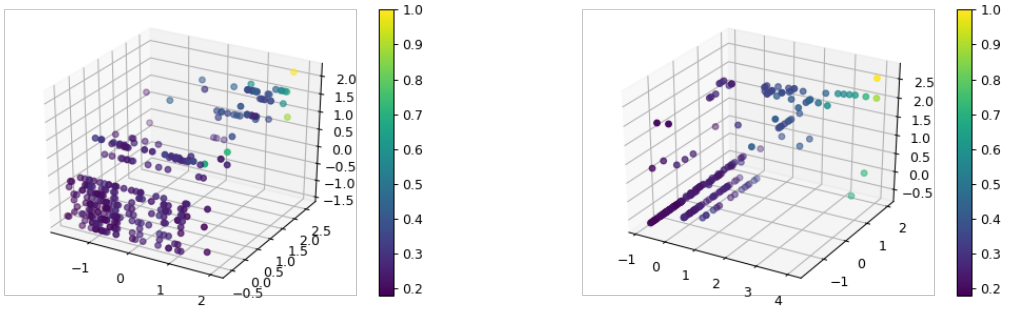
Figura 5: Gràfics sense relacions aparents

S'observa doncs com sembla que només els atributs DateofHire i Position obtenen alguna relació amb combinació dels altres atributs amb el Salary.
Es conclou que per a predir el Salary només és imprescindible els atributs DateofHire i Position.
Es procedeix a estudiar el comportament però en comptes de 2 a 2 fent 3 a 3.

[2]El conjunt de gràfics complets es troba a 9.2

4.4.2 Representació 3 a 3 en R4

Per a poder representar les dades amb 4 atributs (el Salary i 3 més dels 5 rellevants), s'ha decidit posar el valor de Salary com el color del punt. Cada eix representa un dels atributs. Els atributs estan estandaritzats, però el Salary està normalitzat (han sigut dividits pel valor màxim que assoleix el Salary) per a evitar que la diferència de colors sigui poc visible. Una mostra dels resultats obtinguts són[3]:



(a) DateofHire & Department & ManagerName

(b) DateofHire & Position & Department

Figura 6: Gràfics del Salary amb els atributs 3 a 3

Dels gràfics ens adonem que no hi existeix cap relació adicional a les ja trobades amb les representacions del Salary fetes anterioemnt a 2 i 4.4.1, i per això deduïm que no hi haurà una gran millora dels resultats una vegada afegim més atributs per al nostre regressor. Igualment s'estudia si la millora que obtenim és considerable o si pel contrari amb un nombre menor de variables s'obté un resultat prou bo.

[3]El conjunt de gràfics complets es troba a 9.2

5 PCA

Per tal de trobar noves relacions que puguin ser significatives a l'hora de predir el *Salary*, es decideix provar d'examinar les dades rellevants mitjançant una PCA.

Els resultats obtinguts d'aquest analísi és la següent:

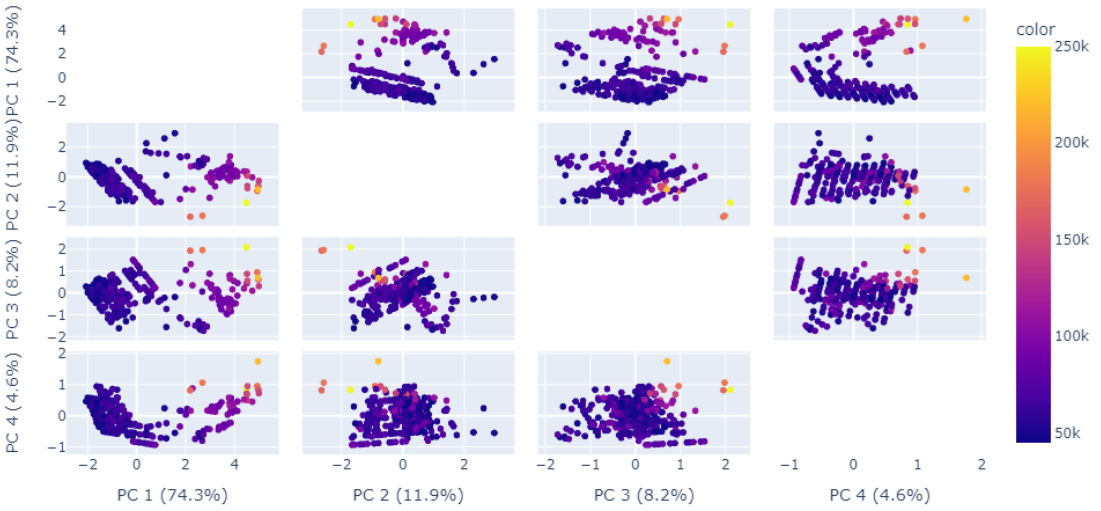


Figura 7: PCA dels atributs rellevant per al Salary

S'observa de l'analísi de la PCA que cap combinació lineal de variables ens dona una bona variança (millor que la que es podria obtenir mitjançant les variables de forma independent).
El millor valor de variança que aconseguim és entre la PC4 i PC1 amb 74.3%. Aquest valor, tot i ser bastant bo, no entra dins del llindar òptim per a fer una PCA (superior o igual al 85%)[4].

Es dedueix, doncs, que es pot obtenir resultats millors utilitzant altres mètodes que no pas una PCA (com per exemple utilitzar transformacions en les dades, fer regressió mitjançant mètodes alternatius, etc.).

[4]El valor del llindar superior a 85% ha sigut extret en la investigació prèvia a l'ús de la PCA on diverses fonts comenten que en cas d'obtenir una molt bona correlació entre algunes dades amb l'atribut objectiu superior a la variança de la PCA és recomanable no invertir temps en desenvolupar l'anàlisi de la PCA i millorar la regressió mitjançant les variables de forma independent

6 Predicció del Salary

Per a predir el *Salary* s'han utilitzat diversos mètodes i transformacions. Primerament, s'ha intentat fer una regressió lineal simple per a totes les dades (incloses les descartades) per si existia alguna relació que havia sigut descuidada.

Posteriorment s'han procedit a estudiar diversos tipus de regresor que s'han cregut útils per a veure amb quin obtenim la millor predicció.

Finalment s'ha procedit a fer un interval de confiança per als regressors que s'han cregut necessaris, per a obtenir la millor precisió possible.

6.1 Regressió Lineal

Per a trobar el millor regresor lineal estudiem, per als atributs rellevants, el seu MSE i el R2 score.

S'utilitzen els resultats obtinguts per comparar-los amb altres regresors lineals.
Per al regresor lineal simple, s'utilitza el mètode *LinearRegressor()* de la llibreria sklearn[5].

IMPORTANT: les dades obtingudes a continuació són mitjançant tots els valors dels que es disposa.

Atribut	**Original**		**Transformada**	
	MSE	R2	MSE	R2
Position	$100 \cdot 10^6$	0.841	$73 \cdot 10^6$	0.816
DateofHire	$414 \cdot 10^6$	0.343	$387 \cdot 10^6$	0.383
Department	$387 \cdot 10^6$	0.387	$333 \cdot 10^6$	0.226
ManagerName	$361 \cdot 10^6$	0.428	$391 \cdot 10^6$	0.353
SpecialProjectsCounts	$468 \cdot 10^6$	0.258	$791 \cdot 10^6$	0.141

Taula 3: MSE i R2 del regressor lineal simple

S'observa com, per a tots els atributs, el MSE és menor quan les dades són estandaritzades, i per això, tot i que el R2 score disminueix significativament en alguns casos, es decideix utilitzar per a tots els regresors l'estandarització de les dades.

IMPORTANT: En aquest cas, com en els següents, les dades estandaritzades són totes menys el *Salary*.

En cas de fer la regressió i validant-la amb conjunts de train i de test, s'obtenen els següents resultats:

Atribut	MSE	R2
Position	$81 \cdot 10^6$	0.857
DateofHire	$272 \cdot 10^6$	0.340
Department	$232 \cdot 10^6$	0.454
ManagerName	$194 \cdot 10^6$	0.539
SpecialProjectsCounts	$293 \cdot 10^6$	0.217

Taula 4: MSE i R2 amb dataset dividit

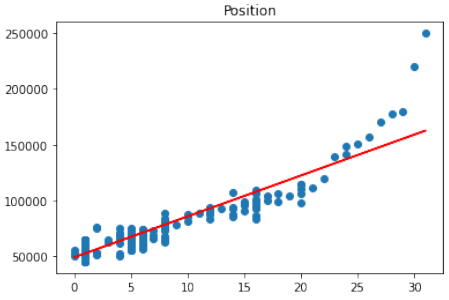


Figura 8: Regressor de Position

Es conclou doncs que, tot i tenir un error inacceptable, el millor regresor lineal simple és mitjançant l'atribut Position[6].
Es dedueix que aquest error és tant elevant a causa de la gran variança del *Salary* entre els treballadors amb *Positions* de molta i poca importància.
Solucionem aquest problema mitjançant transformacions al dataset per, així, obtenir millors precisions.

[5]La informació sobre les llibreries i funcions utilitzades es troba a 2.1
[6]La resta de gràfics obtinguts amb el MSE i R2 corresponents per aquesta anàlisi es troben a 9.3.1

6.2 Regressió lineal amb escala logarítmica

Observem de totes gràfiques generades (com per exemple 2, 4.4.1, 13 i les classificades a 9.3.1) que el valor de l'atribut *Salary* creix de manera exponencial, pel que procedim a fer la regessió lineal de l'apartat anterior (6.1) però aplicant la transformació logarítmica a les dades del *Salary*:

$$\sum_{i}^{n} w_i x_i = \log y \Longrightarrow y = \exp\left(\sum_{i}^{n} w_i x_i\right)$$

Mitjançant aquest nou model recalculem el MSE i el R2 score de les variables més rellevants[7].

Atribut	MSE	R2
Position	0.010	0.862
DateofHire	0.045	0.404
Department	0.042	0.440
ManagerName	0.036	0.520
SpecialProjectsCounts	0.051	0.321

Taula 5: MSE i R2 amb dataset dividit

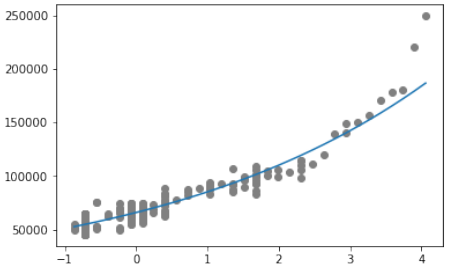


Figura 9: Regressor de Position

S'observa una millora notòria tant en l'error comès com en l'R2 score a l'hora de predir el valor del *Salary*.

Com totes les variables tenen relació exponencial amb *Salary*, aplicarem la transformació logarítmica a tots els nostres regresors per així obtenir una millor predicció.

A partir d'ara el valor amb el que compararem els resultats següents serà:

- MSE[8]: 0.01
- R2 score: 0.862

Per mirar de millorar la predicció, provem altres tipus de regresors, com la *BayesianRidge* o *Lasso*.

[7] El conjunt sencer dels gràfics es troben a 9.3.2

[8] **IMPORTANT:** El càlcul de l'error y R2 score s'ha efectuat amb el valor predit, el logaritme de y, no amb e^y, on y és el valor predit.

6.3 Regressió Lineal amb BayessianRidge

Procedim a estudiar els resultats obtinguts per al mètode de regressió lineal del *BayesianRidge()*[9] i comparant-los amb els valors anteriorment obtinguts.

Atribut	MSE	R2
Position	0.010	0.862
DateofHire	0.045	0.404
Department	0.042	0.440
ManagerName	0.036	0.520
SpecialProjectsCounts	0.051	0.321

Taula 6: MSE i R2 amb dataset dividit

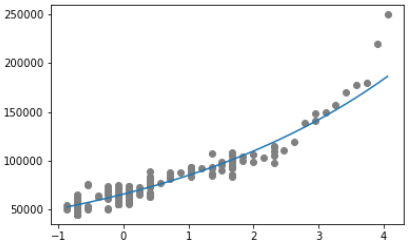


Figura 10: Regressor de Position

S'observa dels resultats obtinguts que no existeix cap diferència (en els primers tres decimals) en les dades obtingudes amb el regresor lineal simple.
Es conclou doncs que el model anterior es manté com el millor per a predir el *Salary*.

6.4 Regressió lineal amb Lasso

Es procedeixen a estudiar els resultats obtinguts per al mètode de regressió lineal del *Lasso()*[10] i comparant-los amb els valors anteriorment obtinguts.

Atribut	MSE	R2
Position	0.010	0.862
DateofHire	0.045	0.404
Department	0.042	0.440
ManagerName	0.036	0.520
SpecialProjectsCounts	0.051	0.321

Taula 7: MSE i R2 amb dataset dividit

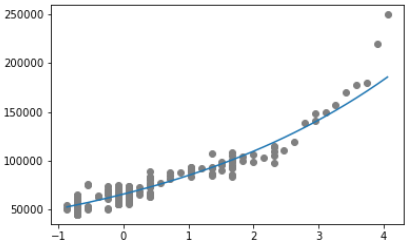


Figura 11: Regressor de Position

S'observa dels resultats obtinguts que no existeix cap diferència (en els primers 3 decimals) en les dades obtingudes amb el regresor lineal simple.
Es conclou doncs que el model anterior es manté com el millor per a predir el *Salary*.

Tot i això, s'ha decidit que és millor utilitzar el mètode *Lasso()* ja que permet escollir el grau de precisió de la recta de regressió mitjançant el paràmetre a[11] i calcula les rectes de regressió més ràpid que el regresor simple.

[9] S'ha introduït com a paràmetre d'entrada $t = 1 \cdot 10^{-12}$, on t és la tolerància.
[10] S'ha introduït com a paràmetre d'entrada $a = 0.001$, on a és el terme de regularització.
[11] Mencionar que s'ha estudiat que a partir del valor $a \leq 0.01$, el regressor només estabilitza els valors, no augmenta significativament la precisió.

6.5 Regressió multilineal

Es procedeix ara a intentar millorar la precisió del regresor afegint-li major complexitat. S'analitzen ara, totes les possibles combinacions lineals del regressor multilineal simple de la llibreria sklearn (*LinearRegressor()*) amb totes les variables rellevants[12].

Els resultats més significatius obtinguts[13], són els següents:

	Atributs	Valor
Millor R2	Position, DateofHire, Department	0.827
Millor MSE	Position, DateofHire, ManagerName	$102 \cdot 10^6$

Taula 8: Millors valors de MSE i R2 obtinuts en la regressió multilineal

Abans d'explicar els resultats cal comentar que els estudis del regresor multilineal s'han dut a terme amb els datatset separat en train i test i que s'han repetit els càlculs 1000 vegades per cada conjunt d'atributs trobats per a evitar valors sobre o infraestimats deguts a fenòmens aleatoris i evitar també possibles overfittings del model.

S'observa com tant l'error MSE com el R2 score no milloren significativament respecte a la millor combinació trobada fins al moment.
Tot i això, es pot observar una certa millora respecte a la mitjana de valors obtiguts inicialment.

Es conclou que el mètode de regressió multilineal pot superar en precissió al millor regresor trobat fins ara, per això es decideix aplicar una transformació al datatset per així intentar millorar la predicció.

6.6 Regressió multilineal amb escala logarítmica

Repetim ara la mateixa transformació que en l'apartat 6.2 per a veure si el regresor multilineal simple és capaç de millorar l'error comès.
Els resultats més significatius obtinguts[14], són els següents:

	Atributs	Valor
Millor R2	Position, DateofHire, ManagerName	0.857
Millor MSE	Position, DateofHire	0.010

Taula 9: Millors valors de MSE i R2 obtinuts en la regressió multilineal

S'observa com tant l'error com el R2 score no arriben a millorar, però si quasi igualar completament, els valors obtinguts pel millor regresor fins ara.
Com hi ha hagut una millora molt significativa respecte al regressor multilineal simple, procedim a intentar millorar els valors obtinguts pel regresor estudiant diferents mètodes.

Cal mencionar que ha aparegut una discrepància entre els atributs trobats a l'apartat anterior i aquest, aquest fenomen segurament és a causa de que els atributs trobats aquesta vegada són en els que la tendència exponencial està més present.

[12] Per a fer aquestes combinacions s'ha utilitzat el mètode explicat a 2.2.7
[13] La resta de resultats obtinguts es troben a 9.3.5
[14] La resta de resultats obtinguts es troben a 9.3.6

6.7 Regressió multilineal amb Lasso

Estudiem ara els millors valors assolits de totes les combinacions d'atributs rellevants mitjançant el mètode de regressió multilineal *Lasso()*[15].
Com els resultats obtinguts per la transformació han millorat el regresor lineal, utilitzem la transformació per al regresor multilineal amb mètode *Lasso()*.
Els resultats més significatius obtinguts[16], són els següents:

	Atributs	Valor
Millor R2	Position, DateofHire, ManagerNamee	0.716
Millor MSE	Position, DateofHire, Department, ManagerName	0.020

Taula 10: Millors valors de MSE i R2 obtinuts en la regressió multilineal Lasso

S'observa com el regresor multilineal amb mètode Lasso no millora ni en R2 ni en MSE al regresor multilineal simple, per la qual cosa s'intenta millorar la precisió amb un altre mètode.

6.8 Regressió multilineal amb BayessianRidge

Estudiem ara els millors valors obtinguts de totes les combinacions d'atributs rellevants mitjançant el mètode de regressió multilineal *BayessianRidge()*[17].
Com els resultats obtinguts per la transformació han millorat el regresor lineal, utilitzem la transformació per al regresor multilineal amb mètode *BayessianRidge()*.
Els resultats més significatius obtinguts[18], són els següents:

	Atributs	Valor
Millor R2	Position, DateofHire	0.857
Millor MSE	Position, DateofHire	0.010

Taula 11: Millors valors de MSE i R2 obtinuts en la tilineal BayessianRidge

S'observa com, mitjançant el mètode *BayessianRidge()*, els millors valors de MSE i R2 score coincideixen en els dos casos amb el mateix conjunt d'atributs.

Comentar també que són els dos millors atributs dels rellevants (com ja hem estudiat a 4.4.1).

Deduïm d'això que els mètodes multilineals intenten assemblar-se al mètode lineal dels atributs Position i DateofHire (sobretot al del Position) incloent-ne algun d'aquests atributs en el seu model de regressió lineal per així asegurar-se tenir una bona predicció.

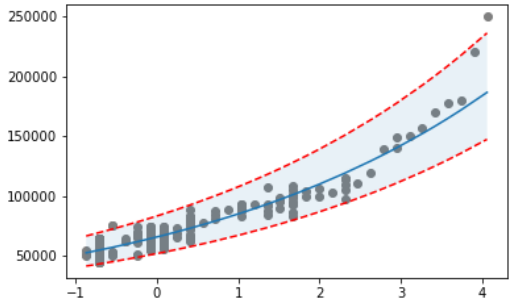
Com que cap dels regressors ha siguit capaç de superar en R2 score o en MSE al regresor de l'apartat 6.4 (Regresor lineal amb mètode Lasso basat en l'aribut Position), es finalitza l'estudi analitzant i polint aquest regresor lineal.

[15] Igual que en seu anàleg lineal el valor del paràmetre $a = 0.01$
[16] La resta de resultats obtinguts es troben a 9.3.7
[17] Igual que en seu anàleg lineal el valor del paràmetre $t = 10^{-12}$
[18] La resta de resultats obtinguts es troben a 9.3.8

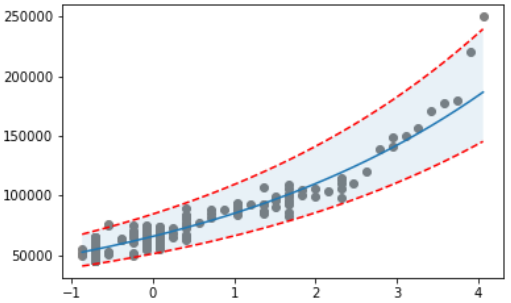
6.9 Versió final del regresor i interval de confiança

Per finalitzar l'estudi dels regresors lineals, es procedeix a estudiar quin és el marge d'error del nostre regresor i estudiar el seu interval de confiança.

Calculant l'interval de confiança del 75%, s'obté el següent llindar:



Regressió amb datatset separat en train i test



Regressió sense datatset separat en train i test

Figura 12: Gràfic de la regressió lineal amb interval de confiança del 75%

S'observa com els intervals de confiança tant amb el datatset separat en train i test com sense separar-lo són pràcticament idèntics.

S'observa com existeixen dos valors (punts) que podrien ser considerats outliners. Aquests es troben a la part superior dreta del gràfic i fa que el regresor augmenti l'error.
Els dos outliners representen el *Salary* dels directors de l'empresa (els CEO).

Com que els sous dels directors de l'empresa és un tema subjecte a canvis, té una dispersió respecte a la resta de dades molt gran i és poc probable de ser necessària de catalogar (ja que en aquest cas es tracta de dos subjectes respecte als 309 restants) es decideix tractar-los com a outliners i eliminar-los per a fer la regresió i l'interval de confiança.

Mostrem ara el regressor sense els outliners i augmentat l'interval de confiança al 95%.

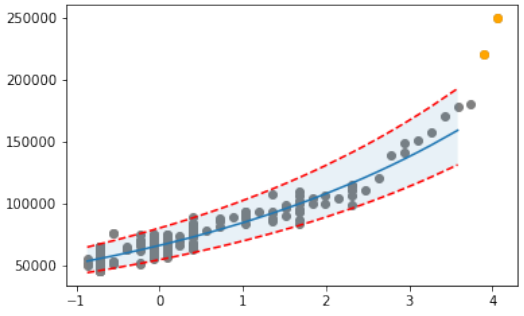


Figura 13: Regressor sense outliners amb interval de confiança del 95%

S'observen en taronja els outliners que no s'han tingut en compte per a fer la regressió lineal, que, com són els últims valors de la gràfica, fa que la regressió s'estanqui quan s'arriba a ells.
Tot i això, s'observa com la tendència de la gràfica s'ajusta molt millor als punts que amb els outliners.
També s'observa clarament que tot i haver augmentat l'interval de confiança, aquest sembla haver-se reduït.
Els resultats que s'obtenen (de MSE i R2 score) són 0.007 i 0.880 (respectivament).

7 Resolució de les preguntes

7.1 Apartat C

1. **Quin és el tipus de cada atribut?**
   Es pot veure el tipus de cada atribut del dataset a la secció 3.3 explicada anteriorment. On indica els tres tipus d'atributs que hi ha.

2. **Quins atributs tenen una distribució Gaussiana?**
   Mitjançant un estudi visual de les gràfiques ,tant a la secció 4.2 com a la 9, s'ha observat que cap dels atributs dels quals disposa el dataset es pot aproximar per una distribució Gaussiana, tot i que, com s'ha explicat anteriorment, sí que s'observen algunes distribucions uniformes i logarítmiques, entre d'altres.

3. **Quin és l'atribut objectiu? Per què?**
   S'ha decidit que l'atribut a predir sigui *Salary*, ja que en fer l'estudi de les correlacions entre els atributs mostrat a l'apartat 4.3, es va veure que és l'atribut amb més correlacions rellevants (sense tenir poca dispersió). Recalcar també que ens semblava dels més interessants a predir, ja que es treballa en el context d'una empresa fictícia.

7.2 Apartat B

1. **Quins són els atributs més importants per fer una bona predicció?**
   S'observa dels gràfics que les millors regressions lineals són amb les variables ja catalogades com a rellevants en l'apartat C: *DateofHire*, *Department*, *Position*, *ManagerName* i *SpecialProjectsCount* (ja que són les que presenten un menor MSE i major R2 score).

2. **Amb quin atribut s'assoleix un MSE menor?**
   La millor dada obtinguda pel que fa a la relació MSE-R2 és la de *Position*, que, sense estandaritzar, té un MSE = 100190785.86004272 i R2 = 0.8411741038500034. Estandaritzat, l'MSE es redueix a $73 \cdot 10^6$.
   Cal recalcar que en la regressió amb les dades normalitzades s'obtenen un MSE i un R2 bastant més petit i gran respectivament, de manera que els valors que obtenim són més fàcils de tractar.

3. **Quina correlació hi ha entre els atributs de la vostra base de dades?**
   Les correlacions de l'atribut *Salary* amb els cinc atributs més rellevants són les següents:

Atribut	**Correlació**
Position	0.917
DateofHire	0.586
Department	0.622
ManagerName	0.654
SpecialProjectsCount	0.508

4. **Com influeix la normalització en la regressió?**
   En el cas del nostre dataset, ens permet tenir uns valors més "manejables", facilitant la visualització de les dades, les relacions entre elles i detectar les distribucions que segueixen.

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

   Veiem que es redueix l'MSE significativament, mentre que augmenta l'R2 score. Observem que inicialment, usant els 5 atributs més rellevants, s'aconsegueix un error de $75 \cdot 10^5$, mentre que usant 3 dels atributs (Position, DateofHire i ManagerName), es redueix a $65 \cdot 10^5$. Per altra banda, l'R2 score augmenta de 0.850 a 0.857. D'aquesta manera millorem la precisió amb la que el regresor predirà el valor objectiu *Salary*.

6. **Si s'aplica un PCA, a quants components es redueix l'espai? Per què?**
   En fer l'estudi aplicant un PCA, com s'ha explicat a la secció 9.2, observem que la màxima variança que són capaços d'explicar els components principals és de tan sols un 74.3%. Decidim així que de totes les combinacions possibles amb els cinc atributs més rellevants, la millor és utilitzant els atributs 'Position', 'DateofHire' i 'ManagerName', ja que és la que presenta una millor relació MSE-R2 score.

7.3 Apartat A

7.3.1 Explicació del regressor lineal programat

S'ha programat un regresor lineal que està capacitat per a calcular relacions lineals i multilineals[19].

Si bé és funcional i ens permet calcular paràmetres de funcions concretes que poden ser útils per al nostre regresor, el cas és que a causa de com esta organitzat el nostre datatset, és completament inecessari i (inclòs) obté major error que amb els mètodes utilitzats de la llibreria *sklearn*.

Per aquest motiu no s'ha utilitzat el nostre regresor personalitzat i s'ha prioritzat millorar els regresors dels que ja hi disposàvem.

Per a observar alguns dels resultats obtinguts pel nostre regresor amb datasets aleatòris accedir a 9.4.

[19]Per a obtenir més informació sobre el nostre regressor accedir a les últimes cel·les del codi adjuntat conjuntament amb aquesta memòria.

8 Anàlisi i conclusions

Resumim els resultats obtinguts del nostre estudi:

- Com que el dataset ha sigut creat de forma fictícia i no conté informació real, hi apareixen moltes incongruències i dificultats a l'hora d'estimar els valors del *Salary*, per aquests motius mètodes com PCA no són prou bons trobant relacions amb combinacions d'atributs.
- Per solucionar el problema anterior s'ha hagut d'estudiar a fons el dataset per a trobar correlacions entre variables i millorar-les per a obtenir bones prediccions. S'ha conclòs que els millors atributs per a fer la regressió eren *Position*, *DateofHire*, *Department*, *SpecialProyectsCount* i *ManagerName*.
- Per millorar la predicció dels regresors s'ha optat per fer una transformació logarítmica de les dades i calcular la recta i les prediccions amb aquesta transformació.
- S'han provat diferents tipus de regresors (tant lineals com multilineals) per a poder trobar la millor regressió. Es conclou que el millor regresor és el Lasso lineal amb la transformada logarítmica predint el *Salary* mitjançant l'atribut *Position*.
- S'ha afegit un interval de confiança del 95% de confiança eliminant els outliners més significatius del conjunt de dades per, així, millorar encara més la predicció.

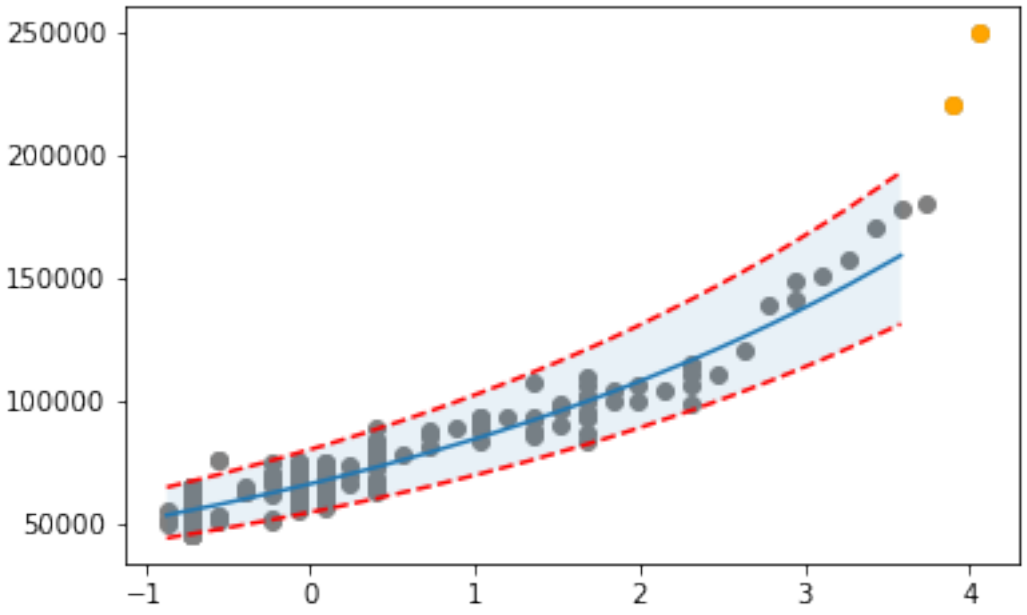


Figura 14: Gràfic del mètode Lasso, amb transformació logarítmica sense outliners i amb variable independent estandaritzada, amb interval de confiança del 95%.

El nostre regesor per a predir el *Salary* en funció de l'atribut *Position* és:

$$\hat{y} = e^{0.24936 \cdot x + 11.095}$$

On $\hat{y}$ és la predicció del *Salary* i x és el valor de l'atribut *Position*[20] després de normalitzar-lo mitjançant la funció 2.2.3.

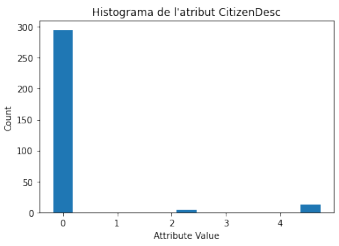
[20] Per a obtenir el valor de la posició es pot utilitzar el diccionari que es genera en la funció 2.2.2

Sabent que un valor d'1 indica un ajust perfecte i, per tant, un model molt fiable per a predir, mentre que un valor de 0 indicaria que el model no aconsegueix ajustar les dades en absolut, assegurem que el nostre regresor funciona de manera correcta comparant la correlació *Salary-Position* (0.92), amb l'R2 Score màxim aconseguit, 0.917.

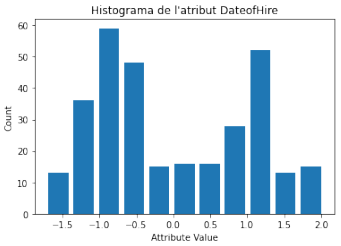
9 Annex

9.1 Histogrames

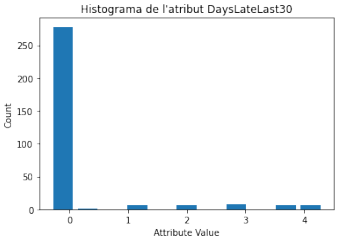
S'adjunten a continuació els histogrames restants mencionats i explicats anteriorment a la secció 4.2.



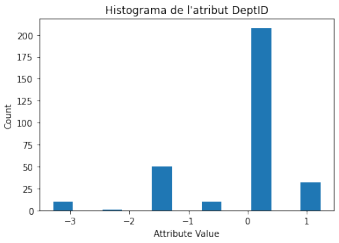
CitizenDesc



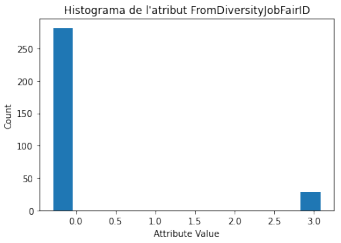
DateofHire



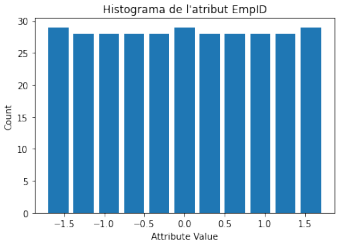
DaysLateLast30



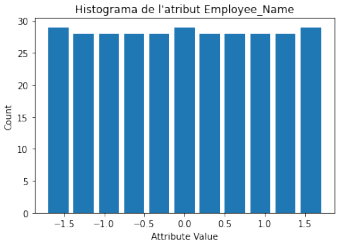
DeptID



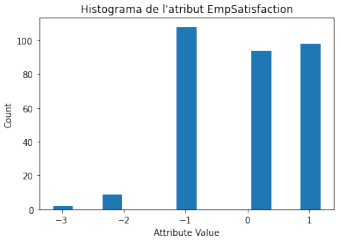
FromDiversityJobFairID



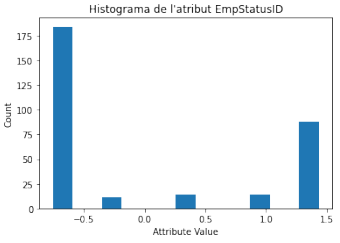
EmpID



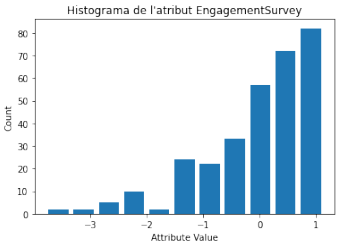
Employee_Name



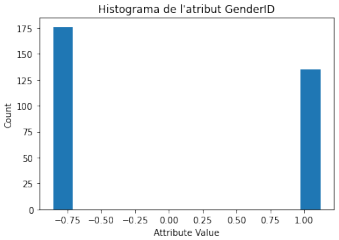
EmpSatisfaction



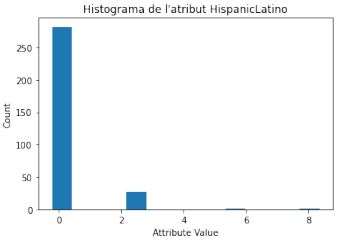
EmpStatusID



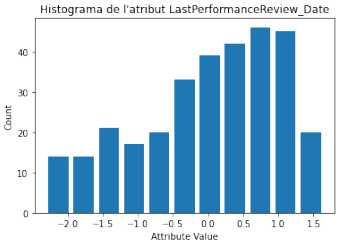
EngagementSurvey



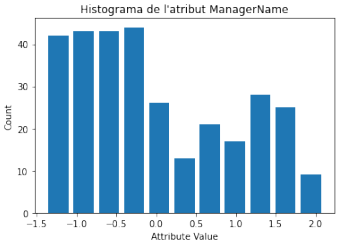
GenderID



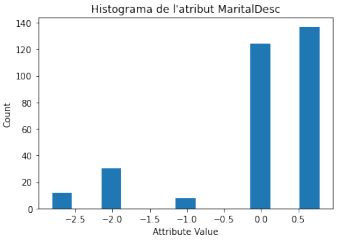
HispanicLatino



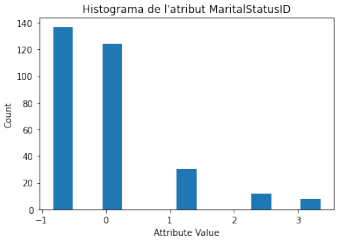
LastPerformanceReview



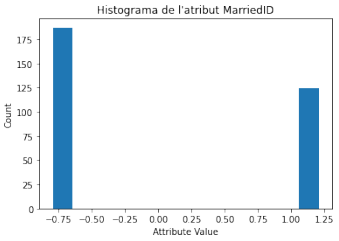
ManagerName



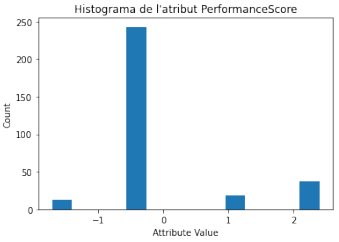
MaritalDesc



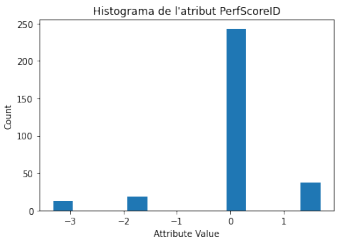
MaritalStatusID



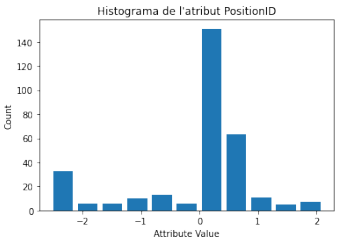
MarriedID



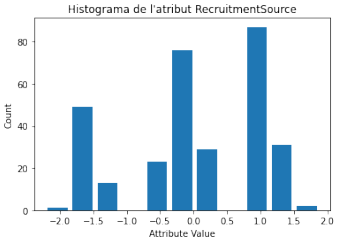
PerformanceScore



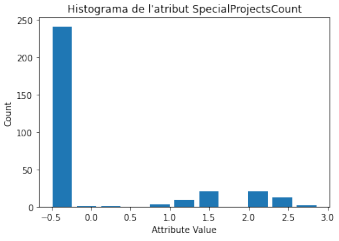
PerfScoreID



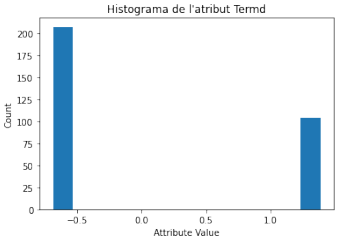
PositionID



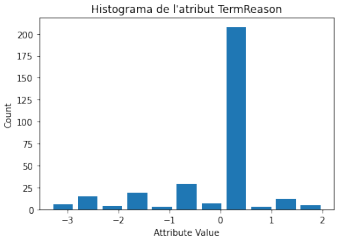
RecruitmentSource



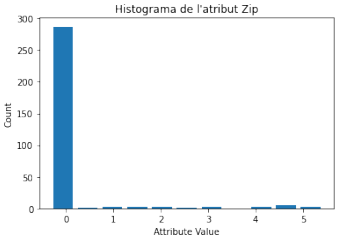
SpecialProjectsCount



Termd



TermReason



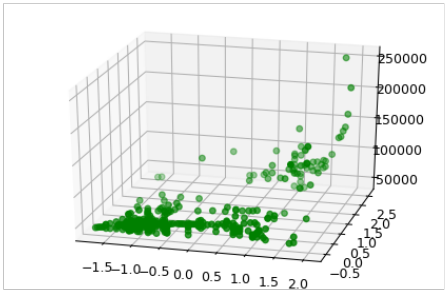
Zip

9.2 PCA

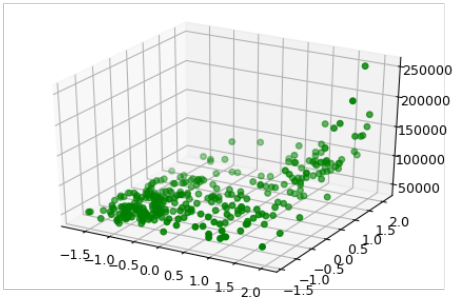
9.2.1 Representació 2 a 2 en R3

Es mostra a continuació el conjunt total de gràfics que representen el Salary respecte a 2 atributs.

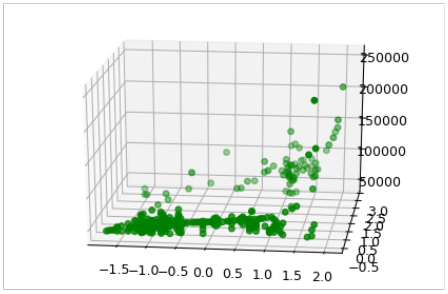
IMPORTANT: No s'ha vist necessària la creació de més gràfics amb l'ús d'altres atributs, ja que els representats són els que més relació tenen amb l'objectiu *Salary*, per la qual s'analitza la tendència del salari usant únicament els més rellevants.



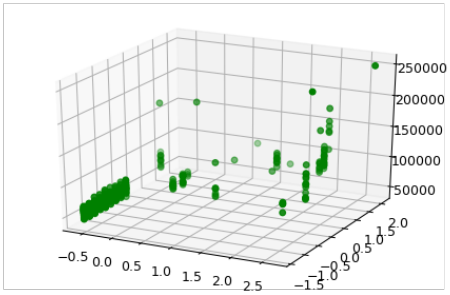
DateofHire & Department



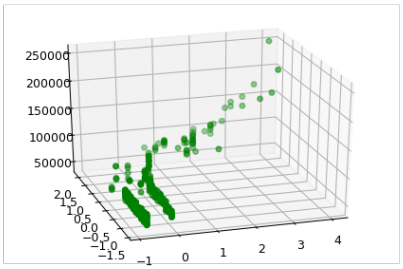
DateofHire & ManagerName



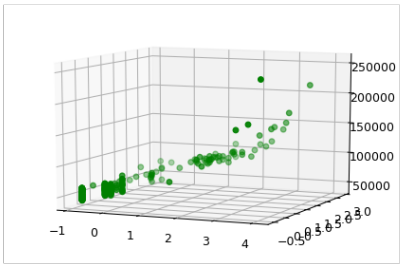
DateofHire & SpecialProjectsCount



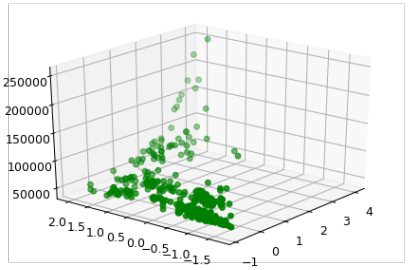
Department & ManagerName



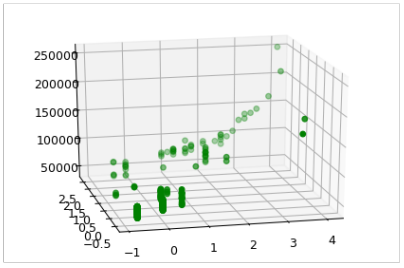
Position & ManagerName



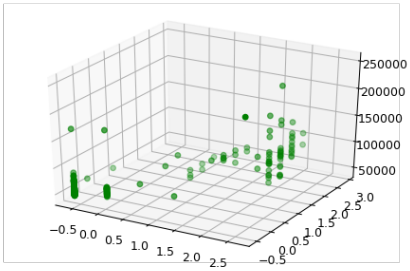
Position & SpecialProjectsCount



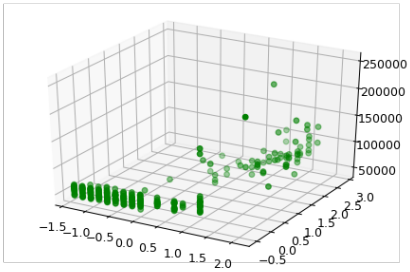
Position & DateofHire



Position & Department



Department & SpecialProjectsCount

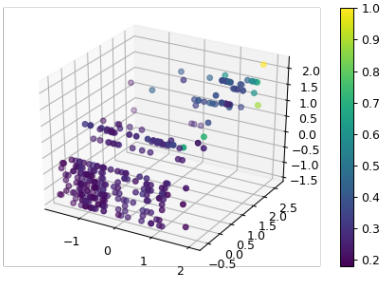


ManagerName & SpecialProjectsCount

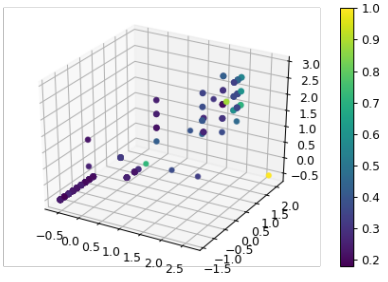
9.2.2 Representació 3 a 3 en R4

Es mostra a continuació el conjunt total de gràfics que representen el Salary respecte a 3 atributs.

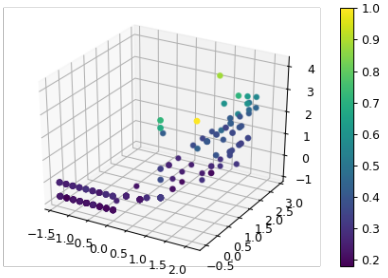
IMPORTANT: No s'ha vist necessària la creació de més gràfics amb l'ús d'altres atributs, ja que els representats són els que més relació tenen amb l'objectiu *Salary*, per la qual cosa s'analitza la tendència del salari usant únicament els més rellevants.



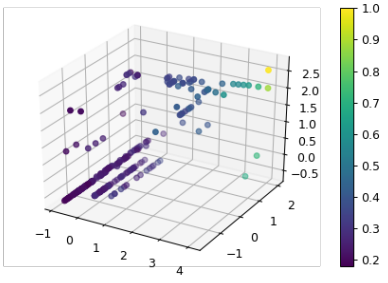
Position & ManagerName



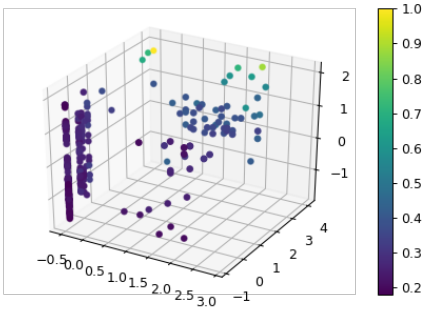
Position & SpecialProjectsCount



Position & DateofHire



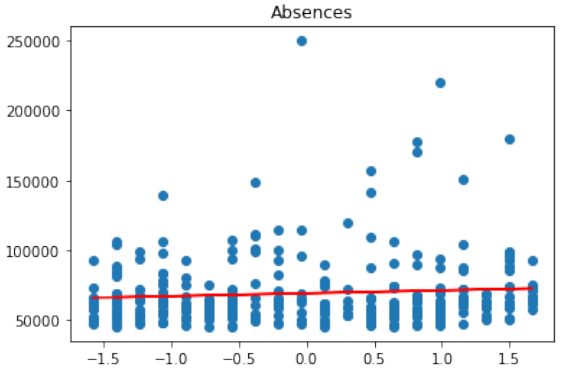
Position & Department



Department & SpecialProjectsCount

9.3 Regressions

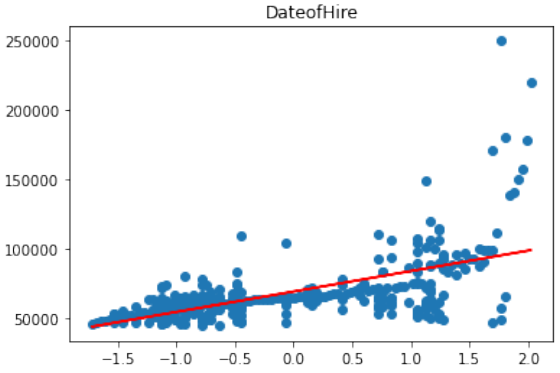
9.3.1 Regressor lineal simple



R2: 0.006786819677949918

MSE: 626540202.0400987

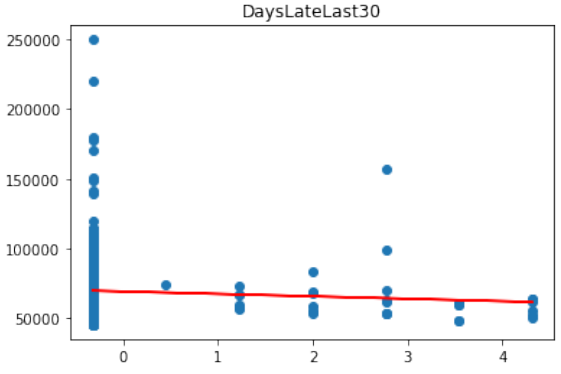
Absences



R2: 0.3434337383583841

MSE: 414176096.7049033

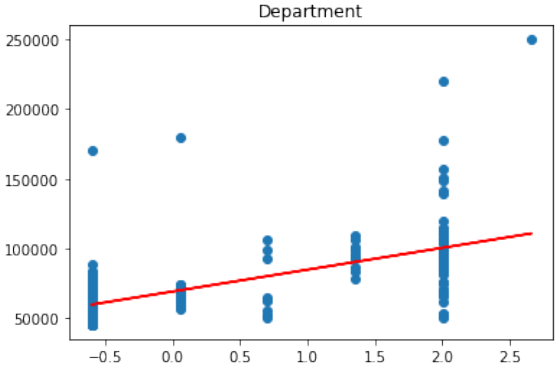
DateofHire



R2: 0.004822283426795915

MSE: 627779473.6930523

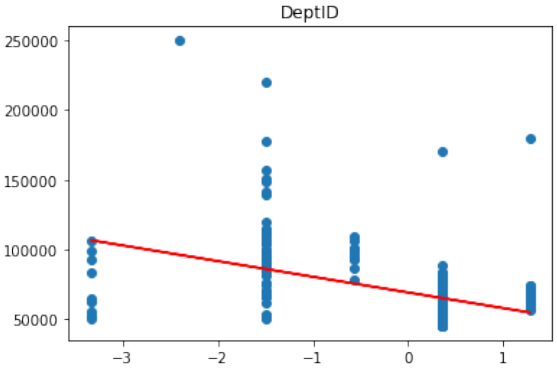
DaysLate30



R2: 0.38665722313325035

MSE: 386909794.34377193

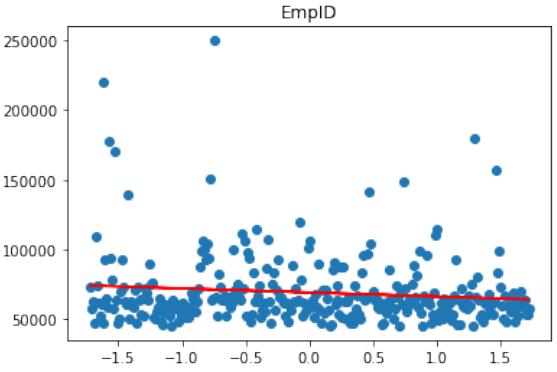
Department



R2: 0.20082220065289724

MSE: 504138517.07700574

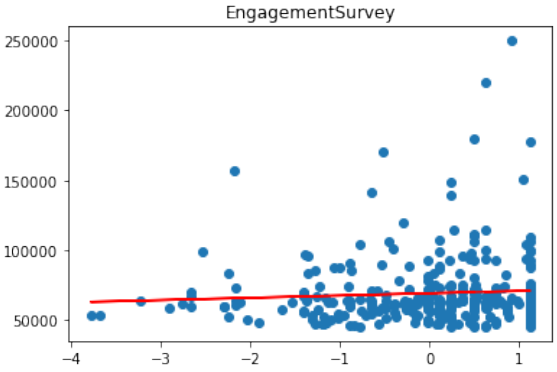
DeptID



R2: 0.013298487327952357

MSE: 622432502.2573311

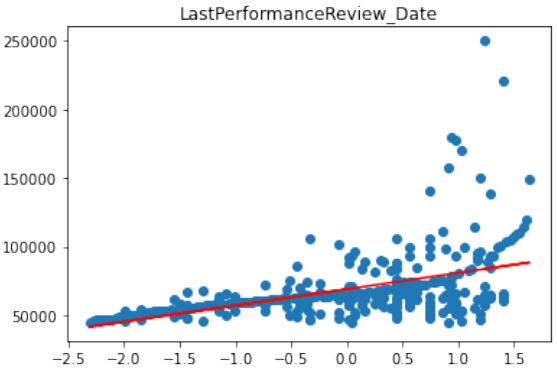
EmpID



R2: 0.004220590387326029

MSE: 628159034.582883

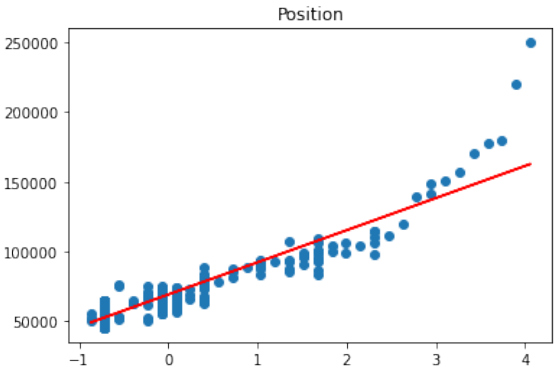
EngagementSurvey



R2: 0.22248708938734219

MSE: 490471840.0394999

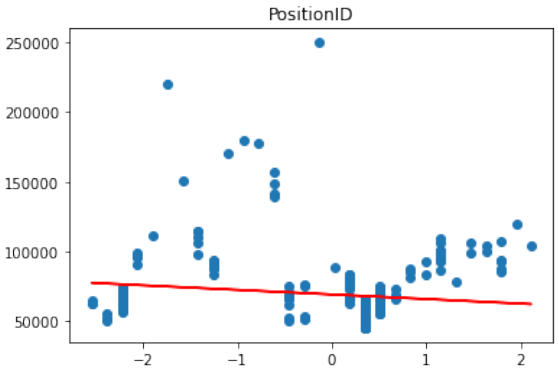
LastPermanceReviewDate



R2: 0.8411741038500034

MSE: 100190785.86004275

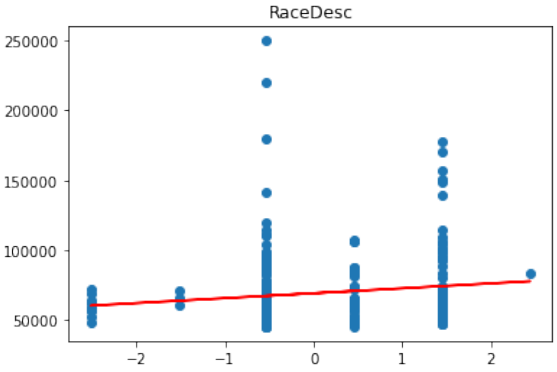
Position



R2: 0.017046821513222787

MSE: 620067972.5629187

PositionID



R2: 0.019675286795809654

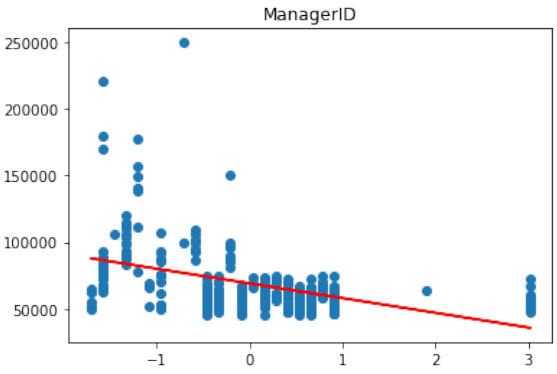
MSE: 618409880.2199701

RaceDesc

R2: 0.0608382328145336

MSE: 592443409.9534619

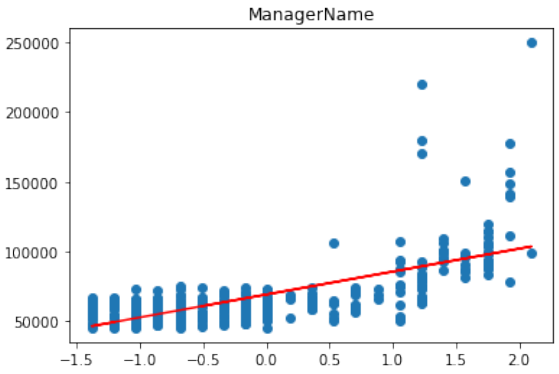
RecruitmentSource



R2: 0.19326151581948225

MSE: 508907959.42527676

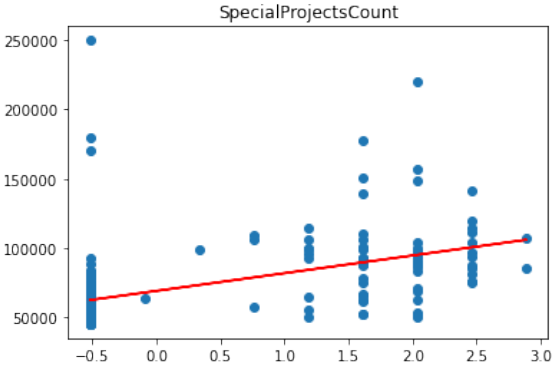
ManagerID



R2: 0.42755630898352837

MSE: 361109772.7375548

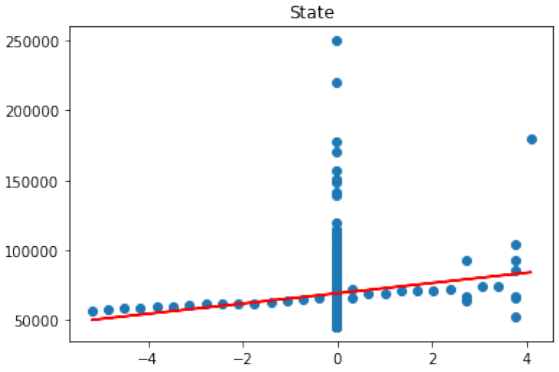
ManagerName



R2: 0.25840233848767225

MSE: 467815729.6762194

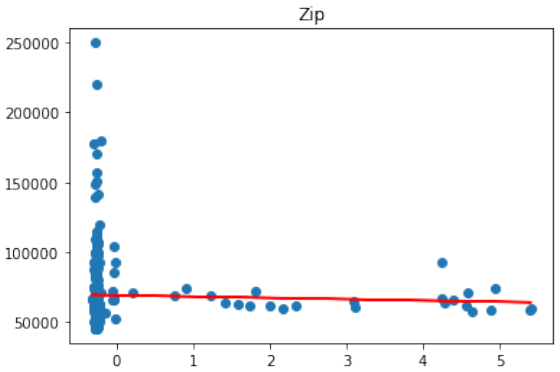
SpecialProjectsCount



R2: 0.021389366583724256

MSE: 617328601.8822374

State

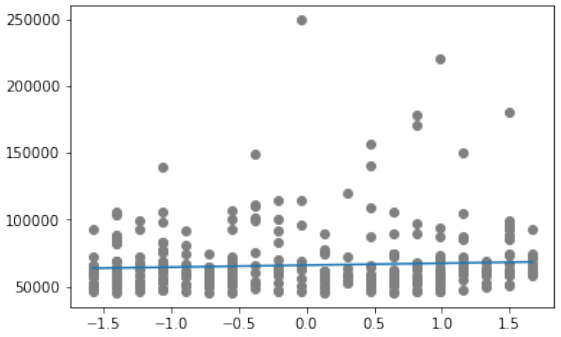


R2: 0.0013869577828778956

MSE: 629946550.878149

Zip

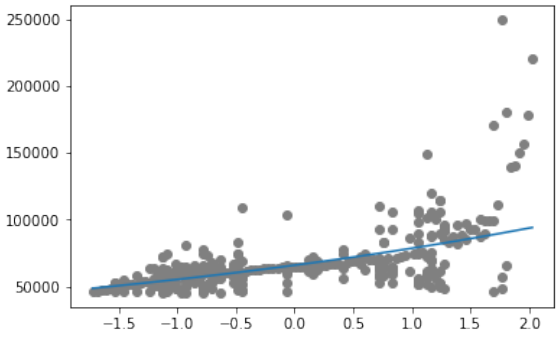
9.3.2 Regressor lineal simple amb la transformació logarítmica



R2: 0.006470672577135184

MSE: 0.07525780628277791

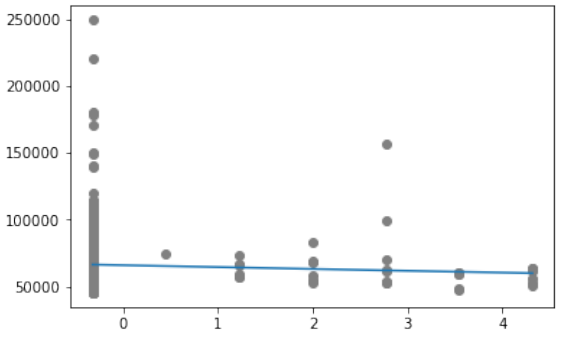
Absences



R2: 0.40479342317953737

MSE: 0.045085675903278666

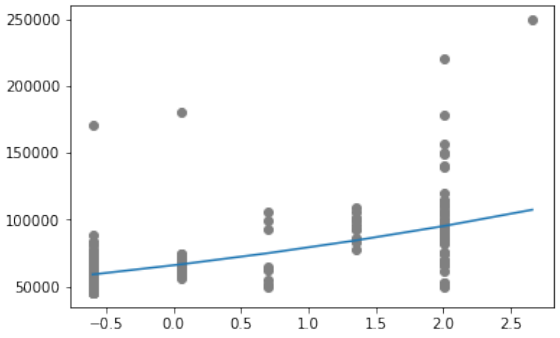
DateofHire



R2: 0.0062521739831206125

MSE: 0.07527435710257503

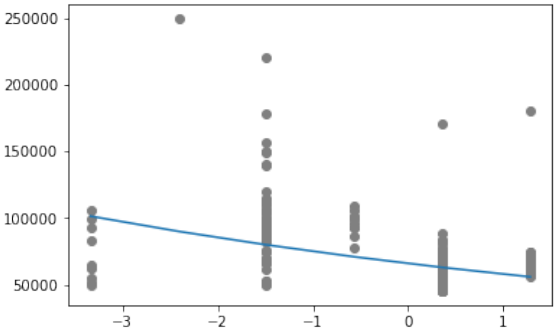
DaysLate30



R2: 0.440123328940548 45

MSE: 0.04240950809387157

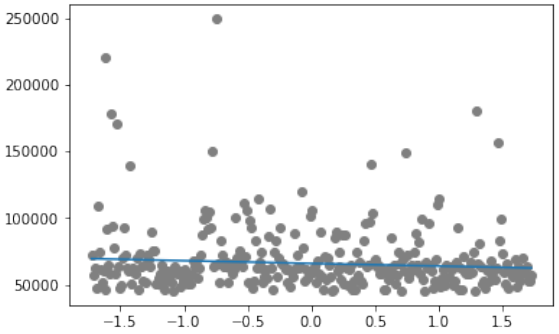
Department



R2: 0.21815180872131623

MSE: 0.059223394919220095

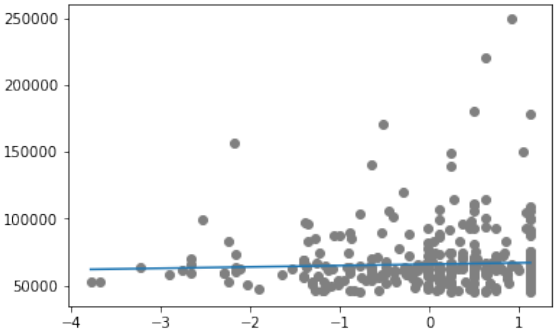
DeptID



R2: 0.012349036643331424

MSE: 0.07481253227632162

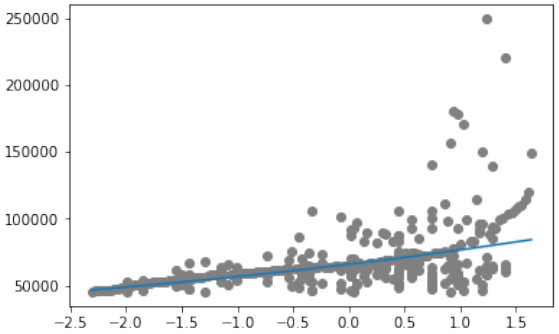
EmpID



R2: 0.0032650201002807355

MSE: 0.07550062787490819

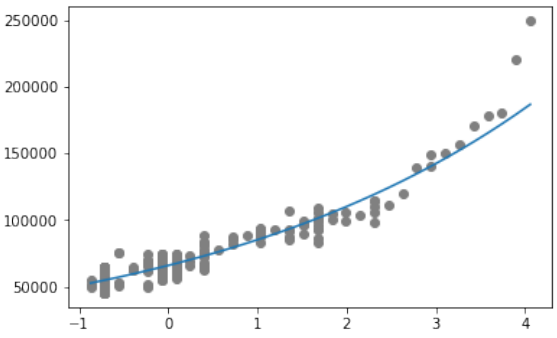
EngagementSurvey



R2: 0.29681793627277564

MSE: 0.053264597302605535

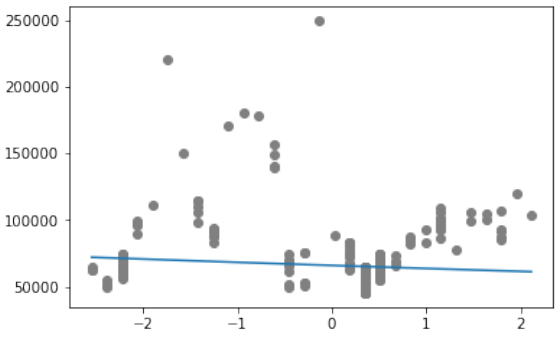
LastPermanceReviewDate



R2: 0.8621574434540897

MSE: 0.010441290590750434

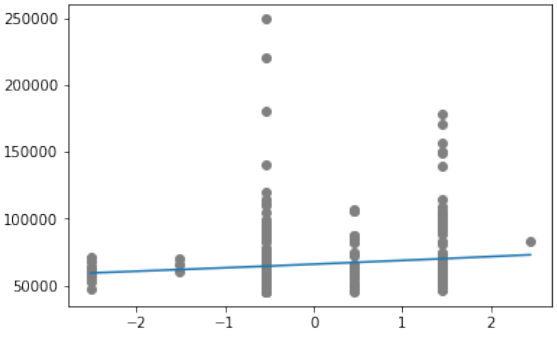
Position



R2: 0.016026158865804163

MSE: 0.07453399781915082

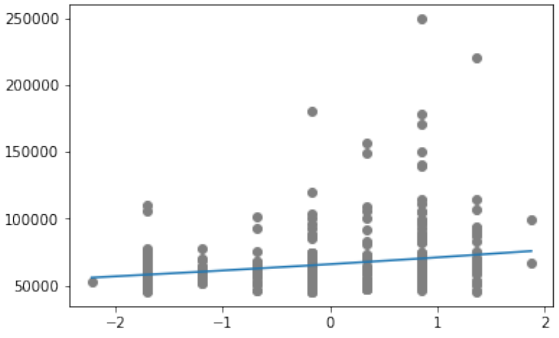
PositionID



R2: 0.022762636952561865

MSE: 0.0740237234378234

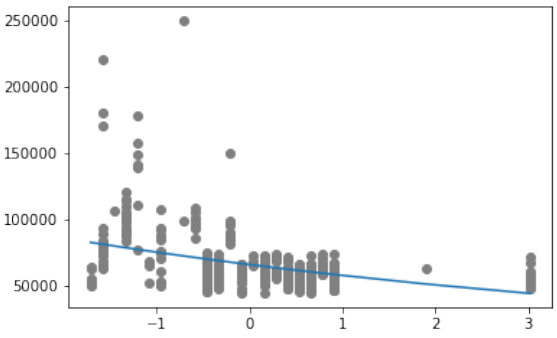
RaceDesc



R2: 0.07227842666435225

MSE: 0.07027300405067168

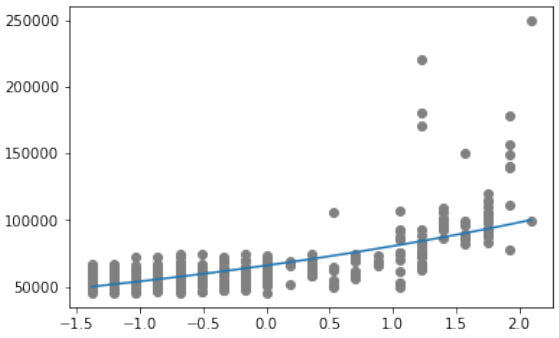
RecruitmentSource



R2: 0.2290743744149576

MSE: 0.05839603299804237

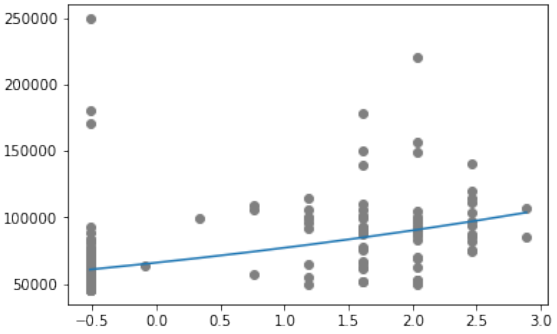
ManagerID



R2: 0.5200852910719191

MSE: 0.03635255366889941

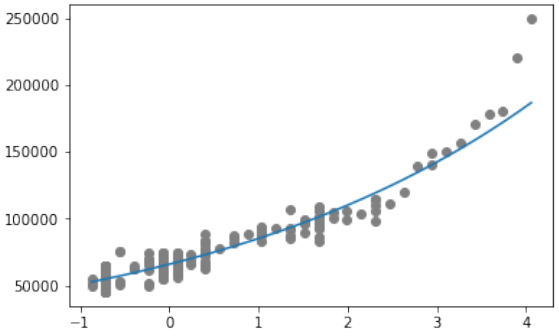
ManagerName



R2: 0.32182279950902926

MSE: 0.05137053026138132

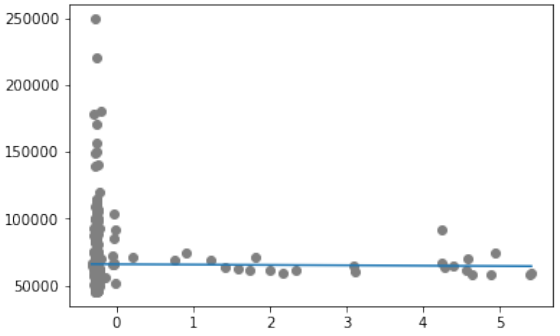
SpecialProjectsCount



R2: 0.026212294653237267

MSE: 0.07376241895106724

State

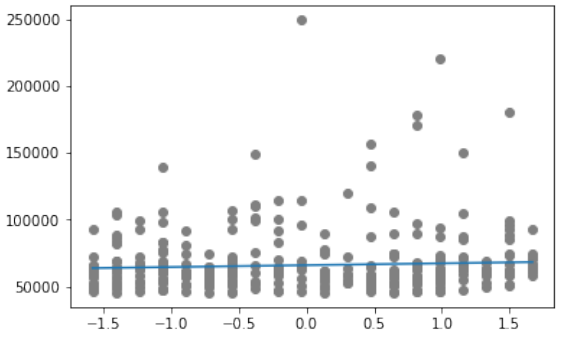


R2: 0.00022770507600045065

MSE: 0.0757306982506967

Zip

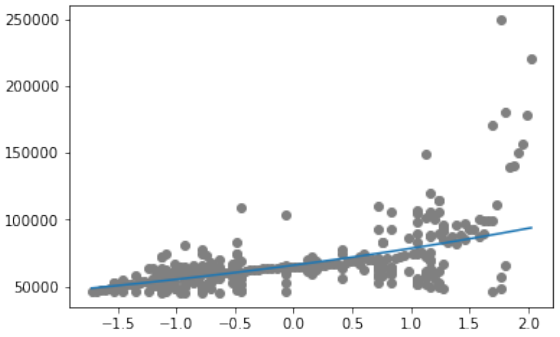
9.3.3 Lasso amb la transformació logarítmica



R2: 0.0064574283130519605

MSE: 0.07525880950858431

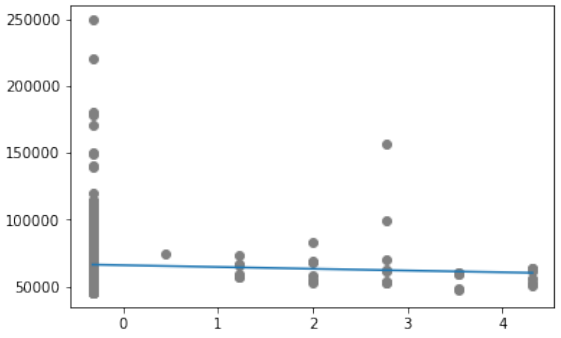
Absences



R2: 0.40478017891545426

MSE: 0.04508667912908511

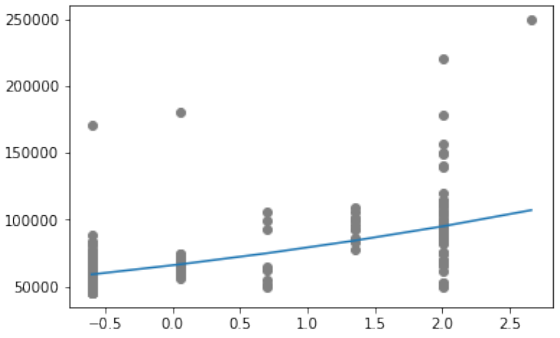
DateofHire



R2: 0.006238929719037611

MSE: 0.07527536032838153

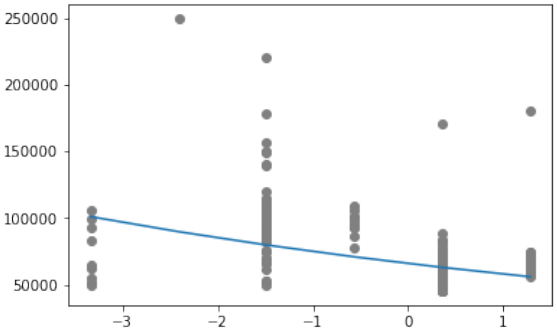
DaysLate30



R2: 0.4401100846764652

MSE: 0.042410511319678004

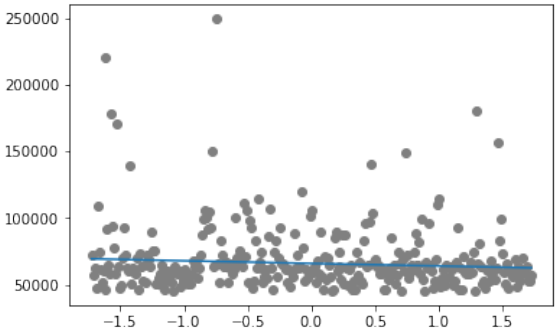
Department



R2: 0.21813856445723245

MSE: 0.0592243981450266

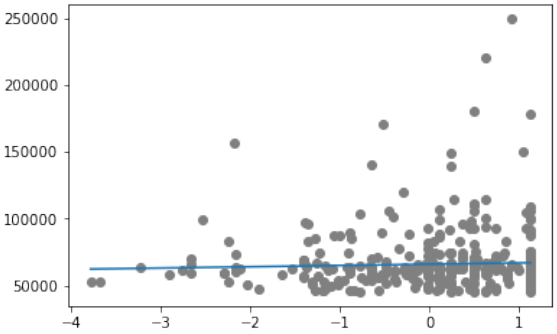
DeptID



R2: 0.012335792379248534

MSE: 0.07481353550212792

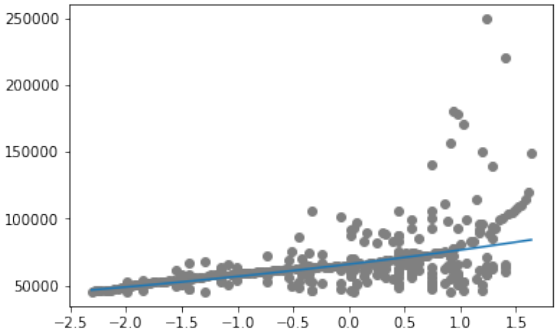
EmpID



R2: 0.0032517758361974014

MSE: 0.07550163110071467

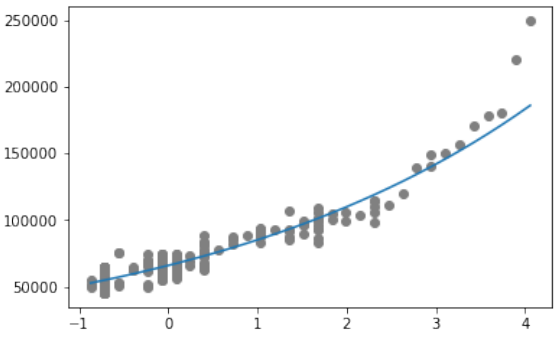
EngagementSurvey



R2: 0.2968046920086922

MSE: 0.053265600528411984

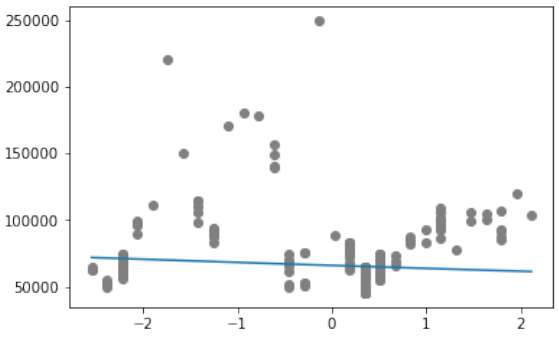
LastPerformanceReviewDate



R2: 0.8621441991900065

MSE: 0.010442293816556885

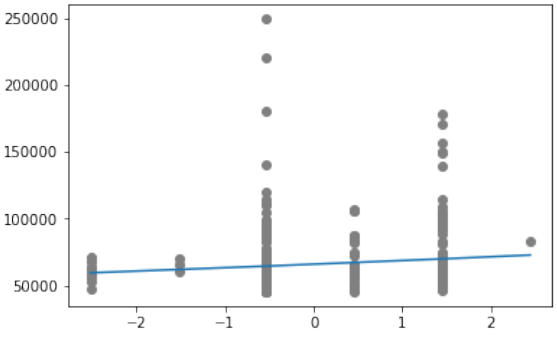
Position



R2: 0.016012914601721162

MSE: 0.07453500104495724

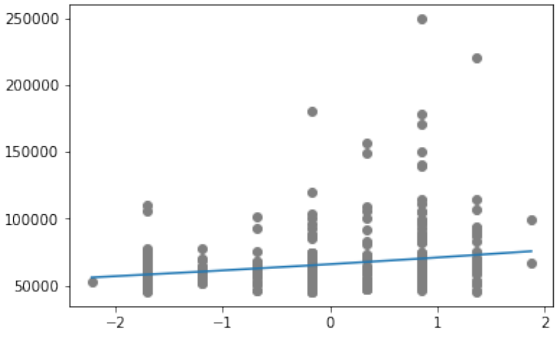
PositionID



R2: 0.02274939268847842

MSE: 0.07402472666362968

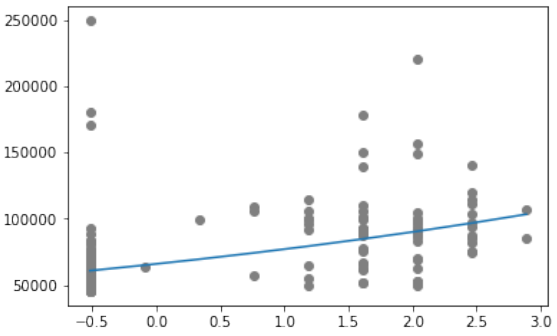
RaceDesc



R2: 0.07226518240026925

MSE: 0.07027400727647805

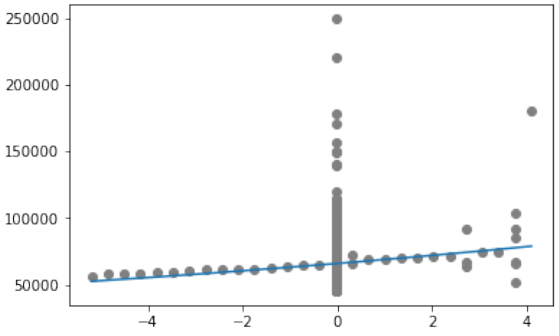
RecruitmentSource



R2: 0.3218095552449459

MSE: 0.05137153348718776

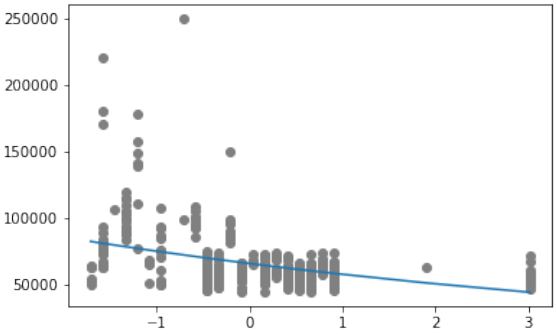
SpecialProjectsCount



R2: 0.026199050389153933

MSE: 0.07376342217687368

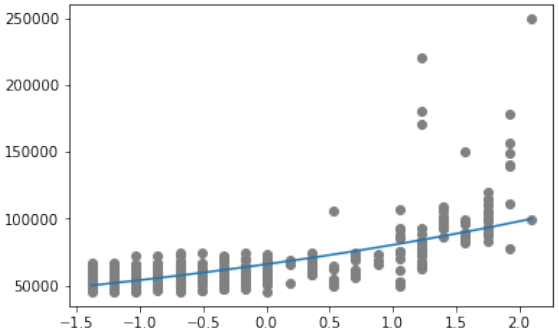
State



R2: 0.22906113015087393

MSE: 0.058397036223848835

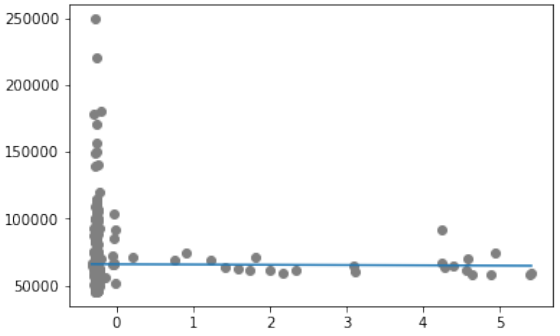
ManagerID



R2: 0.520072046807836

MSE: 0.036353556894705874

ManagerName

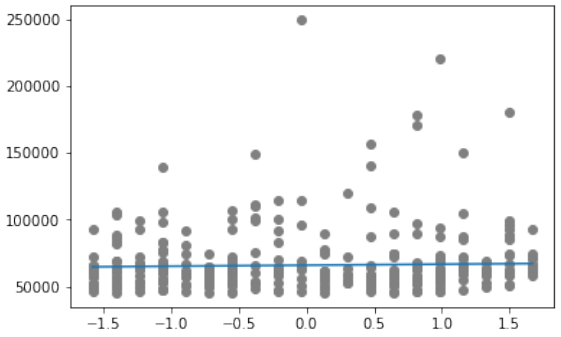


R2: 0.00021446081191722755

MSE: 0.0757317014765031

Zip

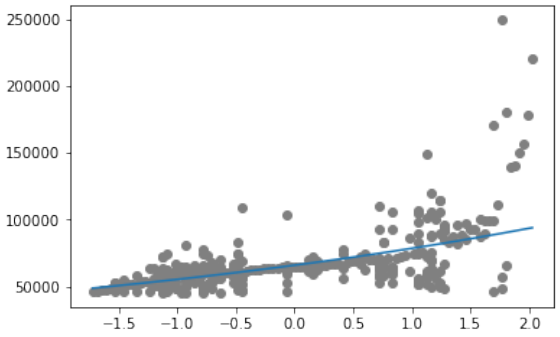
9.3.4 BayessianRidge amb la transformació logarítmica



R2: 0.004931413269374962

MSE: 0.07537440201438339

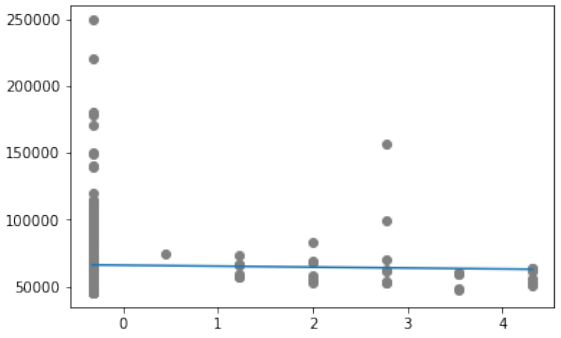
Absences



R2: 0.40478431726731656

MSE: 0.045086365657429864

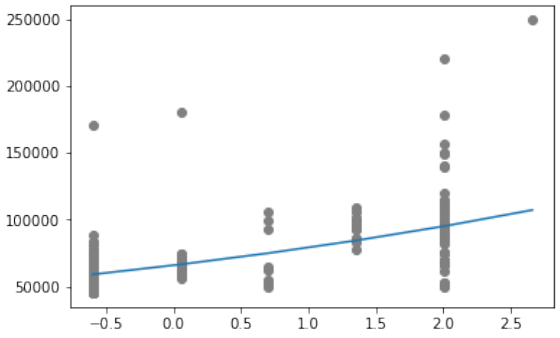
DateofHire



R2: 0.004663493986173828

MSE: 0.0753946963498982

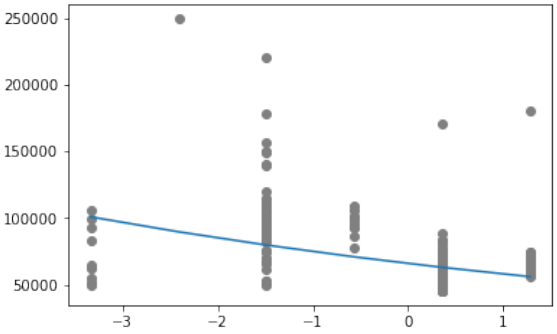
DaysLate30



R2: 0.44011591862896604

MSE: 0.04241006940975646

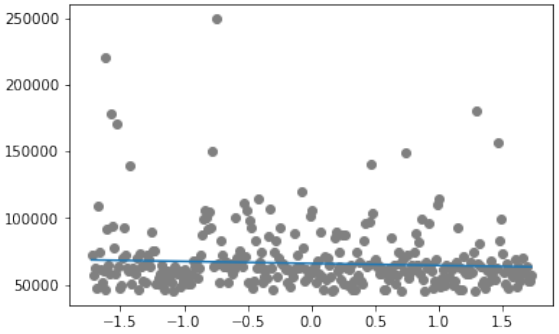
Department



R2: 0.2181226574660562

MSE: 0.059225603066942224

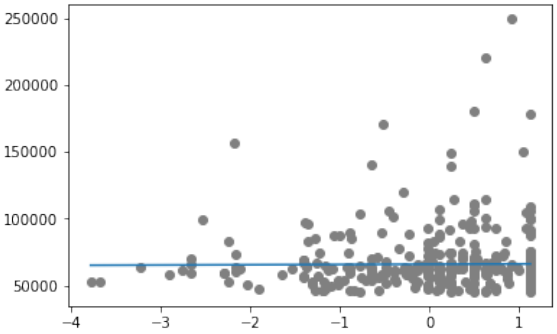
DeptID



R2: 0.011533387534307327

MSE: 0.07487431602134785

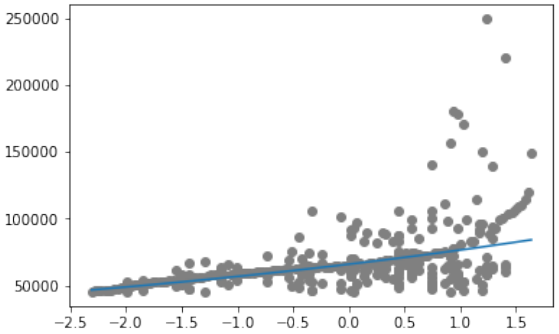
EmpID



R2: 0.001134034024054098

MSE: 0.07566204569408011

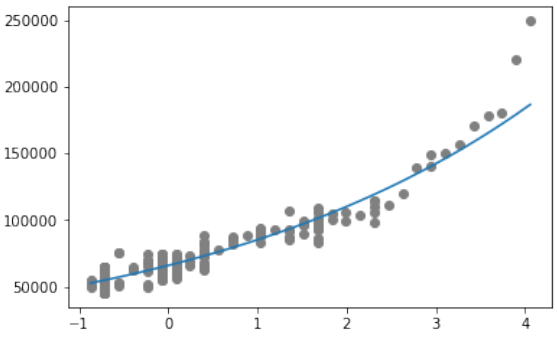
EngagementSurvey



R2: 0.29680060439234135

MSE: 0.05326591015695642

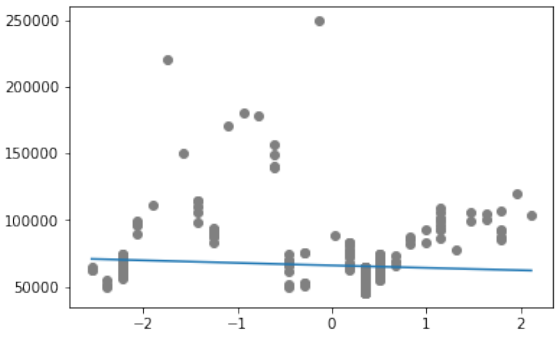
LastPerformanceReviewDate



R2: 0.8621572141391264

MSE: 0.010441307960887996

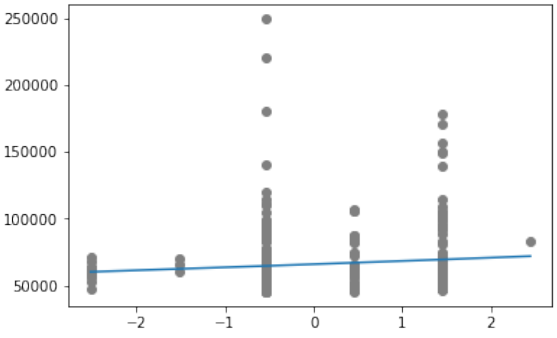
Position



R2: 0.01540069429209201

MSE: 0.07458137547618211

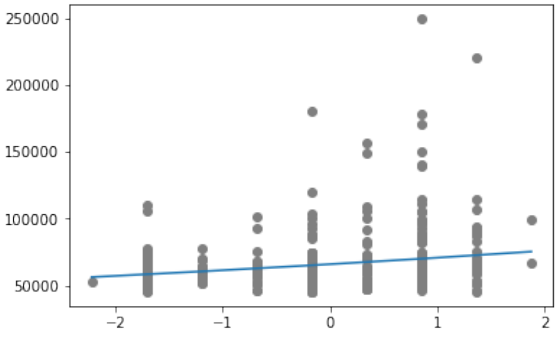
PositionID



R2: 0.0223274215310183

MSE: 0.07405669011226543

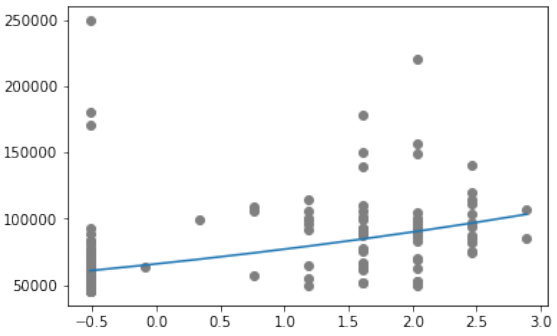
RaceDesc



R2: 0.07215461516975186

MSE: 0.07028238251713365

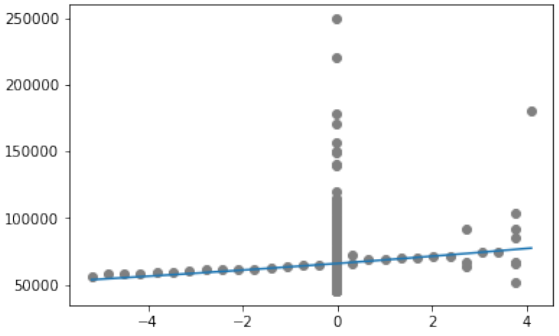
RecruitmentSource



R2: 0.32180793070560587

MSE: 0.051371656542706694

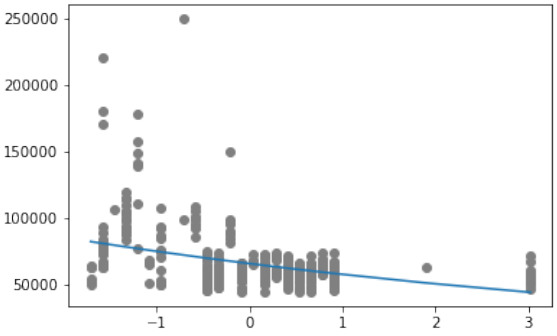
SpecialProjectsCount



R2: 0.02583682302913648

MSE: 0.07379086015554037

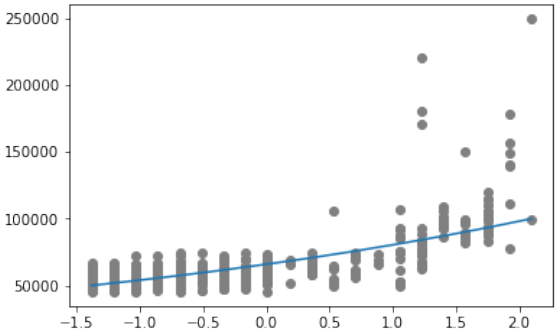
State



R2: 0.22904738304544214

MSE: 0.058398077538854835

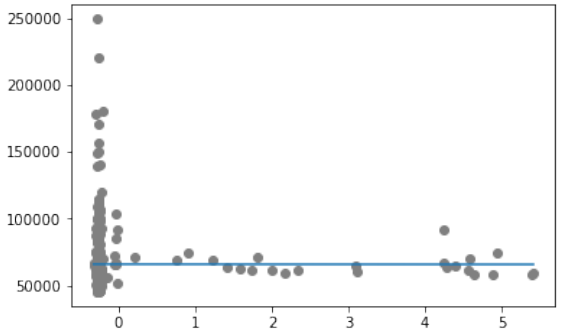
ManagerID



R2: 0.5200806833332823

MSE: 0.03635290269563891

ManagerName



R2: 3.883286169337197e-05

MSE: 0.07574500493307051

Zip

9.3.5 Regressió multilineal

Mostra del output del programa, per a l'output complet consultar el codi adjuntat.

```
['Position', 'DateofHire']
Mean squeared error:  105404799.54876395
R2 score:  0.8275056871982004
Coef:  [22235.6976501  1297.9132474]
Intercept:  69017.36571365858
--------------------------------------

            [···]

--------------------------------------
['Position', 'DateofHire', 'Department', 'ManagerName', 'SpecialProjectsCount']
Mean squeared error:  121985220.57901783
R2 score:  0.8046125176646411
Coef:  [23734.74740658  1958.13126168  2013.75626643 -2378.93585933
 -2685.22514853]
Intercept:  68977.85317116327
--------------------------------------
MILLOR R2 SCORE A:  ['Position', 'DateofHire', 'ManagerName'] 0.8275600914537563
MILLOR ERROR A:  ['Position', 'DateofHire', 'ManagerName', 'SpecialProjectsCount']
                 102709332.28481683
```

9.3.6 Regressió multilineal amb transformació logarítmica

Mostra del output del programa, per a l'output complet consultar el codi adjuntat.

```
['Position', 'DateofHire']
Mean squeared error:  0.010022008910896293
R2 score:  0.8563943243144865
Coef:  [0.23590029 0.03308796]
Intercept:  11.097657857934589
--------------------------------------

            [···]

--------------------------------------
['Position', 'DateofHire', 'Department', 'ManagerName', 'SpecialProjectsCount']
Mean squeared error:  0.010532039539133566
R2 score:  0.8501786812730818
Coef:  [ 0.23628708  0.03340143 -0.02047883  0.01055037  0.00918509]
Intercept:  11.097395616323494
--------------------------------------
MILLOR R2 SCORE A:  ['Position', 'DateofHire', 'ManagerName'] 0.8575794932710843
MILLOR ERROR A:  ['Position', 'DateofHire'] 0.010022008910896293
```

9.3.7 Lasso multilineal amb transformació logarítmica

Mostra del output del programa, per a l'output complet consultar el codi adjuntat.

```
['Position', 'DateofHire']
Mean squeared error:  0.021452992310037416
R2 score:  0.7132818071734579
Coef:  [0.15511766 0.        ]
Intercept:  11.097741043347556
--------------------------------------

            [···]

--------------------------------------
['Position', 'DateofHire', 'Department', 'ManagerName', 'SpecialProjectsCount']
Mean squeared error:  0.021220503098127275
R2 score:  0.7122942881409924
Coef:  [0.15546191 0.         0.         0.         0.        ]
Intercept:  11.097674174726995
--------------------------------------
MILLOR R2 SCORE A:  ['Position', 'DateofHire', 'ManagerName']
                    0.7162151968345063
MILLOR ERROR A:  ['Position', 'DateofHire', 'Department', 'ManagerName']
                 0.020889003305314513
```

9.3.8 BayessianRidge multilineal amb transformació logarítmica

Mostra del output del programa, per a l'output complet consultar el codi adjuntat.

```
['Position', 'DateofHire']
Mean squeared error:  0.010032885386522068
R2 score:  0.8579434816618091
Coef:  [0.23555831 0.03324445]
Intercept:  11.097708518996102
--------------------------------------

            [···]

--------------------------------------
['Position', 'DateofHire', 'Department', 'ManagerName', 'SpecialProjectsCount']
Mean squeared error:  0.010398309068287022
R2 score:  0.8512454410542898
Coef:  [ 0.23426653  0.03370605 -0.016519    0.01062965  0.00639307]
Intercept:  11.097525592309028
--------------------------------------
MILLOR R2 SCORE A:  ['Position', 'DateofHire'] 0.8579434816618091
MILLOR ERROR A:  ['Position', 'DateofHire'] 0.010032885386522068
```

9.4 Descens del gradient

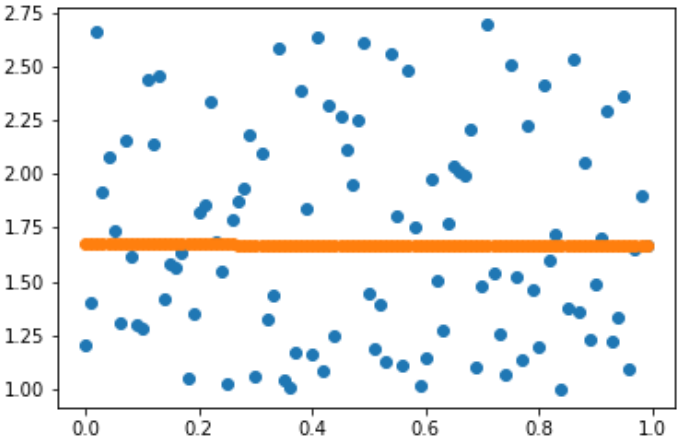


Figura 15: Primer exemple del regresor lineal programat amb un dataset aleatòri

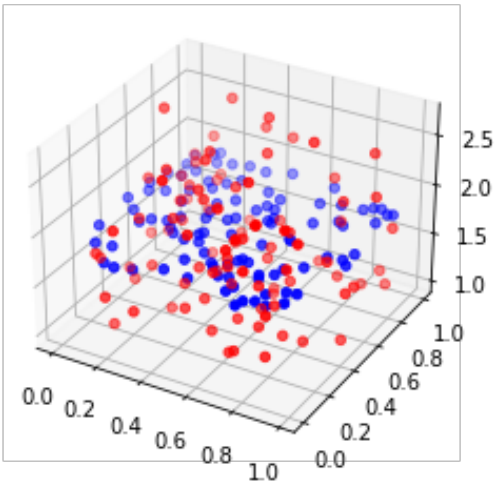


Figura 16: Segon exemple del regresor lineal programat amb un dataset aleatori