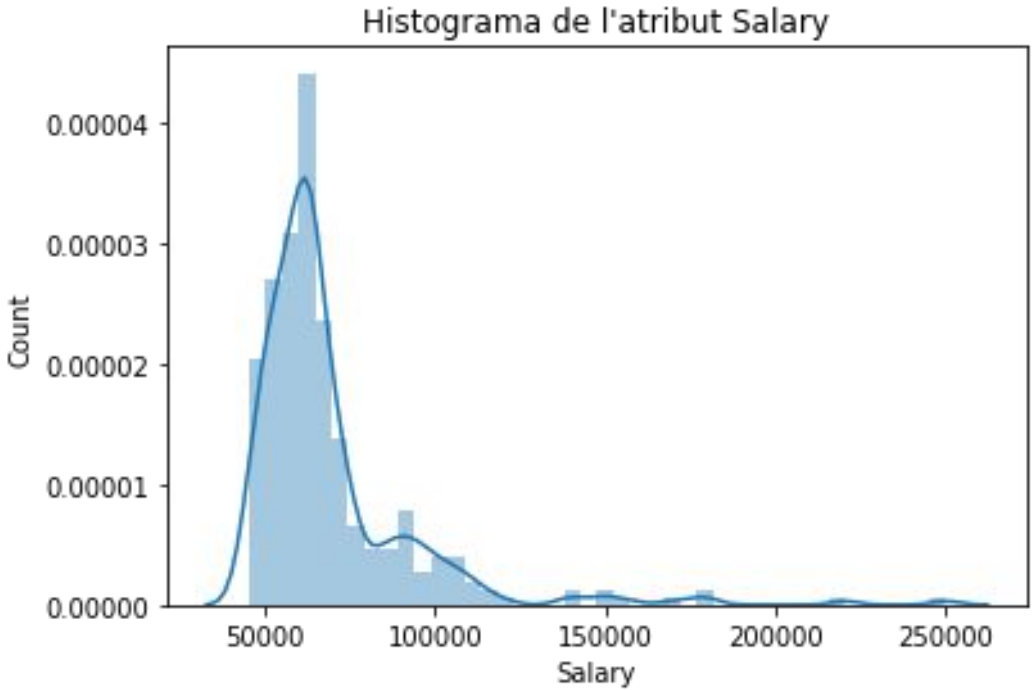
Pràctica 1 APC
Regressió

Àlex Correa Orri 1564967
Júlia Pumares Benaiges 1566252

PREPARACIÓ DATASET

El nostre dataset conté dades dels treballadors d'una empresa. Volem determinar el salari dels treballadors a partir dels altres atributs.

- Hem tret els outliers de l'atribut salary.
- Hem passat atributs categòrics a binaris.
- Hem borrat atributs repetits.
- No tenim cap atribut que segueixi una distribució normal.
- Hem estandarditzat les dades

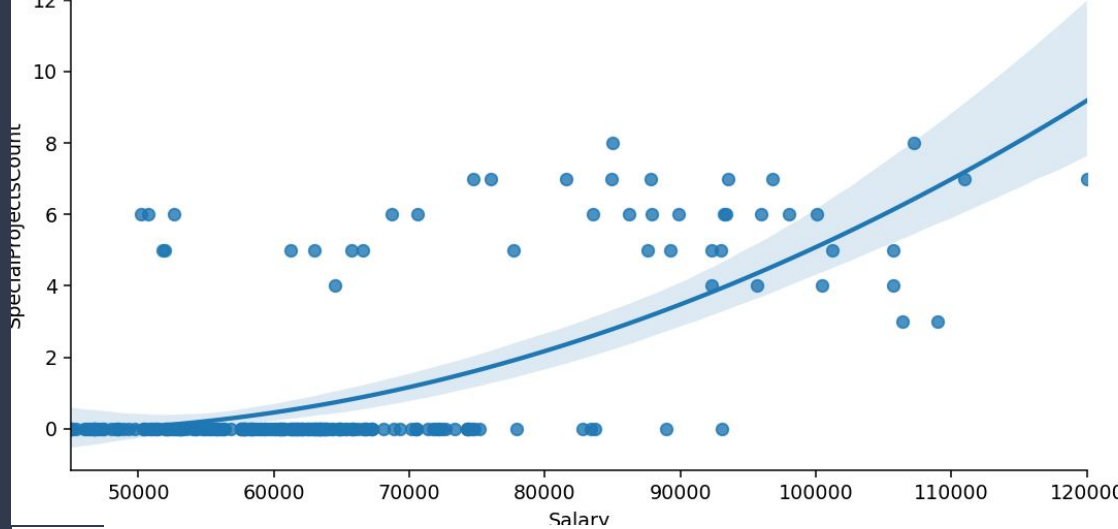


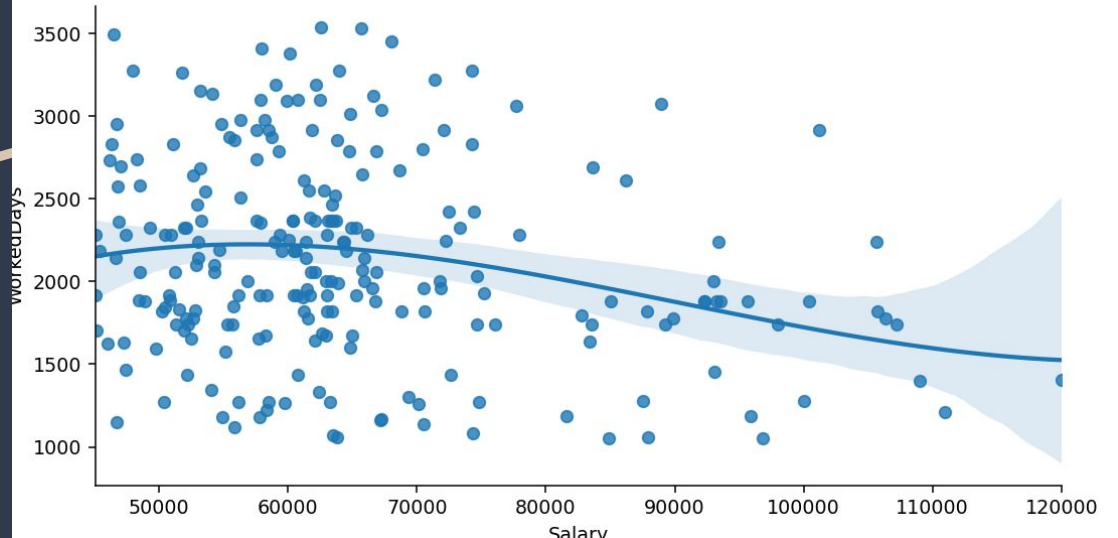
Correlacions entre els atributs i el salary

Els atributs més correlacionats amb el salary són:

- Special Projects Count (0.49)
- Worked Days (-0.16)

La resta d'atributs donaven correlacions inferiors a 0.06 en valor absolut.



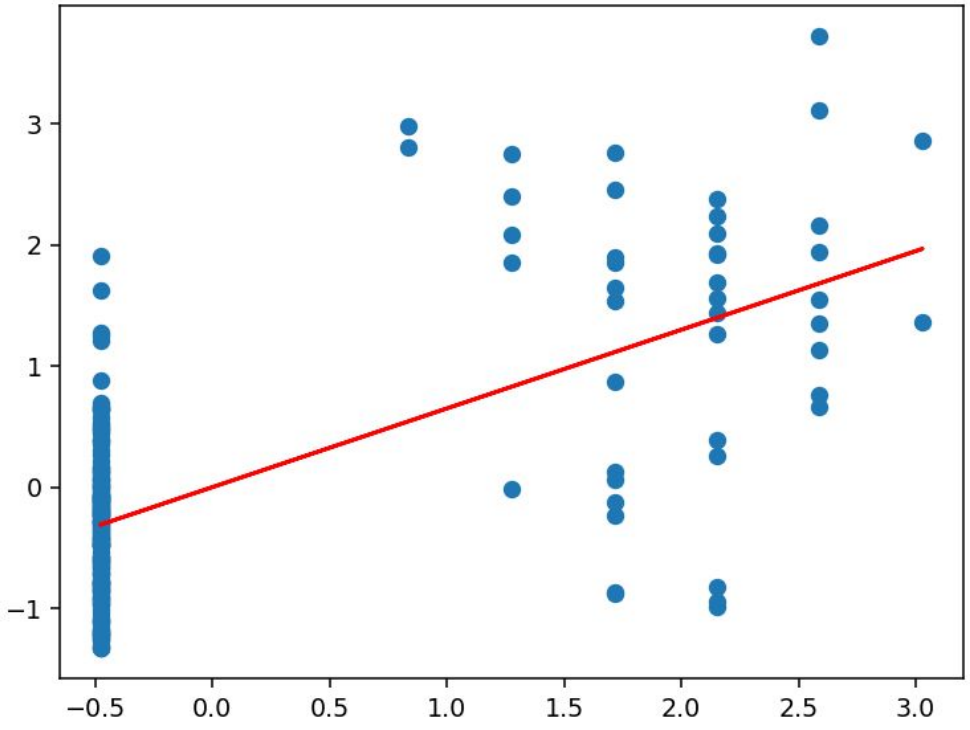


Regressió lineal amb un atribut

Hem fet una regressió amb cada atribut i el que donava menys error ha sigut *Special Projects Count*.

Error: 0.2952

També hem provat de fer servir un polinomi de grau 2 amb 1 sol atribut. L'error ha sortit bastant similar (0.2939).



Descens del gradient: regressió amb 2 atributs

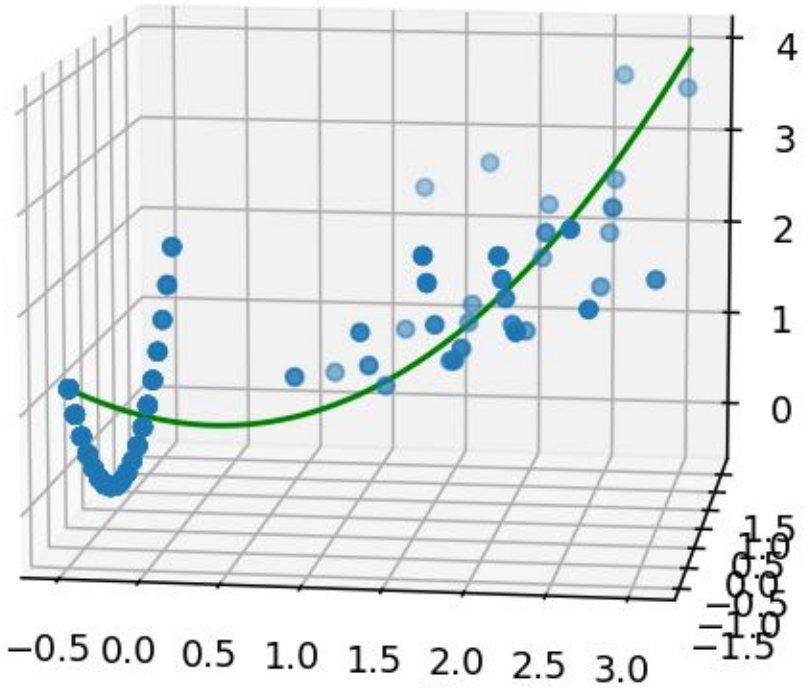
x_1: Special Project Counts

x_2: Abscences

$f = 0.65x_1 + 0.015x_2^2$

error: 0.29499

L'error no ha millorat gaire respecte quan feiem servir 1 sol atribut, veiem que x_1 influeix molt en el resultat mentres que x_2 no ho fa gaire. Amb x_1 *Special Project Counts* i canviant x_2 per tots els altres atributs hem obtingut un error proper a 0.29.



Conclusions

L’atribut més determinant per predir el *salary* és el *Special Projects Count*. Hem trobat que aquest era l’atribut més correlacionat amb el *salary* i el que ens donava mínim error en les diferents regressions que hem provat.