1 - SQL tasks

1. Report what is the average salary of men and women in the company, for all time?
Answer :
For women : **61432.56347245**
For men : **61464.61142705**

2. What year and month did the company hire the most people?
Answer : **03/1985 (** number of people hired **: 3431)**

3. What is the expected percentage increase in salary that an employee can expect after working in the company for 3 years?
Answer : **8.095 %**

4. Which employee had the biggest increase in salary during his time in the company?
Answer : (I understand it in two different ways, so I answered both cases)

- The employee that had the biggest increase in salary in term of percentage in comparison with his starting salary is the employee **484934,** his salary increased by : **129.40 %** in comparison with his starting salary.
- The employee that had the biggest increase in salary in term of how much did his salary increase is the employee **43145,** his salary increased by : **53875** in comparison with his starting salary.

5. Which month had the biggest difference between the lowest and highest salary in the company?
Answer : **2002/05**

6. Let's say that every salesperson sells 500 products per month. Each sale is making the company 300 EUR in profit. What is the profit or loss of the company for each month. Assume that the only profit are products sold by salespeople and the only cost the salaries of all employees.
Answer :

result	**year**	**month**
-417687	**1985**	**1**
-44728884	**1985**	**2**
-11135736	**1985**	**3**
33673616	**1985**	**4**
71358847	**1985**	**5**
114504625	**1985**	**6**
152531887	**1985**	**7**
194406737	**1985**	**8**
237667187	**1985**	**9**

283886045	1985	10
328534457	1985	11
372704031	1985	12
413710174	1986	1
376254207	1986	2
398345094	1986	3
447743232	1986	4
477701903	1986	5
528519594	1986	6
558500362	1986	7
600113630	1986	8
647584853	1986	9
685994058	1986	10
726646717	1986	11
764840235	1986	12
816803463	1987	1
780442871	1987	2
791218802	1987	3
844730642	1987	4
871268670	1987	5
925149200	1987	6
946022360	1987	7
988348933	1987	8
1037770182	1987	9
1066419305	1987	10
1107995216	1987	11
1144349302	1987	12
1198964574	1988	1
1159156301	1988	2
1176274375	1988	3
1226352186	1988	4
1252663805	1988	5
1310605159	1988	6
1324345204	1988	7
1369576467	1988	8
1428075847	1988	9
1450525543	1988	10
1491150559	1988	11
1519311292	1988	12
1573192702	1989	1
1548033628	1989	2
1547253607	1989	3
1595681495	1989	4

1626691170	1989	5
1690278046	1989	6
1700419870	1989	7
1742255865	1989	8
1804224232	1989	9
1822639417	1989	10
1866965362	1989	11
1893204167	1989	12
1952023989	1990	1
1926554556	1990	2
1915184944	1990	3
1969755453	1990	4
1993357426	1990	5
2060883148	1990	6
2063189515	1990	7
2095917915	1990	8
2167013264	1990	9
2180180816	1990	10
2228898334	1990	11
2256094249	1990	12
2312007025	1991	1
2294186144	1991	2
2274509183	1991	3
2330169040	1991	4
2339231176	1991	5
2412896135	1991	6
2416799966	1991	7
2439606571	1991	8
2515316761	1991	9
2526860702	1991	10
2568967571	1991	11
2591295314	1991	12
2651397646	1992	1
2619492795	1992	2
2608695853	1992	3
2672346675	1992	4
2680991146	1992	5
2757398246	1992	6
2749246649	1992	7
2774315186	1992	8
2862277594	1992	9
2860496337	1992	10
2907039287	1992	11

2926979976	1992	12
2972973066	1993	1
2974832971	1993	2
2933677177	1993	3
2997622402	1993	4
3001648282	1993	5
3076767483	1993	6
3072809061	1993	7
3096986658	1993	8
3180154918	1993	9
3178053081	1993	10
3215409625	1993	11
3227656222	1993	12
3278480484	1994	1
3287426271	1994	2
3237883803	1994	3
3305976827	1994	4
3299813111	1994	5
3384124894	1994	6
3381484660	1994	7
3408775791	1994	8
3498823600	1994	9
3493710749	1994	10
3538224515	1994	11
3553075182	1994	12
3606473140	1995	1
3620066233	1995	2
3560790569	1995	3
3630288555	1995	4
3624021817	1995	5
3700267968	1995	6
3691347802	1995	7
3706742539	1995	8
3807359858	1995	9
3787318594	1995	10
3832343786	1995	11
3838411675	1995	12
3887525126	1996	1
3872621038	1996	2
3836575059	1996	3
3906523706	1996	4
3891822968	1996	5
3966385379	1996	6

3944422046	1996	7
3963597284	1996	8
4067969815	1996	9
4047115421	1996	10
4088305399	1996	11
4076449022	1996	12
4125247280	1997	1
4165014690	1997	2
4073591605	1997	3
4146674308	1997	4
4123527670	1997	5
4203134481	1997	6
4175645844	1997	7
4200634121	1997	8
4303162465	1997	9
4281664096	1997	10
4327892390	1997	11
4318749269	1997	12
4366112183	1998	1
4410022342	1998	2
4317270752	1998	3
4400540403	1998	4
4369969845	1998	5
4446675671	1998	6
4412562011	1998	7
4429112953	1998	8
4536076424	1998	9
4506740221	1998	10
4549456776	1998	11
4533864948	1998	12
4585425192	1999	1
4632529048	1999	2
4530543955	1999	3
4604600739	1999	4
4576823422	1999	5
4659210798	1999	6
4614981247	1999	7
4638358381	1999	8
4749943288	1999	9
4717350651	1999	10
4765434428	1999	11
4746689989	1999	12
4793105364	2000	1

4832677190	2000	2
4735195595	2000	3
4766313590	2000	4
4689292648	2000	5
4731697168	2000	6
4651688466	2000	7
4615244192	2000	8
4688096926	2000	9
4610899566	2000	10
4611629411	2000	11
4545242281	2000	12
4549747150	2001	1
4639590466	2001	2
4496005254	2001	3
4533988539	2001	4
4454168893	2001	5
4494706927	2001	6
4408776820	2001	7
4378061091	2001	8
4454107242	2001	9
4378699510	2001	10
4378701487	2001	11
4316659313	2001	12
4318640951	2002	1
4399008526	2002	2
4260039913	2002	3
4298980806	2002	4
4213599221	2002	5
4259083316	2002	6
4179821131	2002	7
5606578520	2002	8

2 - Data science task

Here you can use any tool of your choosing.
Please provide us with the code and the explanation of your reasoning.

1. Let's say a Senior Engineer is looking to join the company in August 2002 He wants to know how much he will be paid in 10 years assuming the company grows at the same rate as before. How would you solve this? Build a model that predicts his salary in 10 years.

Solution :

The goal here is to predict the salary in 10 years of a senior engineer that will join the company at August 2002 .
According to what are we going to predict this salary ?
Well, the response to this question is : depending on what informations we have about this employee. In the question we have just the date he will join the company and after how many years he wants to predict his salary.
So here, I see two solutions :
1) A solution that will predict his entry salary and his salary after 10 years.
2) A solution that will predict how much his initial salary (that should be given) will increase.

I choose to work on the first one since it includes the evolution of the salary in terms of experience.

To build a model for this problem, we should start by **building a table with the data we will need from our SQL tables**, **analyse the data we have**, **train our model** and then **generate the predictions**.

1 - Create a table from the sql tables we have on the database :

The first thing that comes to my mind is to create a table that contains only employees who joined the company as senior engineers.
This tables contains the following features :
Emp_no : Employee number
Gender : Employee gender
Hire_date : The date we hired the employee
Experience: how many years the employee has been in the company
Salary : the employee's salary.

So, our table : **SE_employees**, contains the salaries of the employees for each year.
(In the file BD.sql you'll find how I created this table. BD.sql should be executed to construct this table).

I also created another table called **SE_employees_entry** that contains the salaries of the employees at their entry. I created it in order to see if there is a correlation between the year of recruitment and the salary.

2 - Data Analysis :

In this step, I tried to draw some graphics to see if I can find any correlation between the features and the salary.
At my notebook, you can see that I tried to plot the salary in terms of the hiring year. I noticed that the salary doesn't increase or decrease in terms of the year. So I conclude that maybe there are other features that can determine the entry salary of a senior engineer.
Also, I plotted the salaries of each employee based on their experience in the company. I noticed that the salary increases most of the time. So, the "***experience***" feature (which represents how many years the employee is in the company) can be a good feature in our model.
I also considered the "***gender***" as a feature that I will give to my model.

3 - Model training :

Before choosing a model for my solution and training it, I did the one hot encoding for the "gender" column, so I can convert this categorical feature to a numerical one.

Then, I split my data into a training data set and a test data set so I can be able to test how close the real data to the predicted one.
I tried to train different models :

- Linear Regression
- Random forest regressor
- K-neighbor regressor
- XGBoost Regressor

I plotted their r2_score, which measure how well the regression line approximates the real data points.

I find that, in fact, I have for all the regressors a low r2_score. And the xgboost have the best score in comparison with the other regressors.
The low score can be explained by the fact that we don't have enough features that participate in the definition of that salary and that train our model.
As I mentioned at first, I noticed a huge variability in the starting salaries and that doesn't necessarily increase or decrease with the hiring years. Maybe other features can be considered in the definition of the starting salary of each employee, like his diploma, his age, his previous experiences and achievements...

4 - Validation :

Even if all the regressors don't have a good r2_score, In the validation part I used the XGBoost regressor to calculate the starting salary of a senior engineer that will join the company in

August 2002, and his salary after 10 years in the company. I considered the both cases where the Senior Engineer is a man or a women.
And those were the results :

emp_no	gender	hire_year	hire_month	experience	Salary estimation
500000	M	2002	8	0	62219
500001	F	2002	8	0	60770
500000	M	2002	8	10	76434
500001	F	2002	8	10	75349

<u>How can we ameliorate the models :</u>

We can tune the parameters of each model so to increase it performance or make it easier to train the model.