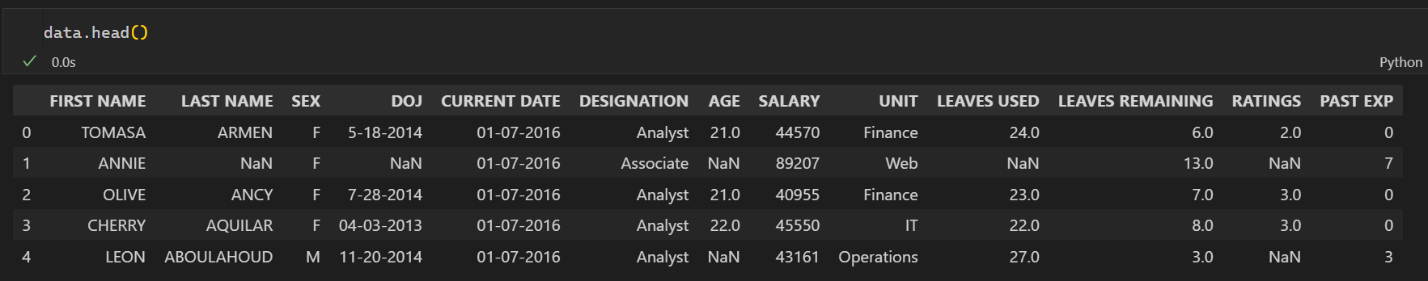
Article

This is with reference to the problem statement given, here are the attributes listed:

* `FIRST NAME`: First name
* `LAST NAME`: Last name
* `SEX`: Gender
* `DOJ`: Date of joining the company
* `CURRENT DATE`: Current date of data
* `DESIGNATION`: Job role/designation
* `AGE`: Age
* `SALARY`: Target variable, the salary of the data professional
* `UNIT`: Business unit or department
* `LEAVES USED`: Number of leaves used
* `LEAVES REMAINING`: Number of leaves remaining
* `RATINGS`: Ratings or performance ratings
* `PAST EXP`: Past work experience

Here is a bird’s eye view of the data.

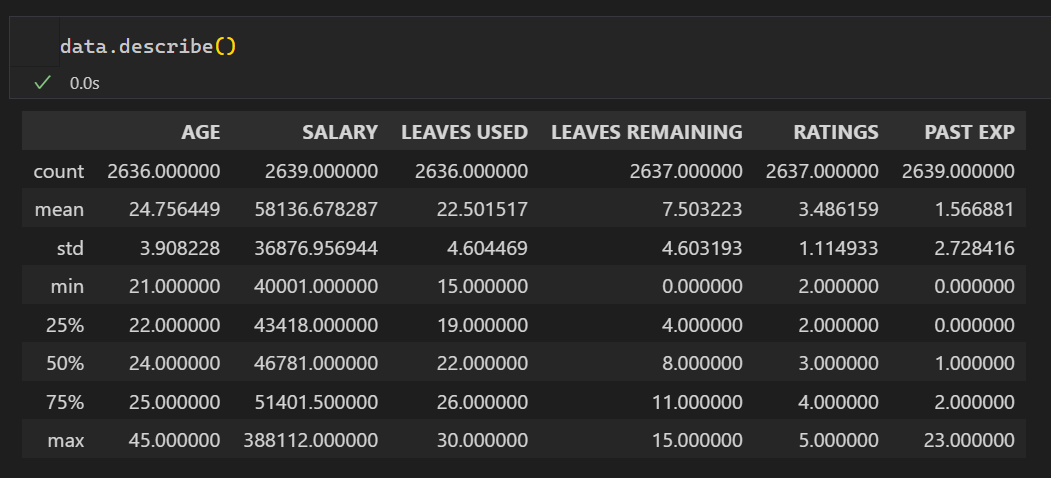


Looking at the data, there are few inferences we can draw:

* Concatenate “FIRST NAME” and “LAST NAME’ column. That way, it would be easier to work with whole names. (Purely Optional)
* Convert “SEX” to Boolean variable. Without loss of generality, set and .
* Observing “DOJ” and “CURRENT DATE”, we can find time period in company, which can serve as an independent attribute in linear regression model.
* “DESIGNATION” column can play a pivotal role in data visualisation. As we would expect higher ups to earn more salary. We could conveniently produce plot using pivot table in excel. More robust approach would be to model visualisations in Power BI.
* “AGE” along with “PAST EXP” can serve as independent variables in linear regression model.
* Same inference can be drawn about “UNIT” column as that of “DESIGNATION”.
* “LEAVES USED” and “LEAVES REMAINING” seems to be dependent on each other. We would expect higher correlation between them. They can also be used as independent attribute in Linear Regression Model.
* “RATINGS” and “PAST EXP” also affect “SALARY”. We would expect higher rated individual with more past experience to earn more.

# EDA

Checking for summary statistics:



Looking at the target variable, Observe standard deviation of about . This can be the benchmark for RMSE score.

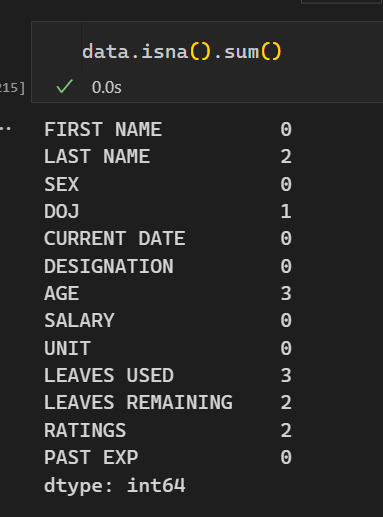
Here’s a comparison. Mathematically,

|  |  |
| --- | --- |
|  |  |

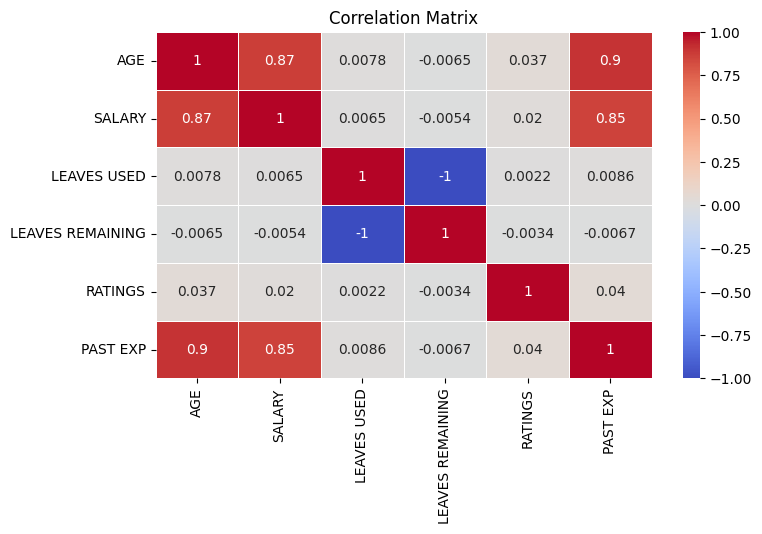
Essentially, RMSE measures the deviation of true values from predicted values, whereas, Standard Deviation measures deviation of true values from mean. Assuming our mean is the best estimate of predicted values,

Implying, a lower RMSE than Std of target variable (SALARY) in our case, corresponds to better fitting model.

Here are the number of NULL values: **(1)**



Here’s the correlation matrix:



As expected, “AGE” and “SALARY” show high correlation. This suggests with higher ages, individuals tend to earn more salary.

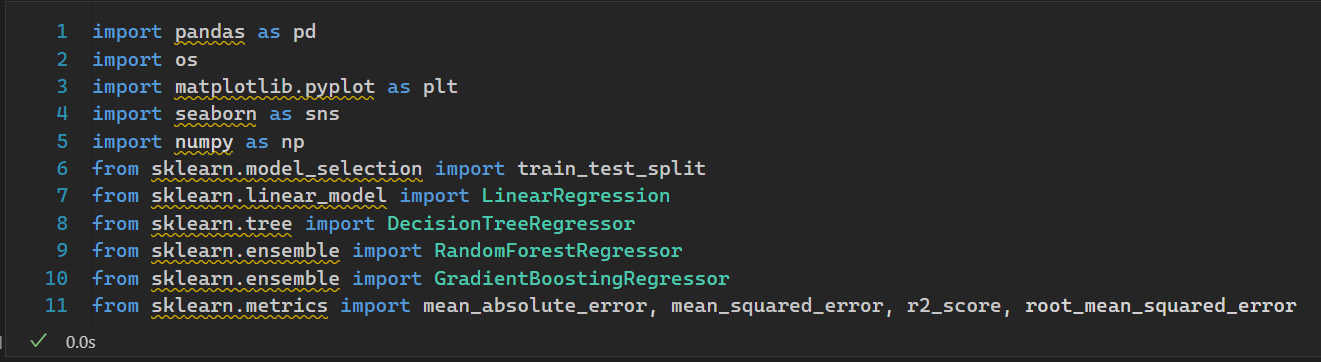
One important observation is that “LEAVES USED” and “LEAVES REMAINING” show strong negative correlation. Implying when working with one, other one would be redundant to use. Also, looking at the data, we intuitively assume these attributes summed up to , **NOT COUNTING** missing values, which were in . That is information loss of about . That’s enough to safely assume both the attributes sums to thirty.

Another inference we can draw is that “AGE” and “PAST EXP” show higher correlation. While building up models, we would not observe significant improvement to performance when considering both compared to considering one at a time. According to **Principle of Parsimony**, it would be wiser to consider one of these attributes at a time. Additional tests need to be done regarding this.

Another important figure is AGE and PAST EXP. The both correspond to correlation of A higher correlation with non – target attributes imply data redundancy. We can use techniques like **Dimensionality Reduction** to treat such behaviours.

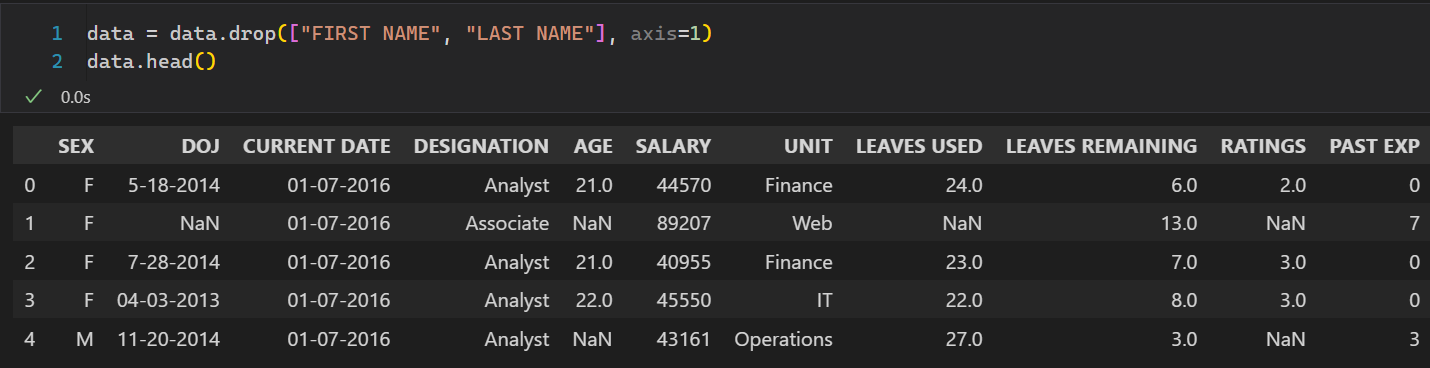
Also, we would expect individuals from same departments to show similar behaviors and work ethics. Implying, they would share similar summary statistics. This is a good lead in filling up NULL values. While performing data transformations, we can segregate our data sets into individuals’ departments. And draw inference from them.

# Required Libraries



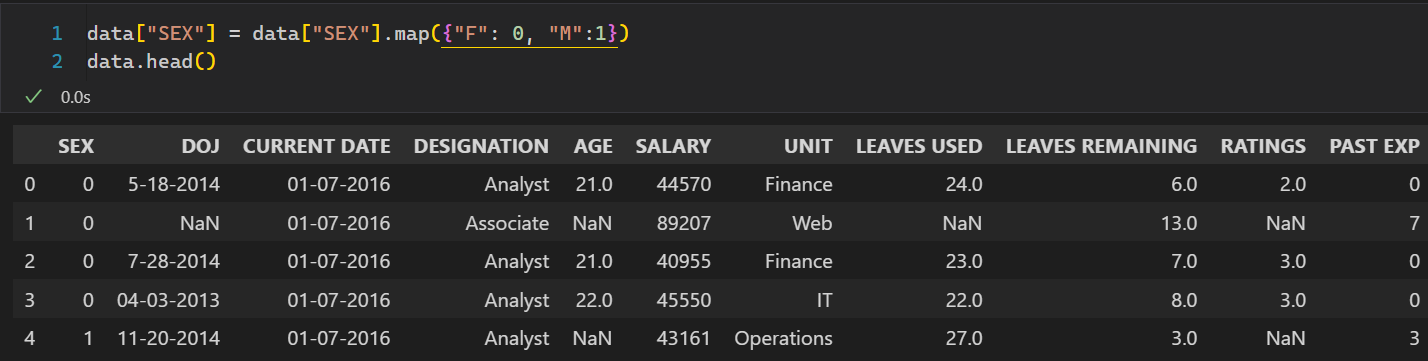
# Feature Engineering

**DROP NAMES**



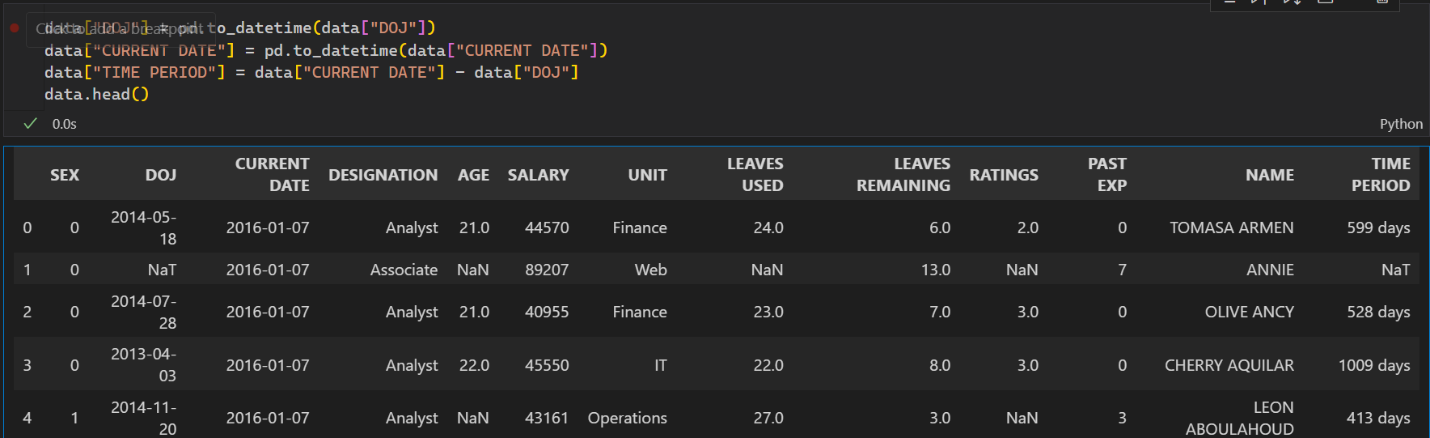
Here we dropped names column as they aren’t useful in data analysis.

**Convert SEX to BOOLEAN**

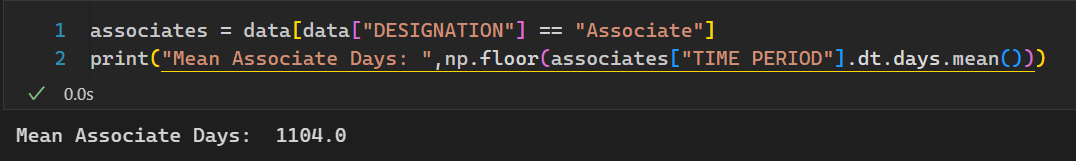


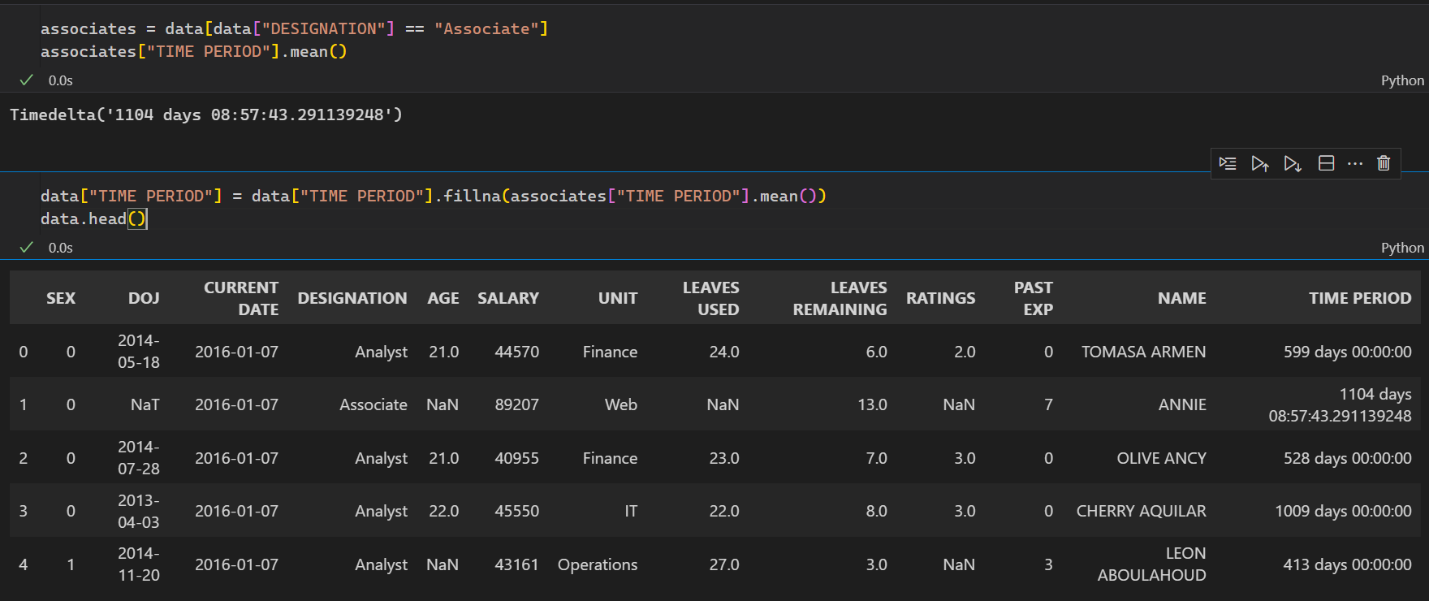
**Evaluate Time Period**

Changing “DOJ” and “CURRENT DATE” to data datatype to conveniently perform date arithmetic. Observe the time periodic in second row is <NA>. We have to meticulously fill the NA value which we believe would be the best fit. Observe that the person is Associate. We would expect every associate in company to roughly join around the same date. Implying, Associates’ time period would be roughly the same. We can conveniently fill up the NULL value according to mean of Associates.



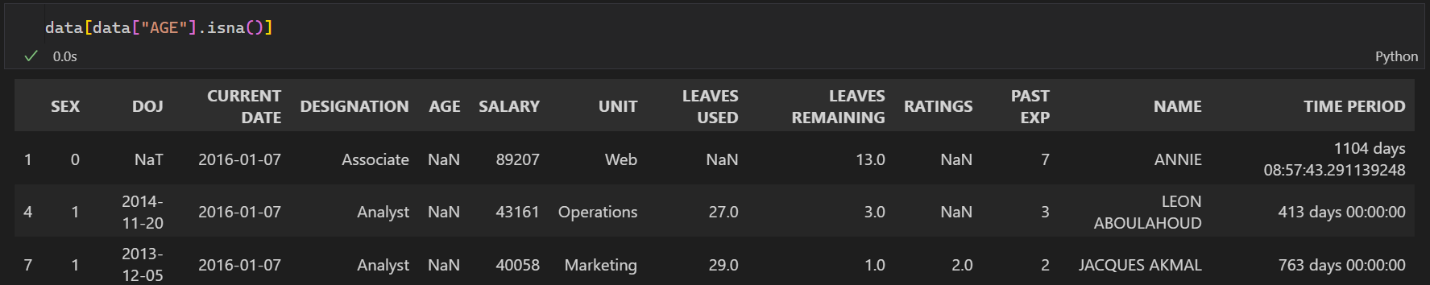
Observe that we dropped “DOJ” and “CURRENT DATE” column as the required information is summarized by “TIME PERIOD”. Next up, extracting Associates’ mean age to fill <NA> value.



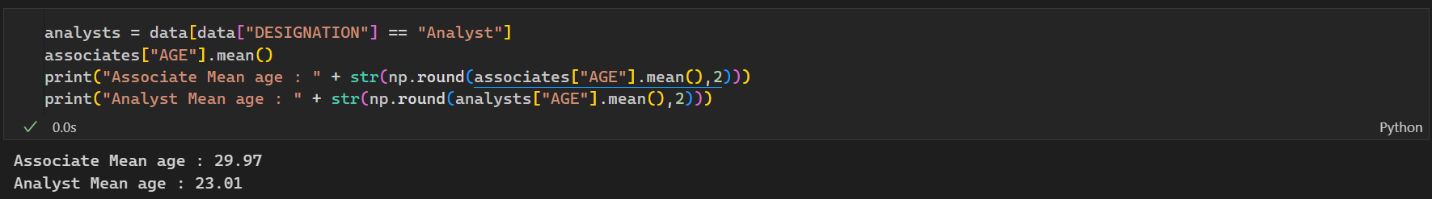


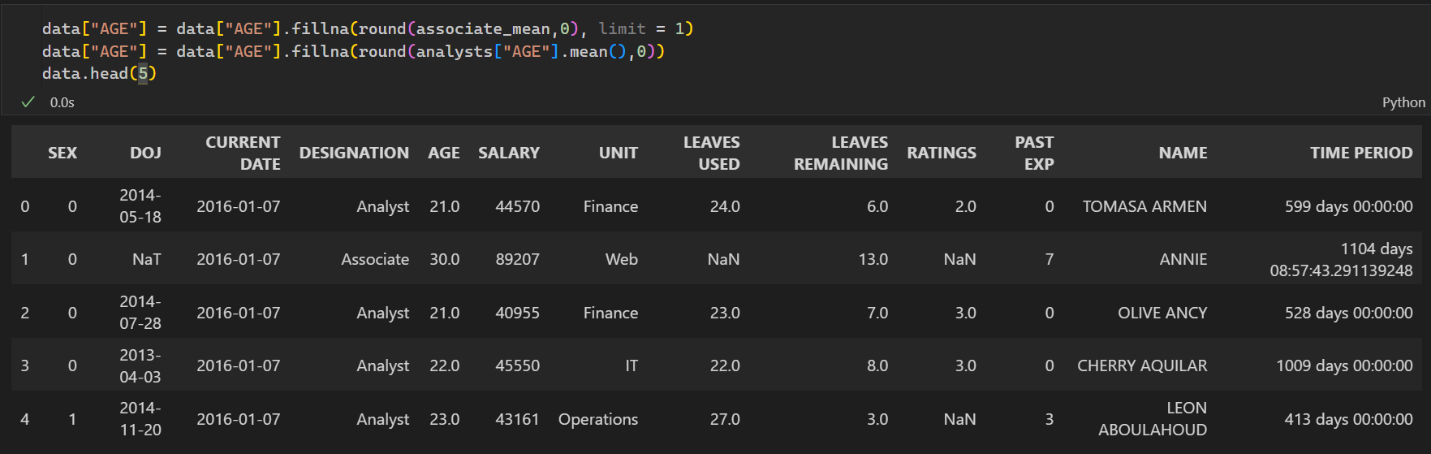
**Treat Age Column**

We would expect people having same designation to be roughly of same age, this is especially true for higher ups. A person at Director’s position is likely to have more experience and age than one at Analyst’s position. Looking at the NA values, we have to treat 3 such individuals, here’s the list of those:

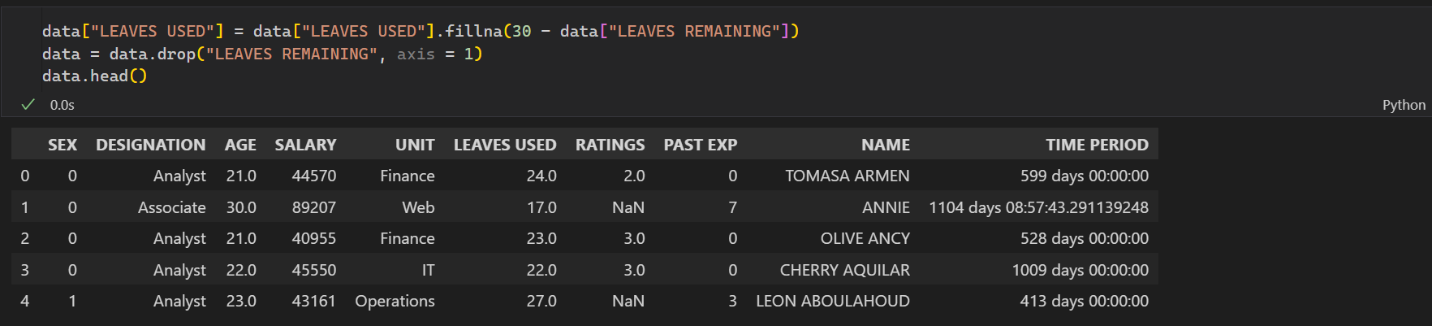


Having filled the NA value for Miss Annie in Time period, we can drop DOJ and CURRENT DATE column. Observe, the AGE missing for these individuals belong to an “Associate” and “Analyst”. Filtering data once again for these two groups.





**Treat LEAVES USED COLUMN and drop LEAVES REAMINING**



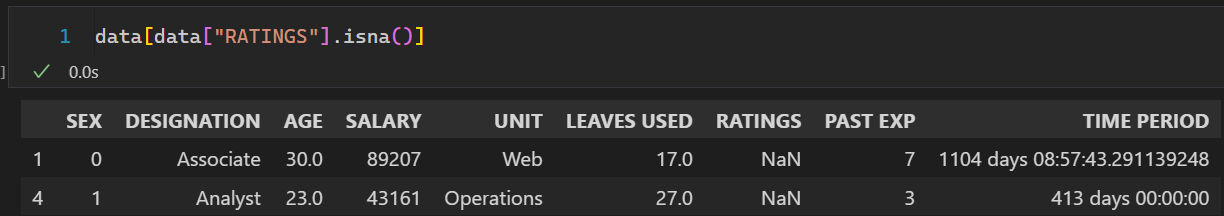
From Exploratory Data Analysis, we inferred a very important result. Both, LEAVES USED and LEAVES REMAINING sum up to 30. Thus, we filled every NA value accordingly.

Here is the updated count of NA values:

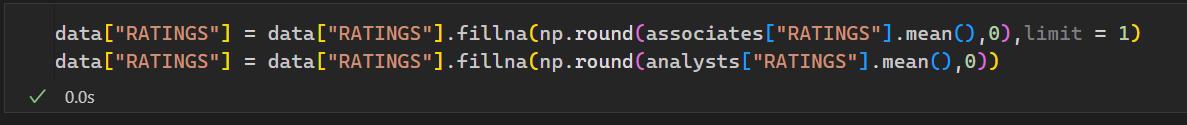


**Treat RATINGS**

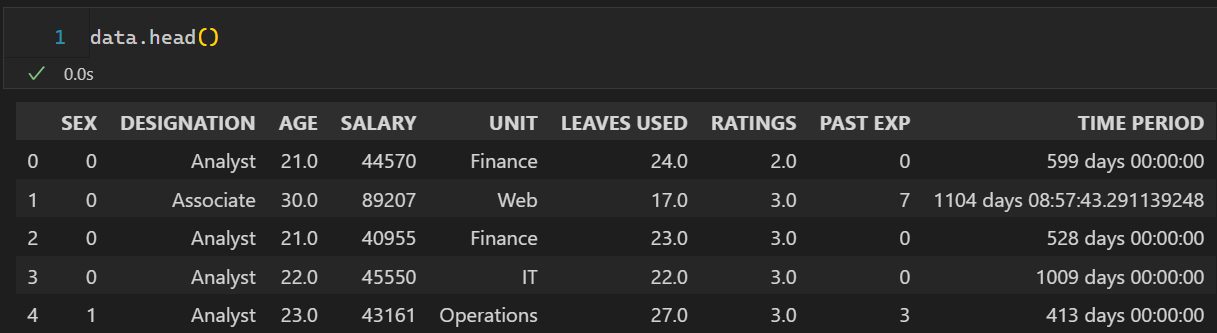
With that, all that remains is to fill RATINGS. Here is the list of people with missing rating.



Conveniently, both of these individuals belong to Associate and Analyst designation respectively. Repeating the same step as AGE column

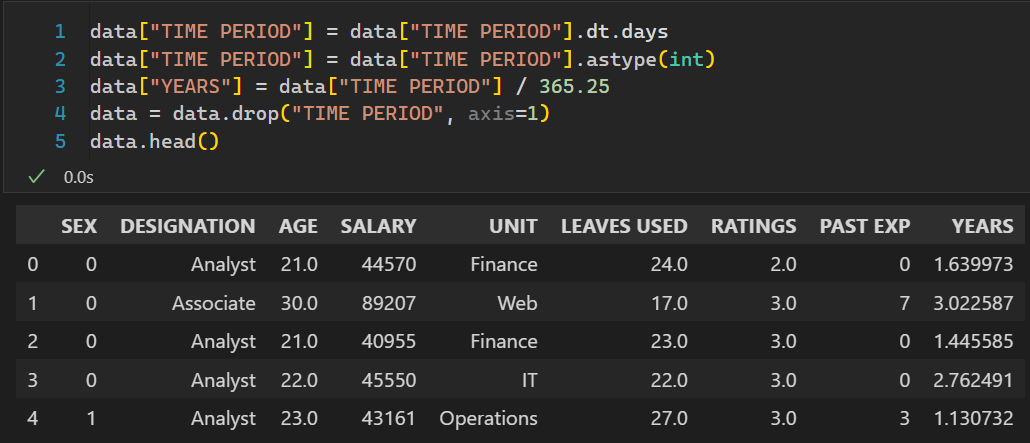


Here’s the updated DATASET.



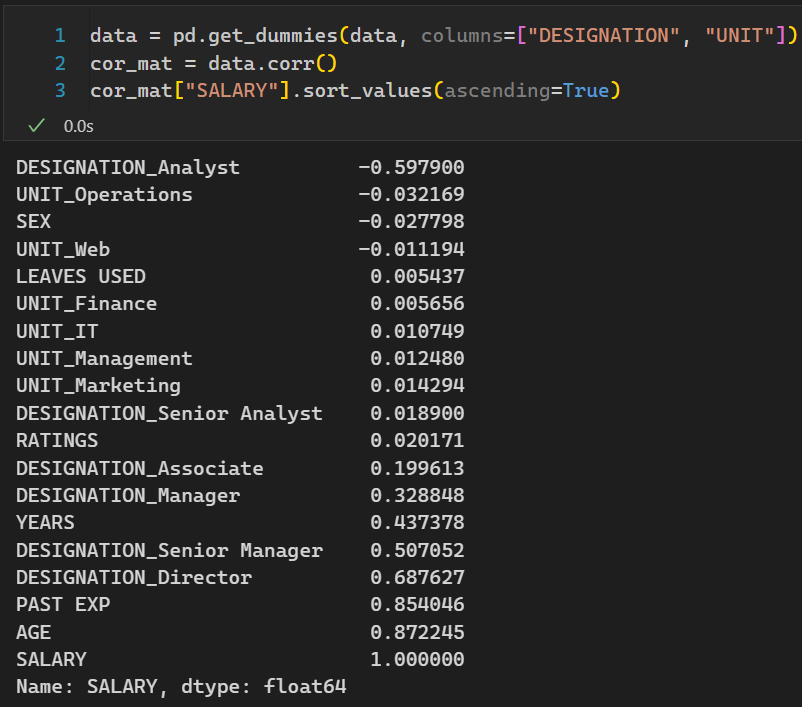
**Treat Time Period**

**Notice TIME PERIOD is listed as DateTime datatype. Ideally, we would be working on Yearly basis. As person having more years in the company would typically earn greated Salary. Thus, we need to convert Days to Years. This can be conveniently achieved as follows.**

****

**Now, all that remains is to treat Categorical variables namely “DESIGNATION” and “UNIT”. A far simpler approach would be to simply drop the columns and perform data modelling. Doing this we might miss out to some important features from the data. One example can be that higher-ups would tend to earn more salary. A Director would be earning more than an Associate would in-turn would earn more than his subordinates as a rule of thumb.**

**Converting Both categorical attributes to Booleans and printing correlation of SALARY with every possible attributes for further analysis.**

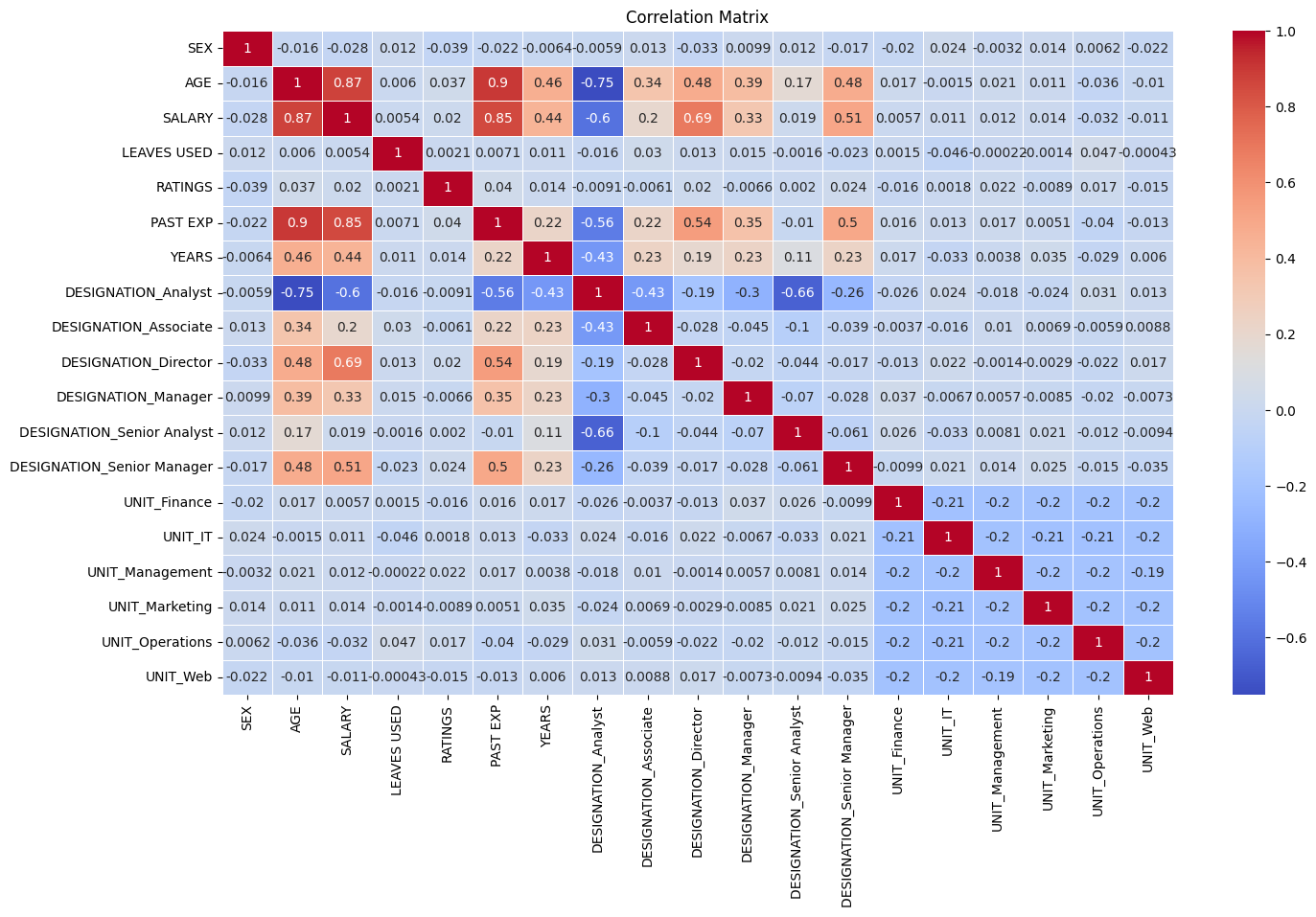
****

**Observe such high attributes. Fitting any more than desired attributes would lead to some notable hurdles in model deployment namely:**

1. **Model Over fitting.**
2. **Model Complexity.**
3. **Increased Computational Load.**

**A lot of these attributes need to be treated with to reduce noise and make model as accurate as possible. Observe every UNIT correspond to weak correlation with SALARY. Having observed that, it would be wiser to drop them. Though additional testing is need to be done.**

**The revised Correlation plot is given below:**

****

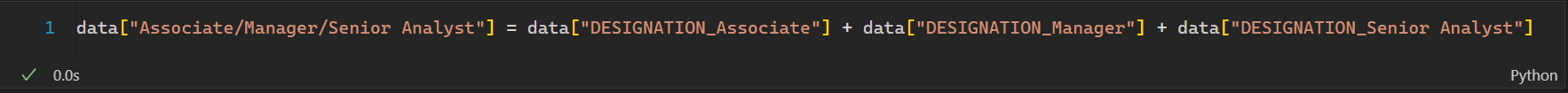
# **Reading Correlation Plot**

A correlation plot is a square matrix, where the entry correspond to correlation between attribute and attribute in our dataset. The intensity of correlation is visualized using color indicator that range from the lowest correlation found to the highest correlation found. For example, Observe the diagonal entries. All are dark red having correlations equal to 1. Each entry corresponds to, say, correlation between or . Note, a higher correlation with target variable implies better predictability, but a higher correlation with non-target variable implies data redundancy.

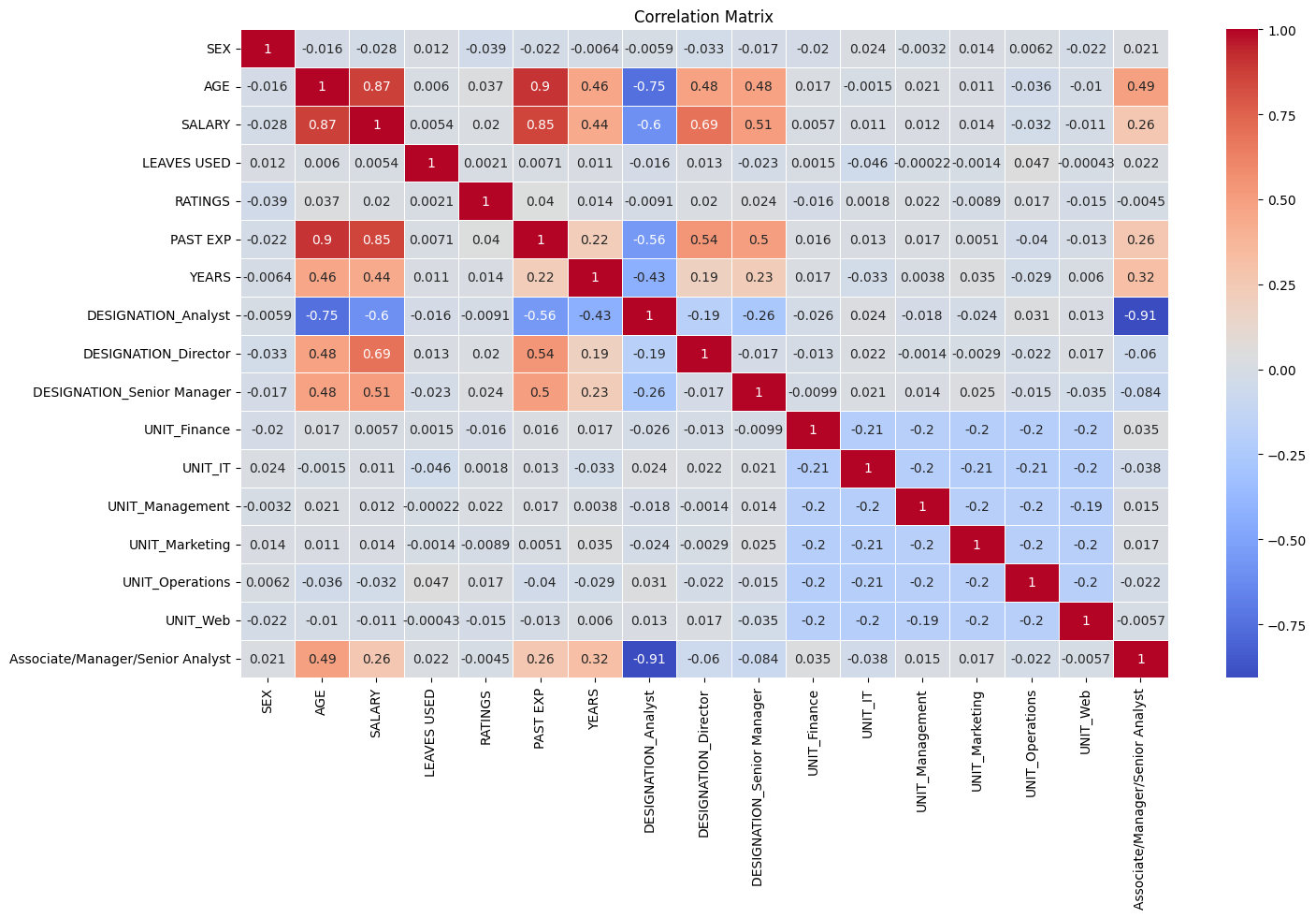
Let’s start with treating DESIGNATIONS first.

**Treat DESIGNATIONS**

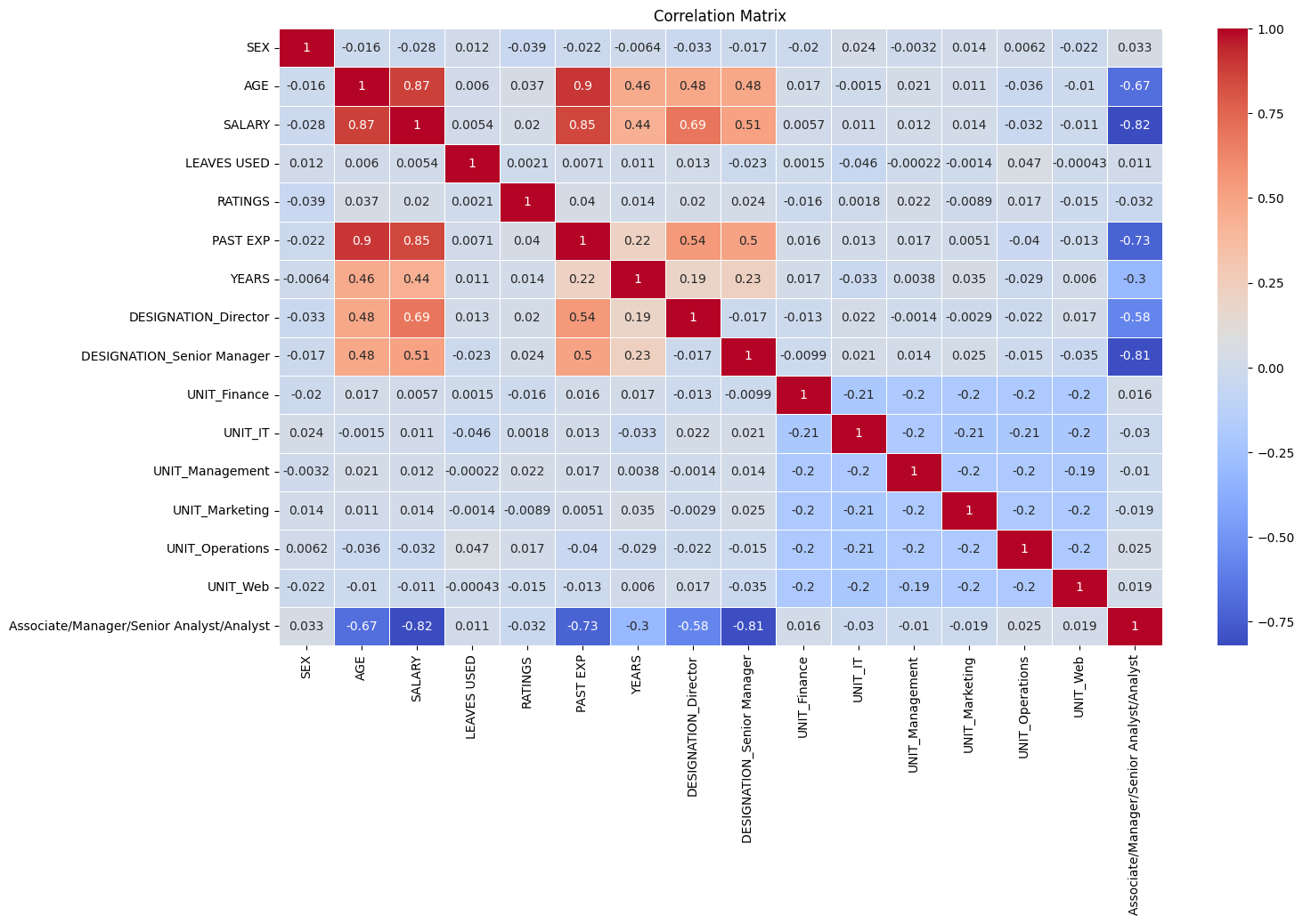
**Immediately we observe Directors, Analyst, and Senior Managers have a high impact on SALARY. On the other hand, Associates and all Managers show weaker correlation to Salary. We can combine to latter two to draw some inference.**

****

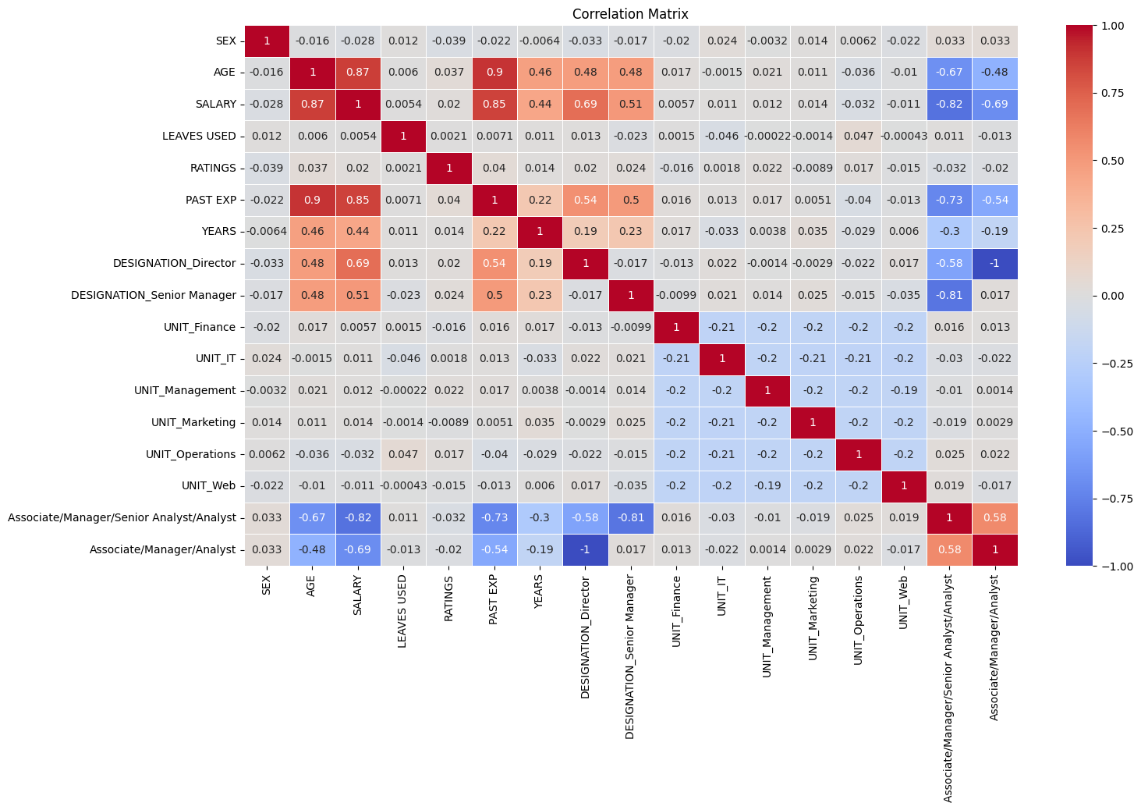
**The code translates to introducing a new column which is the join of Associates, Manager, Senior Analyst, all of whom had weak correlation with the salary. Here’s the updated Correlation Plot:**

****

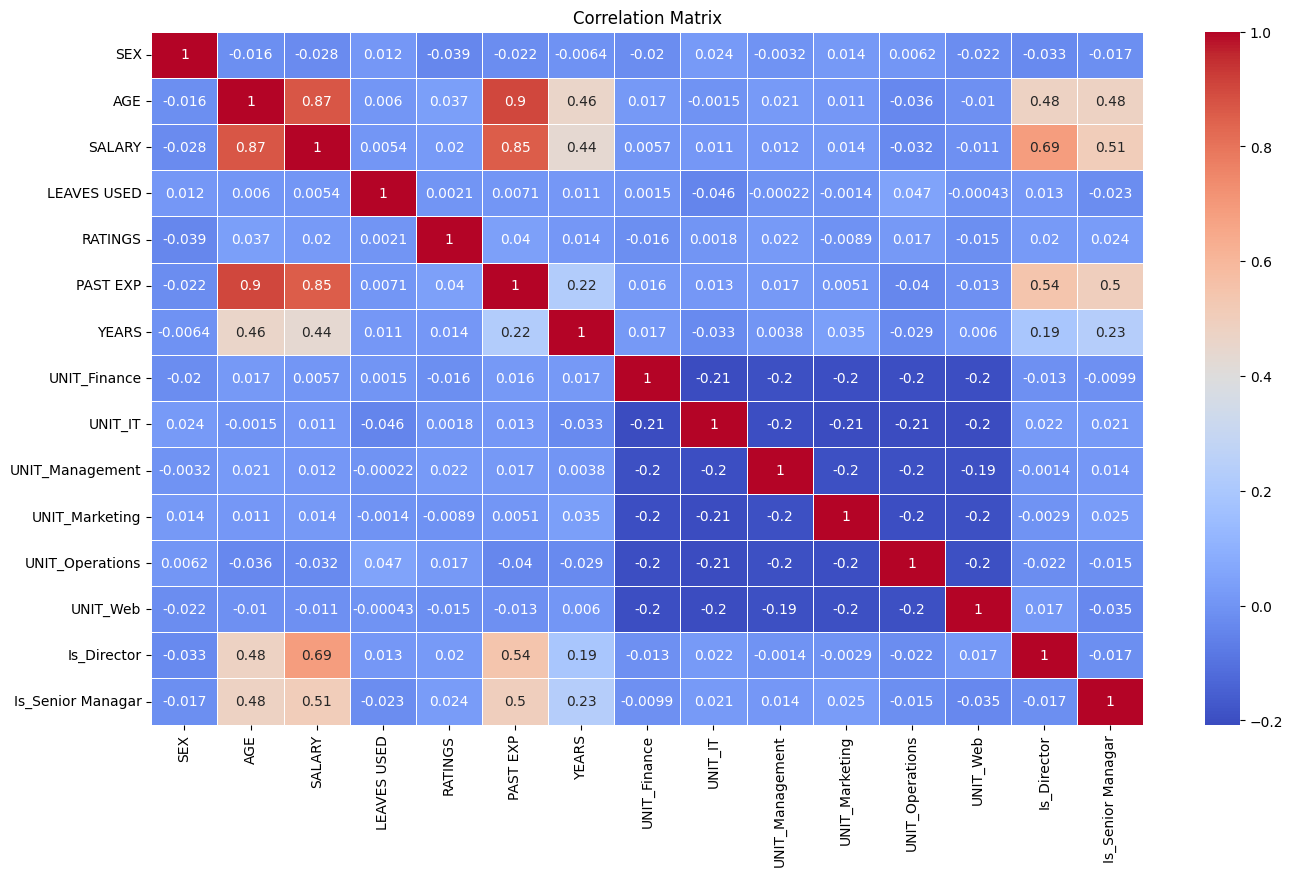
**Immediately we observe very high negative correlation between the newest column and Analysts. This is consistent with the expectation that the person who is neither an Associate, nor a Manager, nor a Senior Analyst would be an Analyst with very high expectation. NOTE: Our correlation with the salary improved drastically after combining the three columns. Next up, we can combine Analysts to draw some more inference. Here’s the updated Correlation Plot.**

****

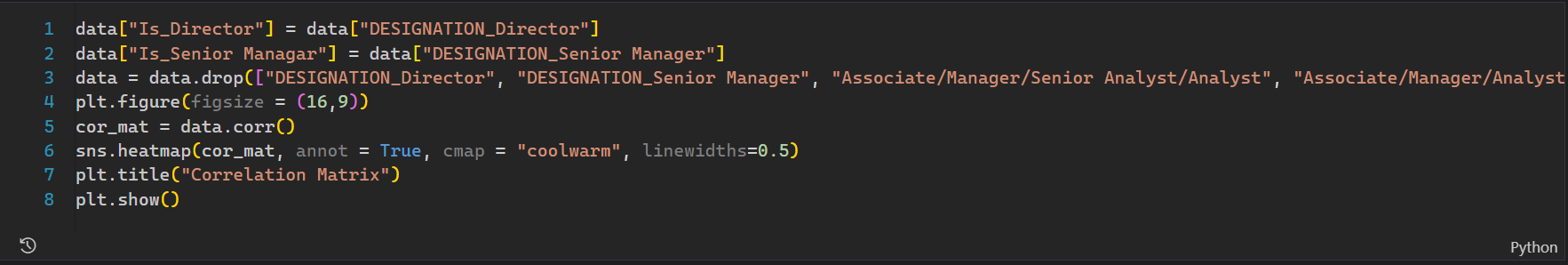
**Observe that we gained a very powerful predictor of SALARY. Though, the newest column correlate with Senior Managers highly. Implying data redundancy. We can combine both of these columns to tackle this situation.**

****

**Having done this, A person who is neither an Associate, nor a manager, nor an Analyst correspond to being a director. That’s true logically and is also depicted by our Correlation plot. We can apply Dimesionality Reduction removing the last two rows and redefining the Director and Senior Manager column.**

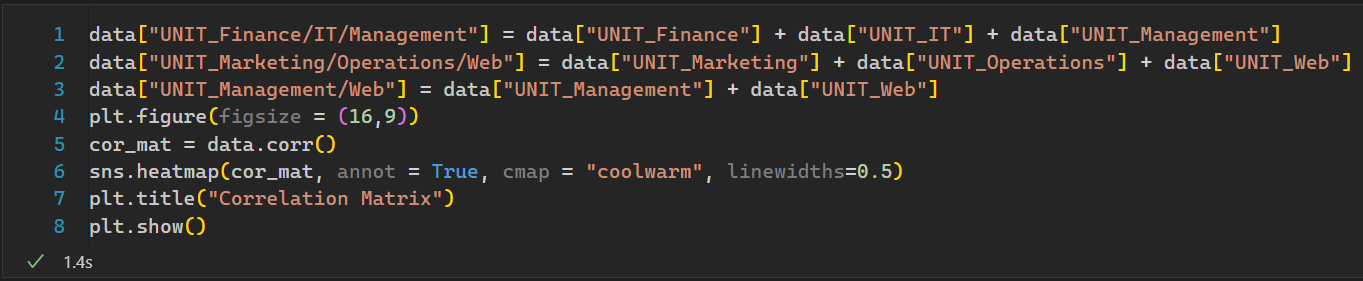
****

**The code for the same is given below:**

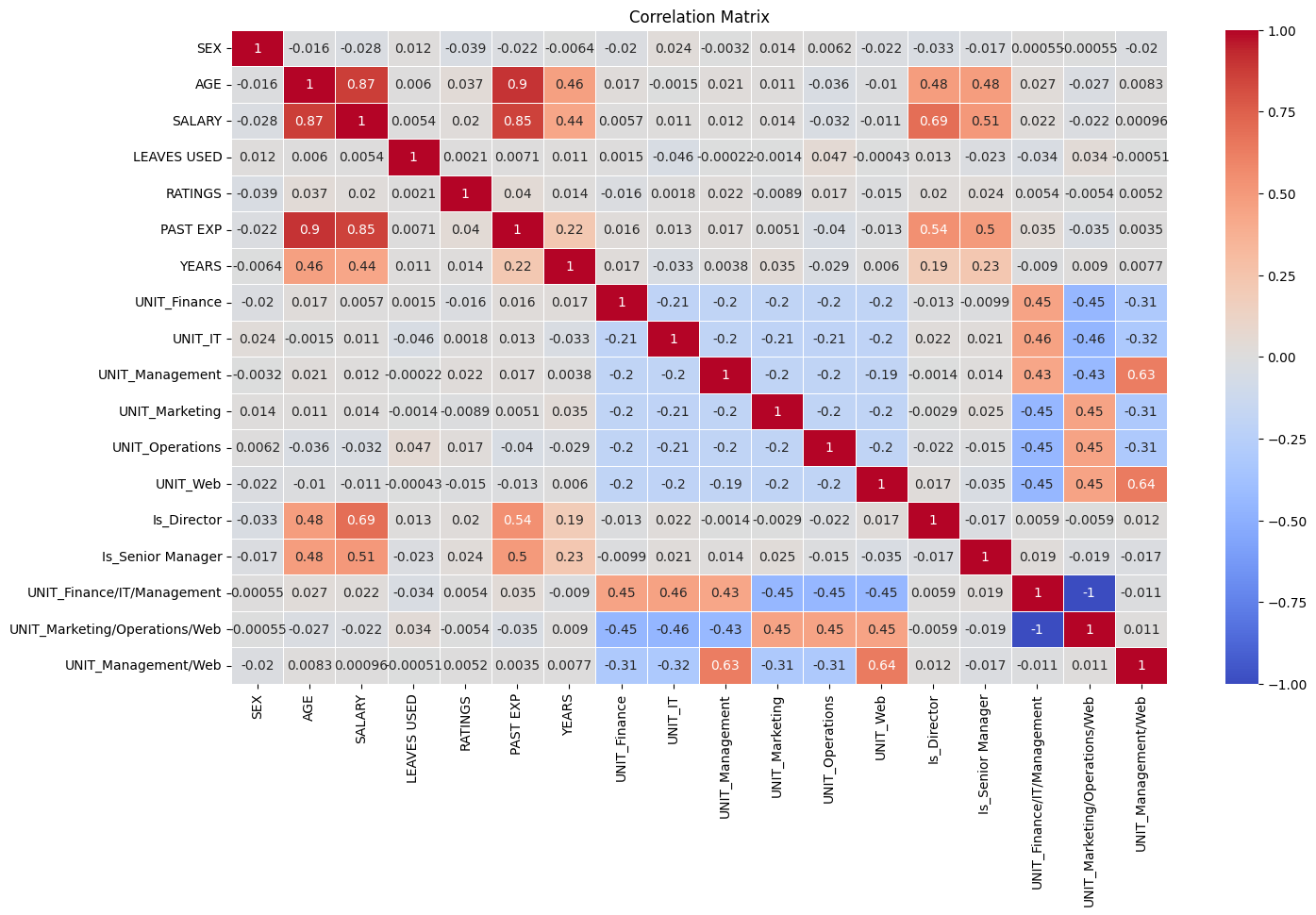
****

**Observe that we have a found a good predictor of Salary column namely, “Is\_Director” and “Is\_Senior Manager”. Both of these correlates to higher Salary.**

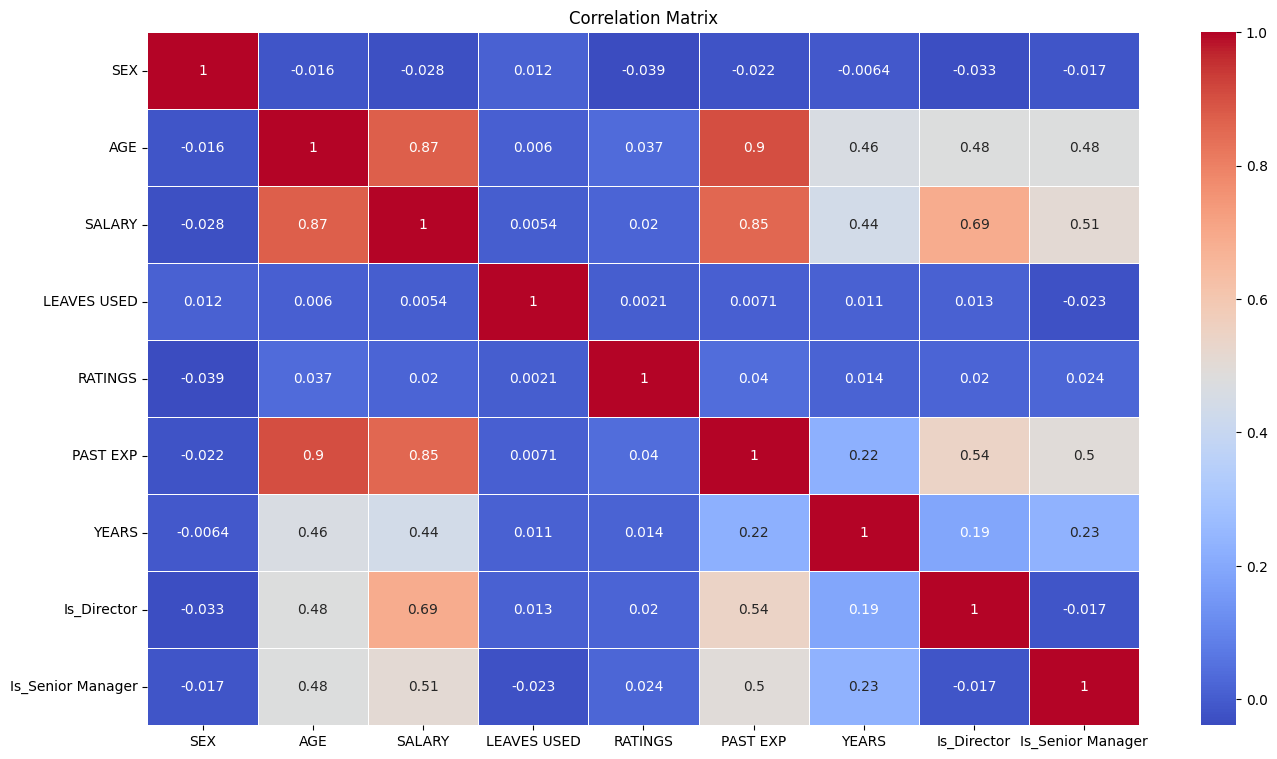
**Next up, we can treat the UNIT DEPARTMENT. Without loss of generality, combining the following columns:**

****

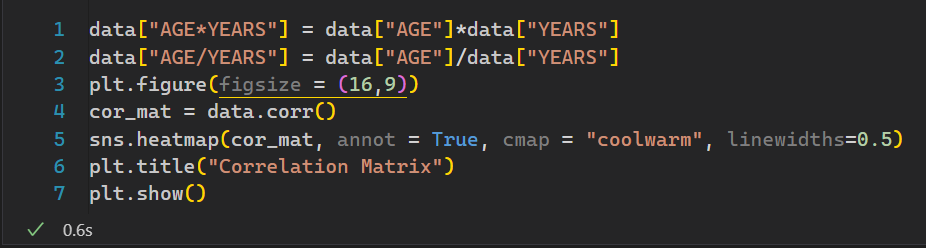
**The Corresponding Plot is given below:**

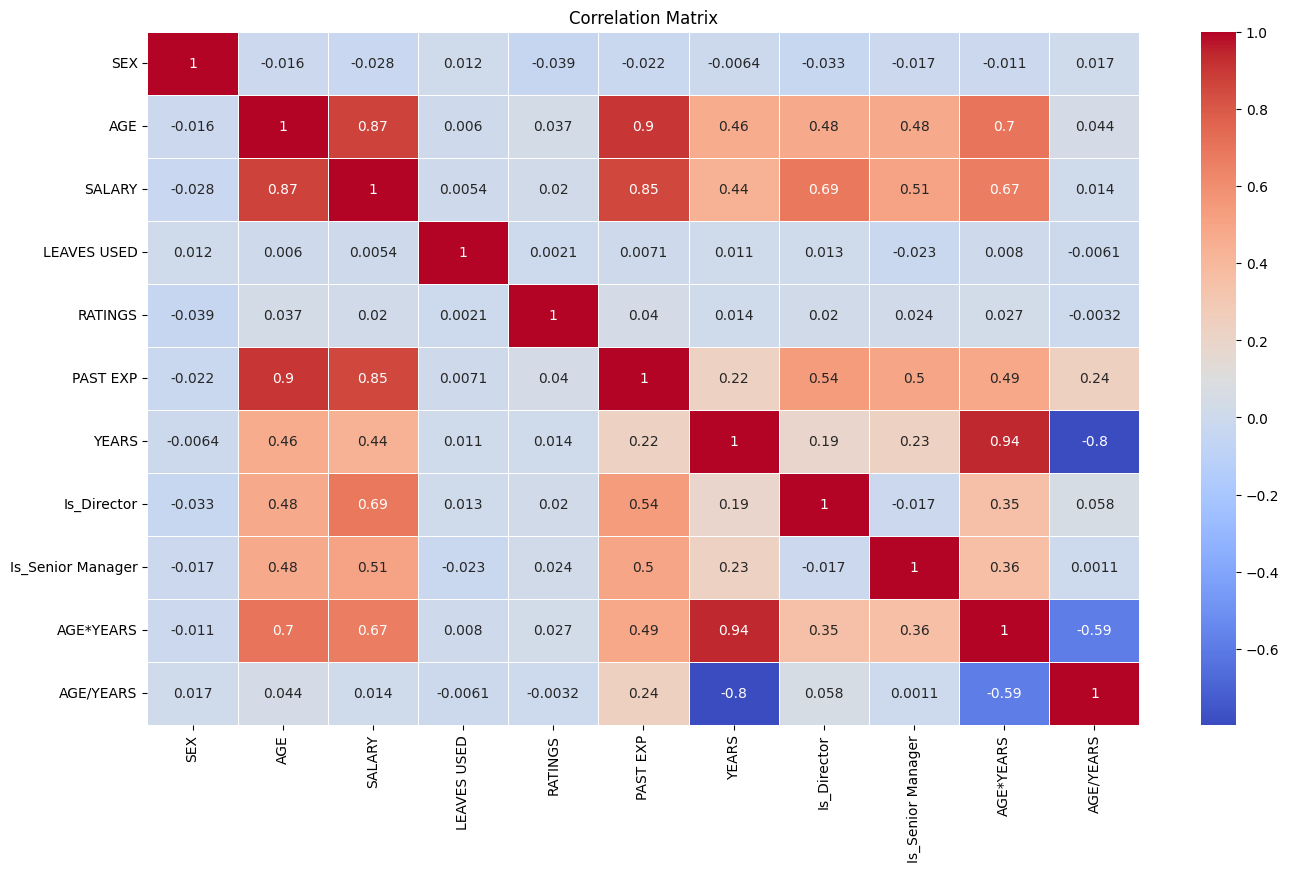
****

**Looking at the plot, we didn’t receive any valuable output from various departments. On the contrary, newer attributes were introduced having higher correlation with non – target variable. This was not intended. Thus, we can drop every UNIT and simplify our dataset.**

****

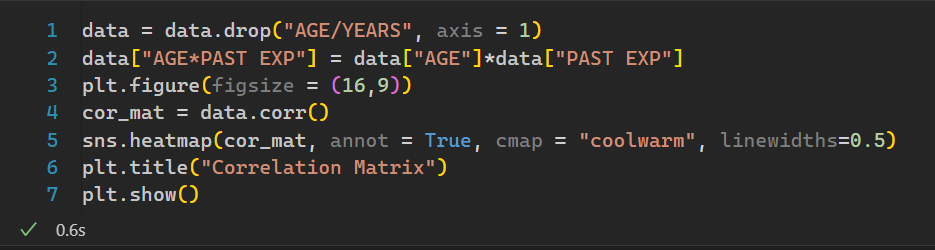
**Don’t get fooled with deep blue color. Every metric implies weak correlation with every other. What we can observe is that AGE and YEAR heavily define how salary alters. Also, they show very high correlation among themselves. We need to combine them into a single attribute to reduce dimesion without much loss of information.**

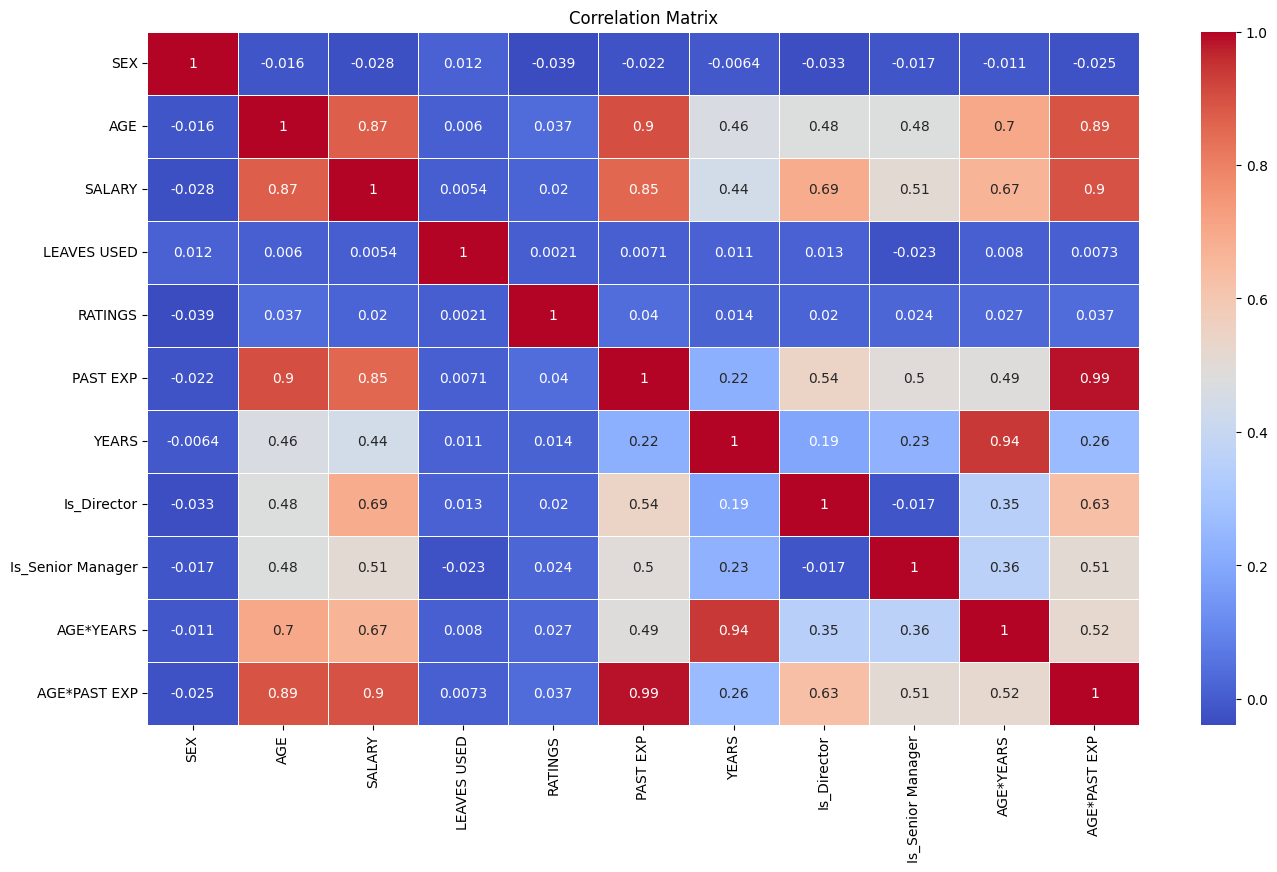
****

****

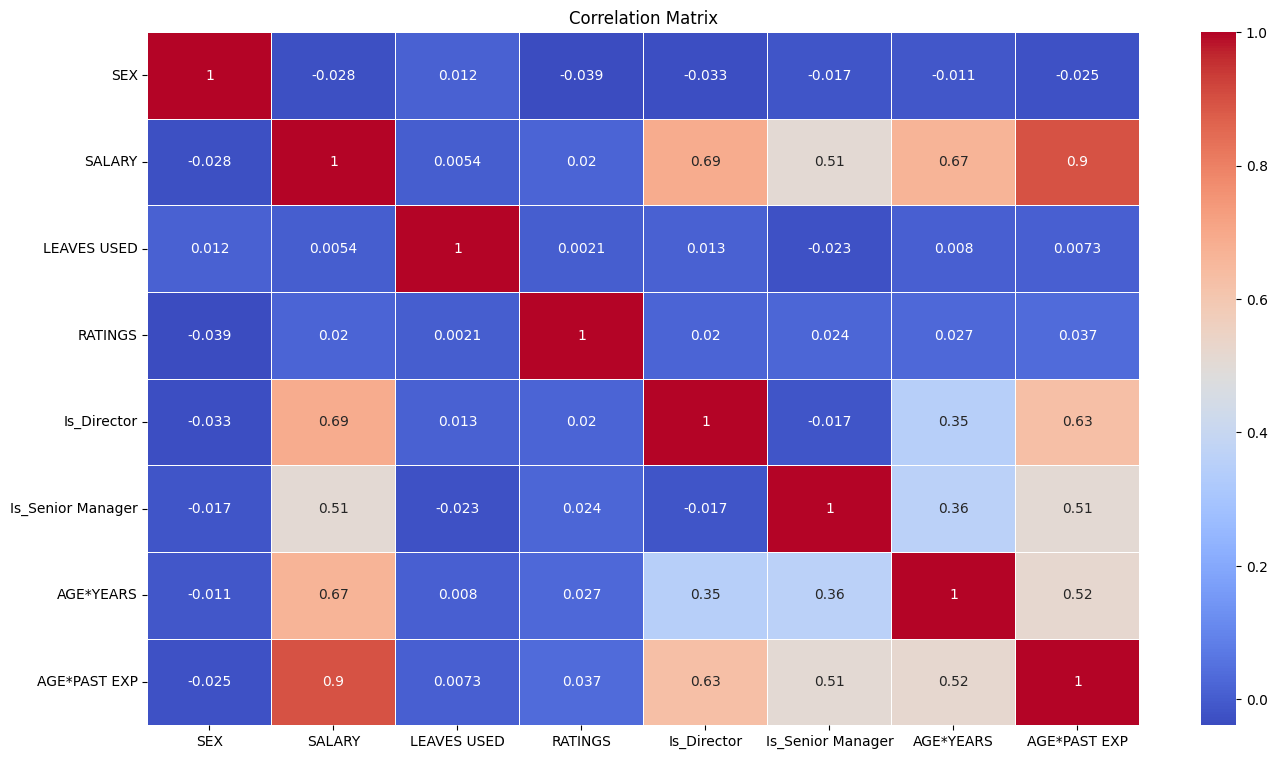
**Observe that AGE/YEAR is clearly not an improvement. Also, it shows very high correlation with YEARS. Since YEARS can be 0, the latest column diverges to infinity too‼ We need to drop it.**

**Also, AGE and PAST EXP heavily define how our salary looks like. We can combine these two.**

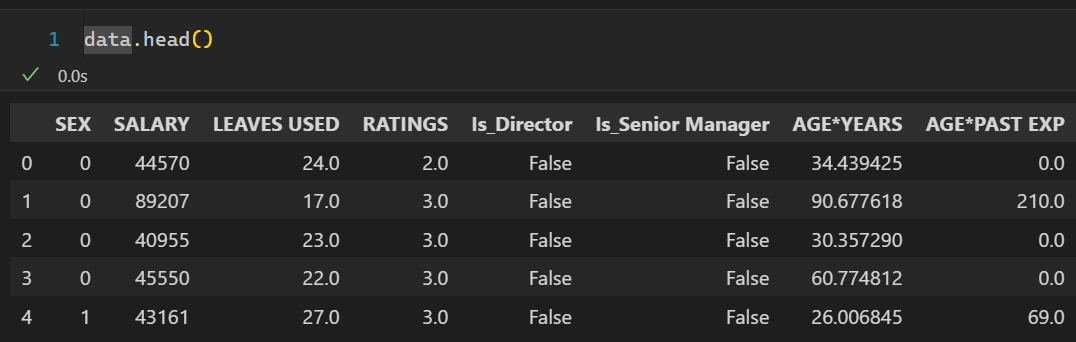
****

****

**Observe the very high correlation between AGE, PAST EXP and AGE\*PAST EXP. Since the last row better predictor of the prior, we can drop AGE and PAST EXP to achieve the optimal model for prediction.**

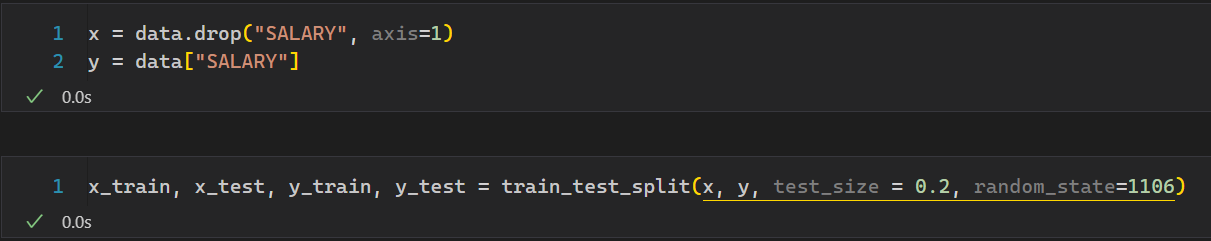
****

**Our Dataset is ready for model deployment. Here’s a bird’s eye view of how it looks like:**

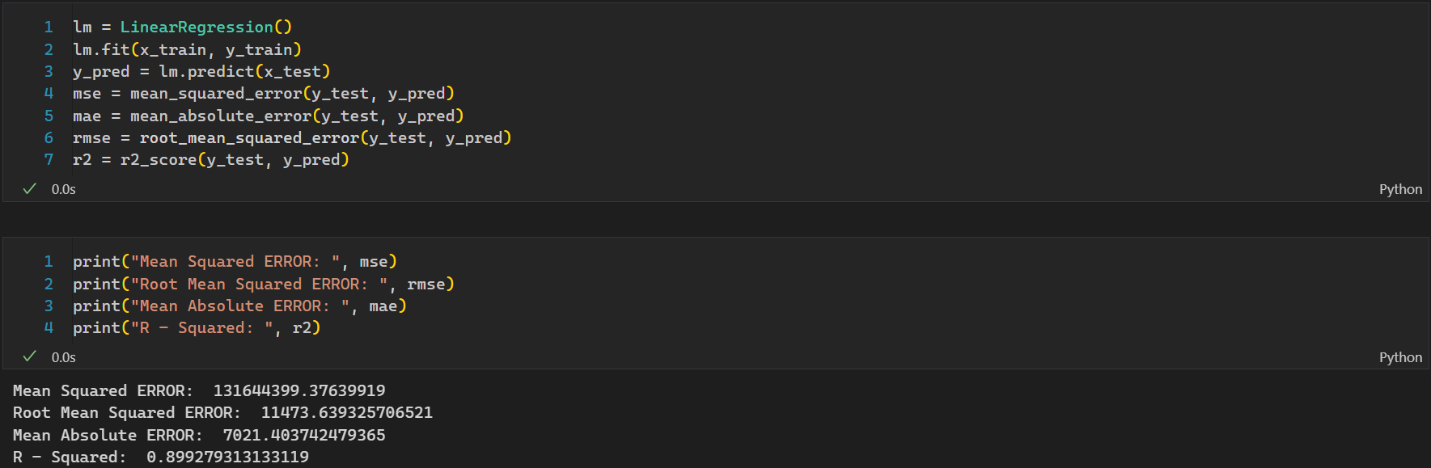
****

# Model Development

## Train Test Split



## LINEAR REGRESSION



MSE is given by,

Don’t get startled by the astoundingly high MSE. That’s squared error. The real information lies behind RMSE. This is given by,

We achieved RMSE of just above . Comparing that with the Standard Deviation of our target variable , That’s a really good score.

MAE is given by,

This tells the absolute deviation of true value from the predicted value. Lower the MAE, Better the fit. Though, the real inference is drawn from R-2 score. Value closer to 1 depicts a better fit. We got really high score from our model . For now, our model seems to be a good fit. Fitting decision tree next.

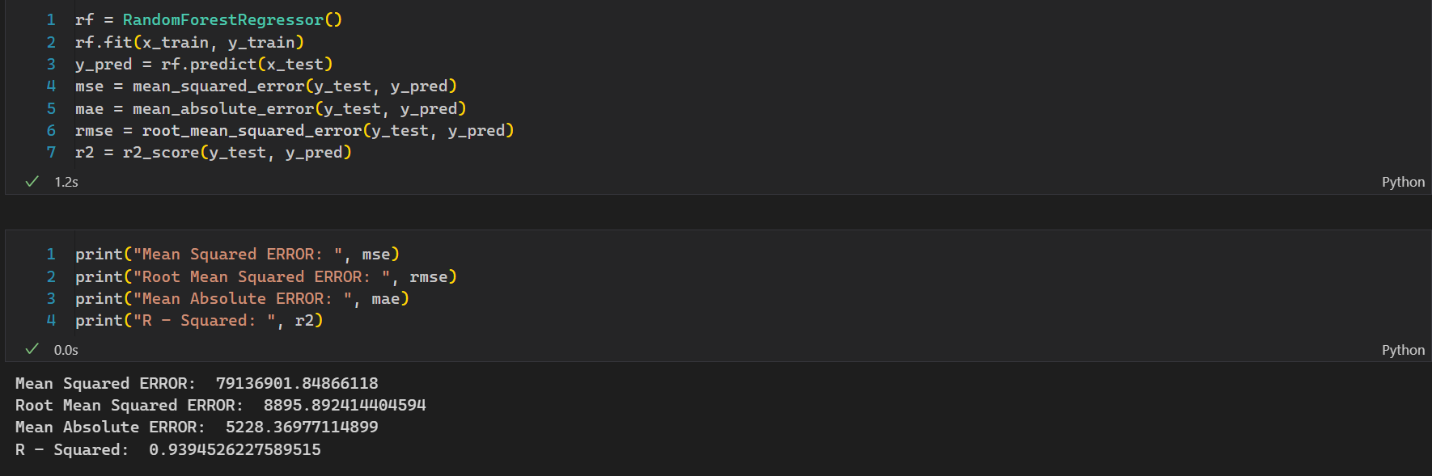
## DECISION TREE

## 

Looking at the score, we can observe:

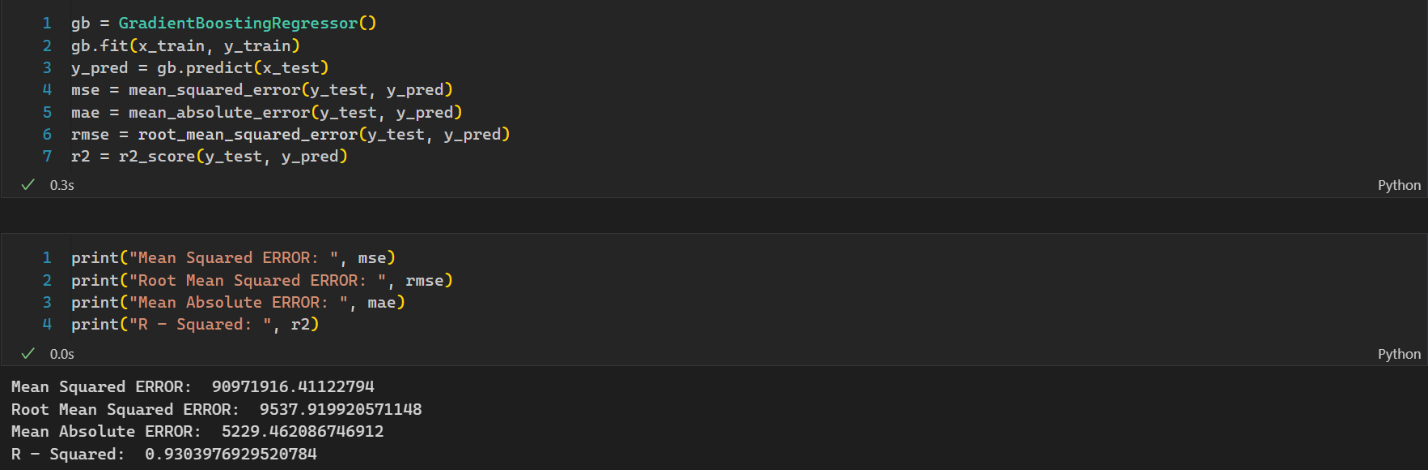
Although having accuracy over , the Linear Regression still lacks behind all 3 key metrics. We can inference Decision Tree is a better fit. Next up, fitting Random Forest.

## RANDOM FOREST



Random Forest seems a better fit than Decision Tree Regressor. A possible reason for that can be difference in algorithm. Decision Tree predicts output based up various decisions, and parameters fitted on the dataset. On the other hand, Random Forest tries to fit multiple forests at once on our training set, increasing accuracy and reducing overfitting. Our Random Forest Regressor is far better fit that the previous two models. Next up, Fitting Gradient Boosting.

## GRADIENT BOOSTING



The Random Forest model has a higher R² (0.9394) compared to the Gradient Boosting model (0.93039). This indicates that the Random Forest model captures a greater proportion of the variance in your data.

# Conclusion

1. **Linear Regressor**:
   * Linear regression assumes a linear relationship between features and the target variable. Random Forest, on the other hand, can handle non-linear relationships by combining multiple decision trees with various split points.
2. **Decision Tree Regressor**:
   * A single decision tree can suffer from overfitting, especially if it's grown too deep. This means it memorizes the training data too well and performs poorly on unseen data. Random Forest addresses this by averaging predictions from an ensemble of decision trees, reducing the impact of any single tree's overfitting.
3. **Gradient Boosting Regressor**:
   * Similar to Random Forest, Gradient Boosting uses an ensemble of decision trees. However, they are built sequentially, with each tree focusing on improving the errors of the previous ones. This can be very effective, but it also makes Gradient Boosting more susceptible to overfitting compared to Random Forest's averaging approach.

Here's how these differences might explain Random Forest's success:

* **Flexibility**: Random Forest can capture complex relationships through its ensemble of trees with various split points, making it adaptable to non-linear data.
* **Reduced Overfitting**: By averaging predictions from multiple trees, Random Forest reduces the influence of any single overfitted tree, leading to better generalization on unseen data.