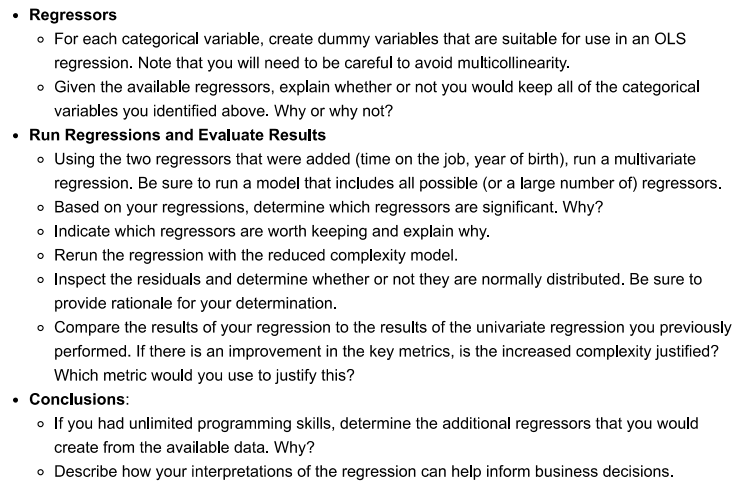
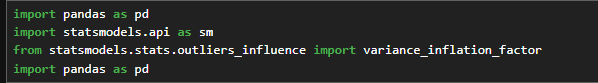


**Requirements**

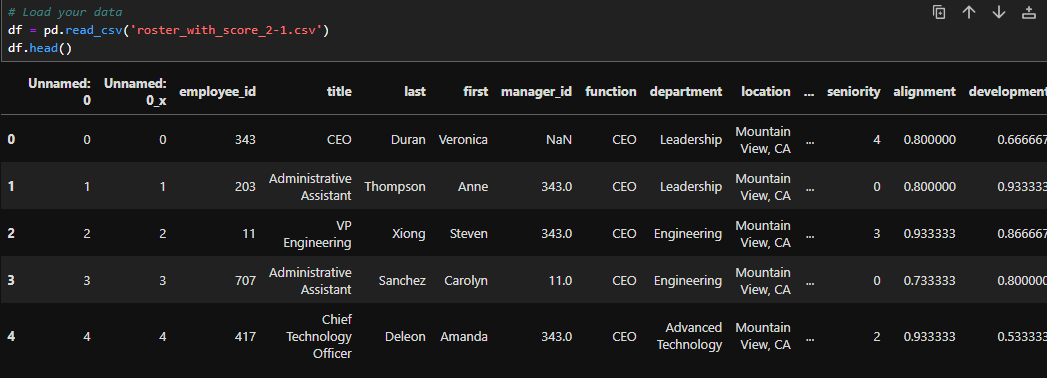


**Explanation**

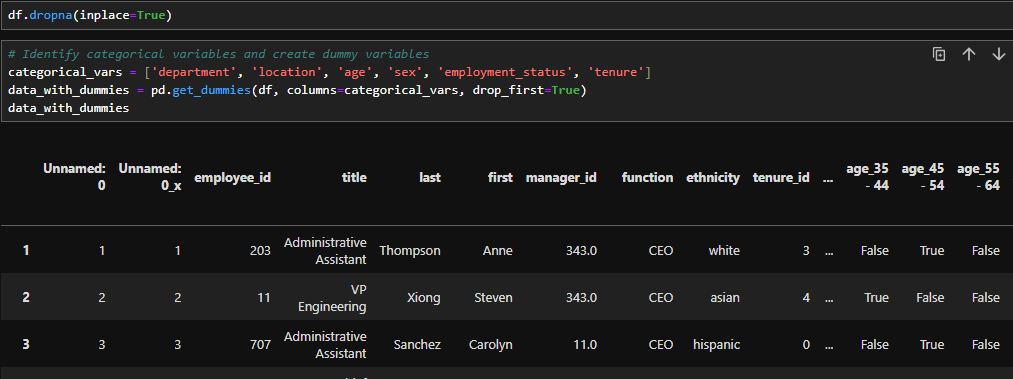
* First of all we need to import libraries that will be needing for our project

****

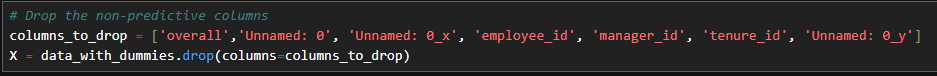
* Then we will load the dataset into data frame to start working on that



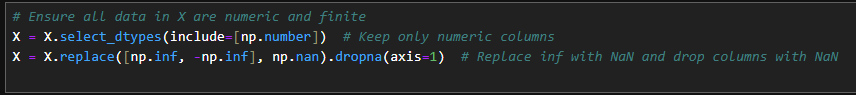
* Next step is to clean or preprocess the dataset like dropping the dataset , creating the dummy variables of categorical values and all other feature engineering on data



* Dropping the column that we don’t need to use



* Ensuring all the variables are finite and in numeric form to successfully work on it.



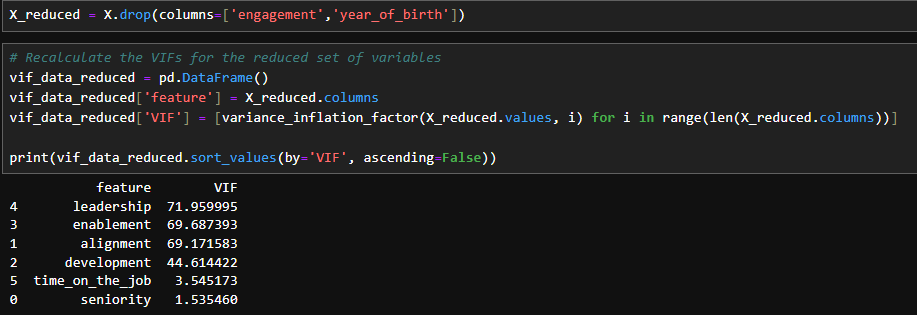
* In next step we are calculating the VIF to in X variable to check the multivariate regression and in results we get some of the highest VIF values for some features.

A screenshot of a computer program

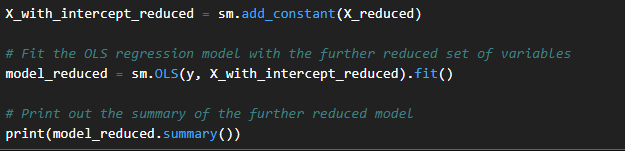
Description automatically generated

**Testing on Different Attributes**

Again dropping some features and finding its VIF to check multivariate regression and here we again got the VIF values but some of are having still high VIF.



Now we fitted the OLD model on above scenario of x\_reduced and prints its summary to see the results.



And the results of x\_reduced we got is

And we find that P>|t| is for senority and time\_on\_the\_job

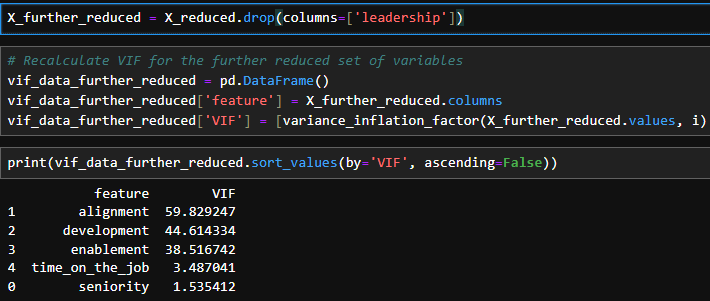
But here the problem is we got R squared is almost about 98% we can use this but this model can overfit on unseen data .

A screenshot of a computer

Description automatically generated

Again dropping the attribute to x\_further\_reduced and check its VIF values.

And we got some reduced VIF values which is preferable.



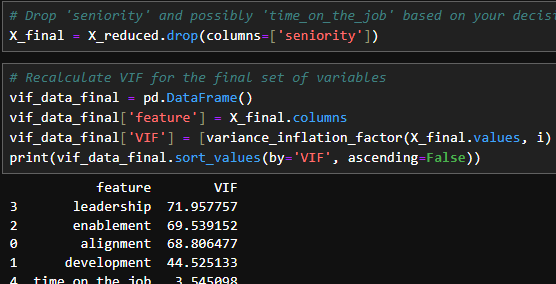
Here we checked the summary of Ols fitted model which giving R squared of 0.86 which is more preferable because it wont overfit the model and having more robust results.

A screenshot of a computer

Description automatically generated

**Dropping Seniority**

Now for checking more I tried to drop one more attribute to see its results

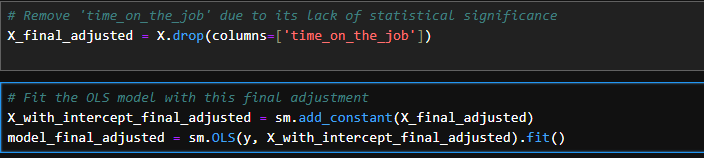


And the summary we got for this is same at 0.98 which can overfit mdoel

A screenshot of a computer

Description automatically generated

**Dropping Time\_on\_the\_job**



And the summary we got for this is

A screenshot of a computer screen

Description automatically generated

Which showing R =1 but condition number is too large so we cannot use that and this can also overfit the model

**Result Comparison of all above**

Here it might be the case for overfitting so the above wiht R squared wiht 0.86 will be good to go In the first model, most predictors are statistically significant with p-values < 0.05, except for 'seniority' and 'time\_on\_the\_job', which suggests they have a meaningful contribution to predicting 'overall'. This significance is a key factor in choosing predictors for a model.

Given these considerations, the first model (R-squared = 0.860) is recommended. It offers a robust, interpretable, and realistic assessment of the relationships between predictors and the dependent variable, making it a better choice for informing decisions or further analysis. However, it's also essential to validate this model on new or hold-out data to ensure its predictions are accurate and reliable in practice.

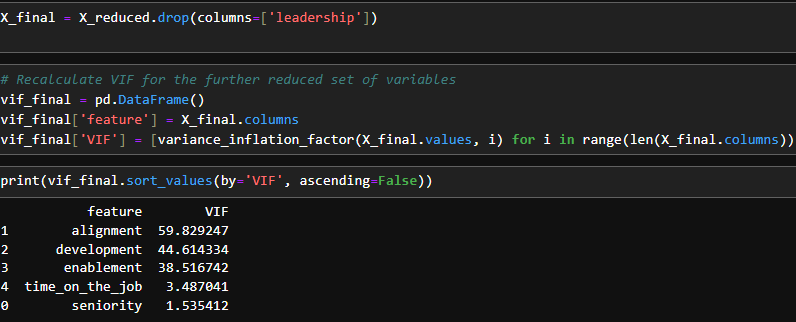
The model with an R-squared of 0.860 is the most appropriate choice for several reasons:

It demonstrates a strong fit without reaching the unrealistic perfection indicated by an R-squared of 1.000, which suggests overfitting.

The variables included show statistical significance, except for 'seniority' and 'time\_on\_the\_job', indicating a meaningful relationship with the dependent variable, 'overall'.

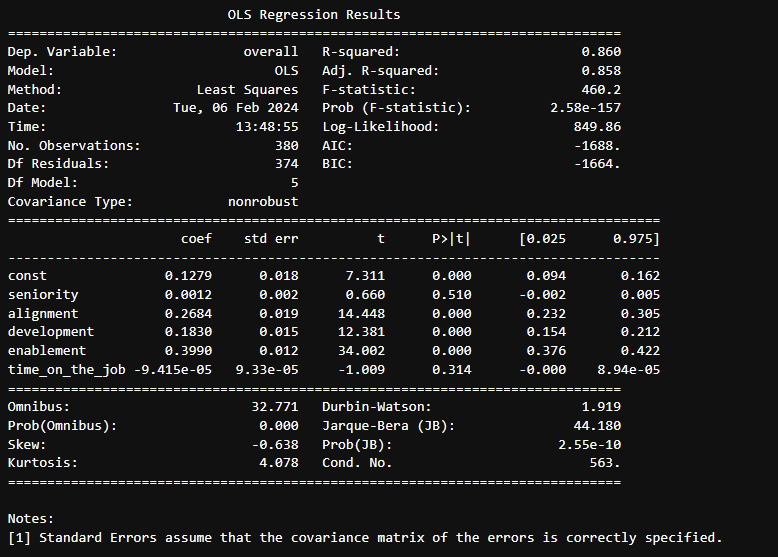
This model avoids the potential overfitting and numerical issues suggested by the perfect fit and extremely high F-statistic in the alternative molection.

**Final Rerun on reduce complexity model after comparing the results**



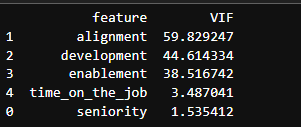
By considering all the constraints the best and robust result of Ols are that have attributes

Const, seniority, alignment, development, enablement, time\_on\_the\_job. Because only from placing these we get good results.

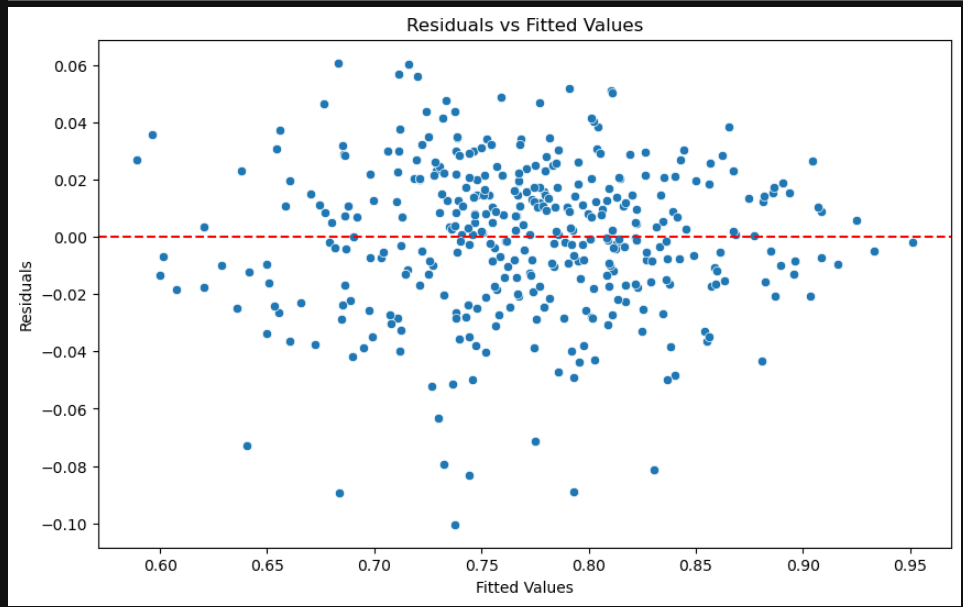
This will work fine than all the models we checked

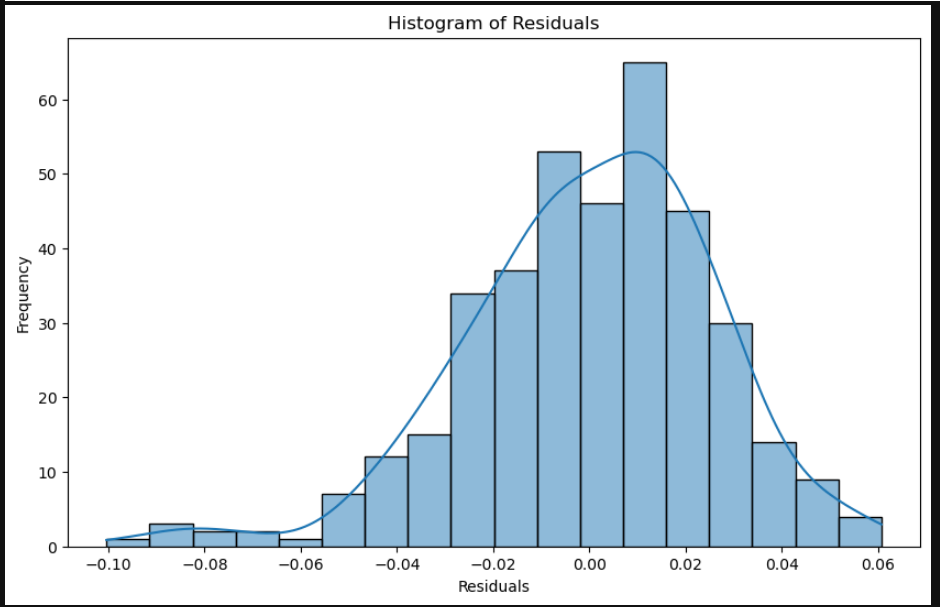
So we can say that the best and robust result have from datatset is considering this

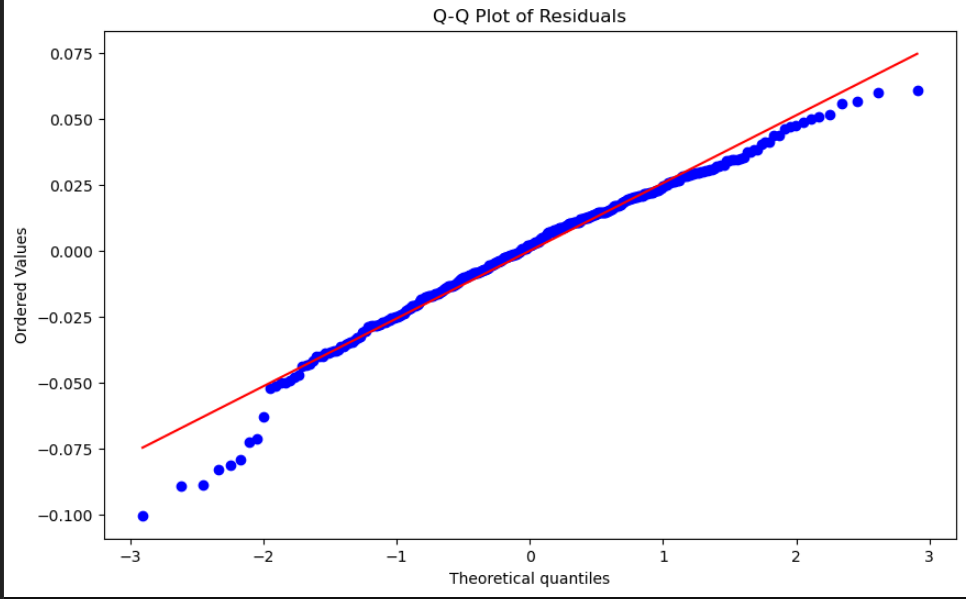
Which is giving R-Squared of 0.86 which show no overfitting



**Visualization**







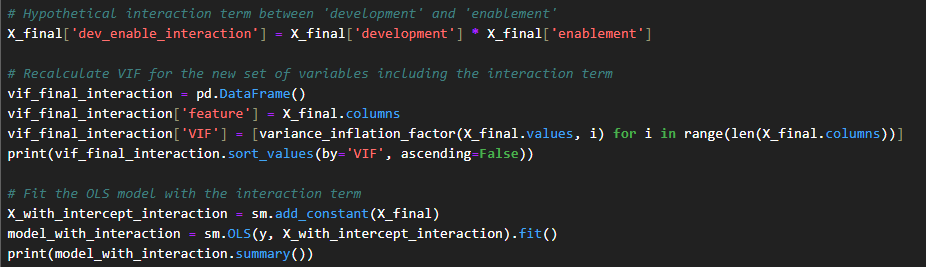
**Histogram of Residuals:** This plot suggests that the residuals are approximately normally distributed, as indicated by the bell-shaped histogram. However, there seems to be a slight skewness to the right. In practice, perfect normality is rare, and this level of skewness may not be a major concern unless it is extreme.

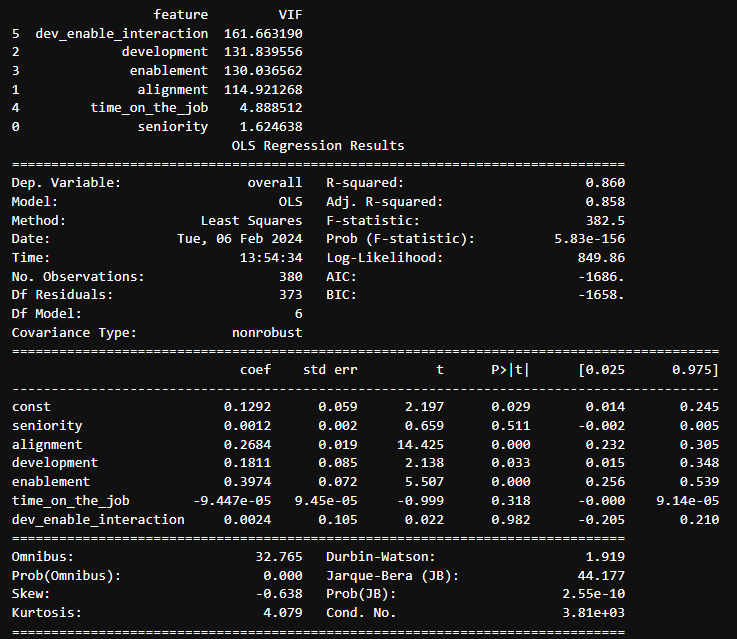
**Q-Q Plot of Residuals:** The Q-Q plot shows that the residuals deviate from the expected normal line at the tails, particularly at the lower tail. This indicates the presence of outliers or that the residuals have heavier tails than a normal distribution. This is common and may not necessarily violate OLS assumptions if not extreme.

**Residuals vs Fitted Values:** This scatter plot aims to detect non-constant variance (heteroscedasticity). The residuals should be randomly scattered around the horizontal line (y=0). There's no clear pattern (like a funnel shape), which is good, but there seems to be some spread. The spread might suggest slight heteroscedasticity, but there is no definitive pattern.

The Durbin-Watson statistic of approximately 1.91 suggests that there is no serious autocorrelation in the residuals, which is a positive indicator for the model.

**Conclusions: Additional:**

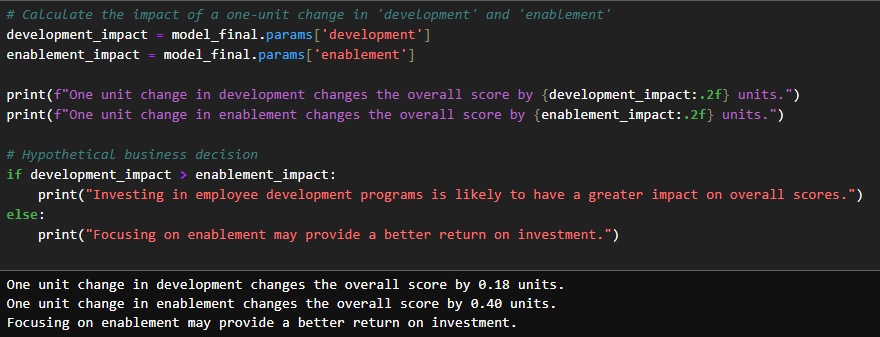




The introduction of interaction term between 'development' and 'enablement' has significantly increased variance inflation factors (vifs) for these variables and this indicating that multicollinearity has become a problem. The VIF for the interaction term itself is extremely high, which is often the case with interaction terms because they are and by definition and products of other variables.

Moreover, the interaction term is not statistically significant (p-value = 0.982), suggesting it does not add value to the model in its current form. The high condition number also indicates potential multicollinearity issues.

**Business Interpretation (Hypothetical):**



* One unit change in development changes the overall score by 0.18 units.
* This means that for each one-unit increase in the 'development' variable, the predicted overall score is expected to increase by 0.18 units, assuming all other variables remain constant.
* One unit change in enablement changes the overall score by 0.40 units.
* Similarly, this interpretation indicates that for each one-unit increase in the 'enablement' variable, the predicted overall score is expected to increase by 0.40 units, assuming all other variables remain constant.
* The statement, "Focusing on enablement may provide a better return on investment," suggests that increasing 'enablement' has a stronger positive impact on the overall score compared to 'development.' In other words, investing resources or efforts in improving 'enablement' is likely to yield a more significant increase in the overall score compared to investing in 'development.'
* This interpretation is based on the magnitudes of the coefficients. Since the coefficient for 'enablement' (0.40) is larger than that for 'development' (0.18), it implies that changes in 'enablement' have a greater influence on the overall score in the context of the regression model.
* However, it's important to note that these interpretations are based solely on the statistical results of the model. In practice, business decisions should consider a broader range of factors, including cost-effectiveness, feasibility, and alignment with strategic goals when deciding where to allocate resources or focus efforts.