CS 4372: Linear Regression Analysis Assignment Report

Sarvesh Gopalakrishnan (NET ID: sxg220257), Sreevasan Sivasubramanian (NET ID: sxs220434)

Link to Github: https://github.com/Sarvesh30/CS4372-Project-1

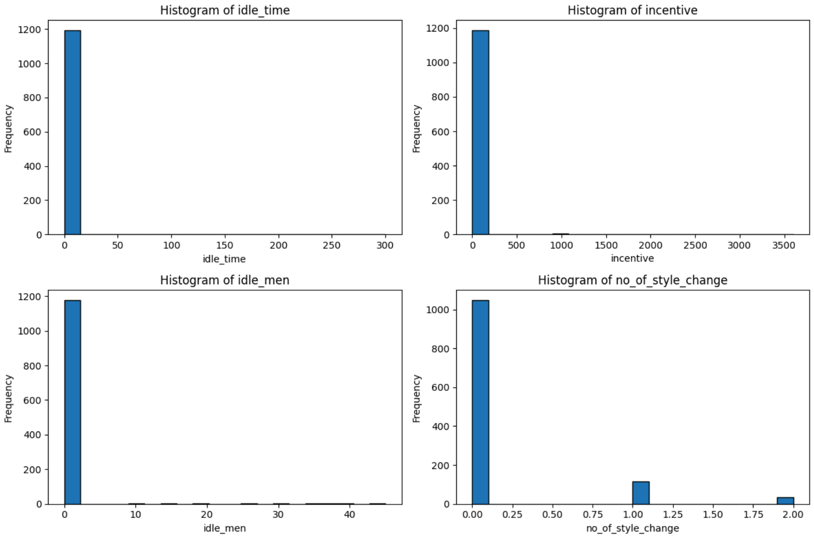
1. Project and Dataset Selection

In this assignment, we selected the *Productivity Prediction of Garment Employees* dataset from the UCI ML Repository: Link to dataset. This dataset consists of the production performance of various employee teams across two major departments (Sewing & Finishing) and four quarterly periods. Our goal in this project was to use the provided metrics such as team compositions, time efficiency statistics, department types, and production values to predict the actual productivity that was delivered by teams on a continuous range from 0 to 1. Additionally, we wanted to analyze which of the 12 teams were most significant to the prediction of actual productivity in the garment manufacturing process.

2. Regression Model Building

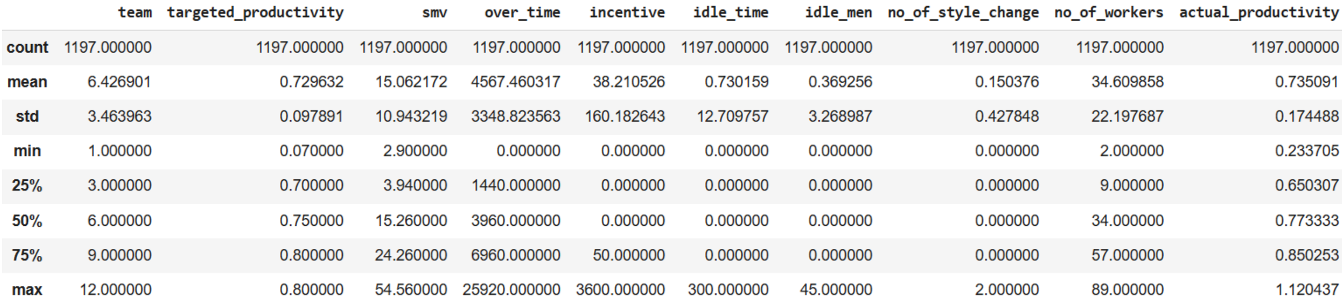
2.1. Pre-Processing

First, common pre-processing tasks such as checking for null values were completed. The wip (work in progress) variable had 506 missing values. Therefore, the wip feature was dropped from the dataset. Moreover, we assessed the distributions of idle_time, incentive, idle_men, and no_of_style change as they were filled with many zero values.



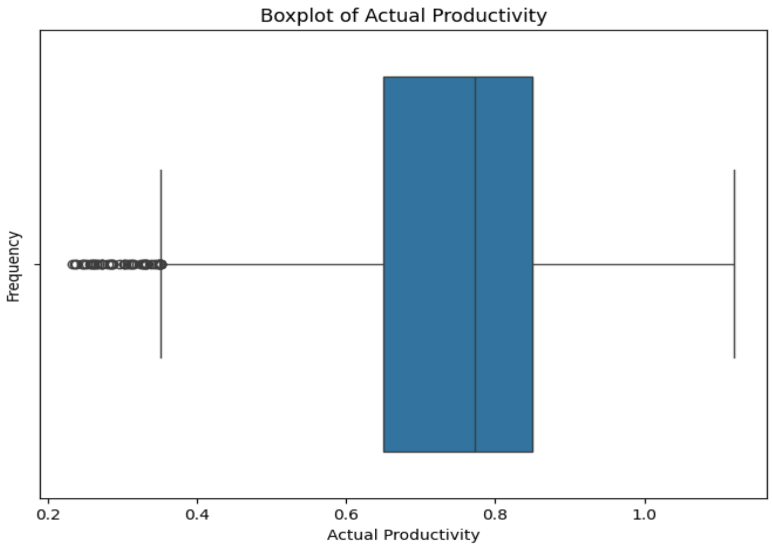
From the histograms above, it is clear that the distributions of idle_time, incentive, idle_men, and no_of_style_change are heavily skewed towards zero. These attributes are low in variance which means that they may not significantly contribute to predicting the actual productivity when building the linear regression models.

We also created a statistical summary for each of the numerical columns to get a better understanding of the variables. In the table below, we wanted to focus on analyzing the statistics for the target variable: actual productivity. It is notable that the values for actual productivity deviate by about 0.174 points on average from the mean productivity of 0.735. This shows that the target variable has measurable variance which makes the target adequate for model building.

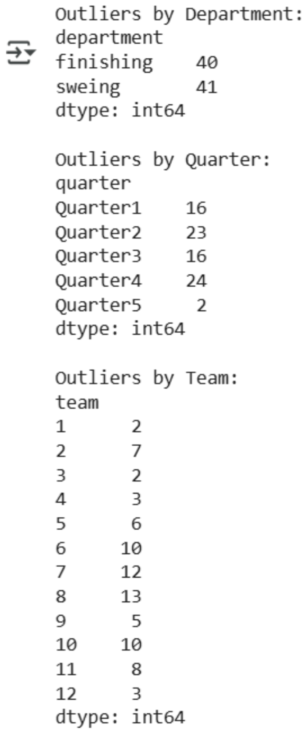


	team	targeted_productivity	smv	over_time	incentive	idle_time	idle_men	no_of_style_change	no_of_workers	actual_productivity
count	1197.000000	1197.000000	1197.000000	1197.000000	1197.000000	1197.000000	1197.000000	1197.000000	1197.000000	1197.000000
mean	6.426901	0.729632	15.062172	4567.460317	38.210526	0.730159	0.369256	0.150376	34.609858	0.735091
std	3.463963	0.097891	10.943219	3348.823563	160.182643	12.709757	3.268987	0.427848	22.197687	0.174488
min	1.000000	0.070000	2.900000	0.000000	0.000000	0.000000	0.000000	0.000000	2.000000	0.233705
25%	3.000000	0.700000	3.940000	1440.000000	0.000000	0.000000	0.000000	0.000000	9.000000	0.650307
50%	6.000000	0.750000	15.260000	3960.000000	0.000000	0.000000	0.000000	0.000000	34.000000	0.773333
75%	9.000000	0.800000	24.260000	6960.000000	50.000000	0.000000	0.000000	0.000000	57.000000	0.850253
max	12.000000	0.800000	54.560000	25920.000000	3600.000000	300.000000	45.000000	2.000000	89.000000	1.120437

To investigate the distribution of the target variable further, a boxplot of the actual productivity was created. The distribution of the actual productivity is skewed left due to data points where the actual productivity is less than 0.4. These data points most likely represent worker faults leading to poor productivity on a day.



We continued to analyze the worker faults by determining the faults with respect to the department, quarter, and team. From the results listed below, it shows that the worker faults are equal among the finishing and sewing departments. The worker faults with respect to the quarter are similar for quarter 1 and quarter 2 (close to 15) and quarter 3 and 4 (close to 25). On the other hand, quarter 5 had significantly less worker faults with only 2. The worker faults with respect to the team were less than 10 except for teams 6, 7, 8, and 10.

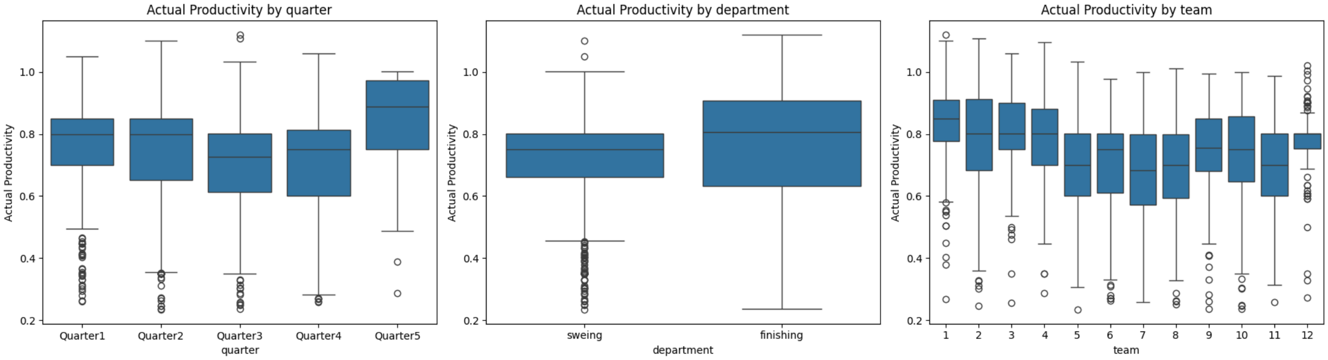


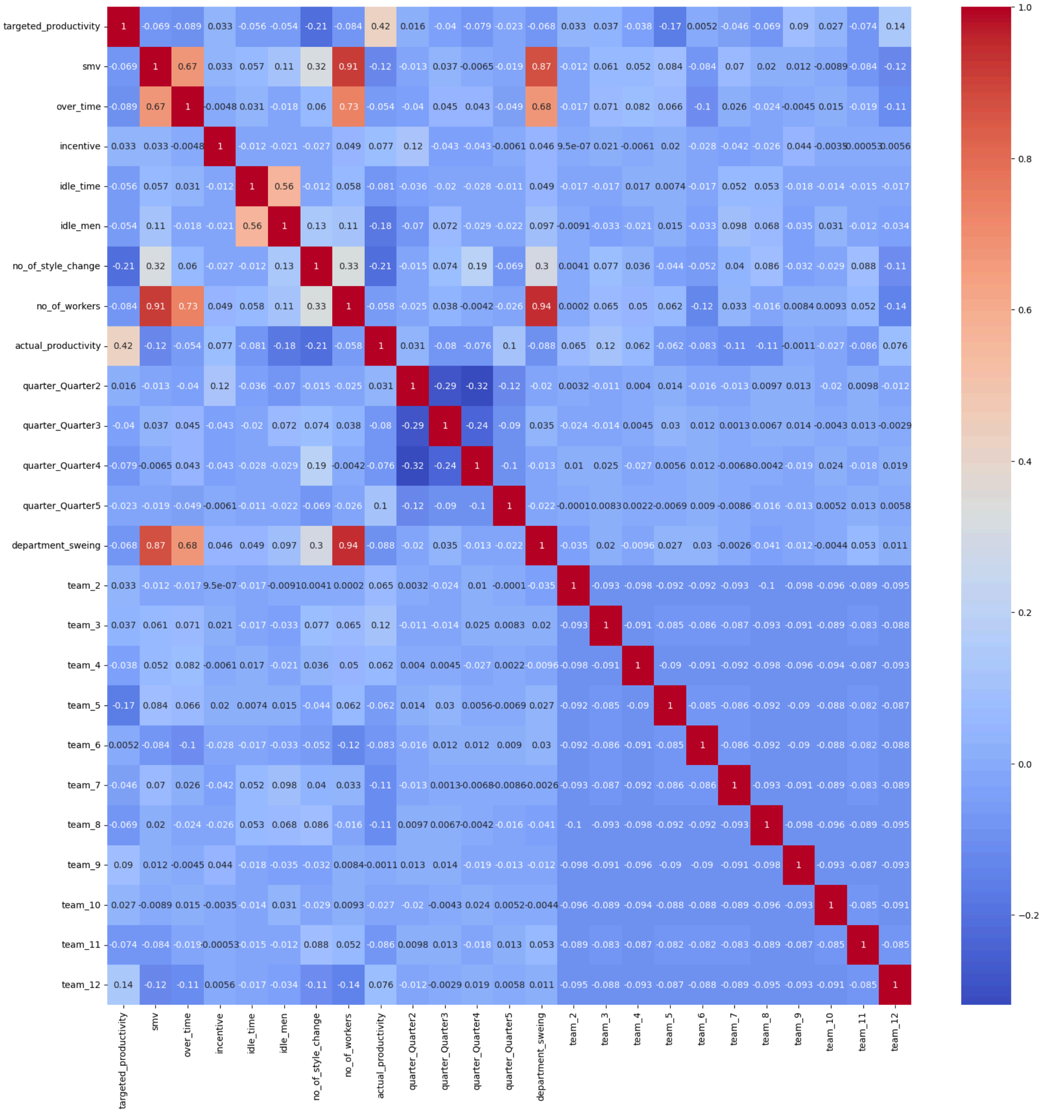
```
Outliers by Department:
department
finishing    40
sweing       41
dtype: int64

Outliers by Quarter:
quarter
Quarter1    16
Quarter2    23
Quarter3    16
Quarter4    24
Quarter5     2
dtype: int64

Outliers by Team:
team
1      2
2      7
3      2
4      3
5      6
6     10
7     12
8     13
9      5
10    10
11     8
12     3
dtype: int64
```

The box plots of actual productivity with respect to the quarter, department, and teams were also visualized. The distributions for each of the boxplots are mostly skewed left due to worker faults in every class. It is interesting to observe that the actual productivities that are less than 0.4-0.5 are outliers for the sewing department, but the actual productivities from 0.4-0.5 are not outliers for the finishing department. This shows that the actual productivity is clustered for the finishing department while it is more variable for the sewing department. Lastly, the actual productivity varies significantly across teams, departments, and quarters as seen by the differing quartiles and medians in the boxplots. Therefore, we converted these categorical variables to dummy encodings as they may be important factors in predicting the actual productivity.





	targeted_productivity	smv	over_time	incentive	idle_time	idle_men	no_of_style_change	no_of_workers	actual_productivity	quarter_Quarter2	quarter_Quarter3	quarter_Quarter4	quarter_Quarter5	department_sweing	team_2	team_3	team_4	team_5	team_6	team_7	team_8	team_9	team_10	team_11	team_12
targeted_productivity	1	-0.069	-0.089	0.033	-0.056	-0.054	-0.21	-0.084	0.42	0.016	-0.04	-0.079	-0.023	-0.068	0.033	0.037	-0.038	-0.17	0.0052	-0.046	-0.069	0.09	0.027	-0.074	0.14
smv	-0.069	1	0.67	0.033	0.057	0.11	0.32	0.91	-0.12	-0.013	0.037	-0.0065	-0.019	0.87	-0.012	0.061	0.052	0.084	-0.084	0.07	0.02	0.012	-0.0089	-0.084	-0.12
over_time	-0.089	0.67	1	-0.0048	0.031	-0.018	0.06	0.73	-0.054	-0.04	0.045	0.043	-0.049	0.68	-0.017	0.071	0.082	0.066	-0.1	0.026	-0.024	-0.0045	0.015	-0.019	-0.11
incentive	0.033	0.033	-0.0048	1	-0.012	-0.021	-0.027	0.049	0.077	0.12	-0.043	-0.043	-0.0061	0.046	9.5e-07	0.021	-0.0061	0.02	-0.028	-0.042	-0.026	0.044	-0.0035	-0.00053	0.0056
idle_time	-0.056	0.057	0.031	-0.012	1	0.56	-0.012	0.058	-0.081	-0.036	-0.02	-0.028	-0.011	0.049	-0.017	-0.017	0.017	0.0074	-0.017	0.052	0.053	-0.018	-0.014	-0.015	-0.017
idle_men	-0.054	0.11	-0.018	-0.021	0.56	1	0.13	0.11	-0.18	-0.07	0.072	-0.029	-0.022	0.097	-0.0091	-0.033	-0.021	0.015	-0.033	0.098	0.068	-0.035	0.031	-0.012	-0.034
no_of_style_change	-0.21	0.32	0.06	-0.027	-0.012	0.13	1	0.33	-0.21	-0.015	0.074	0.19	-0.069	0.3	0.0041	0.077	0.036	-0.044	-0.052	0.04	0.086	-0.032	-0.029	0.088	-0.11
no_of_workers	-0.084	0.91	0.73	0.049	0.058	0.11	0.33	1	-0.058	-0.025	0.038	-0.0042	-0.026	0.94	0.0002	0.065	0.05	0.062	-0.12	0.033	-0.016	0.0084	0.0093	0.052	-0.14
actual_productivity	0.42	-0.12	-0.054	0.077	-0.081	-0.18	-0.21	-0.058	1	0.031	-0.08	-0.076	0.1	-0.088	0.065	0.12	0.062	-0.062	-0.083	-0.11	-0.11	-0.0011	-0.027	-0.086	0.076
quarter_Quarter2	0.016	-0.013	-0.04	0.12	-0.036	-0.07	-0.015	-0.025	0.031	1	-0.29	-0.32	-0.12	-0.02	0.0032	-0.011	0.004	0.014	-0.016	-0.013	0.0097	0.013	-0.02	0.0098	-0.012
quarter_Quarter3	-0.04	0.037	0.045	-0.043	-0.02	0.072	0.074	0.038	-0.08	-0.29	1	-0.24	-0.09	0.035	-0.024	-0.014	0.0045	0.03	0.012	0.0013	0.0067	0.014	-0.0043	0.013	-0.0029
quarter_Quarter4	-0.079	-0.0065	0.043	-0.043	-0.028	-0.029	0.19	-0.0042	-0.076	-0.32	-0.24	1	-0.1	-0.013	0.01	0.025	-0.027	0.0056	0.012	-0.0068	-0.0042	-0.019	0.024	-0.018	0.019
quarter_Quarter5	-0.023	-0.019	-0.049	-0.0061	-0.011	-0.022	-0.069	-0.026	0.1	-0.12	-0.09	-0.1	1	-0.022	-0.0001	0.0083	0.0022	-0.0069	0.009	-0.0086	-0.016	-0.013	0.0052	0.013	0.0058
department_sweing	-0.068	0.87	0.68	0.046	0.049	0.097	0.3	0.94	-0.088	-0.02	0.035	-0.013	-0.022	1	-0.035	0.02	-0.0096	0.027	0.03	-0.0026	-0.041	-0.012	-0.0044	0.053	0.011
team_2	0.033	-0.012	-0.017	9.5e-07	-0.017	-0.0091	0.0041	0.0002	0.065	0.0032	-0.024	0.01	-0.0001	-0.035	1	-0.093	-0.098	-0.092	-0.092	-0.093	-0.1	-0.098	-0.096	-0.089	-0.095
team_3	0.037	0.061	0.071	0.021	-0.017	-0.033	0.077	0.065	0.12	-0.011	-0.014	0.025	0.0083	0.02	-0.093	1	-0.091	-0.085	-0.086	-0.087	-0.093	-0.091	-0.089	-0.083	-0.088
team_4	-0.038	0.052	0.082	-0.0061	0.017	-0.021	0.036	0.05	0.062	0.004	0.0045	-0.027	0.0022	-0.0096	-0.098	-0.091	1	-0.09	-0.091	-0.092	-0.098	-0.096	-0.094	-0.087	-0.093
team_5	-0.17	0.084	0.066	0.02	0.0074	0.015	-0.044	0.062	-0.062	0.014	0.03	0.0056	-0.0069	0.027	-0.092	-0.085	-0.09	1	-0.085	-0.086	-0.092	-0.09	-0.088	-0.082	-0.087
team_6	0.0052	-0.084	-0.1	-0.028	-0.017	-0.033	-0.052	-0.12	-0.083	-0.016	0.012	0.012	0.009	0.03	-0.092	-0.086	-0.091	-0.085	1	-0.086	-0.092	-0.09	-0.088	-0.082	-0.088
team_7	-0.046	0.07	0.026	-0.042	0.052	0.098	0.04	0.033	-0.11	-0.013	0.0013	-0.0068	-0.0086	-0.0026	-0.093	-0.087	-0.092	-0.086	-0.086	1	-0.093	-0.091	-0.089	-0.083	-0.089
team_8	-0.069	0.02	-0.024	-0.026	0.053	0.068	0.086	-0.016	-0.11	0.0097	0.0067	-0.0042	-0.016	-0.041	-0.1	-0.093	-0.098	-0.092	-0.092	-0.093	1	-0.098	-0.096	-0.089	-0.095
team_9	0.09	0.012	-0.0045	0.044	-0.018	-0.035	-0.032	0.0084	-0.0011	0.013	0.014	-0.019	-0.013	-0.012	-0.098	-0.091	-0.096	-0.09	-0.09	-0.091	-0.098	1	-0.093	-0.087	-0.093
team_10	0.027	-0.0089	0.015	-0.0035	-0.014	0.031	-0.029	0.0093	-0.027	-0.02	-0.0043	0.024	0.0052	-0.0044	-0.096	-0.089	-0.094	-0.088	-0.088	-0.089	-0.096	-0.093	1	-0.085	-0.091
team_11	-0.074	-0.084	-0.019	-0.00053	-0.015	-0.012	0.088	0.052	-0.086	0.0098	0.013	-0.018	0.013	0.053	-0.089	-0.083	-0.087	-0.082	-0.082	-0.083	-0.089	-0.087	-0.085	1	-0.085
team_12	0.14	-0.12	-0.11	0.0056	-0.017	-0.034	-0.11	-0.14	0.076	-0.012	-0.0029	0.019	0.0058	0.011	-0.095	-0.088	-0.093	-0.087	-0.088	-0.089	-0.095	-0.093	-0.091	-0.085	1

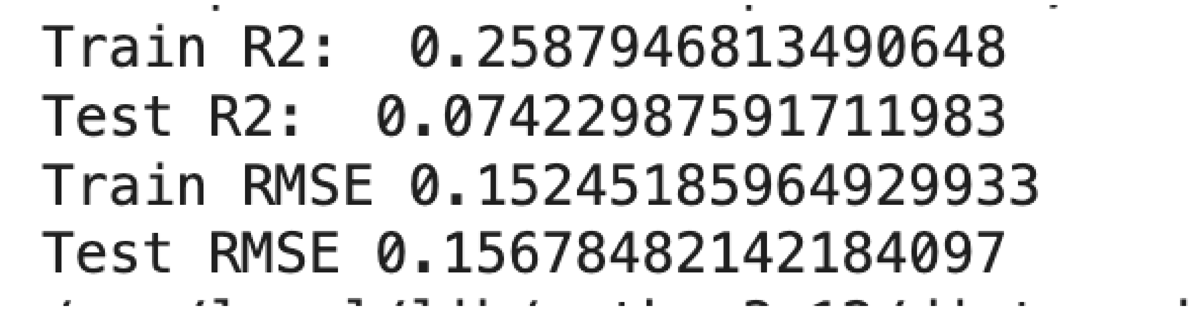
To determine the relationships between all the columns, we utilized a correlation matrix. This matrix shows the strength and direction of relationships between variables in a pairwise manner.

For instance, a positive value means that the variables rise together and negative values mean that there is an inverse relationship. We were able to analyze that the target productivity had a moderate, positive correlation with actual productivity. Out of all numerical features, the target productivity had the strongest relationship with actual productivity. Additionally, idle_men and no_of_style_change had a weak, negative, correlation with actual productivity. We also looked out for variables that were heavily correlated to each other to reduce the occurrence of multicollinearity in the model. Smv had a strong, positive correlation with no_of_workers and over_time, so it was dropped. Over_time and no_of_workers also had a strong, positive correlation and no_of_workers had a weak correlation with the actual productivity, so it was dropped as well. Our final model consisted of the targeted productivity, over_time, incentive, idle_men, and quarter/department/team dummy encodings.

Finally, we standardized our data to ensure that all features were on the same scale and created a 80/20 training and testing split to build the models to predict actual productivity.

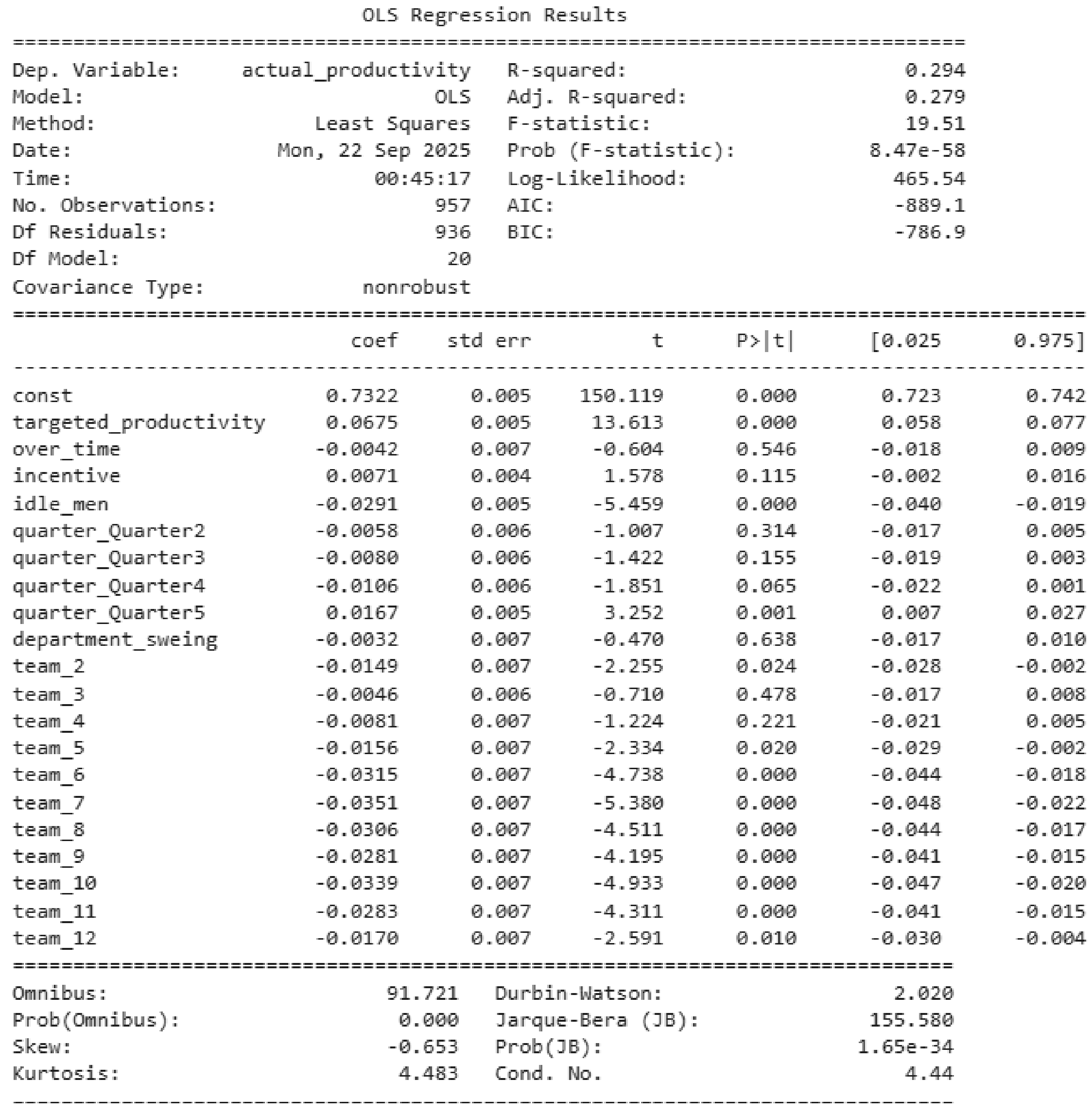
2.2 Model Construction

For model construction, we built a SGDRegressor with max iterations being 20000, random state being 42 and the tol=1e-4. We set up a dictionary with the different hypertuning parameters and then utilized a grid search to test different sets of parameters to help determine the best parameters for performance from this model. After that, we made predictions for the training and test data and then displayed the performance metrics for the best model. Our best parameters ended up being {'alpha': 0.1, 'eta0': 0.05, 'learning_rate': 'invscaling', 'max_iter': 15000}, and we also displayed the R^2 and RMSE values for both the test and train splits.



```
Train R2:  0.2587946813490648
Test R2:  0.0742298759171983
Train RMSE 0.1524518596492993
Test RMSE 0.1567848214218409
```

We also built an OLS model as well using Statsmodels and trained that model as well. Our results from that model are that the R^2 value was 0.294 and the F-statistic was 19.51.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     actual_productivity   R-squared:                       0.294
Model:                             OLS   Adj. R-squared:                  0.279
Method:                  Least Squares   F-statistic:                     19.51
Date:                 Mon, 22 Sep 2025   Prob (F-statistic):           8.47e-58
Time:                         00:45:17   Log-Likelihood:                 465.54
No. Observations:                  957   AIC:                            -889.1
Df Residuals:                      936   BIC:                            -786.9
Df Model:                           20
Covariance Type:             nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                     0.7322      0.005    150.119      0.000       0.723       0.742
targeted_productivity     0.0675      0.005     13.613      0.000       0.058       0.077
over_time                -0.0042      0.007     -0.604      0.546      -0.018       0.009
incentive                 0.0071      0.004      1.578      0.115      -0.002       0.016
idle_men                 -0.0291      0.005     -5.459      0.000      -0.040      -0.019
quarter_Quarter2         -0.0058      0.006     -1.007      0.314      -0.017       0.005
quarter_Quarter3         -0.0080      0.006     -1.422      0.155      -0.019       0.003
quarter_Quarter4         -0.0106      0.006     -1.851      0.065      -0.022       0.001
quarter_Quarter5          0.0167      0.005      3.252      0.001       0.007       0.027
department_sweing        -0.0032      0.007     -0.470      0.638      -0.017       0.010
team_2                   -0.0149      0.007     -2.255      0.024      -0.028      -0.002
team_3                   -0.0046      0.006     -0.710      0.478      -0.017       0.008
team_4                   -0.0081      0.007     -1.224      0.221      -0.021       0.005
team_5                   -0.0156      0.007     -2.334      0.020      -0.029      -0.002
team_6                   -0.0315      0.007     -4.738      0.000      -0.044      -0.018
team_7                   -0.0351      0.007     -5.380      0.000      -0.048      -0.022
team_8                   -0.0306      0.007     -4.511      0.000      -0.044      -0.017
team_9                   -0.0281      0.007     -4.195      0.000      -0.041      -0.015
team_10                  -0.0339      0.007     -4.933      0.000      -0.047      -0.020
team_11                  -0.0283      0.007     -4.311      0.000      -0.041      -0.015
team_12                  -0.0170      0.007     -2.591      0.010      -0.030      -0.004
==============================================================================
Omnibus:                       91.721   Durbin-Watson:                   2.020
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              155.580
Skew:                          -0.653   Prob(JB):                     1.65e-34
Kurtosis:                       4.483   Cond. No.                         4.44
==============================================================================
```

2.3 Result Analysis

From our results with the SGDRegressor, we can see that our training R^2 value was 0.25 and our test R^2 was 0.07. R^2 is a statistic that ranges from 0 to 1 and indicates whether the proportion of variance in the dependent variable is explained by the independent variable. The ideal R^2 value should be above 0.5 although we see that our value is below that in regards with both train and test. The model was built with the assumptions of linearity so our data might not be the best to capture a linear relationship with actual productivity leading to low performance metrics. The RMSE value, or the root mean squared error, is a way to measure how far the model's predictions are from the actual values. The RSME value for training data was 0.152 and for test data was 0.157. This indicates that on average, the model's predictions deviate from the actual productivity values by 0.15 points on the 0-1 productivity scale. The RMSE values indicate that the model is not overfitting because they are close to each other. The target range is from 0-1 so our values are moderate but could be better. The model overall doesn't generalize well and has weak performance.

For the OLS Regressor, the R^2 value is 0.294 meaning that about 29.4% of the variance in the target value, actual_productivity, was explained by the model. The test R^2 value was 0.1596. This indicates that the model performance is weak as there is only a small fraction of variance and since the train R^2 is higher than the test, the model fits the training data better so there is slight overfitting. The RMSE values for the training set is 0.1488 and for the testing set is 0.1494. This indicates that the model isn't necessarily overfitting and that it generalizes well but the model is weak.

The adj. R^2 value, which accounts for the number of predictors by adding a penalty for including variables that do not significantly reduce the error, is 0.279 which means that around 27.9% of the variance in actual productivity is explained after adjustment.

The F-statistic(19.51) shows that the overall regression model is statistically significant. The Prob(F-statistic 8.47e-58) is around 0, showing that the model is jointly significant. The other statistics that talk about likelihood, such as the Log-Likelihood(465.54) where a higher value indicates a better likelihood fit, the AIC(-889.1), where there is a relatively good fit for the number of parameters, and the BIC(-786.9), which is also low and no penalty for extra predictors, all indicate that this model has good likelihood for the data.

The coefficient table shows an estimated regression coefficient for each row. The coef value helps determine the expected change in the actual productivity for a one unit increase in a selected predictor, while holding all the other predictors constant. The targeted_productivity with a coef of 0.0675 had the highest effect where one unit increase is associated with a 0.0675 rise in actual_productivity.

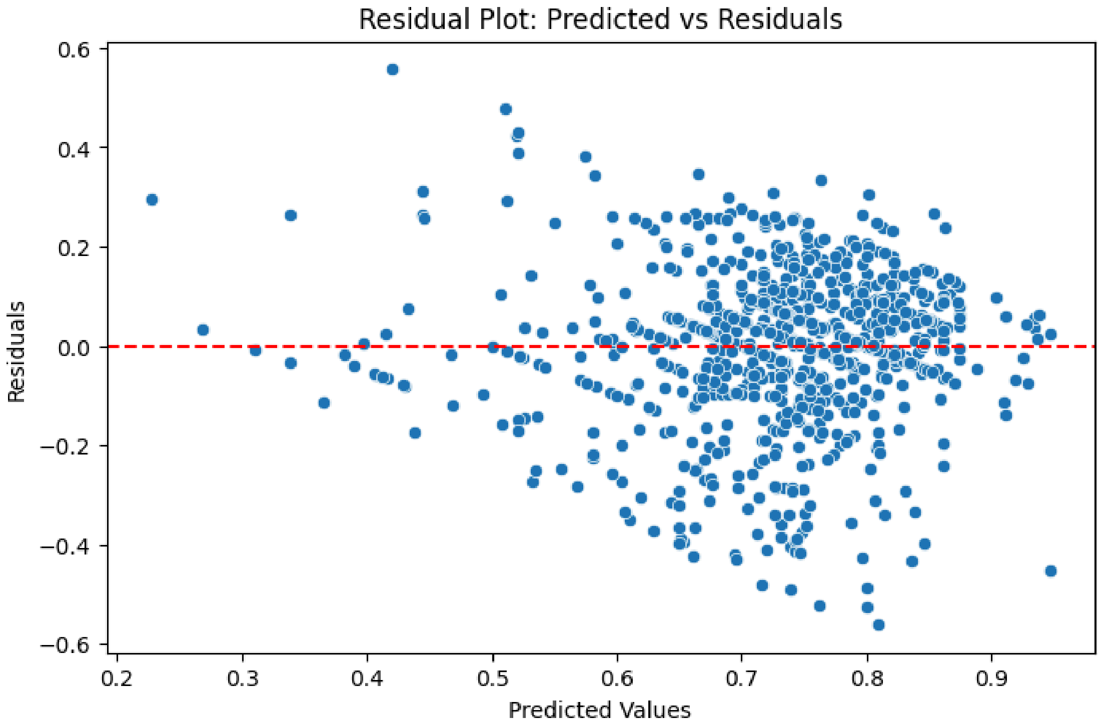
The std err shows the standard error of the coefficient estimate. The smaller the standard error, the more confident we are with the coefficient estimate. In the model summary, the standard errors for all features is significantly low (around 0.005-0.007) showing that the coefficient estimates are precise.

The t statistic is a ratio between the coef and the std err and shows whether a particular coefficient is significantly different from zero when holding all other predictors constant. The larger the t value, the greater the statistical significance of the predictor having a real effect on actual productivity. P > |t| shows whether the p value is statistically significant. Predictors which have a p-value < a = 0.05 are statistically significant in predicting actual productivity. To the right of the p-values, the 95% confidence intervals are also provided for each predictor. As long as zero is not included in the 95% confidence interval for a predictor, the predictor is statistically significant in the model. The predictors that are not significant in this model are over_time, incentive, quarter 2/3/4, department sewing, and team 3/team 4 as their p-values are greater than the significance level (a = 0.05).

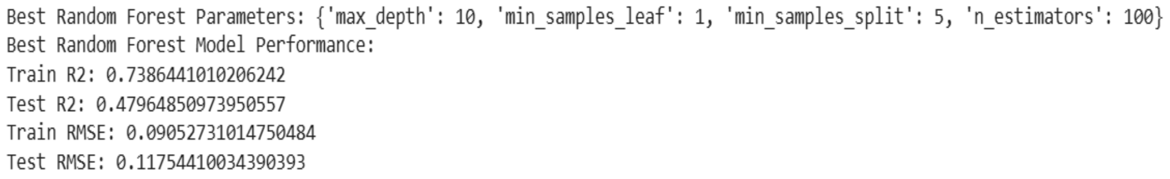
3. Final Results

From the result analysis, it is clear that the linear regression models do not perform well when predicting the actual productivity. The test R2 ranges from 0.07 to 0.30 for the SGDRegressor and OLS models and the testing RMSE is around 0.15 for both models. After obtaining these results, we went further by fitting linear regression models with different subsets of features by dropping statistically insignificant variables. This did not significantly improve our results as the R2 and RMSE were still around 0.30 and 0.15. The poor performance can be explained by the “outliers” of the dataset as worker production in a real-world setting most likely does not follow a linear relationship with time efficiency and production values. This can be explained by unexpected worker faults that can occur, leading to unpredictable drops in actual productivity.

We created a residual plot from the results of the OLS to assess linearity. From the residual plot, we were able to see that there is a slanted pattern in the residuals as the predicted values increase. To assume linearity, the points on the residual plot should be randomly scattered with no significant pattern.



One of our main objectives of the project was to build a model that could predict the actual productivity given the features in the dataset. Hence, we tried to train our dataset on a random forest regressor to see if we could improve the performance of the R^2 and RMSE values. We decided to use decision trees as they help with capturing non-linear relationships between the features and target. The random forest model improved the Test R^2 to 0.48 and the test RMSE to 0.12. In the future, we want to explore a larger hypertuning grid for random forest and train/test on different models such as XGBoost and AdaBoost to improve the prediction of actual productivity.



```
Best Random Forest Parameters: {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 100}
Best Random Forest Model Performance:
Train R2: 0.7386441010206242
Test R2: 0.47964850973950557
Train RMSE: 0.09052731014750484
Test RMSE: 0.11754410034390393
```