Productivity Prediction of Garment Industry Employees

Team Titans: Wilkister Mbaka, Sharon Olago, Margaret Gathoni and Gozzo Evrard.

Links: Powerpoint, Github, Colab

1. Business Understanding

Business Overview

The garment industry deals with the production of different types of clothing. It is a trillion-dollar industry on the global scale (reference), and is a labor- and capital-intensive industry. Workforce productivity in general refers to a measurement of goods and/or services produced by a group of workers within a given time period. High productivity levels by a team translate to higher profits, and are an indication of efficient use of the inputs in the production process. Worker productivity in the garment industry should therefore be maximized as it is a key factor to its success.

Low workforce productivity in a company can lead to reduced profitability, increased conflict, high employee turnover, and a lack of motivation. It is therefore important for a company to track its levels of productivity and investigate the factors that influence this.

Business Objective

General Objective

Create a model that can predict the level of productivity of employees in the garment industry.

Business Success Criteria

- Visualizing the relationships between actual productivity and the predictor variables.
- Building a model that can predict employee productivity.
- Identifying the top factors influencing the productivity level of employees.

Assessing the Situation

1. **Resource Inventory**
    a. Datasets:

i. Garment Worker Productivity Dataset[link]

b. Software(Github, Google Collaboratory, Ms Word,JIRA,Canva,Tableau)

2. **Assumptions**
    a. The data provided is correct and up to date
3. **Constraints**
    a. There are no constraints

Data Mining Goals

Our data mining goals for this project are as follows:

- Investigate the relationships between level of productivity and the predictor variables.
- Determine the most and least productive day of the week.
- Build a model that can predict employee productivity.
- Determine the top factors influencing the productivity level of employees.

2. Data Understanding

Data Understanding Overview

For this data analysis project, we are using the availed dataset by the company. These datasets are

★ **Garment Worker Productivity Dataset** - This sample contains data from January 1 to March 11, 2015 from the garment factory about worker productivity for different departments and teams.

Data Description

We have one dataset available for this project. A detailed description of the datasets is provided as follows:

❖ **Garment Worker Productivity Dataset** - This dataset provides information on the garment factory on a daily basis and work to be done and the productivity achieved in each team. It consists of **15 columns** and **1197 rows.**

Below is a brief description of what the column names represent and their definitions.

1. **date** : Date in MM-DD-YYYY
2. **day** : Day of the Week
3. **quarter** : A portion of the month. A month was divided into four quarters
4. **department** : Associated department with the instance
5. **team_no** : Associated team number with the instance
6. **no_of_workers** : Number of workers in each team
7. **no_of_style_change** : Number of changes in the style of a particular product

8. **targeted_productivity** : Targeted productivity set by the Authority for each team for each day.
9. **smv** : Standard Minute Value, it is the allocated time for a task
10. **wip** : Work in progress. Includes the number of unfinished items for products
11. **over_time** : Represents the amount of overtime by each team in minutes
12. **incentive** : Represents the amount of financial incentive (in BDT) that enables or motivates a particular course of action.
13. **idle_time** : The amount of time when the production was interrupted due to several reasons
14. **idle_men** : The number of workers who were idle due to production interruption
15. **actual_productivity** : The actual % of productivity that was delivered by the workers. It ranges from 0-1.

Verifying Data Quality

The Autolib dataset had the following missing values:

- The work in progress column(wip) had 506missing values

3. Data Preparation

These are the steps followed in preparing the data

1. **Loading Data**

   Loaded the dataset into google colab and using pandas created dataframes.

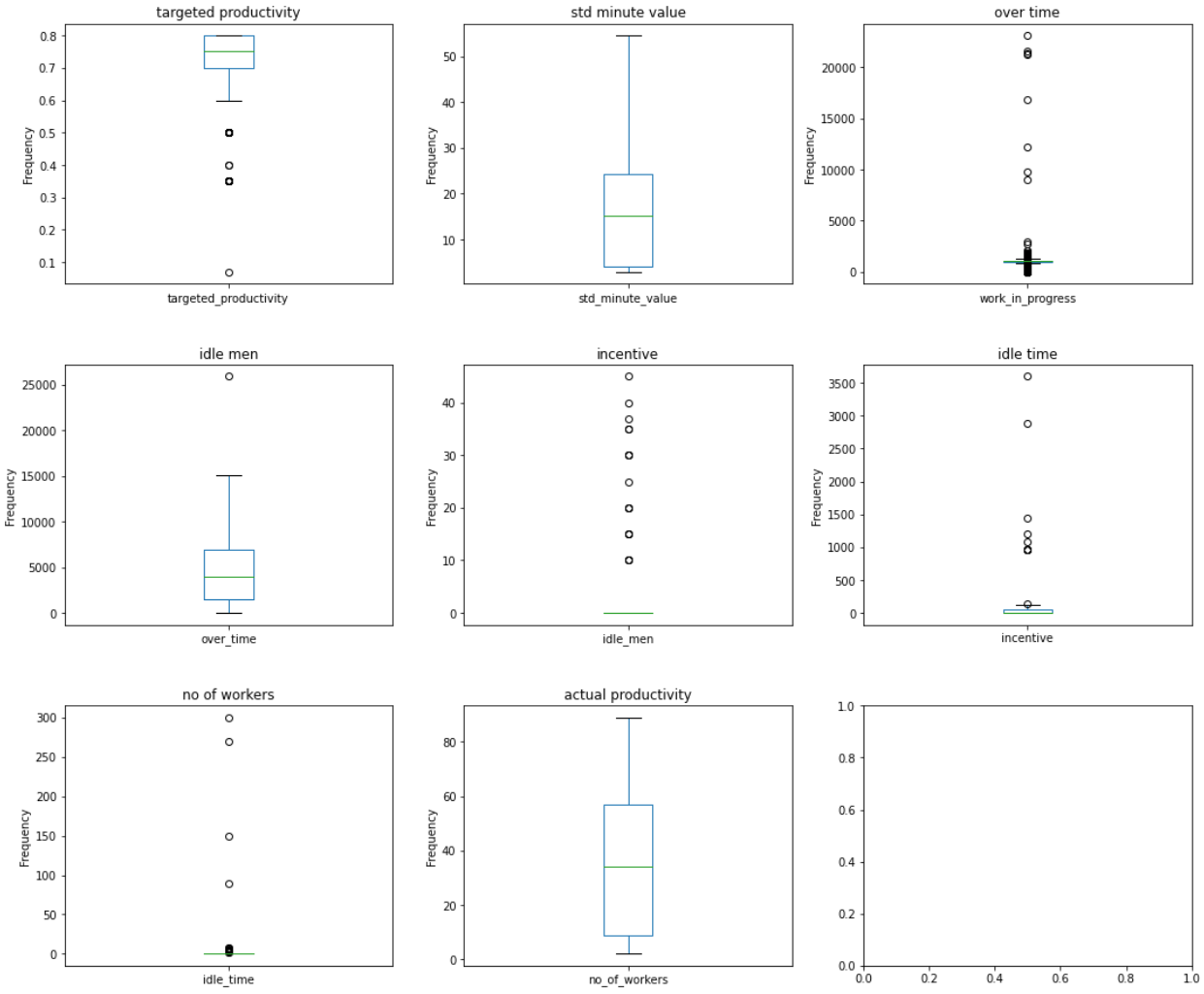
2. **Cleaning Data**

   The following actions were taken in cleaning the Autolib dataset

   a. Renamed shortened columns ie **wip** as **work_in_progress** and **smv** as **std_minute_value**
   b. There were missing values in the **work_in_progress** column which were imputed with the median value of the column
   c. There were no duplicates in the dataset
   d. The **no_of_workers** column had decimal numbers which is an anomaly because people cannot be fractions. The numbers were truncated to integers.
   e. In the **quarter** column, a month was divided into 4 quarters. However, there is a 'Quarter5' value observed which only consisted of 3 days 29, 30, 31. The 'Quarter5' value was changed to quarter 4.
   The word quarter was dropped from the column values and the numbers changed from strings to integers.
   f. Also under the **department** column, the spelling of sewing is wrong (sewing) and finishing appears as a unique value twice due to

whitespace. They were all replaced with the correct spelling and no white space.

g. The **date** column was a string and was converted to a datetime datatype.
h. Checked for outliers in the numerical columns



There were outliers in **targeted productivity, overtime, incentive, idle time, idle men, and actual productivity** columns. They will not be dropped as they are likely due to natural variability in the workflow of different teams, where some teams perform significantly above or below average in terms of time, pending work, and productivity. Also, the performance of a particular team can vary on different days, with some days being significantly above or below average.

3. Merging of the Datasets

No need for merging because there is only one dataset.

4. **Deriving New Attributes**

   No new Attributes were derived from the dataset.

4. Analysis

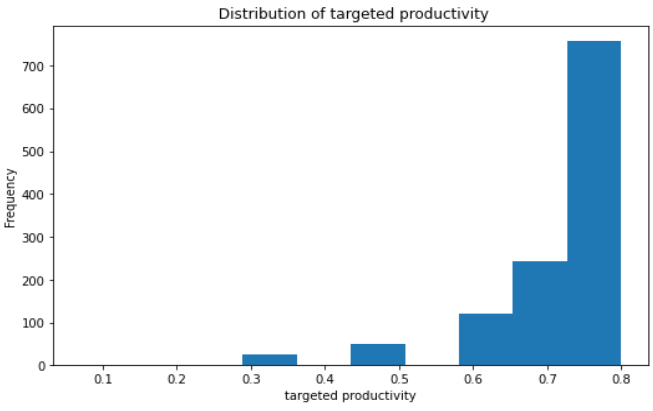
Univariate Analysis

Statistical Analysis of the Numerical Columns

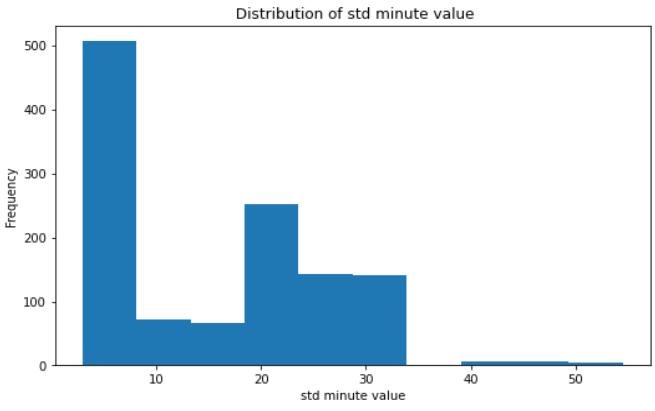
	targeted_productivity	std_minute_value	work_in_progress	over_time	idle_men	incentive	idle_time	no_of_workers	actual_productivity
count	1197.000000	1197.000000	1197.000000	1197.000000	1197.000000	1197.000000	1197.000000	1197.000000	1197.000000
mean	0.729632	15.062172	1126.437761	4567.460317	0.369256	38.210526	0.730159	34.551378	0.735091
std	0.097891	10.943219	1397.653191	3348.823563	3.268987	160.182643	12.709757	22.152559	0.174488
min	0.070000	2.900000	7.000000	0.000000	0.000000	0.000000	0.000000	2.000000	0.233705
25%	0.700000	3.940000	970.000000	1440.000000	0.000000	0.000000	0.000000	9.000000	0.650307
50%	0.750000	15.260000	1039.000000	3960.000000	0.000000	0.000000	0.000000	34.000000	0.773333
75%	0.800000	24.260000	1083.000000	6960.000000	0.000000	50.000000	0.000000	57.000000	0.850253

max	0.800000	54.560000	23122.000000	25920.000000	45.000000	3600.000000	300.000000	89.000000	1.120437

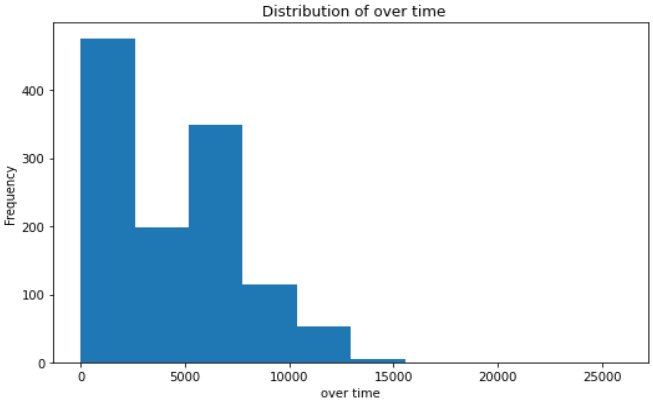
Histograms for Numerical Columns



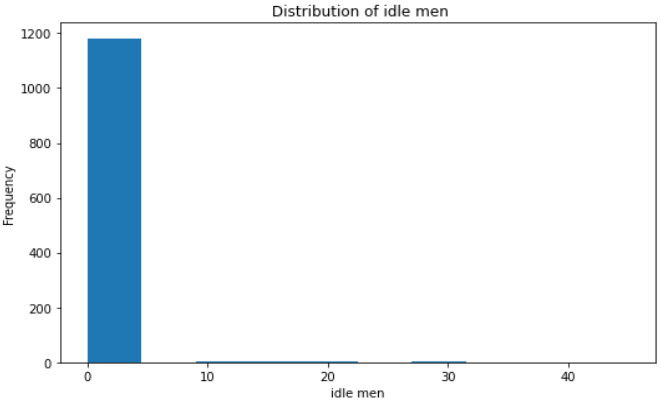
Most 'targeted_productivity' values fell in the 0.73 to 0.8 bin



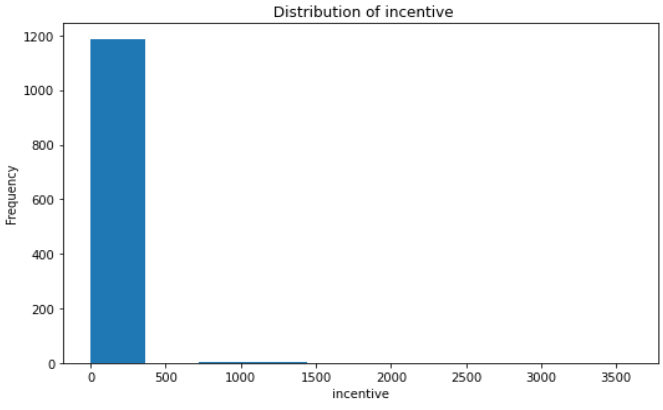
Most 'std_minute_value' values fell in the 2.9 to 8.07 bin



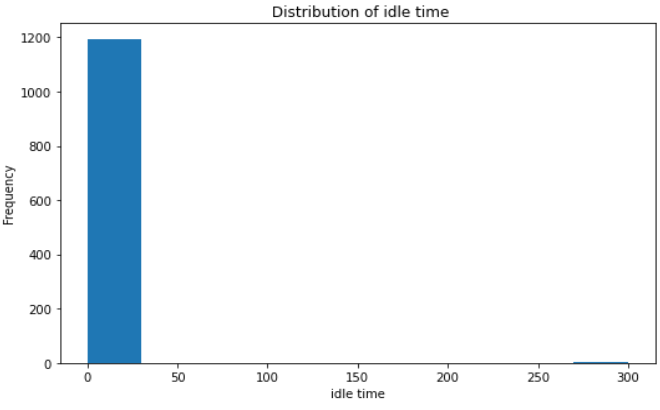
Most 'over_time' values fell in the 0.0 to 2592.0 bin



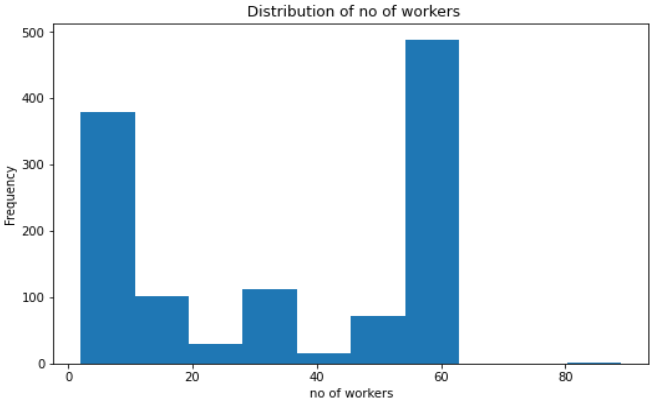
Most 'idle_men' values fell in the 0 to 4 bin



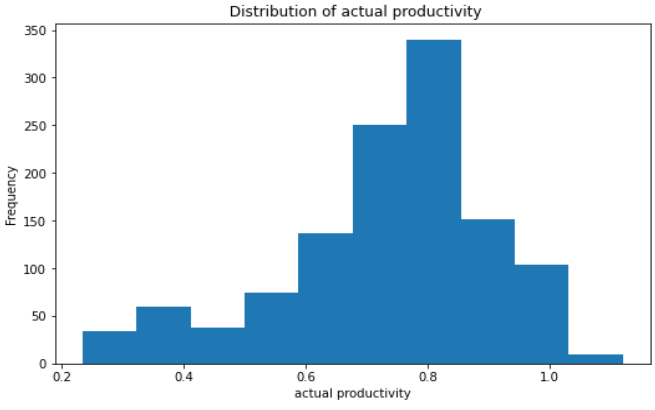
Most 'incentive' values fell in the 0.0 to 360.0 bin



Most 'idle_time' values fell in the 0.0 to 30.0 bin



Most 'no_of_workers' values fell in the 54 to 62 bin

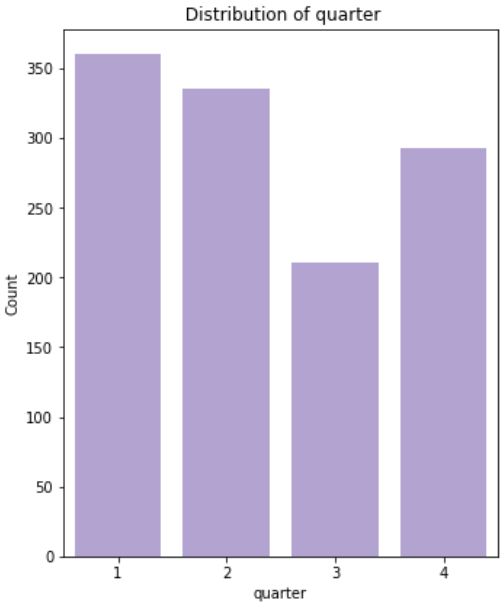


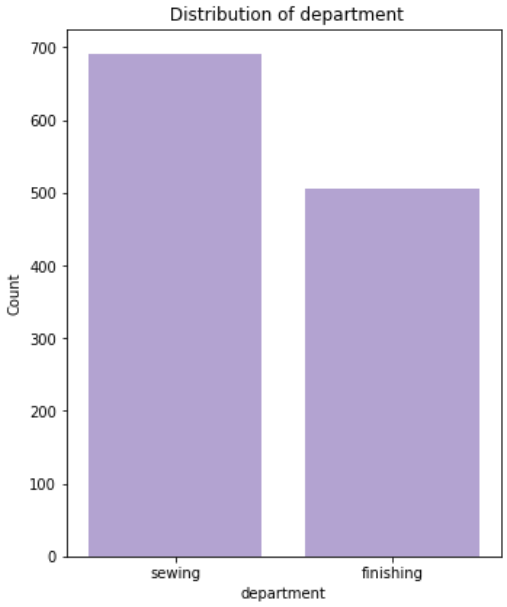
Most 'actual_productivity' values fell in the 0.77 to 0.85 bin

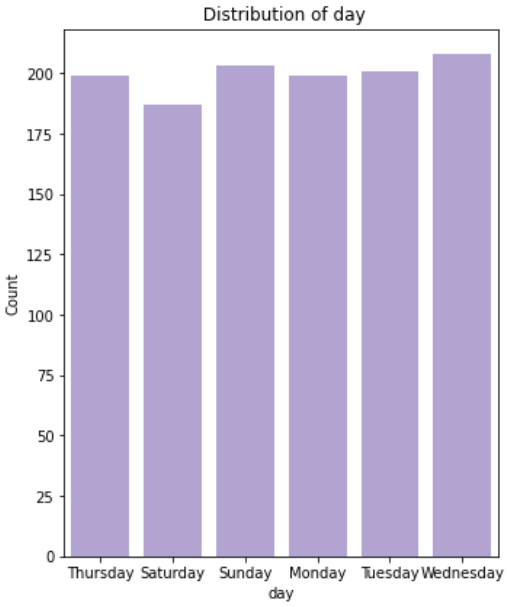
Observations:

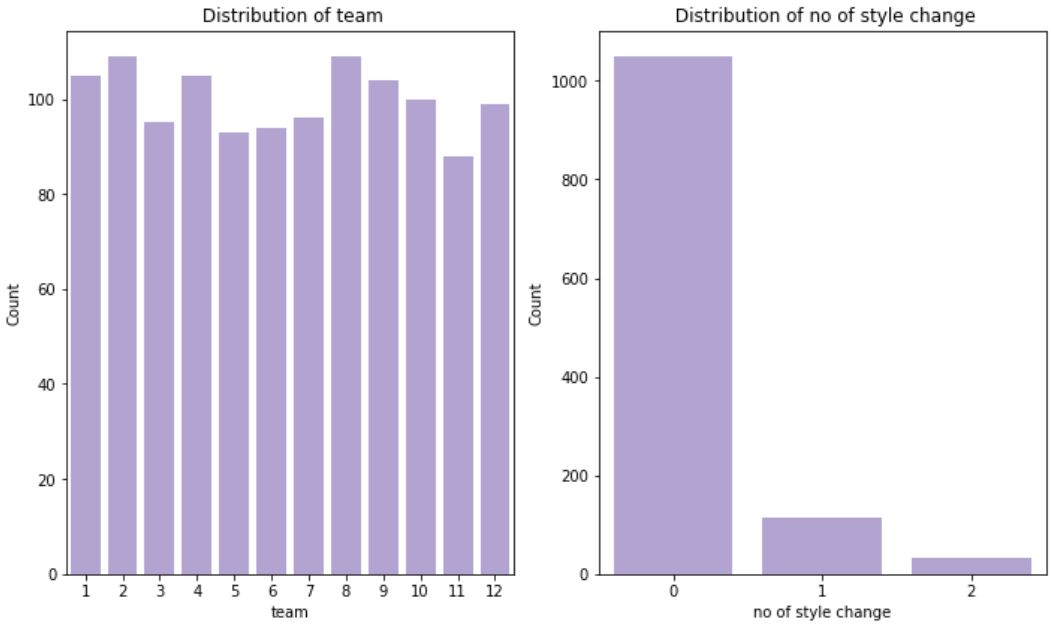
- Most **idle time** values fell in the 0 to 30 bin.
- Most workers fall in the 54 to 62 bin.
- Most **idle men** fell in the 0 to 4 bin.
- Most **incentive** values fell in the 0 to 360 bin
- Most **over time** values fell in the 0 to 2592 bin
- Most **targeted productivity** values fell in the 0.73 to 0.8 bin
- Most **standard minute value** values fell in the 2.9 bin to the 8.07 bin

Count Plots for Categorical Columns







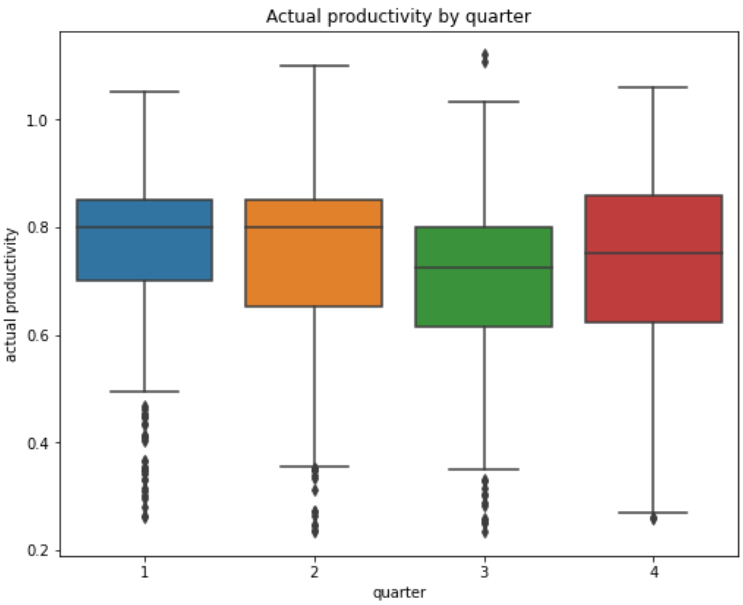


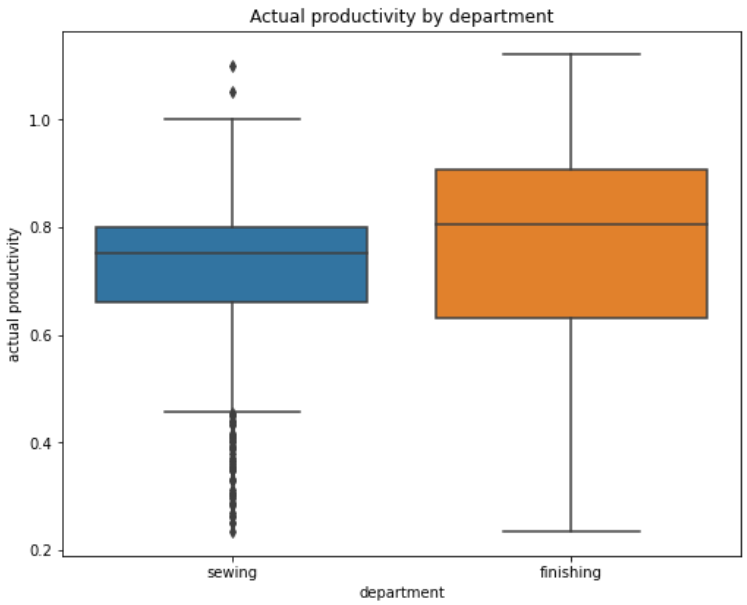
Observations:

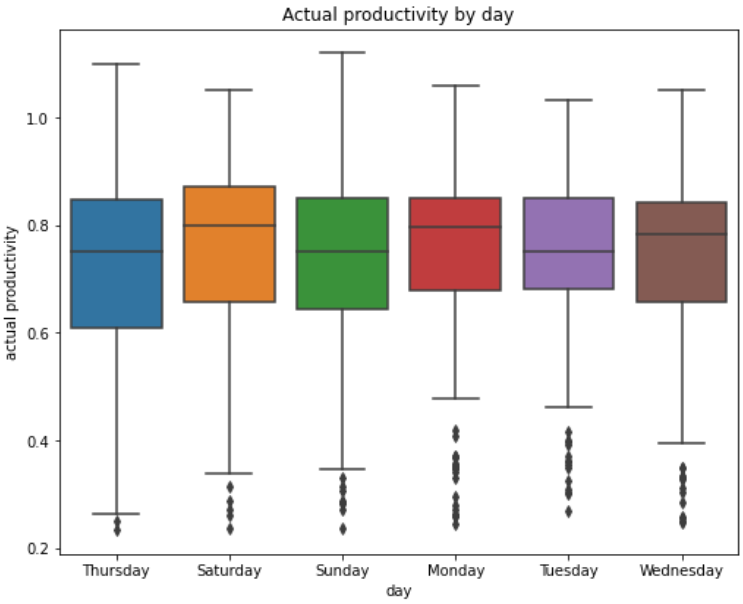
- Quarter one had the most records in the dataset.
- There were more records related to the sewing department than the finishing department.
- The day of the week with the most records is Wednesday.
- Teams 2 and 8 appeared more frequently than other teams in the dataset.
- For most records, the number of changes in the style of a particular product was 0.

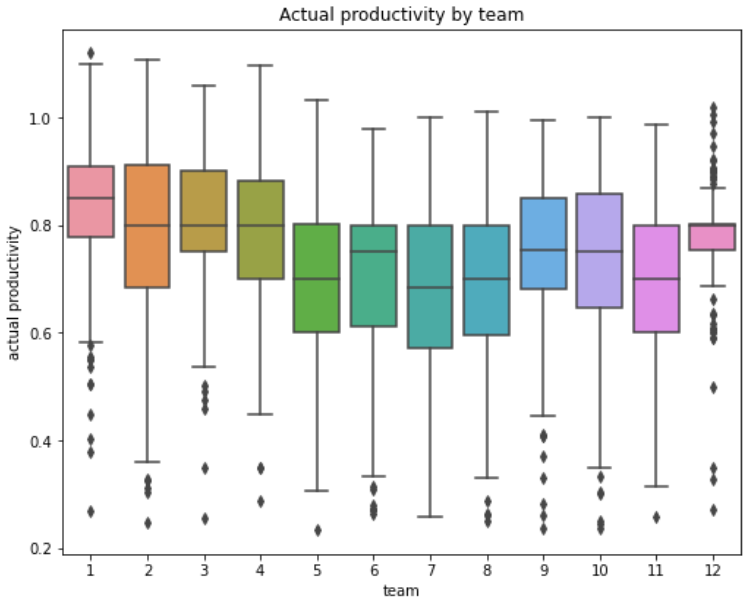
Bivariate Analysis

Actual productivity by day of week, department, quarter of the month, and team.





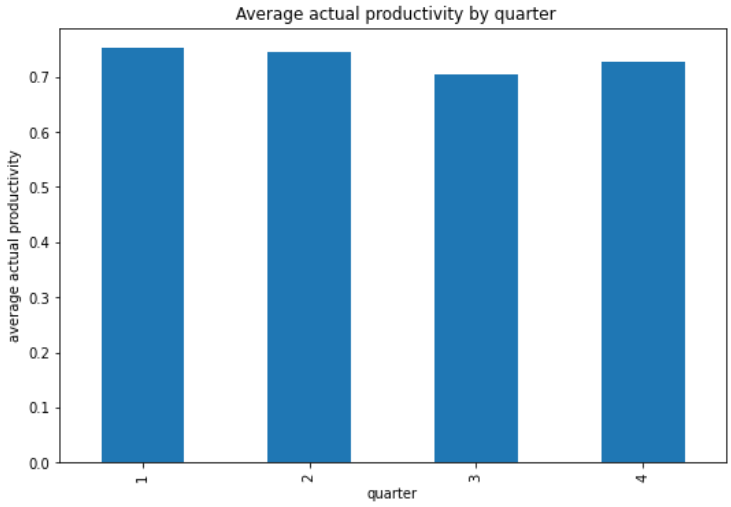


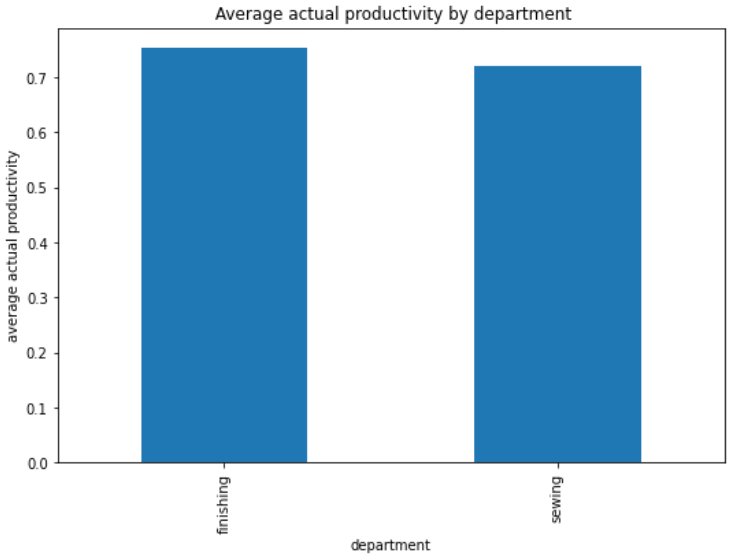


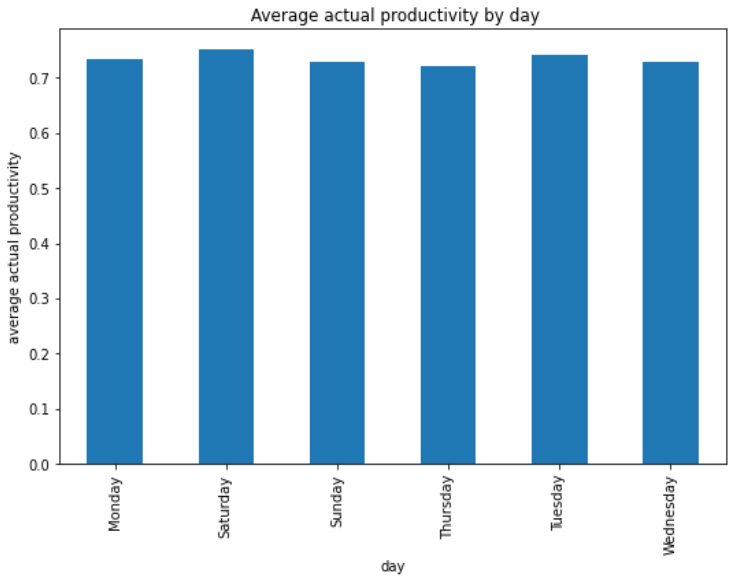
The median actual productivity is highest:

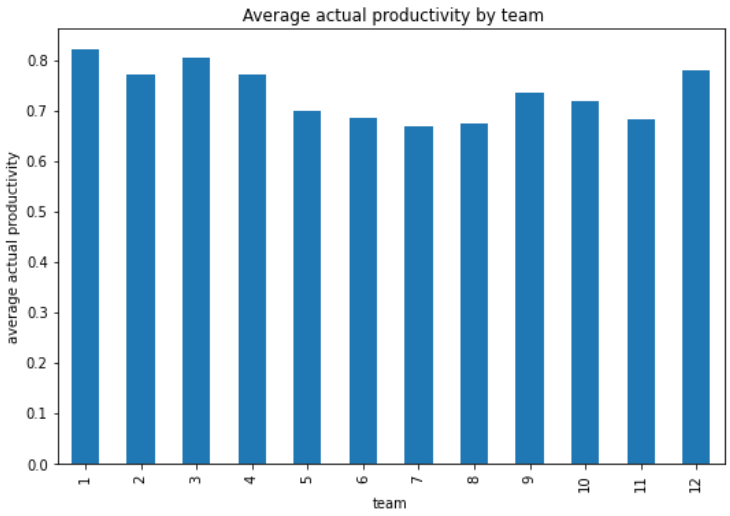
- Quarters: during the first 2 quarters of the month.
- Department: in the "finishing" department
- Day: on Saturdays
- Team: in team 1

Average actual productivity by day of week, department, quarter of the month, and team.





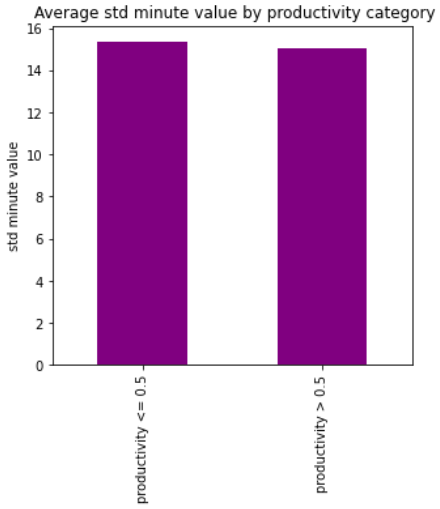


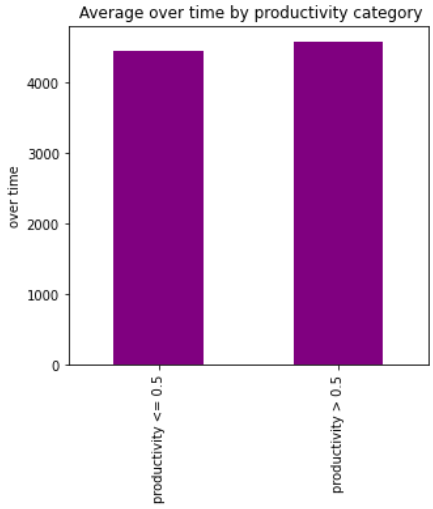


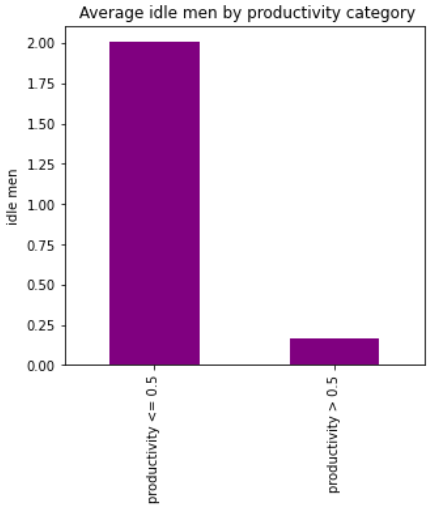
The mean actual productivity is highest:

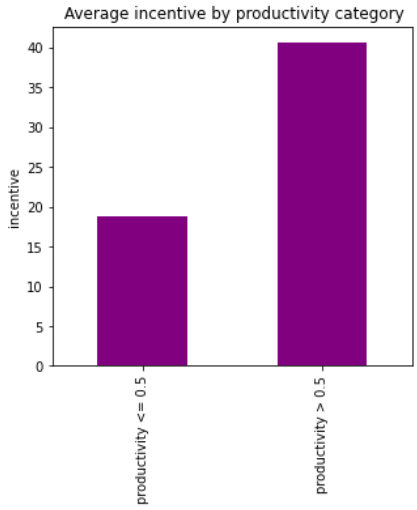
- Quarters: during the first 2 quarters of the month.
- Department: in the "finishing" department
- Day: on Saturdays
- Team: in team 1

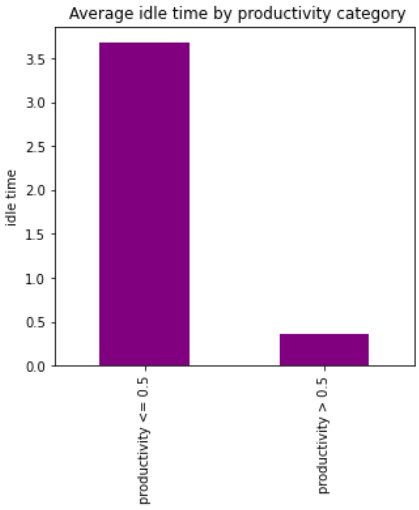
Average standard minute value, over time, idle men,incentive, idle time and no of workers by productivity category (>0.5 or not)

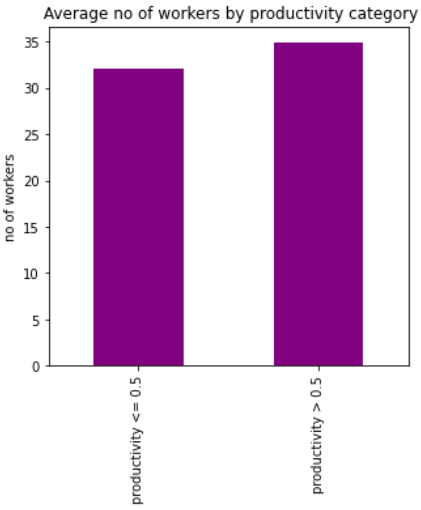








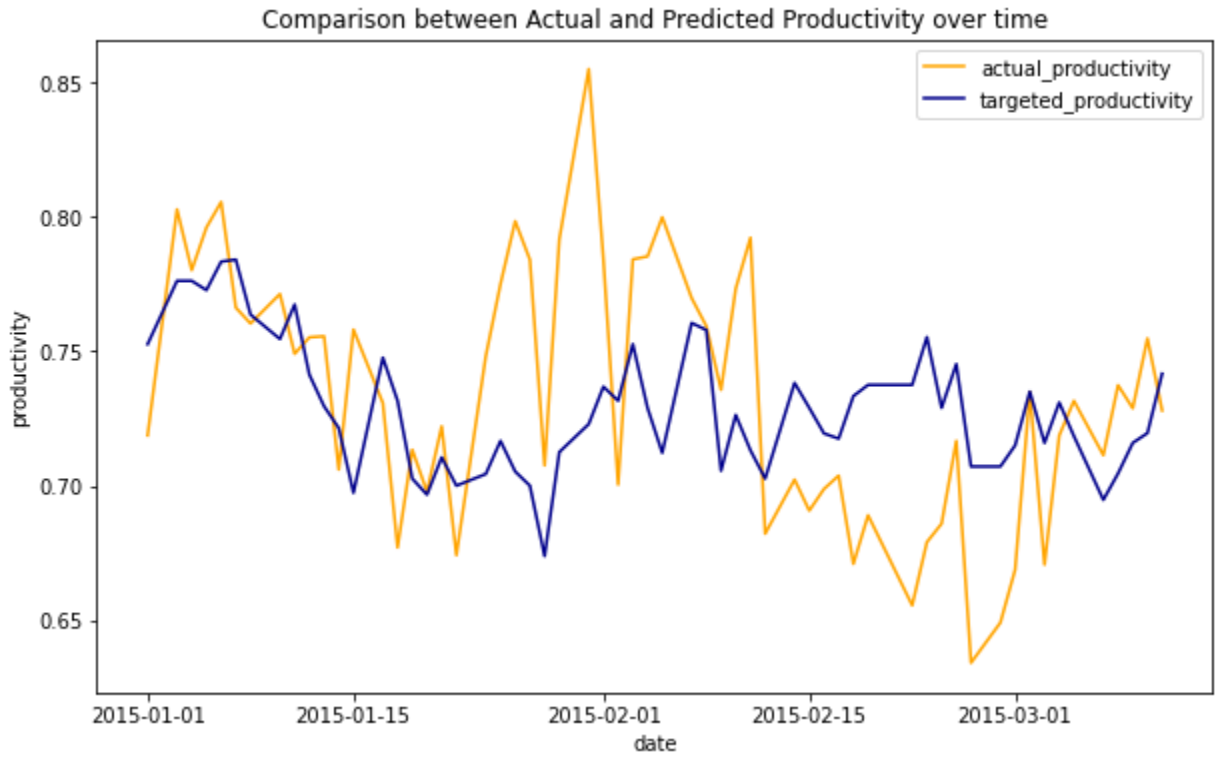




Values higher in the category where actual productivity levels fall between are <= 0.5 include average standard minute value, average idle time and average idle men.

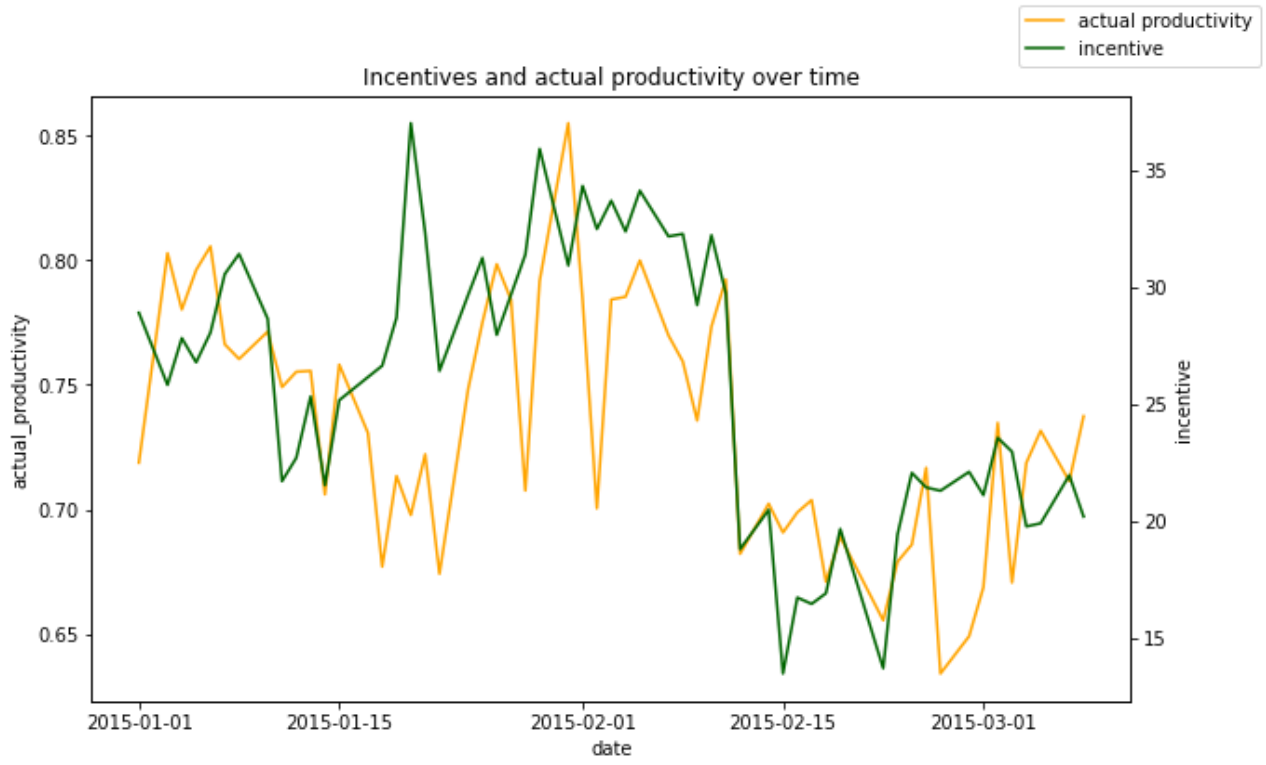
Values higher in the category where actual productivity levels are > 0.5 include average incentive, average number of workers, and average overtime.

Comparison between Actual and Predicted Productivity over time

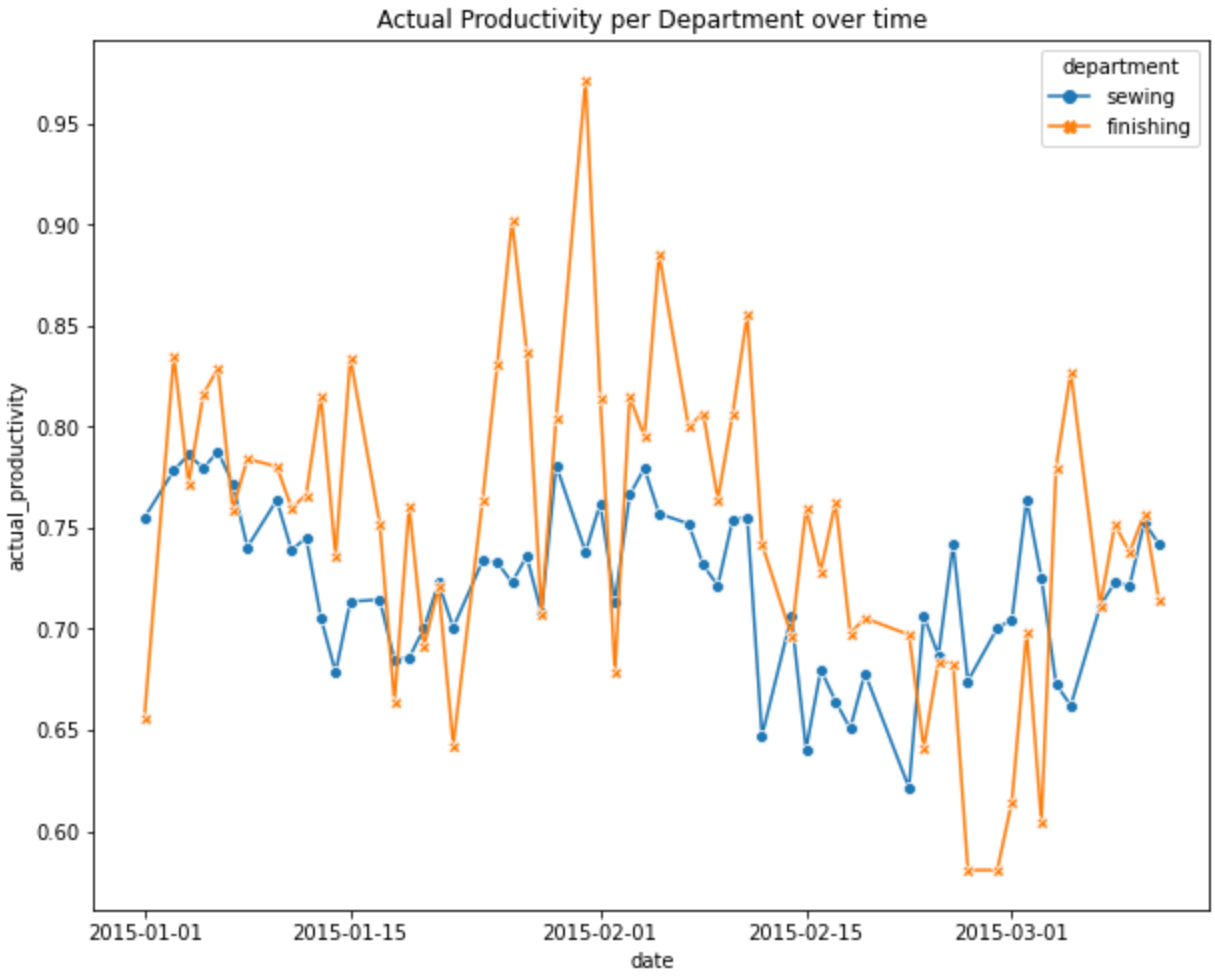


Actual productivity exceeded generally targeted productivity except mid February to beginning of March where targets were not being met.

Comparison between Actual Productivity and incentives over time

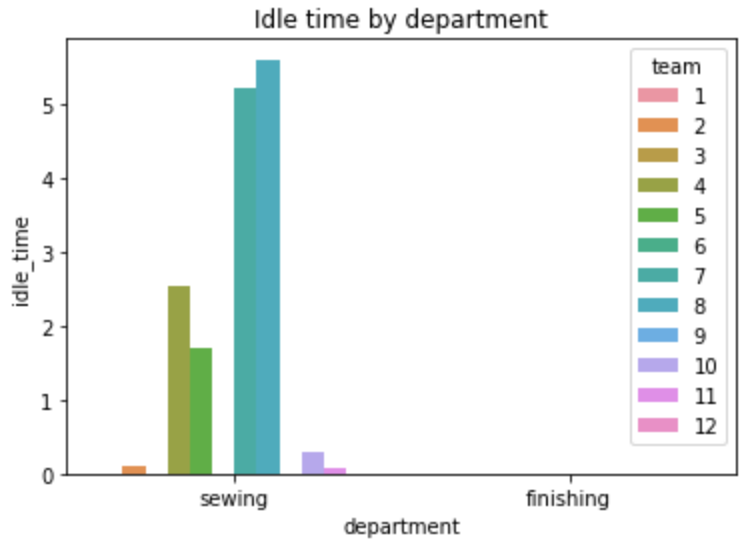


Actual Productivity per Department over time



The finishing department generally had higher actual productivity than the sewing department.

Idle time by department

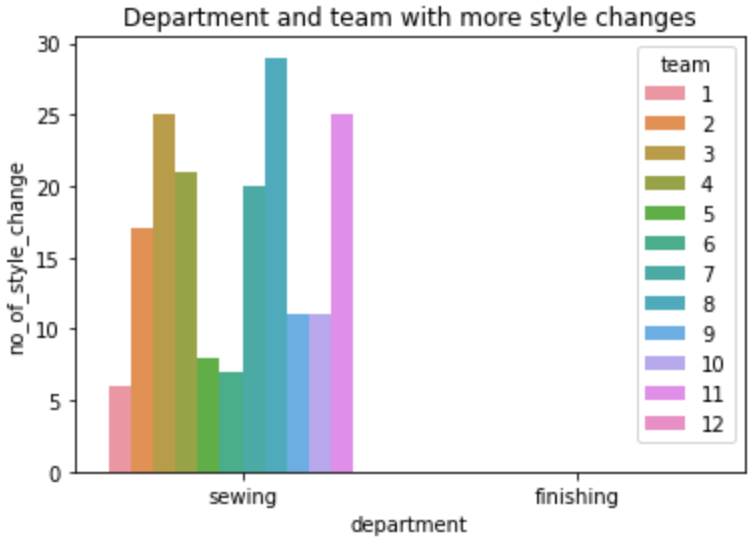


Only the sewing department had idle time, with team 8's sewing division recording the highest idle time.

Idle men by department

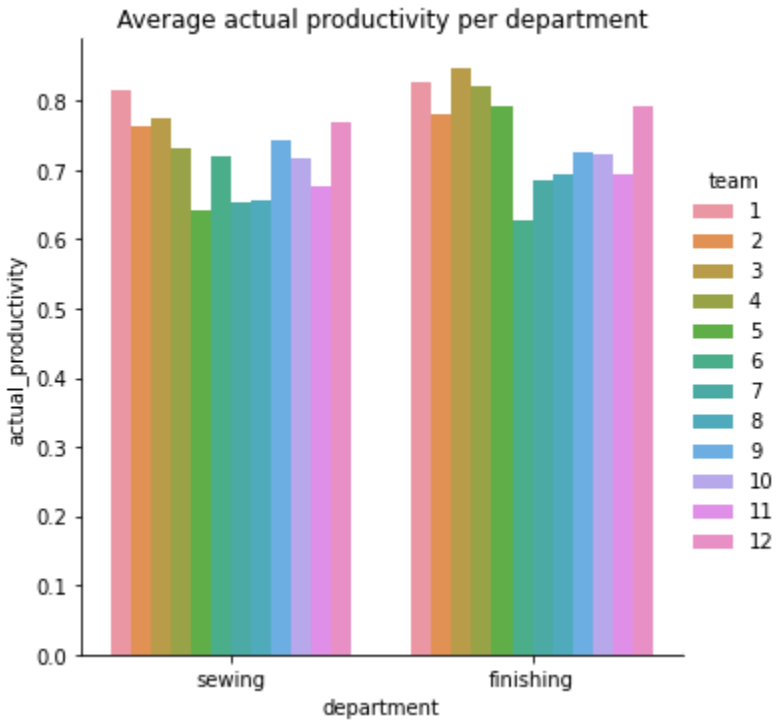
Only the sewing department had idle men, with team 7's sewing division recording the highest number of idle men.

Department and team with more style changes



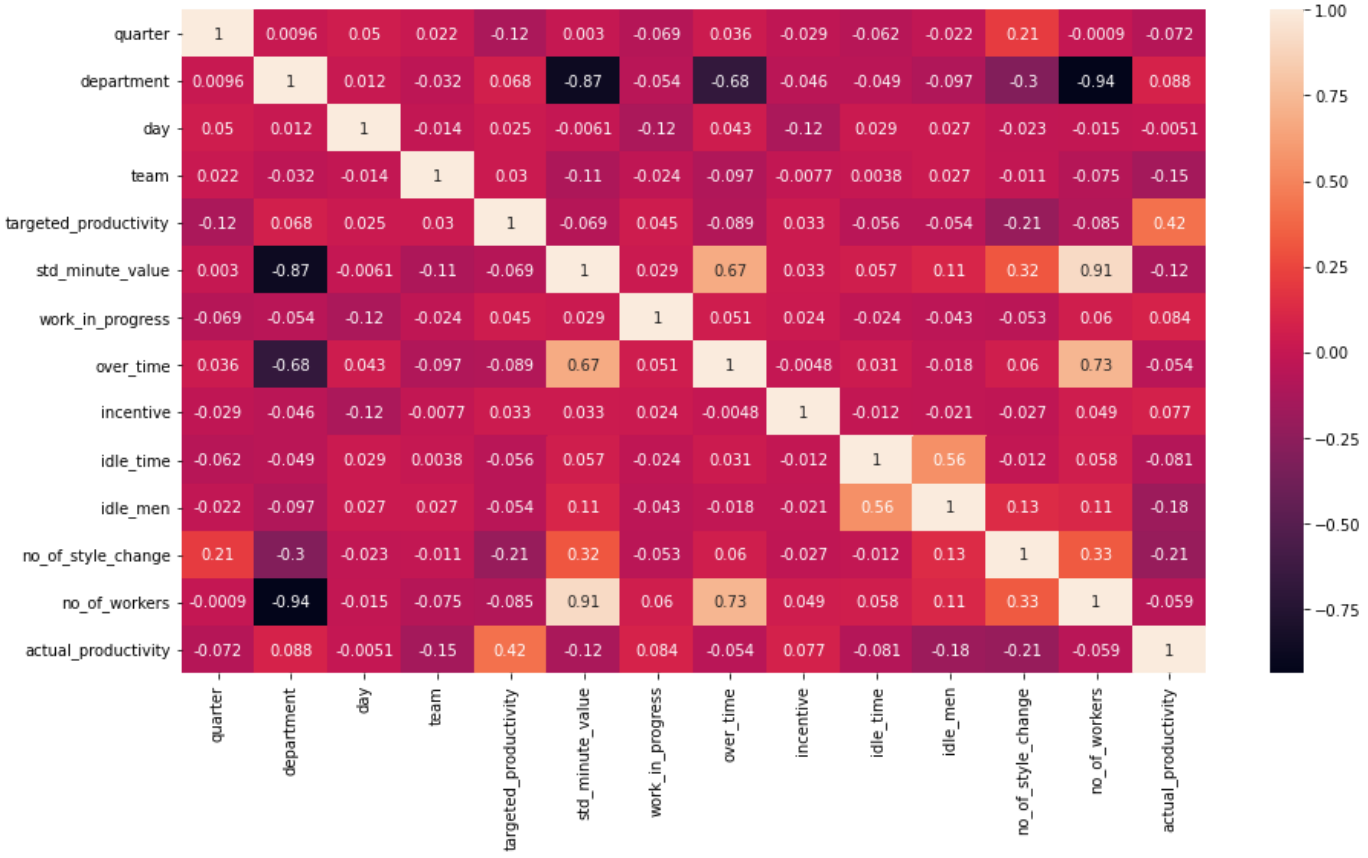
Only the sewing department had style changes, with team 8's sewing division recording the highest number of style changes.

Average actual productivity per department



Team 1's sewing division had the highest average actual productivity in the sewing department, while in the finishing department team 3's finishing division had the highest.
Team 5's sewing division had the lowest average actual productivity in the sewing department, while in the finishing department team 6's finishing division had the lowest.

Correlation Heatmap



-Actual productivity, the target column, has the highest correlation (moderate positive correlation of 0.42) with targeted productivity.

-Strong positive correlations:

- standard minute value has strong positive correlations with number of workers and overtime

- Over time has a strong positive correlation with number of workers, standard minute value

- Idle time and idle men have a strong positive correlation

5. Modeling

Linear regression

- The baseline linear regression model was built. It had a root mean squared error of 0.143616
- Its assumptions were tested. The data showed heteroscedasticity, the residuals were not normally distributed and multicollinearity was present in the data.

Lasso regression

- A grid search was performed to identify the best alpha hyper parameter value, and the root mean squared error of the model built was 0.143396

Ridge regression

- A grid search was performed to identify the best alpha hyper parameter value, and the root mean squared error of the model built was 0.143575

Elastic net regression

- A grid search was performed to identify the best alpha and l1 ratio hyper parameter values, and the root mean squared error of the model built was 0.143528

KNN

- The first knn model was built with an arbitrary value of 5 neighbours. The root mean squared error was 0.144635
- After hyperparameter tuning, 7 neighbours were used and the root mean squared error was 0.136820

Random forest

- The first model was built with arbitrary hyperparameter values and the root mean squared error was 0.125023
- After tuning hyperparameters, the root mean squared error was 0.123642

Gradient boosting

- The first model was built with arbitrary hyperparameter values and the root mean squared error was 0.135371
- After tuning hyperparameters, the root mean squared error was 0.123004
- The third model is explained below after neural networks. This model was the best, with an rmse of 0.119481.

Neural networks

- The first model was built with arbitrary hyperparameter values and the root mean squared error was 0.143735
- After tuning hyperparameters, the root mean squared error was 0.135856

As the gradient boosting regressor with mean squared error of 0.123004 was the best out of the different model types tested to that point, feature importance was determined using this model.

The feature ranking was as follows:

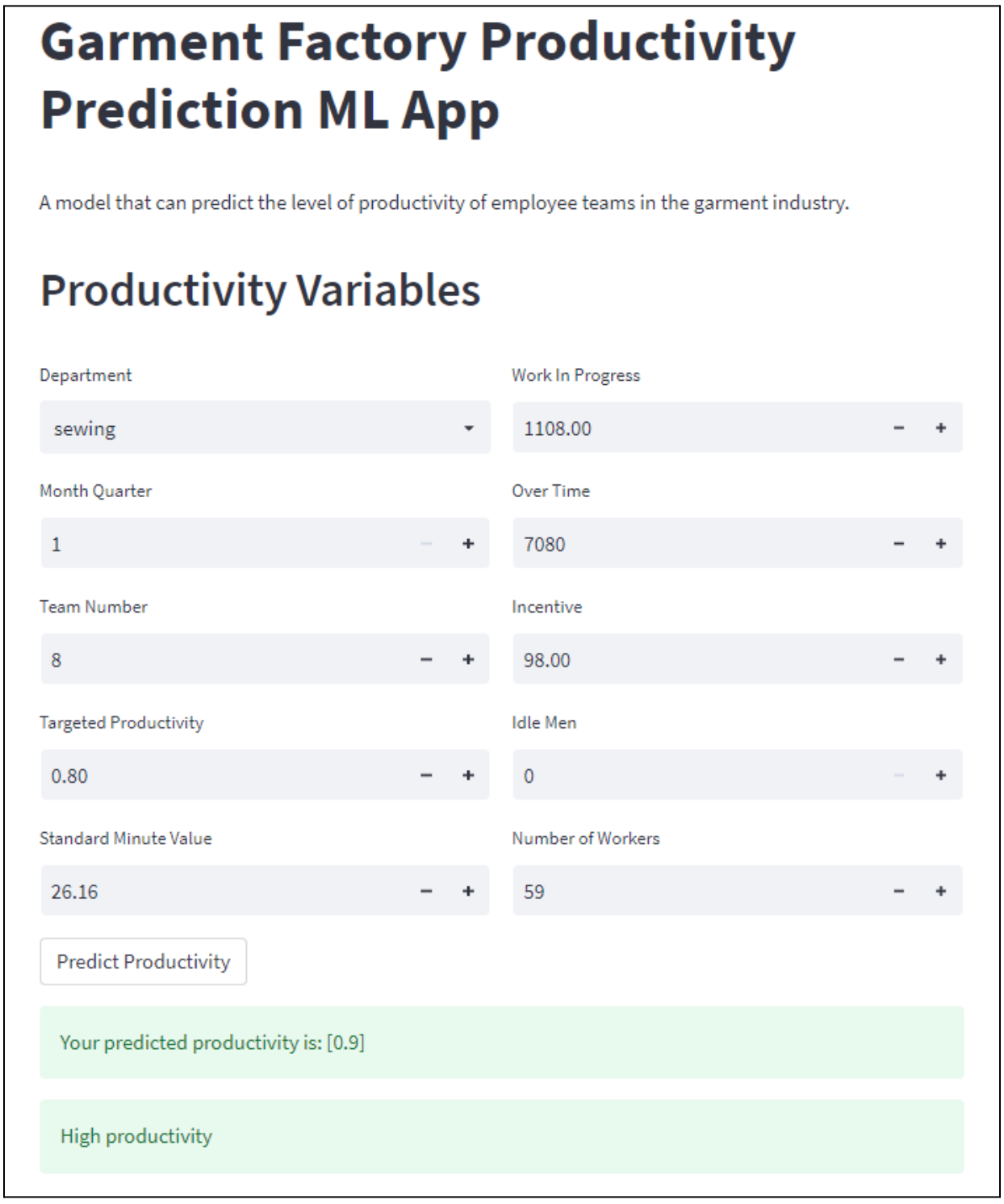
- Variable: targeted_productivity, Importance: 0.28
- Variable: std_minute_value, Importance: 0.2
- Variable: incentive, Importance: 0.17
- Variable: no_of_workers, Importance: 0.15
- Variable: team, Importance: 0.06
- Variable: work_in_progress, Importance: 0.05
- Variable: over_time, Importance: 0.04
- Variable: idle_men, Importance: 0.03
- Variable: quarter, Importance: 0.02
- Variable: department, Importance: 0.01
- Variable: day, Importance: 0.0
- Variable: idle_time, Importance: 0.0
- Variable: no_of_style_change, Importance: 0.0

The features with importance of 0.0 (day, idle time, number of style changes) were dropped and another gradient boosting regressor built with select features. Hperparameters were tuned, and the resulting model had an rmse of 0.119481, which is better than the previous benchmark (tuned gradient boosting regressor with all features - 0.123004). This model is therefore the best out of all tested models.

6. Deployment

The best model (Gradient Boosting) was successfully deployed via streamlit and we are now able to predict productivity given the various selected predictors.

This is the interface of the deployed model.



7. Conclusion

The main objectives of the project were achieved:

1. A model that predicts the level of productivity of employees in the garment industry was built.
   - The best model - a gradient boosting regressor (tuned learning rate : 0.1, max_depth 3, and n_estimators: 300) trained with these features: targeted productivity, standard minute value, incentive, number of workers, team, work in progress, over time, idle men,quarter, and department). The rmse of this model was 0.11948.

2. The top factors influencing the productivity level of employees were identified. The top 3 include targeted productivity, standard minute value, and incentive.

3. The relationships between level of productivity and the predictor variables were investigated and visualized through exploratory analysis.

8. Recommendations

1. The predictive model to be used in determining the productivity level of garment industry employees should be the gradient boosting regressor trained on select features, as it performed best out of all models tested.

2. The management of each team in the industry should set clear, high, achievable goals each day. Targeted productivity was identified as a key feature during modelling, and it had a moderate positive correlation with actual productivity.

3. Incentives motivate teams to work harder. The average incentives were higher in the category where productivity was greater than 0.5. Offering favourable incentives therefore aids in boosting productivity.

4. Team 5, 7, and 8's sewing divisions performed lowest in the sewing department. In the finishing department, lowest performing teams were 6, 7, and 8. These teams have the lowest average incentives when compared to other teams in their respective departments, with teams 6 and 7's finishing divisions having as low as 0 average incentives. The managements of these teams should consider offering more favourable incentives to motivate the teams. Additionally, in sewing, teams 7 and 8 recorded the

highest idle time and men, with team 5 having the fourth highest. Idle time may be due to machinery breakdown, therefore this should be investigated by management.