**Garment Employee Productivity Prediction**

**Dataset:**

Link: <https://archive.ics.uci.edu/dataset/597/productivity+prediction+of+garment+employees>

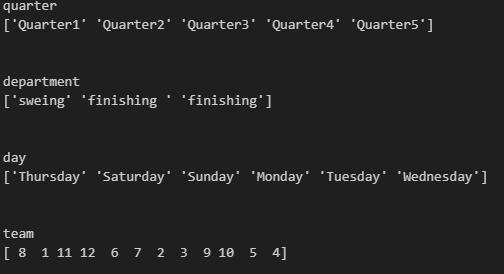
|  |  |  |
| --- | --- | --- |
| **Column Name** | **Default Data type** | **Description** |
| **date** | Object | Date in MM-DD-YYYY |
| **day** | Object | Day of the Week |
| **quarter** | Object | A portion of the month. A month was divided into four quarters |
| **department** | Object | Associated department with the instance |
| **team\_no** | Integer | Associated team number with the instance |
| **no\_of\_workers** | Float | Number of workers in each team |
| **no\_of\_style\_change** | Integer | Number of changes in the style of a particular product |
| **targeted\_productivity** | Float | Targeted productivity set by the Authority for each team for each day |
| **smv** | Float | Standard Minute Value, it is the allocated time for a task |
| **wip** | Float | Work in progress. Includes the number of unfinished items for products |
| **over\_time** | Interger | Represents the amount of overtime by each team in minutes |
| **incentive** | Float | Represents the amount of financial incentive (in Bangladeshi Taka) that enables or motivates a particular course of action |
| **idle\_time** | Float | The amount of time when the production was interrupted due to several reasons |
| **idle\_men** | Integer | The number of workers who were idle due to production interruption |
| **actual\_productivity** | Float | The actual % of productivity that was delivered by the workers |

Insights:

* team\_no should be changed into categorical data type
* no\_of\_workers should be changed into integer type
* wip (work in progress) should be changed into integer

**EDA**

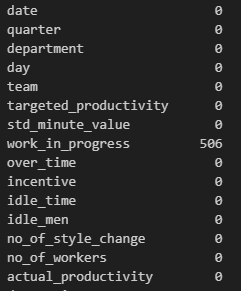
Unique values in the categorical columns



Insights:

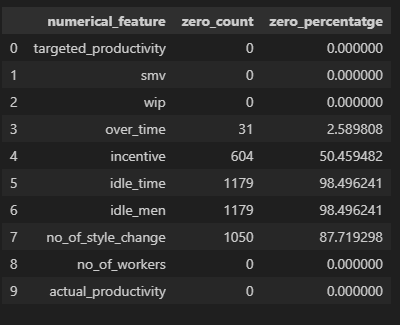
* A month was divided into 4 quarters. However, there is a 'Quarter5' value observed. (This is obvious since some months contains 5 weeks).
* Also, under department spelling of sewing is wrong and finishing appears as a unique value twice due to an extra whitespace.
* Teams based on the department may increase the readability of the team in each of the departments.
* There is no Friday in the day column.

**Addressing missing values**

Insight:

* Null values are present due to the type of the department. All the finishing department WIP values set as NaN since there is no work in progress after finishing the product. Instead of NaN we can replace '0' indicating that there is no work in progress product. Other than that there are no null values.

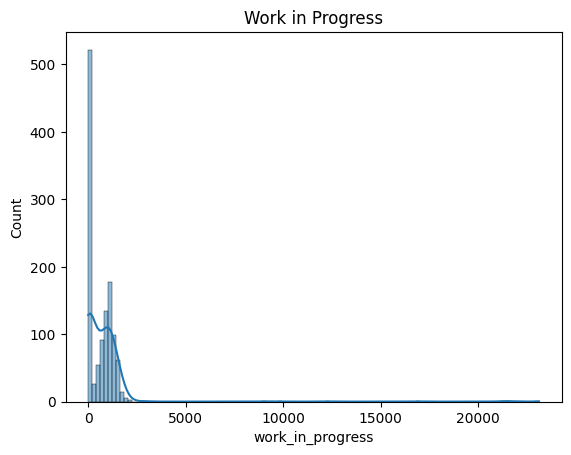
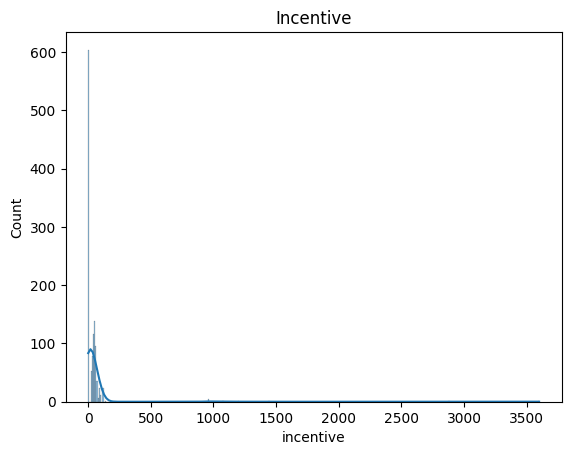
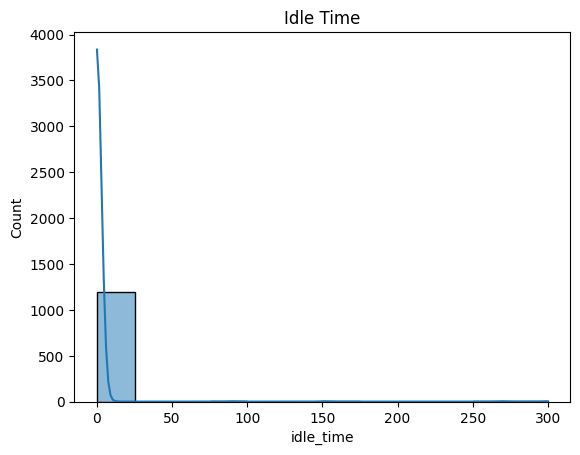
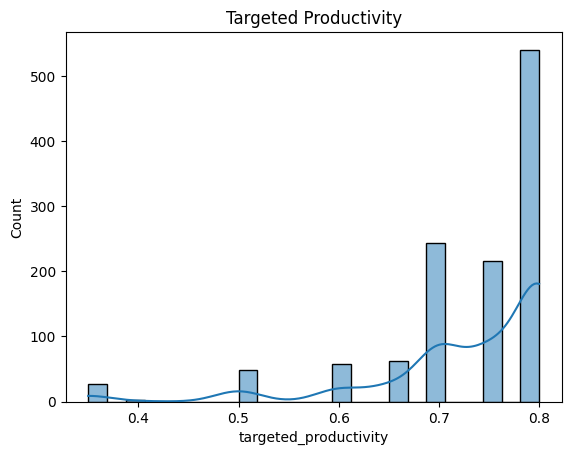
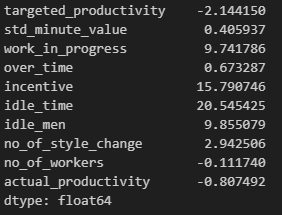
**Zero Value Counts**



Insights:

* In reality, it is very rare to have a production interruption and changing style of a garment due any kind of reason. Due to that reason we can expect the high percentage to 0 values in idle\_time, idle\_men and no\_of\_style\_change variables.

**Checking Skewness of the numerical variables**

********distributions 

Columns having highly skewed distributions (-1 > skew > 1):

* targeted\_productivity (negative)
* work\_in\_progress (positive)
* idle\_men (positve)
* incentive (positive)
* idle\_time (positive)

Columns having Moderately skewed distributions

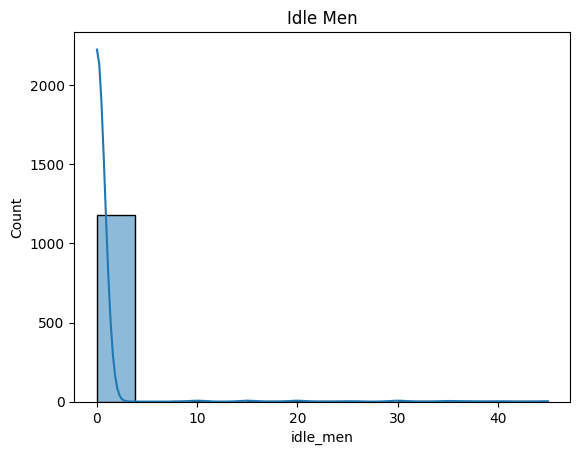
(-1 < skew < -1/2 or 1/2 < skew < 1):

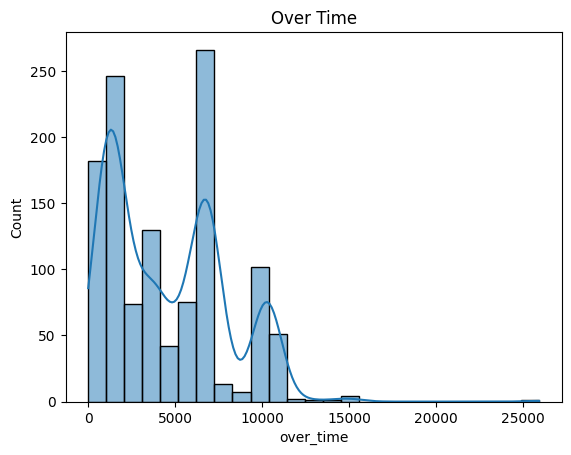
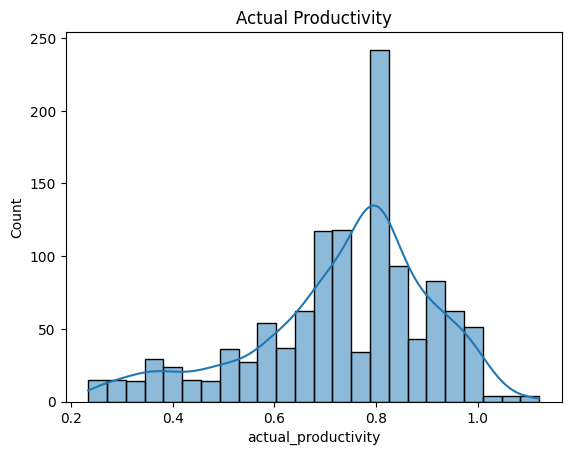
* over\_time (positive)
* actual\_productivity (negative)

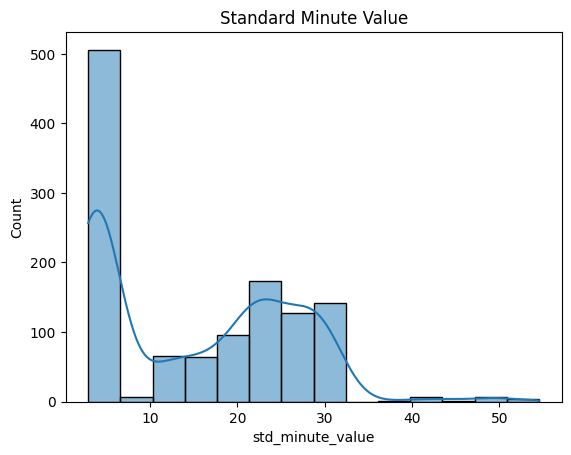
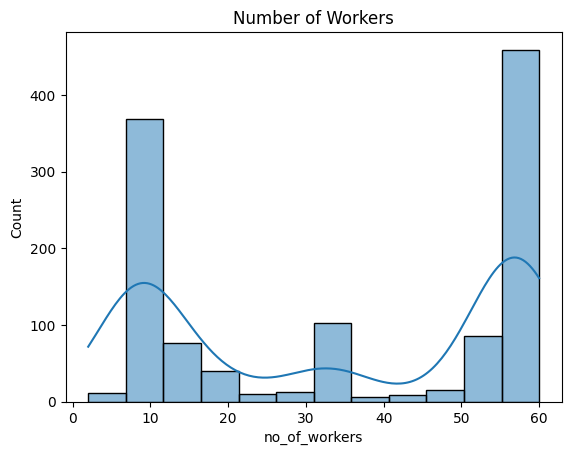
Columns which are approximately symmetric distributions

(-1/2 < skew < 1/2):

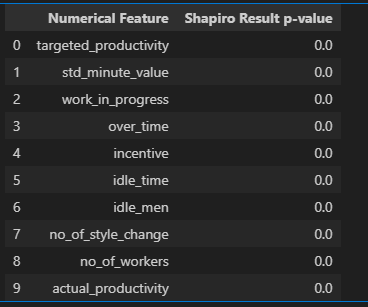
* std\_minute\_value
* no\_of\_workers

****

****

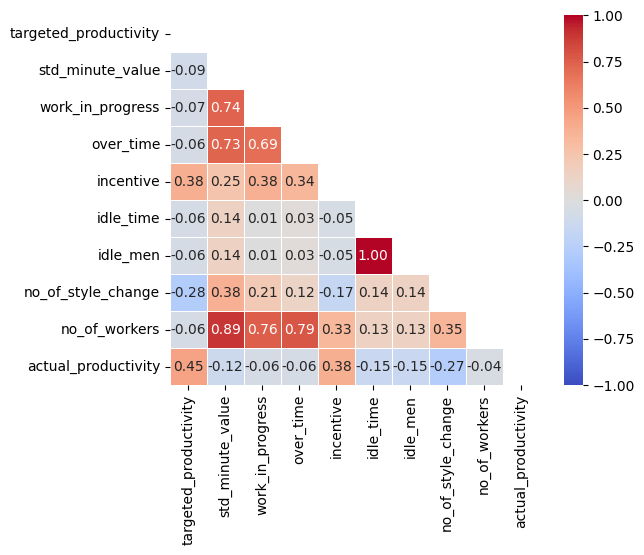
****

Checking the normality (Applying Shapiro Wilk test)

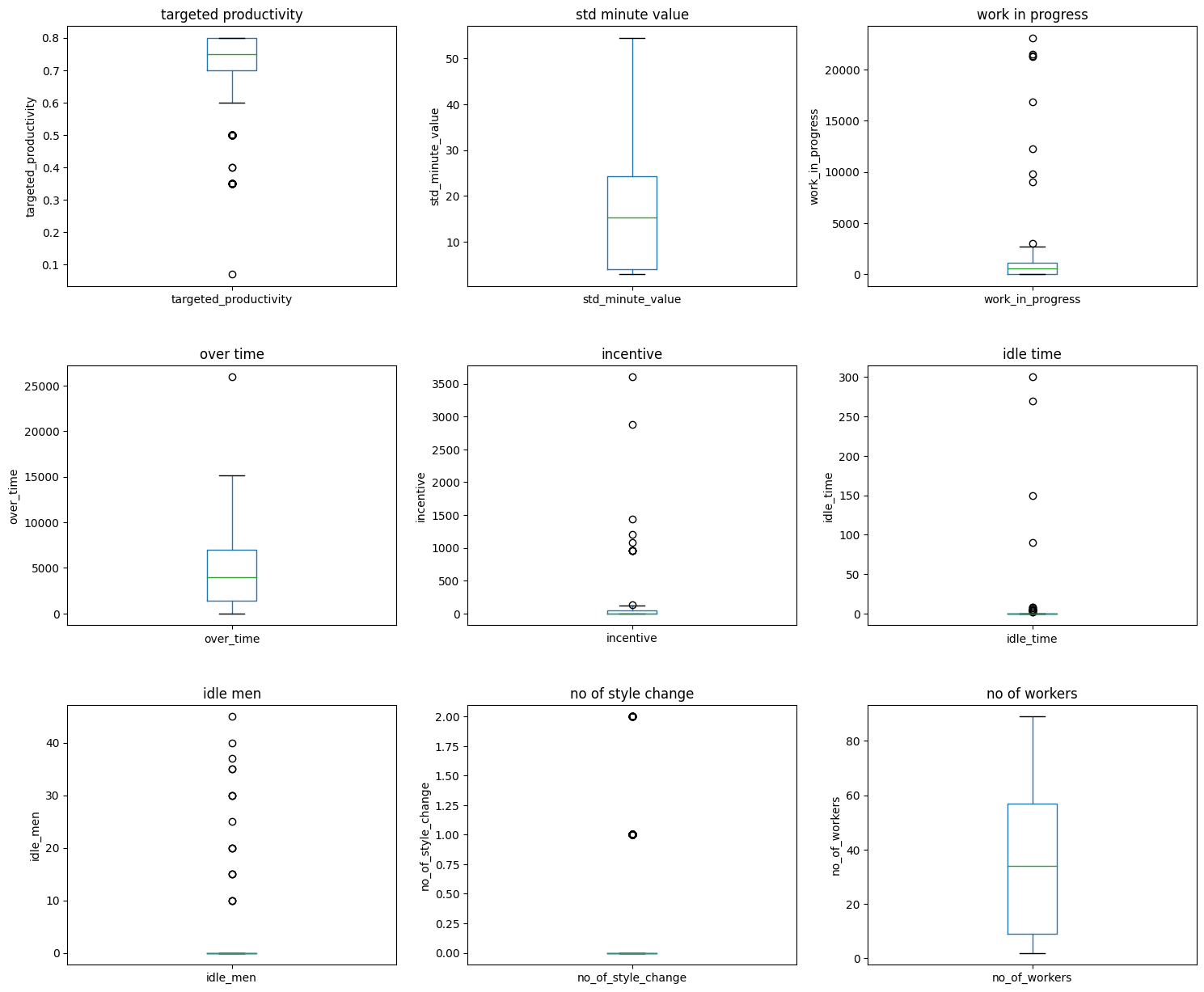


Insights:

* Since all the p-value are less than .05, we reject the null hypothesis of the Shapiro-Wilk test.
* This means we have sufficient evidence to say that the sample data does not come from a normal distribution

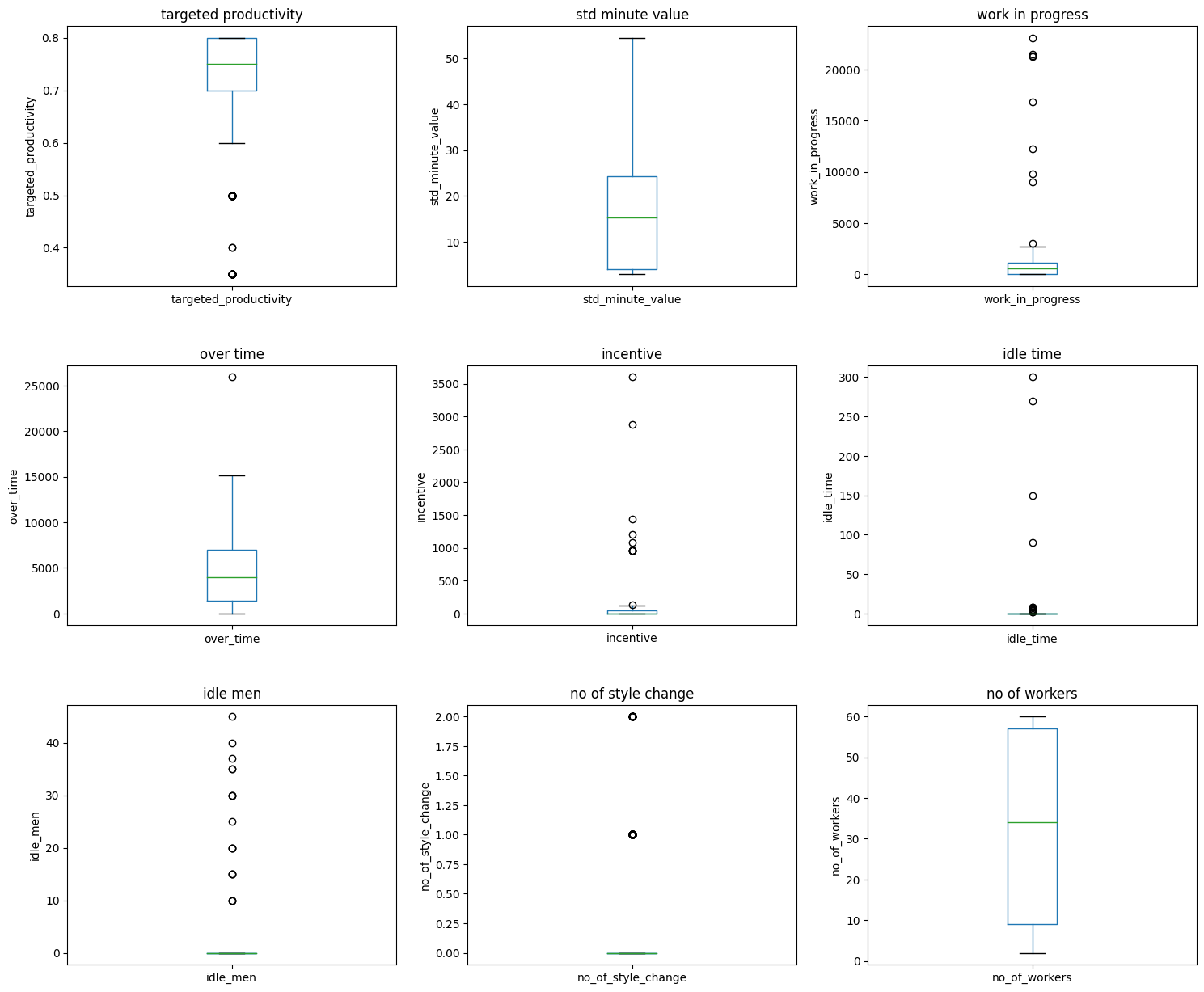


Spearman’s Correlation Plot



Insights:

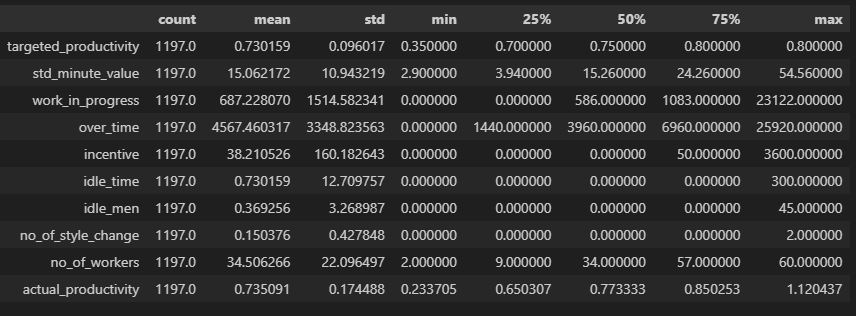
* Outlier of the targeted\_productivity might be a recoring error since most of the targeted productivities are around 0.7. So that it can be changed in to 0.7.
* Outlier in the overtime is practical ((25920 / 60) / 54 = 8) Each employee can work 8 hours overtime.
* Cannot determine anything about the actual\_productivity, idle\_time, incentive and idle\_men since they cannot be determine properly as an error or an actual value.
* Actual\_produtivity can go beyond 1 as well.

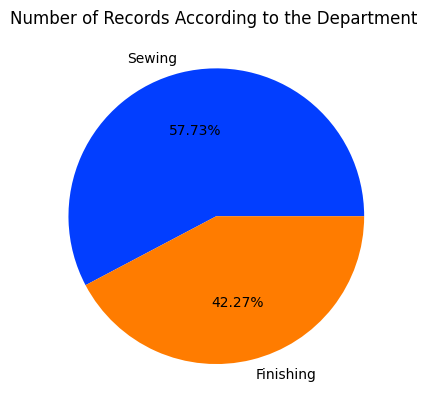
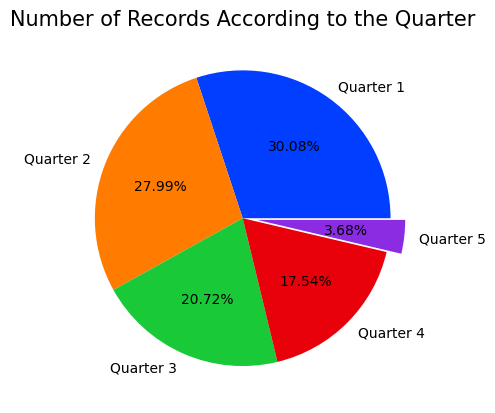


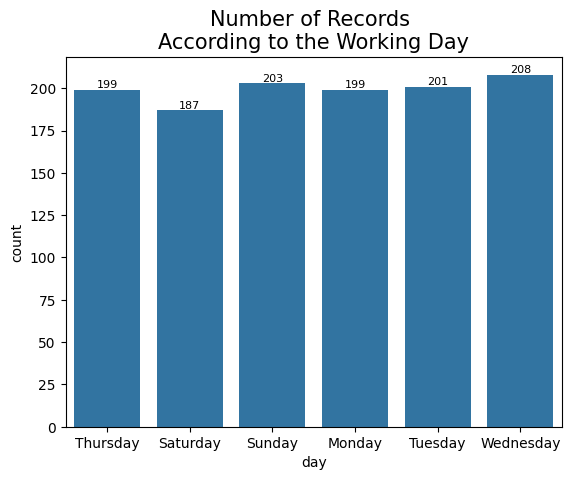
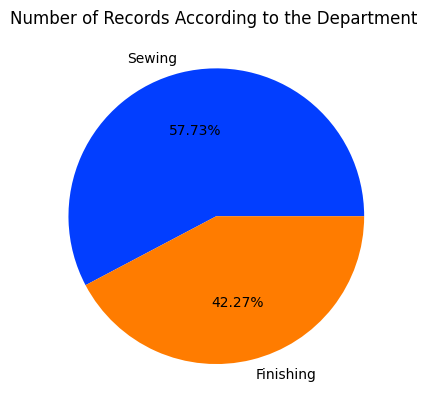
Insights:

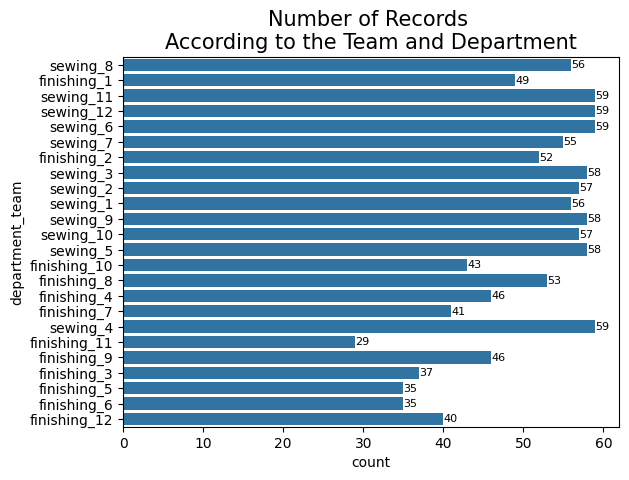
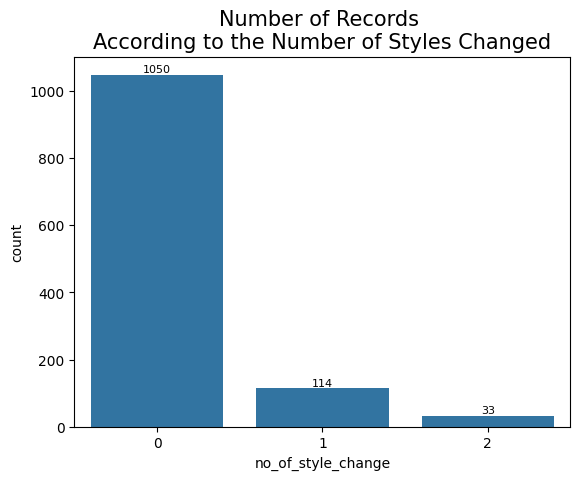
* There were outliers in targeted productivity, overtime, work in progress, incentive, idle time, idle men, and actual productivity columns. They will not be dropped as they are likely due to natural variability in the workflow of different teams, where some teams perform significantly above or below 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.

**Descriptive Statistics for the numerical Variables**

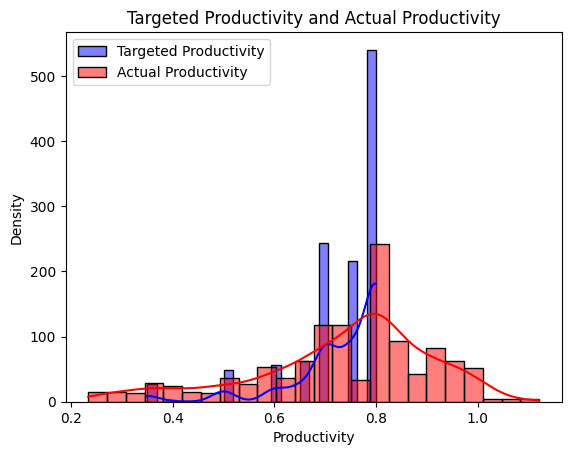


**Categorical Variables**

****

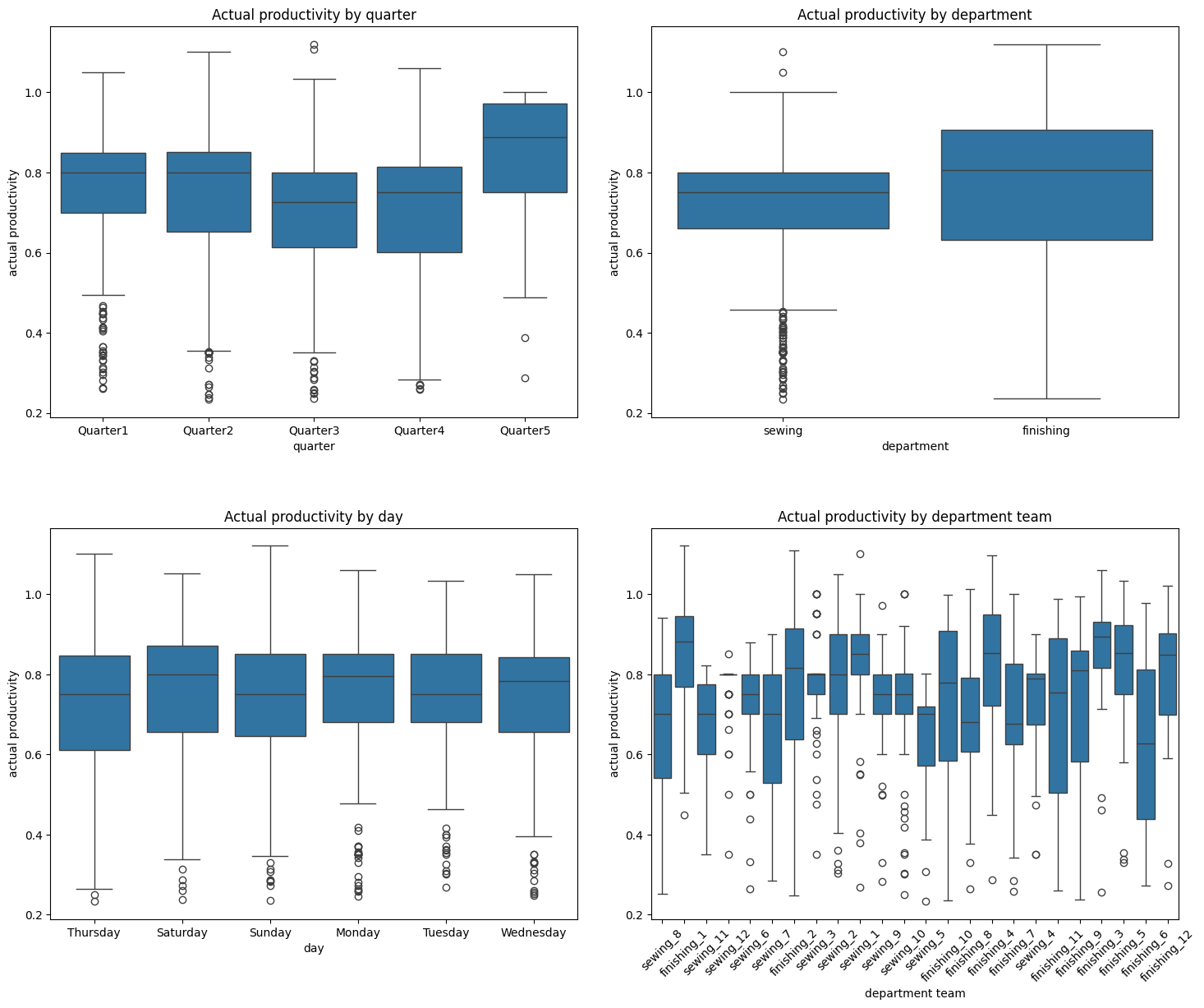


**Bivariate Analysis**

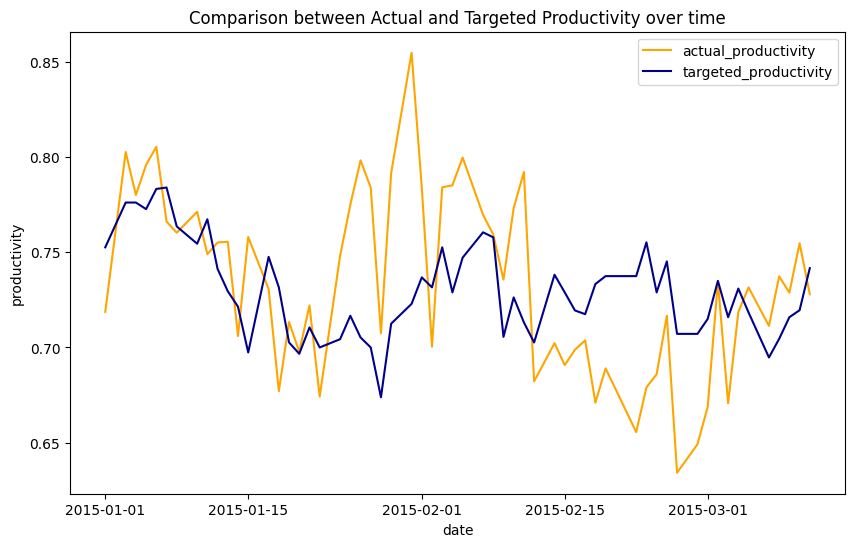
****

Insights**:**

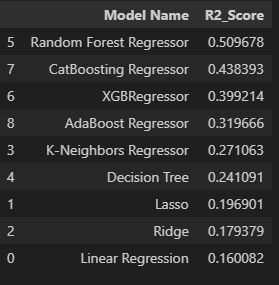
* Most of the targeted productivities lie around 0.6 to 0.8 region.
* Most of the actual productivities lie around 0.7 to 0.9 region.
* Actual Productivity by day of week, department, Quarter of the month, team



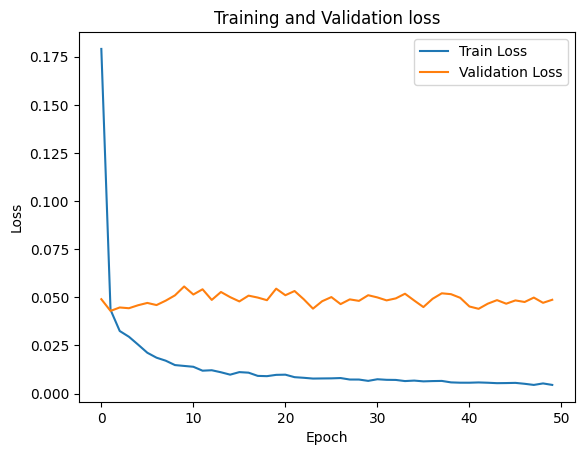
Actual Productivity by day of week, department, Quarter of the month, team



**Actual Productivity Prediction Models**



Neural Network (In progress)



For the test set

