THREADING PRODUCTIVITY: DATA ANALYSIS FOR PREDICTING PRODUCTIVITY OF GARMENT EMPLOYEES

Group 11

Samudika Wanasinghe - 16075
Meedum Keerthisiri - 16220
Pasindu Madusanka - 16333

Abstract

The garment industry is a highly competitive and labor-intensive sector where workforce productivity plays a crucial role in operational efficiency and profitability. This study analyzes a dataset on garment employee productivity from Kaggle, focusing on understanding key factors that influence performance. A descriptive analysis is conducted using numerical and visual methods to explore the characteristics and distribution of various attributes. Additionally, the study examines the relationship between potential influencing factors and productivity levels to identify significant contributors. The findings provide valuable insights into workforce efficiency, helping businesses optimize resource allocation and improve productivity management strategies.

Content

- List of Figures
- List of tables
- Introduction
- Description of the Question
- Description of the Dataset
- Data Cleansing and Preprocessing
- Main Results of the Descriptive Analysis
- Suggestions for a quality advanced analysis
- References
- Appendix including python code and technical details

List of Figures

- Fig1 - Distribution of Actual Productivity
- Fig2 - Boxplot of Actual Productivity and Targeted Productivity with Quarter of the month
- Fig3 - Boxplot of Actual Productivity and Targeted Productivity with Days of the week
- Fig4 - Line Plot of Actual Productivity per Department
- Fig5 - Box Plot of Actual Productivity per Department
- Fig6 - Stacked Bar Chart of percentage of productive work done by each department
- Fig7 - Scatter Plot of Actual Productivity vs wip
- Fig8 - Scatter Plot of Actual Productivity vs smv
- Fig9 - Scatter Plot of Actual Productivity vs no.of workers
- Fig10 - Scatter Plot of Actual Productivity vs overtime
- Fig11 - Box Plot of Actual Productivity vs team
- Fig12 - Scatter Plot of Actual Productivity vs Incentive
- Fig13 - Scatter Plot of Actual Productivity vs Incentive (<300)
- Fig14 - Scatter Plot of Actual Productivity vs Idle time , Idle time and No of Style changes
- Fig15 - PCA loading plot and PCA loading score
- Fig16 - Loading Plot of Comp1 and Comp2
- Fig 17 - Correlation Heatmap

List of Tables

- Table 1 - Attribute Information

Introduction

The garment industry is a vital part of the global economy, involving the design, production, and distribution of clothing to meet consumer demand. With its fast-paced nature and evolving trends, maintaining efficiency is essential for sustained growth. This industry relies heavily on labor and technology, making workforce productivity a crucial factor in overall performance. Understanding the factors that influence productivity can help businesses improve operations, reduce inefficiencies, and maximize output. This study analyzes garment employee productivity using data-driven methods to identify key patterns and provide insights that can support better decision-making and workforce management strategies.

Description of the Question

This study has two main objectives:

1. Identify key factors influencing productivity – By analyzing workforce characteristics, operational variables, and workflow interruptions, we aim to pinpoint the most critical elements that impact productivity levels in garment manufacturing.
2. Gain insights from visual analysis to improve productivity– By examining trends and patterns in graphical representations, we seek to uncover actionable insights that can help optimize workflow, reduce inefficiencies, and enhance overall employee performance.

Through this approach, we aim to translate data into meaningful observations that support better decision-making and drive productivity improvements in the garment industry.

Description of our Dataset

The “Productivity Prediction of Garment Employees” dataset, available on Kaggle, provides detailed information on factors affecting employee productivity in the garment manufacturing sector. The dataset covers the period from January 1, 2015, to March 4, 2015, and includes 1,197 observations with 15 attributes.

Attribute Information :

Variable	Description	Type of Variable
date	The date in MM-DD-YYYY format.	Categorical-Ordinal
day	The day of the week when the data was recorded.	Categorical-Ordinal
quarter	A portion of the month, where each month is divided into four quarters	Categorical-Ordinal

department	The department associated with the data instance.	Categorical-Nominal
team_no	The team number corresponding to the data instance.	Categorical-Nominal
no_of_workers	The number of workers in each team.	Numerical-Discrete
no_of_style_change	The number of times the style of a particular product has been changed.	Numerical-Discrete
targeted_productivity	The productivity target set by the authority for each team, for each day.	Numerical-Continuous
smv	Standard Minute Value, which represents the allocated time for completing a task.	Numerical-Continuous
wip	Work in progress, referring to the number of unfinished items for products	Numerical-Discrete
over_time	The amount of overtime worked by each team, measured in minutes	Numerical-Discrete
incentive	The time during which production was interrupted for various reasons.	Numerical-Continuous
idle_time	The time during which production was interrupted for various reasons.	Numerical-Continuous
idle_men	The number of workers who were idle due to production interruptions.	Numerical-Discrete
actual_productivity:	The actual percentage of productivity achieved by workers, ranging from 0 to 1.	Numerical-Continuous

Table 1 - Attribute Information

Data Cleaning and Preprocessing

To ensure data quality and consistency, several preprocessing steps were applied to the dataset:

- **Duplicate Check**: The dataset was examined for duplicate entries, and none were found.
- **Handling Missing Values**: A total of 506 missing values were identified in the Work in Progress (WIP) variable, all from the Finishing Department. Since WIP is typically absent in this department during report submission, these missing values were replaced with 0.
- **Outlier Detection**: Outliers were observed in Targeted Productivity, Overtime, Work in Progress, Incentive, Idle Time, Idle Men, and Actual Productivity. These outliers were retained as they likely reflect natural workflow variations across different teams.
- **Data Type Corrections**:
    - The Number of Workers variable contained decimal values, even though worker count should be a whole number. This was corrected by converting the variable to an integer type.
    - The Date variable was converted to datetime format for better analysis.
- **Quarter Adjustment**: An anomaly was found in the Quarter variable, where some records were labeled as Quarter 5, even though the dataset description specifies only four quarters per month. These records (dated January 29 and 31) were reassigned to Quarter 4 since January cannot be evenly divided into four quarters.
- **Text Standardization**:
    - The spelling of "sewing" in the Department column was corrected.
    - Extra spacing in "finishing" was removed to maintain consistency..
- **Dataset Splitting**: The dataset was divided into training and test sets. The training set contained 957 observations, which were used for descriptive analysis.
- **Productivity Classification :** A new column was created to classify records as Productive or Non-Productive.The classification was based on Actual Productivity compared to Targeted Productivity:
    - Productive: If Actual Productivity ≥ Targeted Productivity.
    - Non-Productive: Otherwise.

These preprocessing steps ensured the dataset was clean, structured, and ready for further analysis.

Main Results of the Descriptive Analysis

Distribution of the Response variable - Actual Productivity

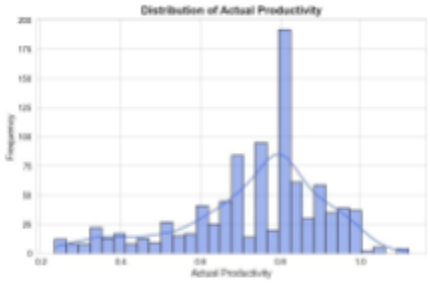

Fig 1 - Distribution of Actual Productivity

The histogram illustrates the distribution of actual productivity among garment employees. It appears to be **left-skewed (negatively skewed)** because the majority of the productivity values are concentrated towards the higher end (around 0.7 to 0.8), with a longer tail extending towards the lower productivity values (0.2 to 0.5). This indicates that most employees perform near or above the average productivity level.

Variation of Actual Productivity based on the Quarter, the Day of the week

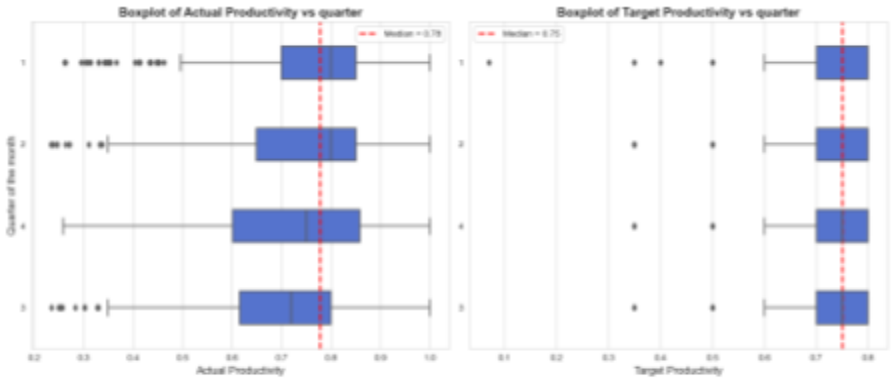

Fig 2 - Boxplot of actual productivity and target productivity with the quarter of month

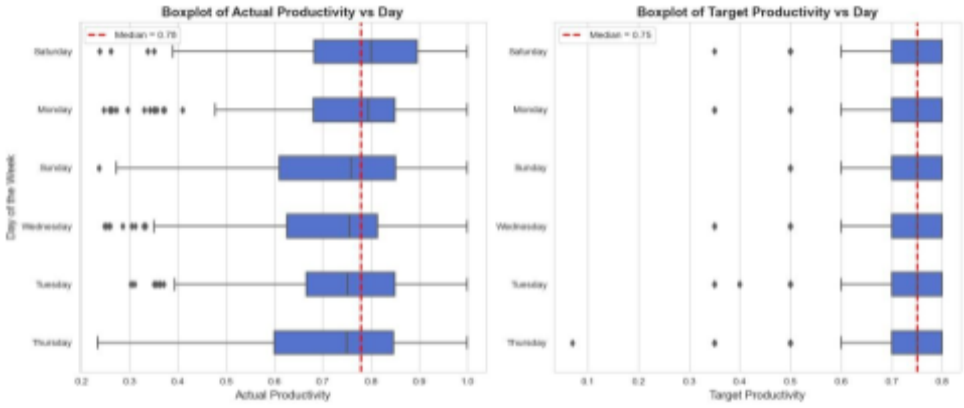

Fig 3 -Boxplot of Actual Productivity and Target Productivity with days of the week

Quarter 1 shows the highest median productivity with less variation, suggesting strong motivation and goal setting at the start of the month. Quarter 3 has the lowest median, indicating a mix of high and low performance, possibly due to mid-month fatigue or disruptions like meetings and

administrative tasks. Quarter 4 shows some recovery but still lags behind Quarters 1 and 2, with a wider spread reflecting both high and low productivity.

On average, employees are meeting and slightly exceeding their target productivity, with a median target of 0.75 and actual productivity at 0.78. The consistent distribution of target productivity across all quarters indicates stable goal setting, while the greater fluctuation in actual productivity suggests that factors like fatigue, motivation, and deadlines are influencing performance. Saturday and Monday show higher actual productivity than median actual productivity.The reason might be that saturdays often have fewer meetings and emails, allowing employees to focus on tasks without distractions.Tuesday & Wednesday display a slight dip in actual productivity, suggesting potential mid-week fatigue or operational inefficiencies.Thursday has a wide productivity range with some low outliers, indicating fluctuations in performance.Steps can be taken to increase productivity on those specific days.Given that actual productivity often exceeds targets, reviewing target-setting methodologies may help maintain realistic and motivating performance benchmarks.

Analysis of Actual Productivity per Department

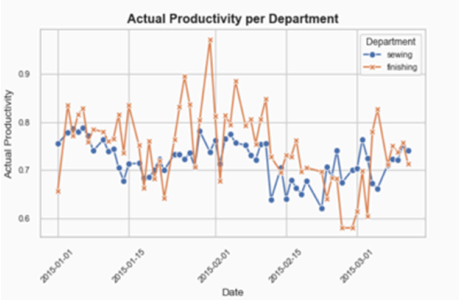

Fig 4 - Line plot of actual productivity per Department

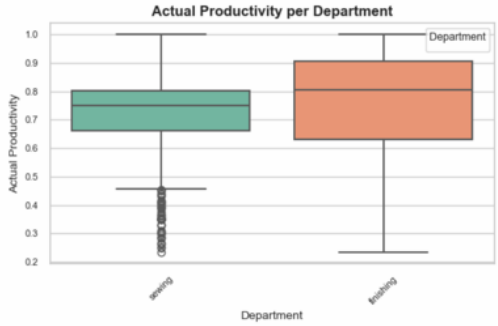

Fig 5 – box plot of Actual productivity with Department

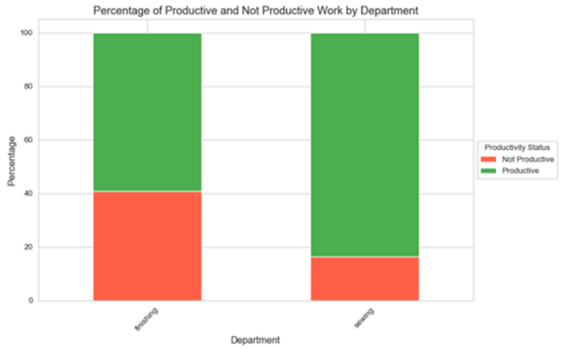

Fig 6 – Percentage of Productive work done by Department

The boxplot in Figure 5 shows that the finishing department has a higher median productivity (0.80) than the sewing department (0.75) but also greater variability, likely due to factors like supply chain delays and batch processing. In contrast, the sewing department maintains more consistent productivity with a compact interquartile range but has lower-end outliers, possibly due to training gaps. The stacked bar chart in Figure 6 confirms that the sewing department has a higher percentage of productive work due to its consistency.The line plot in Figure 4 illustrates fluctuations in productivity from January to March 2015. The sewing department maintains steadier levels, while the finishing department experiences sharp peaks and drops. These trends align with the boxplot in Figure 5, highlighting that while the finishing department can achieve high productivity under optimal conditions, its inconsistency reduces overall efficiency, as reflected in the stacked bar chart in figure 6.

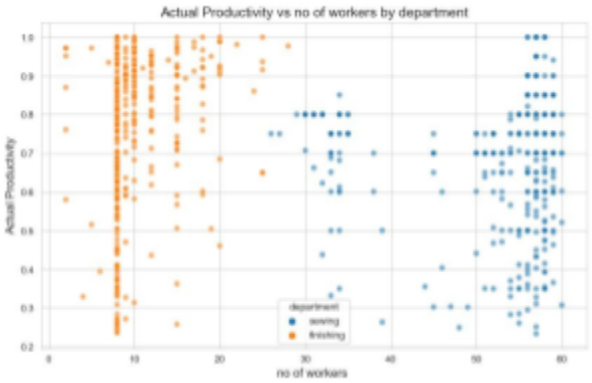

Fig6 : Scatter plot of Actual productivity vs no.of workers

Fig10 : Scatter plot of Actual Productivity vs overtime

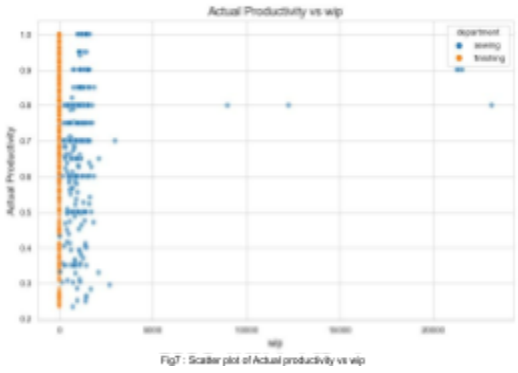

Fig7 : Scatter plot of Actual productivity vs wip

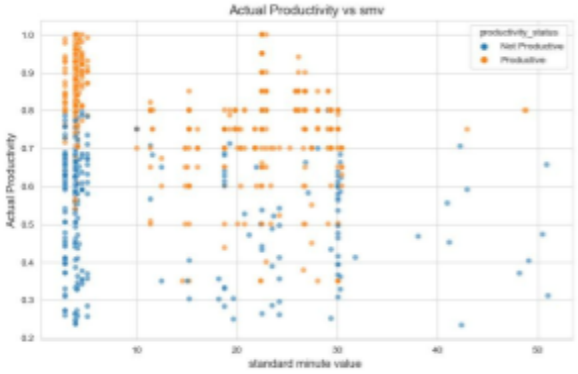

Fig8 : Scatter plot of Actual Productivity vs smv

Figure 7 shows that the Finishing Department has almost no Work in Progress (WIP), while the Sewing Department exhibits varying WIP levels. However, there is **no clear correlation between WIP and Actual Productivity**, as productivity remains scattered across different WIP values. Figure 8 highlights that the Finishing Department has a narrow SMV range, indicating repetitive tasks, whereas the Sewing Department's wider SMV range reflects a more diverse workload. Overall, **no clear trend links SMV to Actual Productivity in either department**. Sewing is more consistent due to its wider range of tasks, continuous workflow (WIP), and less dependency on other processes, while Finishing faces fluctuations due to batch processing, minimal WIP, and reliance on Sewing for inputs. The Finishing Department has fewer workers (mostly below 20), while the Sewing Department has a wider range, reaching up to 60 workers. In the Finishing Department, a **slight positive correlation** is observed between the number of workers and productivity, suggesting that adding more workers may improve efficiency. However, in the Sewing Department, there is **no clear correlation** between worker count and productivity, indicating that other factors, such as workflow efficiency and task complexity, may play a bigger role in determining productivity levels.

According to figure 10 ,the overtime range for the finishing department is relatively small, meaning employees in this department do not work excessive overtime.But the sewing department shows a much wider range of overtime, with some employees working significantly more overtime.For both departments, the **data does not show a strong trend** where overtime directly affects actual productivity.

Analysis of Productivity vs Teams

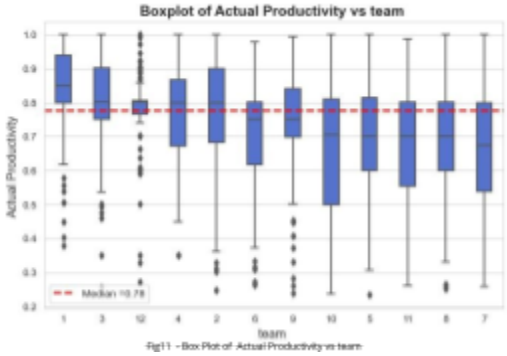

Fig11 - Box Plot of Actual Productivity vs team

The boxplot analysis reveals a clear distinction between the best and worst performing teams. Team 1 is the top performer with the highest median productivity, followed by Team 12 and Team 3, which are on the same level. Team 1 stands out for its consistently high productivity, with a higher upper quartile and only a few lower outliers. Team 12 exhibits steady performance with a narrow IQR but has outliers at both ends, indicating a mix of high and low performers. Team 3, although a strong performer, shows more variability and has several outliers, indicating greater fluctuation in individual team member productivity. On the other hand, Teams 10, 5, 11, 8, and 7 fall below the overall median of 0.78, with Team 7 having the lowest median productivity. Teams 10, 5, and 11 share the same low median productivity, but Team 10 stands out with the highest variability, followed by Team 11 and Team 7. The large spread in these teams' performance levels suggests significant inconsistency, with some members performing well while others lag behind. Team 7, with the lowest median, represents the weakest performer, and its performance is marked by high variability, but less so than Teams 10 and 11.These performance differences could result from factors like team dynamics, skill variability, motivation levels, leadership effectiveness, and alignment of roles. Teams with high variability may face issues like inconsistent contributions, while top-performing teams benefit from clear roles, strong collaboration, and effective leadership.

Effects of Incentives Pay on Productivity

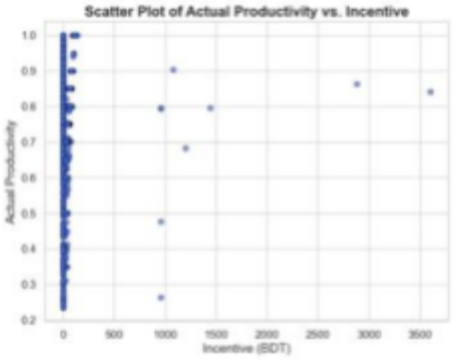

Fig12 - Scatter Plot of Actual Productivity vs Incentive

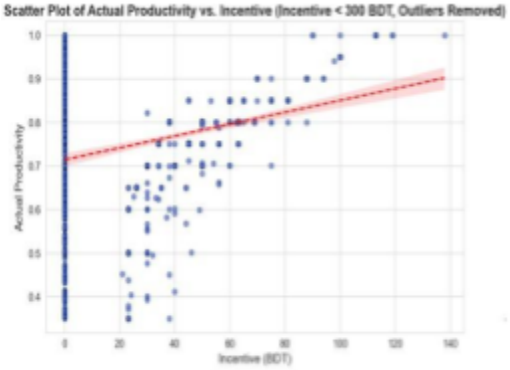

Fig13 - Box Plot of Actual Productivity vs Incentive (< 300)

The first plot illustrates the relationship between actual productivity and incentive amounts,with a considerable spread. The second plot focuses on incentives below 300 BDT (Bangladesh Taka), with outliers removed, revealing a slight positive trend within this limited incentive range. This suggests that smaller incentives may have a more consistent positive impact on productivity, while the effect of larger incentives is more variable.

Analysis of Actual Productivity with Interruptions: Idle Time, Idle Men, and Style Changes

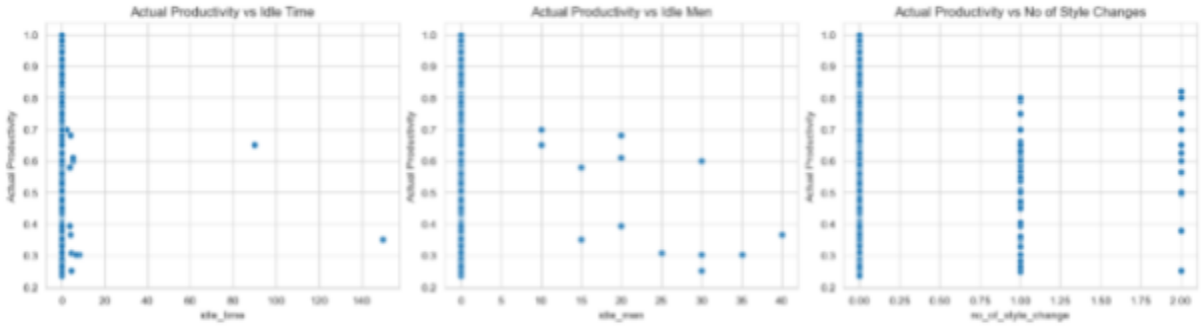

Fig14 - Scatter Plot of Actual Productivity vs Idle time , Idle Men,No.of Style changes

From the three scatter plots above, it is evident that the garment factory operates with minimal interruptions. The idle time is largely concentrated around zero, with only a few observations showing idle times above zero. Similarly, the number of idle men is mostly zero, indicating minimal interruptions. The range of style changes, which is from 0 to 2, is mostly clustered around 0, showcasing that the garment factory underwent minimal style changes. This suggests that the productivity downturn is not due to interruptions, as these factors are minimal and do not show a significant relationship with actual productivity.

Principal Component Analysis

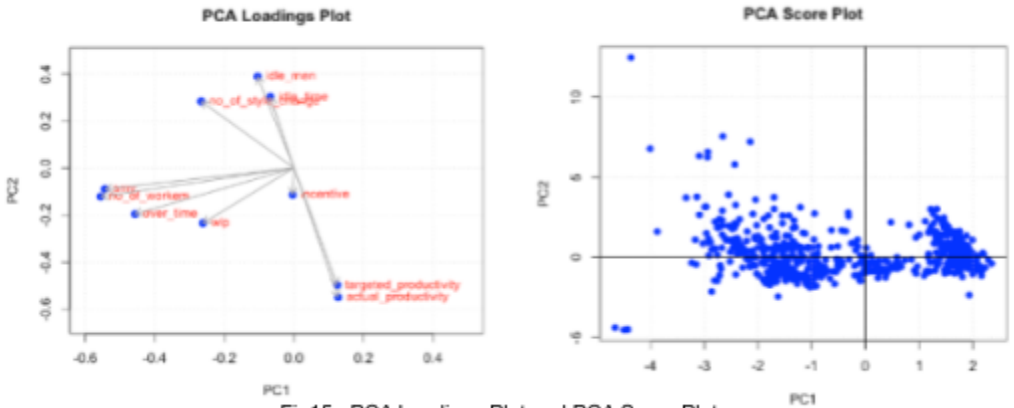

Fig15 : PCA Loadings Plot and PCA Score Plot

Principal Component Analysis (PCA) was used to reduce the dimensionality and to see relationships between variables and observations in a 2D space defined by the first two principal components (PC1 and PC2). For PCA we only use the quantitative variables in our dataset because PCA can be performed only with quantitative variables. The loadings plot shows that variables like targeted_productivity and actual_productivity are close to each other. This means they are positively correlated. On the other hand, variables like idle_time and idle_men are in the opposite

direction, meaning they are negatively correlated with targeted_productivity and actual_productivity**.** Variables like over_time**,** smv**,** and no_of_workers are roughly at a 90-degree angle to targeted_productivity and actual_productivity**,** meaning they are not correlated. The score plot shows how observations are placed along PC1 and PC2. There are no clear clusters or patterns, but a few outliers can be seen.However, the first two components together explain only 45.78% of the total variance in the data. Because of this, we cannot make strong conclusions about variable correlations or data clusters based only on these components.

Partial Least Squares

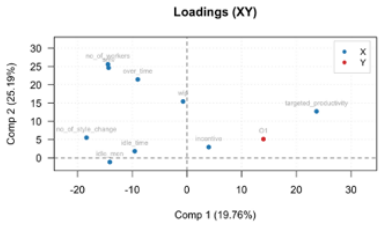

Fig 16 – Loading plot Comp1 and Comp2

Partial Least Squares (PLS) is used to see how much of the variance in actual_productivity(Y) can be explained by other quantitative variables(X) while reducing the number of dimensions to two, because If we can explain a reasonable amount of variability using just two dimensions, we can create meaningful plots that capture the overall structure of the data in a 2D space. However, in this case, the first two components explain less than 24% of actual_productivity(Y), meaning they do not provide a strong enough summary of the relationship between X and Y. (the percentages in the plot are the percentage of the variance of X explained by 2 components, not the percentage of the variance of Y explained by 2 components).

Suggestions for Advanced Analysis

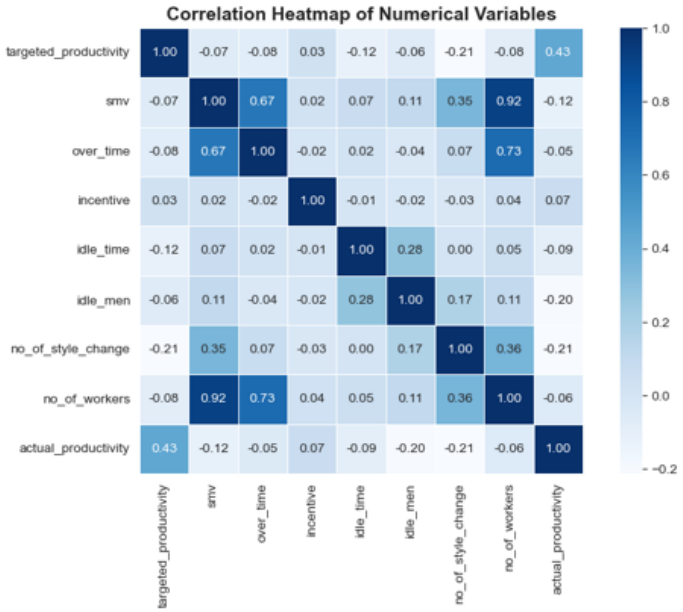

Fig 17 – Correlation Heatmap

The dataset contains correlated variables as shown in the heatmap , suggesting multicollinearity and in addition to that most predictor variables may be irrelevant in explaining the response variable. Therefore multiple linear regression may not be the most suitable and alternative methods such as **Ridge Regression, Lasso Regression, or Partial Least Squares (PLS) Regression** may be more appropriate. These methods help address **multicollinearity** and reduce the impact of irrelevant predictors. However, due to the presence of outliers in the dataset, methods like Ridge and Lasso may still struggle with accuracy, as they don't specifically address outliers.Tree-based algorithms, such as **regression trees**, **random forests**, and **XGBoost**, are less affected by outliers, which makes them a good choice when dealing with datasets that include extreme values. These models are particularly effective in identifying and modeling complex, nonlinear relationships between the predictor variables and the response, leading to potentially more accurate predictions of actual productivity for garment workers compared to traditional regression models.

Conclusion

The analysis of actual productivity in the garment factory highlights key patterns in employee performance. Overall, productivity is generally high, with most employees performing near or above the target. However, fluctuations exist across different time periods, departments, and teams.
Quarter 1 exhibits the highest median productivity with lower variability, suggesting strong motivation at the start of the month. Productivity declines in Quarter 3, likely due to mid-month fatigue or operational disruptions, before showing slight recovery in Quarter 4. Additionally, productivity is higher on Saturdays and Mondays, potentially due to fewer distractions, while mid-week dips indicate fatigue or inefficiencies

Department wise, the Sewing Department demonstrates more consistent productivity due to a steady workflow, whereas the Finishing Department experiences greater fluctuations, likely influenced by batch processing and supply chain delays. Despite having fewer workers, the Finishing Department shows a slight positive correlation between worker count and productivity, whereas the Sewing Department does not, suggesting that efficiency is driven by process stability rather than workforce size.

Team performance analysis reveals notable disparities. Team 1 emerges as the top performer with consistently high productivity and minimal variability. Team 12 and Team 3 follow closely but show some fluctuations. In addition, Teams 10, 5, 11, 8, and 7 fall below the overall median, with Team 7 recording the lowest productivity. High variability teams likely struggle with inconsistent contributions and role alignment, whereas top-performing teams benefit from strong collaboration and leadership. Moreover, interruptions such as idle time, idle workers, and style changes are minimal, indicating that productivity downturns are not caused by disruptions. Instead, factors like motivation, workload distribution, and operational processes play a more significant role in determining overall efficiency. Addressing these areas, especially mid-week fatigue, department-specific inefficiencies, and team-level inconsistencies can help optimize performance across the factory.

Appendix

- Link for the dataset :
  https://www.kaggle.com/datasets/ishadss/productivity-prediction-of-garment-employees
- Link for the Google Collab NoteBook :
  - https://colab.research.google.com/drive/11MNG7oRs1Mzukd5JK2uVUJ5AjpCDxeKW?usp=sharing
  - https://colab.research.google.com/drive/1yKPvLfRLGdU4dCMweZeIX-EXsvE4lOC1?usp=sharing
  - https://colab.research.google.com/drive/1o58_vOorvX3K1Cj02_8rNV1DkENGQfe5?usp=sharing

References

- https://colab.research.google.com/drive/1I-JUr3dMi7Y-5u6WpbaQO7JL4DH4fH-i?usp=chrome_ntp
- https://www.geeksforgeeks.org/principal-component-analysis-pca/
- https://en.wikipedia.org/wiki/Partial_least_squares_regression#:~:text=Partial%20least%20squares%20(PLS)%20regression,model%20by%20projecting%20the%20predicted