STITCHING SUCCESS

GROUP 11

PREDICTING GARMENT WORKER PRODUCTIVITY THROUGH DATA ANALYTICS AND MACHINE LEARNING

Prepared for :

Data Analysis Project 1

Prepared by :

Ruwinda Rowel - s15654
Vihanga Anjana - s15627
Sithmi Pehara - s15494
Darshi Yashodha - s15584

Abstract

The garment industry focuses on making various clothing types, representing a trillion-dollar global business. It heavily relies on both labor and capital, being labor- and capital-intensive. Workforce productivity measures goods and services produced by a worker group within a specific timeframe, correlating with increased profits and efficient input use. Maximizing worker productivity in the garment industry is crucial for success. This project aims to conduct a descriptive analysis on a garment workforce productivity dataset sourced from Kaggle. The exploration involves a comprehensive examination of each variable using graphical and numerical methods to understand the nature and distribution of attribute values. Additionally, the project investigates the relationships between predictor variables and the actual productivity response variable, aiming to identify influential predictors that characterize productivity levels. The findings of this descriptive analysis shed light on the critical implications of workforce productivity on overall organizational success, providing actionable insights for improvement.

Contents

List of Figures

List of Tables

Introduction

The garment industry, a dynamic and multifaceted sector within the broader realm of fashion and textiles, plays a pivotal role in shaping global consumer culture. From designing and manufacturing to distribution and retail, the garment industry encompasses a vast and interconnected network of processes that bring clothing from concept to the closets of millions. Characterized by constant innovation, rapidly changing trends, and a complex supply chain, this industry is a powerhouse of creativity, technology, and entrepreneurship. As a keystone of the global economy, the garment industry stands at the intersection of artistry and commerce, continuously adapting to meet the demands of a diverse and discerning global audience. We conduct this analysis to enhance operational efficiency and inform data-driven strategies for workforce productivity improvements.

Description of the Question

The productivity prediction of garment employees is a critical aspect of optimizing operational efficiency within the textile and apparel industry. Leveraging advanced technologies such as data analytics, machine learning, and industrial engineering methodologies, businesses aim to forecast and enhance the performance of their workforce. By analyzing historical data, identifying key performance indicators, and implementing predictive models, companies can gain valuable insights into factors influencing productivity, enabling proactive measures to improve workflow, resource allocation, and overall output. This approach not only contributes to increased efficiency and cost-effectiveness but also empowers the industry to adapt to changing market dynamics while ensuring a conductive and supportive environment for its workforce.

In the pursuit of optimizing productivity in the garment industry, our objectives are twofold.

1. By scrutinizing how various elements interact, we seek to unearth valuable insights and discern any emerging trends that may significantly impact productivity. This investigative approach enables us to understand the complex dynamics within our industry.
2. Identify influential predictor variables that serve as key drivers of productivity. That is to pinpoint the critical factors that play a pivotal role in shaping efficiency levels.

Together, these objectives underscore our commitment to enhancing productivity through a comprehensive and data-driven exploration of the factors shaping our industry.

Description of Dataset

The “Productivity Prediction of Garment Employees” dataset, sourced from the UCI Machine Learning repository and made available on Kaggle, presents a comprehensive collection of attributes relevant to the garment manufacturing process and the productivity of the employees. The dataset spans from January 1, 2015, to March 4, 2015. The dataset comprises of 1197 observations and 15 variables.

No.	Attribute	Type of variable	Comments
1	date	Categorical-Ordinal	Date in MM-DD-YYYY
2	quarter	Categorical-Ordinal	A portion of the month. A month was divided into four quarters
3	department	Categorical-Nominal	Associated department with the instance
4	day	Categorical-Ordinal	Day of the Week
5	team	Categorical-Nominal	Associated team number with the instance
6	targeted_productivity	Numerical-Continuous	Targeted productivity set by the Authority for each team for each day.
7	smv	Numerical-Continuous	Standard Minute Value, it is the allocated time for a task
8	wip	Numerical-Discrete	Work in progress. Includes the number of unfinished items for products
9	over_time	Numerical-Discrete	Represents the amount of overtime by each team in minutes
10	incentive	Numerical-Continuous	Represents the amount of financial incentive (in BDT) that enables or motivates a particular course of action
11	idle_time	Numerical-Continuous	The amount of time when the production was interrupted due to several reasons
12	idle_men	Numerical-Discrete	The number of workers who were idle due to production interruption
13	no_of_style_change	Numerical-Discrete	Number of changes in the style of a particular product
14	no_of_workers	Numerical-Discrete	Number of workers in each team
15	actual_productivity	Numerical-Continuous	The actual % of productivity that was delivered by the workers. It ranges from 0-1

Table 1 :- Description of Variables

Data Preprocessing

- The data set was checked for duplicates. There were no duplicates.
- Checked for missing values and identified 506 missing values in the "work in progress" variable, all of which belong to the Finishing Department. Normally report submission mandates an absence of work in progress in the Finishing Department, we have assigned all these missing values a value of 0.
- Checked for outliers. Outliers have been observed in the columns related to targeted productivity, overtime, work in progress, incentive, idle time, idle men, and actual productivity. However, these outliers will not be removed, as they may be attributed to the inherent variability in the workflow among different teams.
- There were decimal values in the "number of workers" variable. However, the number of workers should be a whole number. We converted that variable into an integer-type variable.
- There was "Quarter 5" in the "Quarter" variable. However, according to the variable description, there should only be four quarters. "Quarter 5" corresponds to the dates of January 29 and 31. Since January cannot be evenly divided into 4 quarters, we assigned the dates of January 29 and 31 to the 4th quarter.
- Corrected the spelling of "sewing" in the department column. Removed the spacing from the word "finishing" in the department column.
- Convert data type of date variable in to datetime format.
- Added a new column including only the date of the job for ease of analysis.
- Then, the dataset was split into training and test sets. The training set consisted of 957 observations. Descriptive analysis was conducted using the training set.

- Settings were adjusted so that Univariate, Bivariate and Multivariate plots can be created and analyzed during the thorough descriptive analysis.

Results of Descriptive Analysis

Distribution of the Response Variable: Actual Productivity

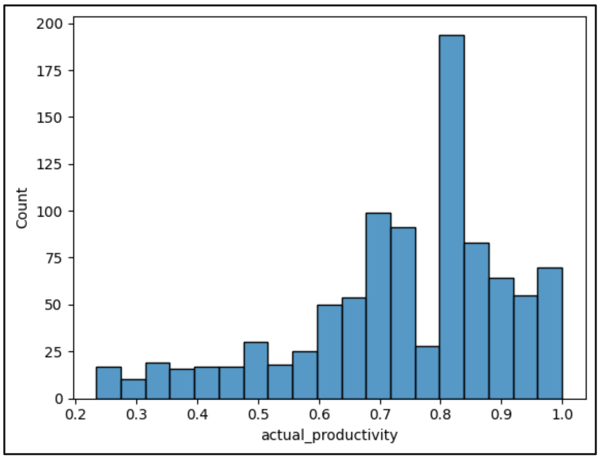


Figure 1 :- Histogram of Actual Productivity

The distribution of the response variable Actual Productivity is mainly centered around the productivity level of 0.8. The response variable can be said to follow a relatively normal distribution with a mean of 0.735897 and median of 0.782448. Given this relative symmetry and the fact most of the data fits with 3 standard deviations of the center, which is the mean we can approximate the distribution to be normal.

Min	Q2	Median	Mean	Q3	Max	Std
0.233705	0.650408	0.782448	0.735897	0.850362	1.000000	0.173719

Table 2 :- Summary Statistics of Actual Productivity

Distribution of Actual Productivity based on the Quarter, the Day, and the Date of the Month.

Observing the Distribution of Productivity based on the workday of the week, the highest mean productivity is achieved jointly relatively on Wednesday, Saturday, and Monday while the lowest is achieved on Sunday, Tuesday and Thursday. Steps can be taken to increase productivity on those specific days. Production based on the Quarters progressively goes down towards the latter quarters. This may be because workers are more motivated to work after the pay day, justifying the increase in productivity in the 1st quarter which gradually decreases towards the latter quarters.

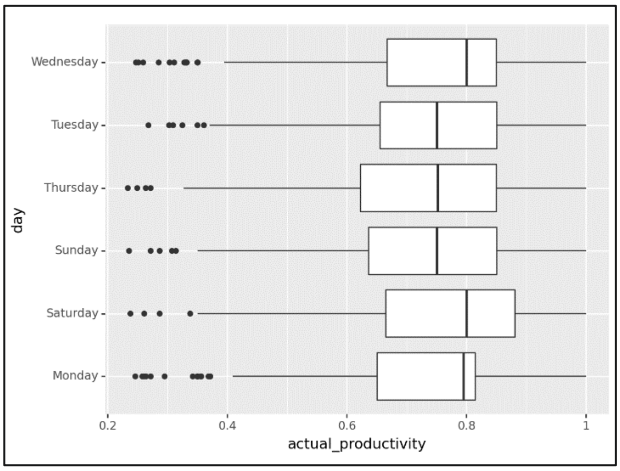


Figure 2 :- Box Plot of Actual Productivity vs Day

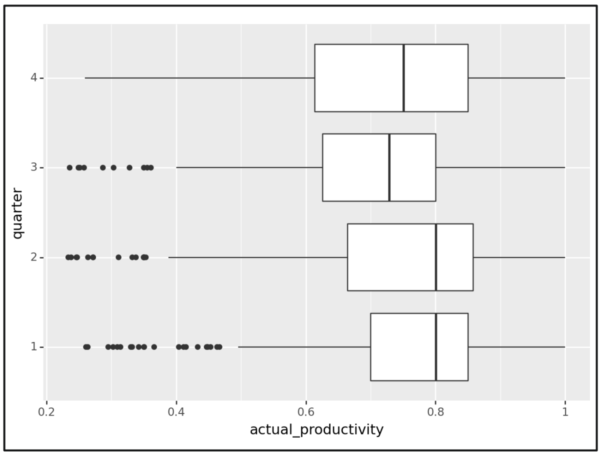


Figure 3: - Box Plot of Actual Productivity vs Quarter

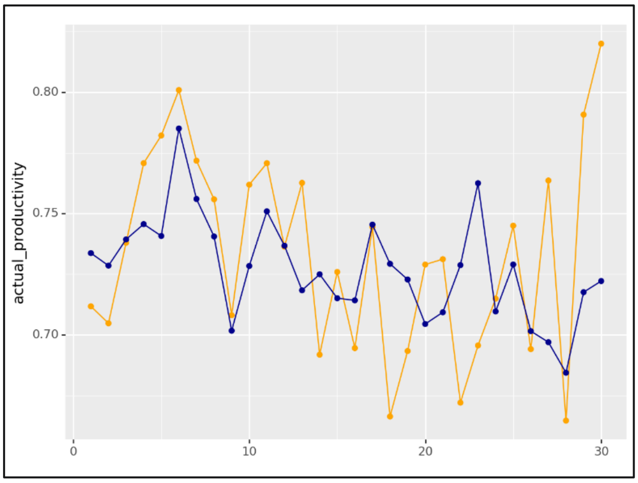


Figure 4: - Line Plot of Actual Productivity and Targeted Productivity vs Day

Observing the distribution of the mean actual productivity levels (Orange Line) of the dates of the month we come across a relatively downward trend. Although it is not perfectly downward, on average productivity goes down towards the end of the month and spikes on the last day of the month, which maybe when payday occurs. Furthermore, we can see that the whole assembly line is achieving most of the targets that are set out to them with rather low variability of the production levels (Orange Line) against the targeted Production Levels (Dark Blue Line) for each days.

Effect of Number of Workers on the Actual Productivity

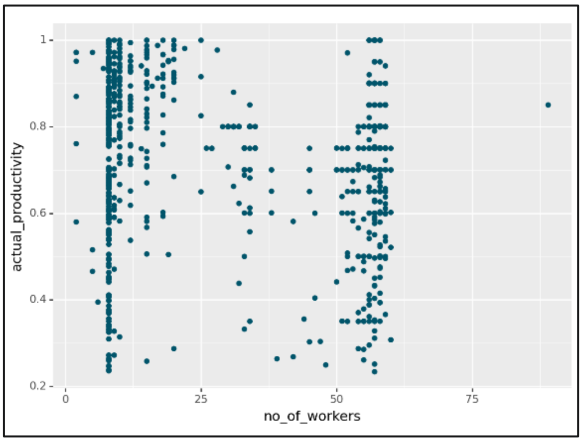


Figure 5: - Scatter Plot of no of Workers vs Actual Productivity

Observing how productivity is scattered according to the no. of workers in a team, there are 2 main clusters. Teams with workers less than 25 and teams with workers more than 50. If you look at both clusters individually, we observe that there is a positive linear relationship between the number of workers and the actual productivity indicating a higher number of workers in a team yield higher productivity. The existence of the 2 cluster though indicates the utilization of team members based on the task at hand. This infers the fact that to achieve more targets a higher number of teams members can be utilized on average.

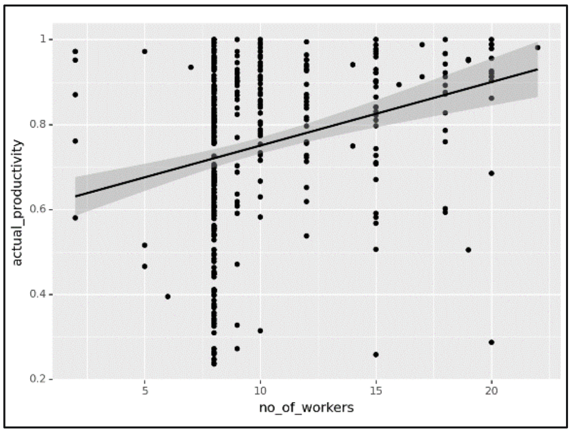


Figure 6: - Scatter Plot of no of Workers<25 vs Actual Productivity.

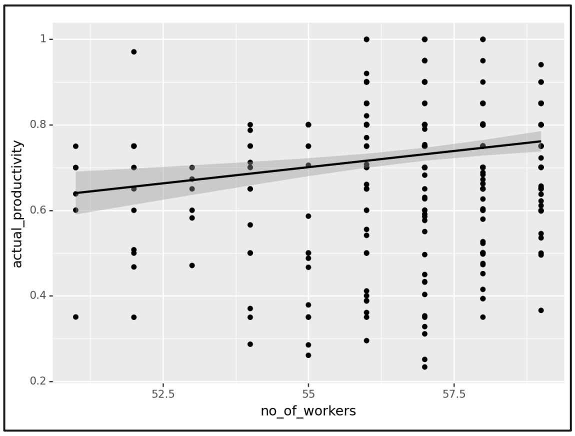


Figure 7: - Scatter Plot of no of Workers >50 vs Actual Productivity

Effects of Incentives and Overtime Pay on Productivity.

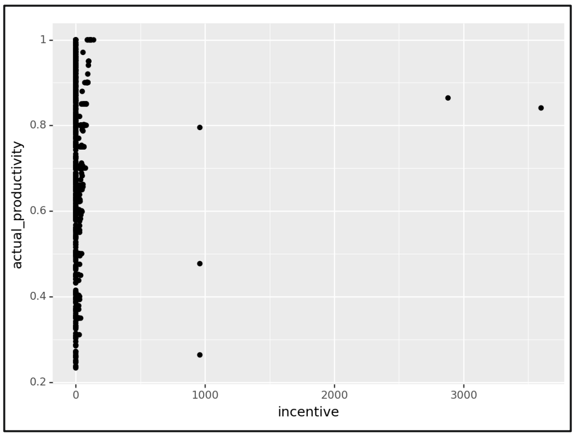


Figure 8:- Scatter Plot of Actual Productivity vs Incentives

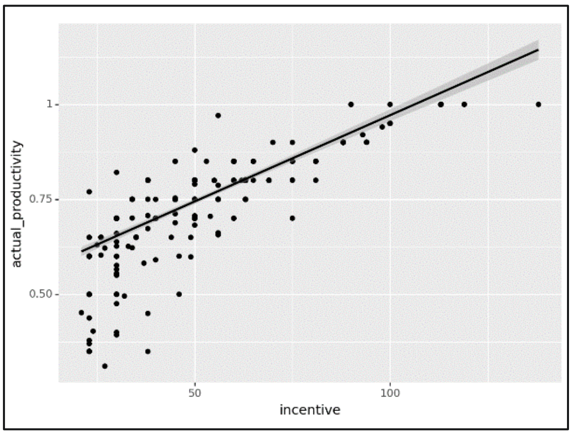


Figure 9:- Scatter Plot of Actual Productivity vs Incentives without Outliers

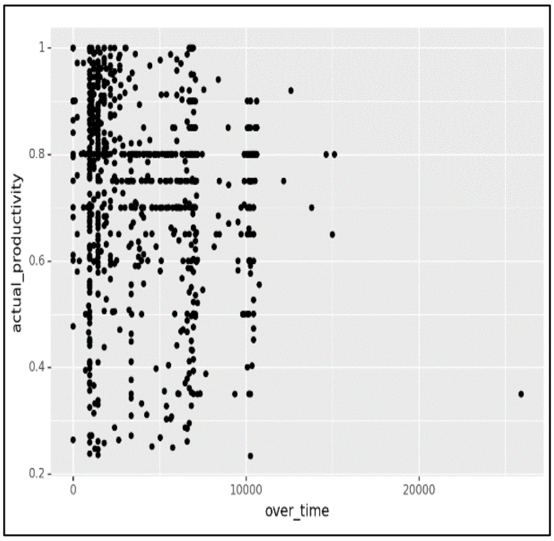


Figure 10:- Scatter Plot of Overtime vs Actual Productivity

Given a there are many outliers, and the data is skewed towards the left, we consider instances where the incentive pay is less than 250. Once we plot productivity based on the incentive, they have received we can observe that there is a positive linear relationship between the incentives and actual productivity. This indicates on average if we increase the incentive pay of workers the average level of productivity will increase. But we do not necessarily see such a relationship between overtime pay and productivity. This may be because overtime was required to fill out targets set out for those teams or individuals and may have had mixed effects on the productivity itself.

Relationship between WIP and Productivity and How Teams Manage WIP

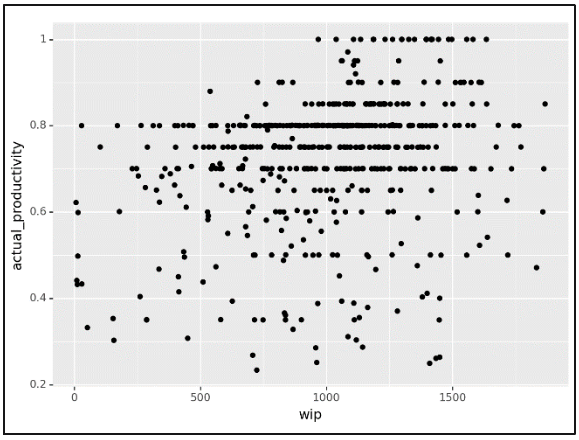


Figure 11:-Scatter Plot of WIP vs Actual Productivity

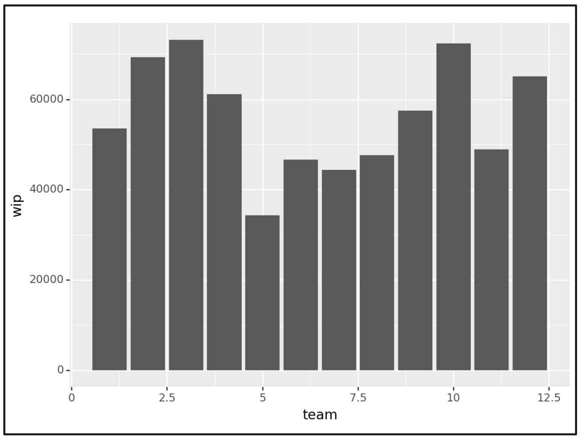


Figure12:- Bar Graph of Team vs WIP

The above plots indicate that there is not much of a linear relationship between the work in progress of an individual and the productivity that can be achieved with a higher or lower level of work-in-progress. Further by observing the bar graph to the right we can see that team 5 is handling their work in progress much better as compared to the other teams while teams 3 and 10 have the least performance when it comes to handling their work in progress.

Relationship between Productivity and Team

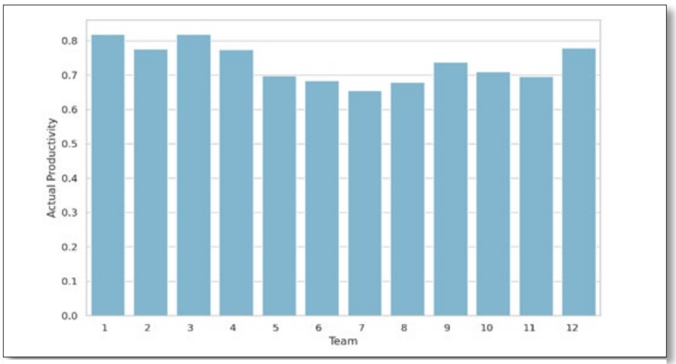


Figure 13:- Bar Graph of Actual Productivity vs Team

The bar graph provides the average actual productivity across twelve distinct teams. According to the graph all teams demonstrate commendable productivity by surpassing the 60% mark. But teams 1 and 3 exhibit the highest productivity of more than 80% while teams 6,7 and 8 exhibit the lowest productivity contribution.

Relationship between Incentive and Team

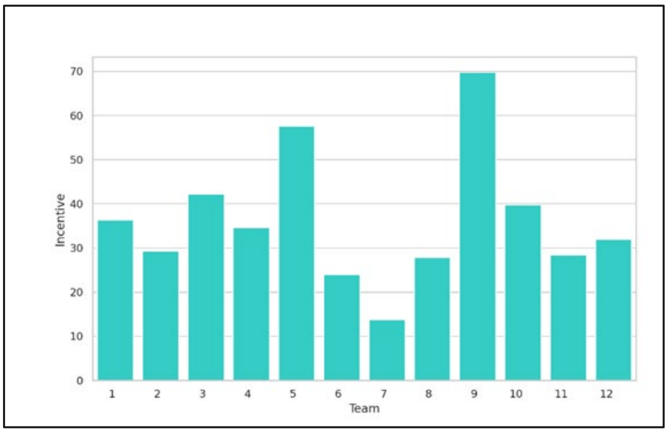


Figure 14:- Bar Graph of Incentive vs Team

The bar graph illustrates Team 9 receiving the highest incentive, while most others receive about half. In contrast, Teams 6, 7, and 8 receive the least. While one might anticipate that such variations in incentives could influence team productivity, a closer look at *Figure 13* contradicts this assumption. Team 9 doesn't rank among the top-performing teams with the highest incentive, while Teams 6, 7, and 8, with the lowest incentives, show the lowest productivity.

Relationship between Productivity and Team with No. of Style Changes

All teams, except for 7, 11, and 12, experienced a notable decline in actual productivity with an increase in the number of style changes. This is likely due to the disruptive nature of style changes, making it challenging for employees to maintain efficiency. However, teams 7 and 11 exhibit the opposite trend, showing higher actual productivity with an increased number of style changes. Team 12, having no style changes, prevents us from making any observations about the impact of style changes on productivity.

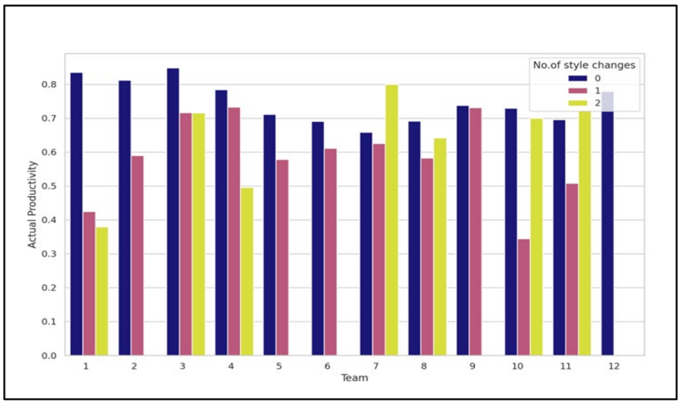


Figure 15: - Bar Graph of Actual Productivity vs Team with No. of Style Changes

Relationship between Productivity and Department

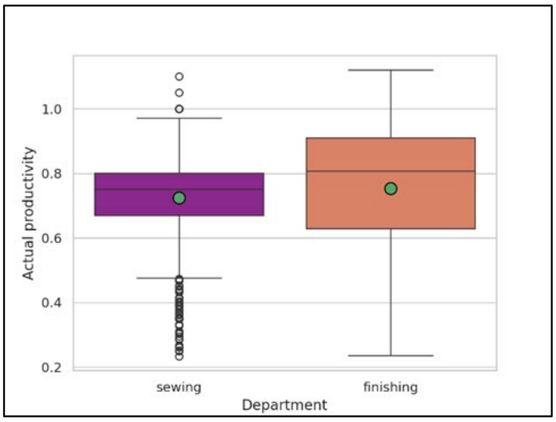


Figure 16: - Box Plots of Actual Productivity vs Department

The sewing department exhibits higher median and mean productivity compared to the finishing department, suggesting generally better performance by sewing department workers. The narrower box length in sewing indicates less variability and a more consistent average productivity, although with more unproductive outliers. In contrast, the finishing department's wider box length and dispersion suggest a broader range of productivity levels. This could reflect differences in task complexity, skill levels, or work organization between the departments.

Relationship Between Productivity and No. of Style Changes with Department

Examining the strip plot, we observe that the sewing department typically encounters 0 to 2 style changes, while the finishing department, in contrast, records no style changes at all. Nevertheless, the sewing department manages to sustain high productivity levels, as is evident in *Figure 16*. Furthermore, it becomes noticeable that increasing style changes in the sewing department appear to be associated with a decrease in productivity. Additionally, each department exhibits considerable variability within the style change category, reflecting diverse performance level

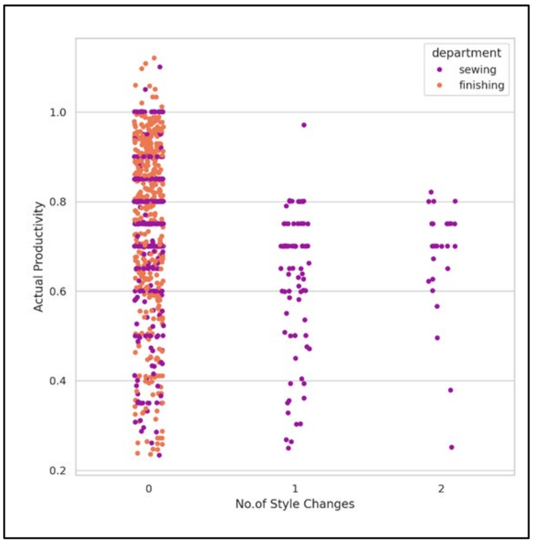


Figure 17:- Strip Plot of Actual Productivity vs No. of Style Changes with Department

Relationship between Median Actual Productivity Vs Number of Workers

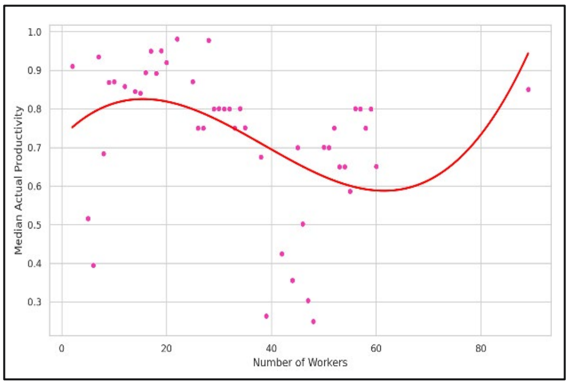


Figure 18:-Scatter Plot of Median Actual Productivity Vs Number of Workers

In this scatter plot, we observe an interesting pattern where the median productivity rate increases with the addition of more workers until reaching a peak. Subsequently, as the number of workers continues to increase, productivity declines to a certain level before rising again. This behavior suggests that initially, more workers lead to increased productivity, but beyond a certain point, the number of workers becomes cumbersome, resulting in decreased productivity. Nevertheless, certain tasks may require an additional increase in the number of workers for improved productivity.

Relationship between Actual Productivity Vs Number of Workers with Department

Upon initial observation, there seems to be a negative correlation between the number of workers and actual productivity when not considering the department. However, upon closer inspection, department-wise analysis reveals a slight positive correlation within the finishing department and no significant relation in the sewing department.

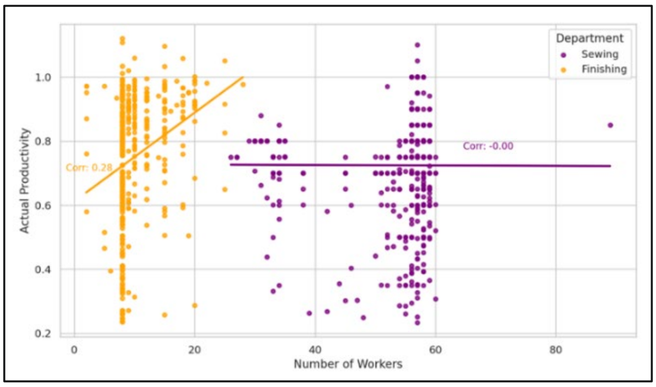


Figure 19:- Scatter Plot of Actual Productivity Vs Number of Workers with Department

Relationship between Idle Time and Idel Men with Date

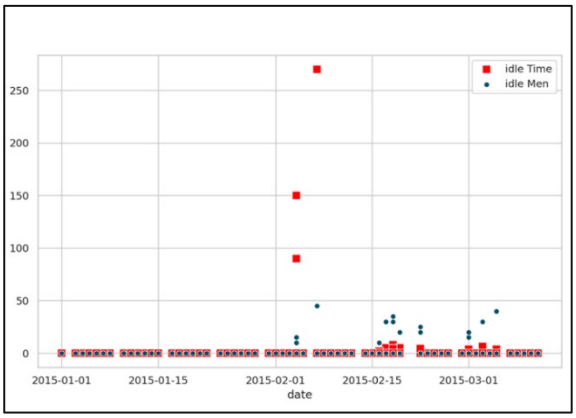


Figure 20:- Scatter Plot of Idle Time and Idel Men with Date

Over the three-month period, the production process has been interrupted around 15 times within 10 days and at each time, at least 10 workers have been idling due to these production interruptions. Moreover, from the graph we can say that the duration of production interruption leads to idle workers. This directly affects the efficiency and utilization of the production process.

Relationship between Idle Time and Actual Productivity

This scatter plot suggests that dates with idle time generally have low mean productivity compared to dates without idle time. However, there is an overlap between the two groups as some days with idle time even have higher productivity than dates without. Additionally, the significant variability in mean productivity within for different dates, even for the same idle time category suggests that other factors beyond idle time are likely to influence productivity.

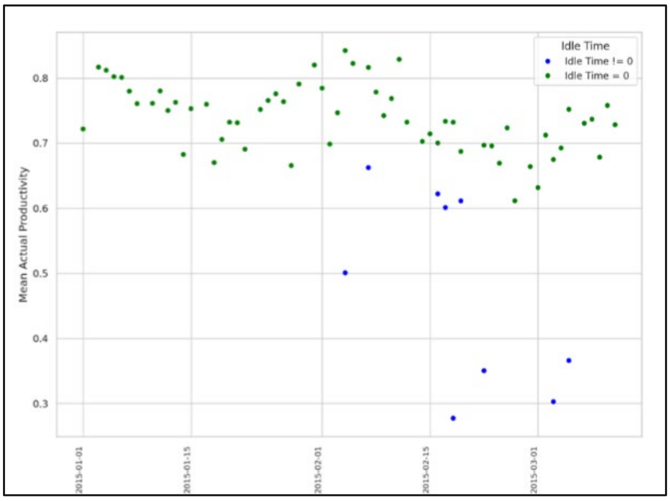


Figure 21:-Scatter Plot of Actual Productivity vs Date with Idle Time = 0 and Idle Time ≠ 0

Relationship between Standard Minute Value and Actual Productivity

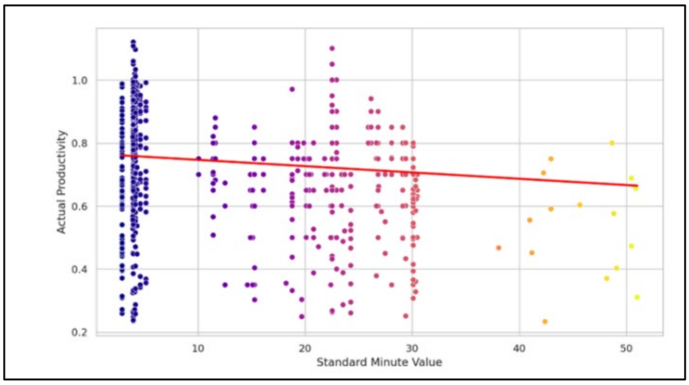


Figure 22:- Scatter Plot of Actual Productivity vs SMV

This graph illustrates a clear trend: as standard minute values (SMV) increase, there is a noticeable decline in actual productivity. This decline may be attributed to higher SMV values, suggesting more challenging or time-consuming tasks. Additionally, the graph reveals distinct clusters for SMV, indicating potential groupings based on similar task complexities or durations.

Principal Component Analysis

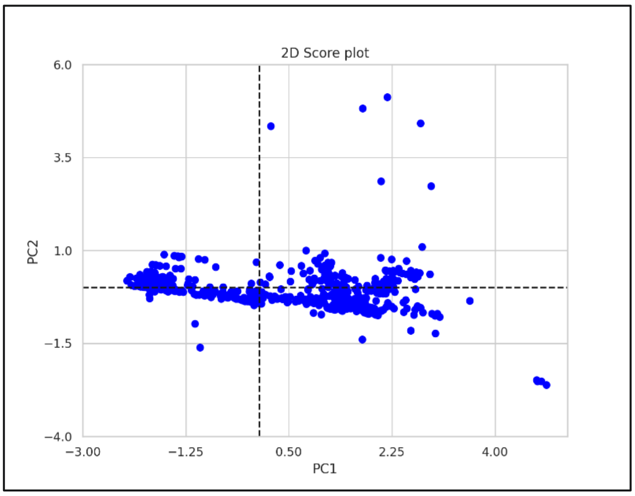


Figure 23:-2D Score Plot of X

Principal Component Analysis was conducted on the data to identify potential clusters among the observations and to visualize the presence of outliers. Before applying PCA, all categorical variables were removed from the dataset, as PCA exclusively works with quantitative data. To present the data in a two-dimensional plane, the dimensionality of x was reduced to two by selecting only two principal components x space. Consequently, the accuracy of the score plot obtained by PCA becomes questionable. The score plot, illustrated below, provides a preliminary indication that there are no significant clusters in the data set. Additionally, some points are observed to lie far from the center of the score plot, indicating the presence of outliers in the dataset.

Partial Least Square Regression

Partial Least Squares Regression was executed on the training set to identify predictors which significantly correlated with each other and illustrate the relationship of each predictor with the response, actual productivity. The plot of loadings of XY suggests strong relationships among some predictors and the response, such as incentives and targeted productivity. Conversely, there are predictors orthogonal to the response, such as the number of workers and standard minutes work. However, the accuracy of the loading plot is also questionable since the first two directions extracted from PLS explain very low variation in both x space and y space.

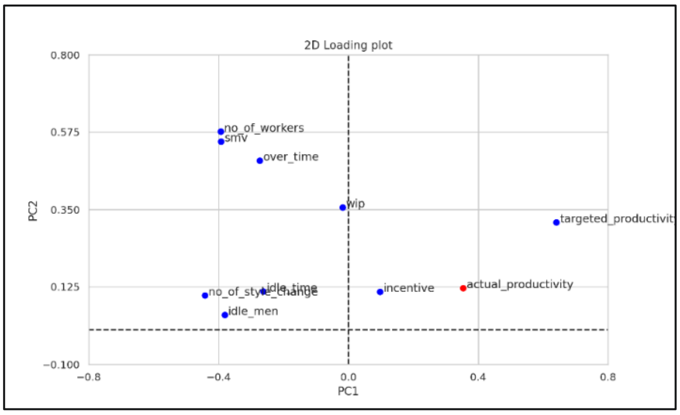
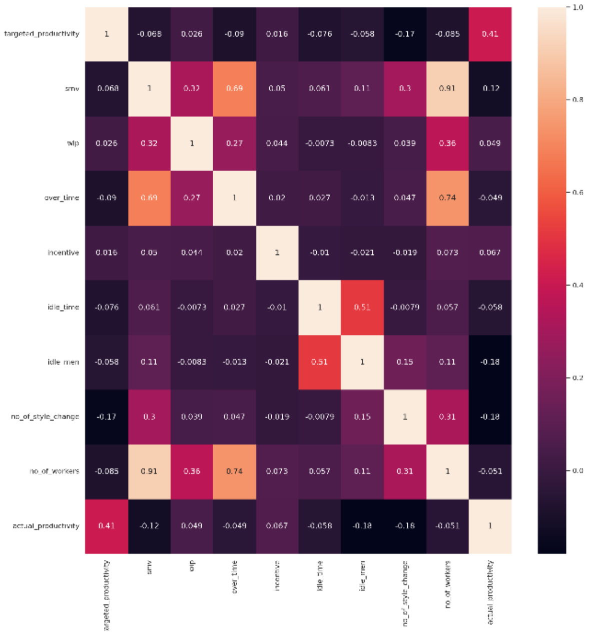


Figure 24:- 2D Loading Plot of XY

Suggestions for Advanced Analysis



The following correlation heatmap reveals that most predictor variables are weakly correlated with the response variable. This could be due to the high presence of outliers. Therefore, the Multiple Linear Regression (MLR) model may not be the most suitable for advanced analysis. It is crucial to employ variable selection methods to identify important variables that explain the variation in the response if we choose to fit MLR. Additionally, some predictors exhibit high correlation, which can impact the MLR model due to multicollinearity issues. To address this, Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) can be employed, though caution is needed due to the presence of a high number of outliers in the dataset.

Figure 25:- Correlation Heat Map

Note that tree-based algorithms are not sensitive to outliers, models such as regression trees, random forests, and XGBoost can be applied to predict the actual productivity of garment workers with higher accuracy. These algorithms are effective in modeling nonlinear patterns between predictors and response variables.

Conclusion

The analysis of the "Productivity Prediction of Garment Employees" dataset reveals several key findings. Firstly, there are no significant differences in productivity across different days of the week. However, quarterly variations indicate higher productivity in Quarter 1 and Quarter 2. Incentives below 250 show a positive correlation with productivity, but beyond a certain threshold, additional incentives may not significantly impact productivity. Team 9 and Team 5 receive the most incentives but do not exhibit the highest productivity, while Teams 6, 7, and 8, receive the least incentives, and show lower productivity. The sewing department generally outperforms the finishing department, with an increase in the number of workers positively influencing productivity in sewing. Style changes negatively impact average productivity, except for Teams 7 and 11, which exhibit an opposing trend with overall lower productivity. Finishing records zero style changes, while sewing, with 0 to 2 style changes, maintains higher productivity. There is a weak correlation between productivity and work in progress (WIP), and Team 3, despite having the highest WIP, demonstrates higher productivity. Productivity and Standard Minute Value (SMV) show a weak negative correlation, revealing three distinct clusters in SMV. SMV and WIP exhibit a positive correlation. Idle time influences productivity, with occurrences on expected days, although the relationship is not always straightforward. We aim to utilize these results thoroughly in the advanced analysis.

Appendix

1. Link for the dataset: https://www.kaggle.com/datasets/ishadss/productivity-prediction-of-garment-employees
2. The Python code used in our project is conveniently accessible through our GitHub repository. You can find the Colab notebook containing all relevant code by following this GitHub link: https://github.com/ruwindarowel/Data-Analytics-and-Machine-Learning-to-Predict-Workers-Productivity/tree/main

References

1. Hkgc, Madbushanka & Appuhamy, Asanka & Ekanayake, Piyal. (2016). Identification of Factors Affecting the Productivity in Apparel Industry in Sri Lanka.
2. Cororaton, Caesar B. (1997) : Productivity Analysis in Garments and Textile Industries, PIDS Discussion Paper Series, No. 1997-09, Philippine Institute for Development Studies (PIDS), Makati City
3. https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/
4. https://seaborn.pydata.org/generated/seaborn.stripplot.html\
5. https://medium.com/@randayandika1/employee-productivity-in-garment-factory-2cdc98de39c0