REGRESSION MODEL TO PREDICT THE PRODUCTIVITY OF GARMENT EMPLOYEES

ST 3008 – GROUP PROJECT

Table of Contents

Table of Figures

Introduction

Most developing nations rely heavily on the garment sector for their income, and it also offers a significant number of job possibilities. Despite the fact that the sector gained popularity with the industrial revolution, there are still many manual operations that demand effort from humans. Tracking these employees' productivity levels and the variables influencing their output is crucial. Therefore, as part of this project, a descriptive study will be conducted to identify and comprehend the elements influencing the productivity of garment employees and a model fitting will be done to predict productivity performance.

The dataset selected for this is available as a public dataset on Kaggle. It includes important attributes of the garment manufacturing process and the productivity of the employees which had been collected manually and been validated by the industry experts.

Literature Review

Employee productivity is influenced by a wide range of factors that can vary based on the nature of the work, the industry, and the organizational context. These factors can be broadly categorized into individual, organizational, and environmental factors. Some of the key factors influencing employee productivity can be listed as follows:

Individual factors	**Organizational factors**	**Environmental factors**	**Economic factors**
Skill and competence	Leadership	Noise levels and distractions	Compensation
Motivation	Communication	Flexibility	Incentives
Health and well-being	Performance management	Work-life balance	
Work ethic	Workplace culture	Workload and complexity	
Time management	Training and development		

Even though those broad categories are the standard factors they are not quantifiable hence the organizations need to identify the factors within the company which comes under those categories which can be measured and then analyzed.

In the dataset considered for this project, those quantifiable variables are the day of the week which could possible affect the productivity, quarter, Associated department with the instance, number of workers, Number of changes in the style of a particular product, targeted productivity set by the department, allocated time for a task, amount of overtime by each team, incentives, amount of time when the production was interrupted due to several reasons and number of workers who were idle due to production interruption.

The same dataset using in this project has been used in several other researches to make predictions techniques like random forest and ensemble machine learning methods. However, as the scope of this project, only multiple linear regression model fitting using three methods namely forward selection, backward elimination and stepwise selection are carried out to obtain the best possible model to predict the productivity of the garment employees.

Descriptive Analysis

ATTRIBUTE INFORMATION : GWP.csv (garments_worker_productivity)

Predictor variables considered for the analysis

Qualitative/ Quantitative	Variable name	Description
Qualitative	day	Day of the Week
	quarter	A portion of the month. A month was divided into four quarters.
	department	Associated department with the instance
Quantitative	no_of_workers	Number of workers in each team.
	no_of_style_change	Number of changes in the style of a particular product.
	smv	Standard Minute Value, it is the allocated time for a task.
	wip	Work in progress. Includes the number of unfinished items for products.
	over_time	Represents the amount of overtime by each team in minutes.
	incentive	Represents the amount of financial incentive (in BDT/Bangladesh Currency) that enables or motivates a particular course of action.
	idle_time	The amount of time when the production was interrupted due to several reasons.
	idle_men	The number of workers who were idle due to production interruption.

Response variable/ Target variable of the analysis

actual_productivity : The actual % of productivity that was delivered by the workers. It ranges from 0–1.

DATA CLEANING

01. Checking the variables

```
> unique(CDGWP$quarter)
[1] "Quarter1" "Quarter2" "Quarter3" "Quarter4" "Quarter5"
> unique(CDGWP$department)
[1] "sweing"     "finishing " "finishing"
> unique(CDGWP$day)
[1] "Thursday"  "Saturday"  "Sunday"    "Monday"    "Tuesday"   "Wednesday"
```

Figure 1: R output - Categories present in the categorical variables

- Category 'finishing' in the Department variable is split into two due to a white space in the response hence those two categories are merged.

02. Missing values

```
> colSums(is.na(GWP))
                date                quarter             department
                   0                      0                      0
                 day                   team targeted_productivity
                   0                      0                      0
                 smv                    wip               over_time
                   0                    506                      0
           incentive             idle_time               idle_men
                   0                      0                      0
   no_of_style_change       no_of_workers   actual_productivity
                   0                      0                      0
```

Figure 2: R output - Missing values before cleaning

- It is observable that if the missing value for the wip variable is removed, the department variable's finishing category will likewise be removed. As a result, we can discern a connection between department and wip variables.

```
> table(GWP$department)

  sweing finishing
     691       506
> cleaned_data <- na.omit(GWP)
> table(cleaned_data$department)

  sweing finishing
     691         0
```

Figure 3: Association of omitting wip missing variables to the department variable

- Now we know all the missing values of 'wip' variable related to finishing department. We know that 'wip' variable stands for "work in progress".

- All items that go to the finishing department after sewing are finished items at the end of the day. So, we can assume that there are no unfinished items in the finishing department at the end of the day. i.e., there are no items in work in progress.

03. 'actual_productivity' variable values ranging over the interval 0-1

In our dataset, we have identified 37 instances out of 1197 observations where the actual productivity values are slightly greater than 1, but remarkably close to this upper limit (1). There can be some reason for this situation.

- These data points likely reflect the inherent precision limitations in the measurement process. In real-world scenarios, measurements are subject to various factors, leading to small variations even in well-controlled conditions. Values that are very close to 1 could be the result of such minor fluctuations.
- On the other hand, the actual productivity calculation may be more accurate (although we don't require that level of accuracy), and it results in exceeding the actual productivity value greater than 1.

Thus, we do not possess any strong logical reason to omit those data points from the data set. Hence those points will also be included in the dataset for the model fitting to preserve the data integrity, real world variation and accuracy of calculations and predictions.

DESCRIPTIVE ANALYSIS

01. Descriptive analysis for qualitative data

a. Distribution of categorical variables

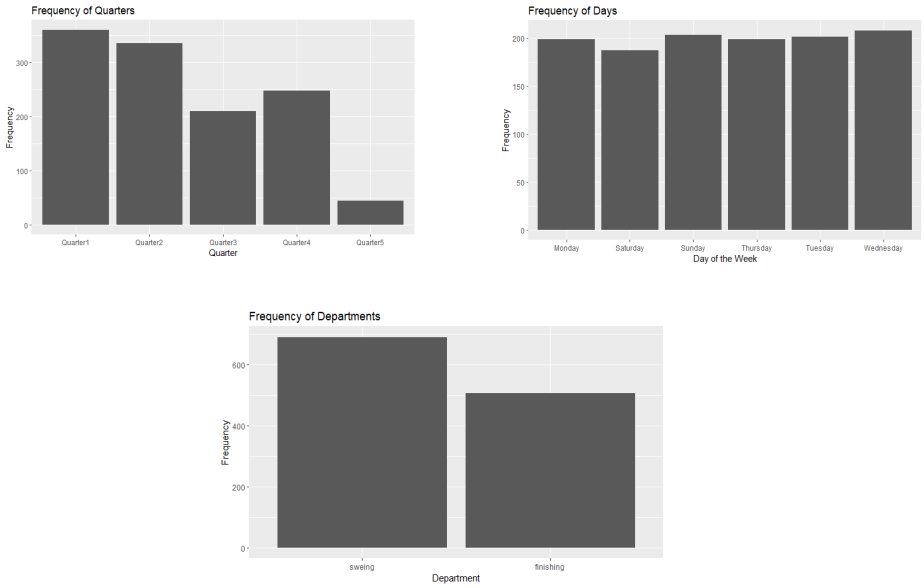


Figure 4: Distribution of categorical variables

b. Association between the response variable and the categorical variables

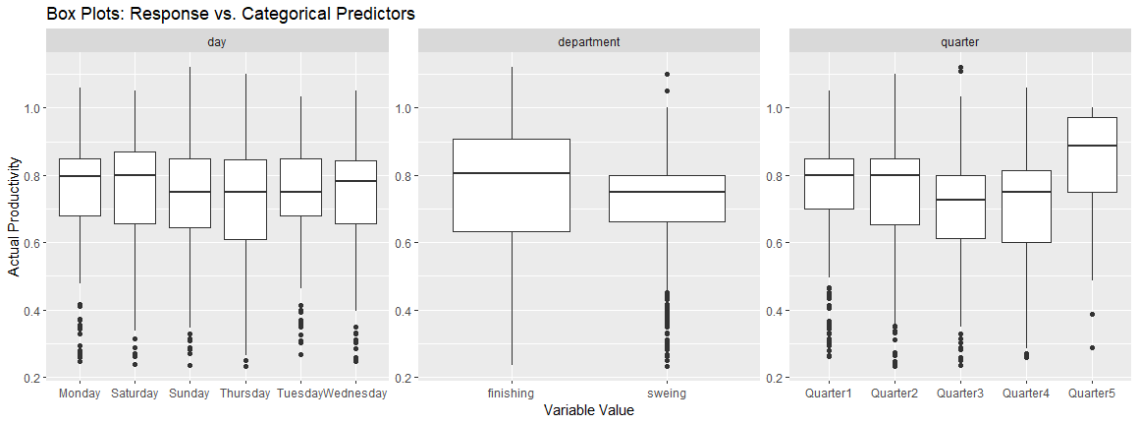
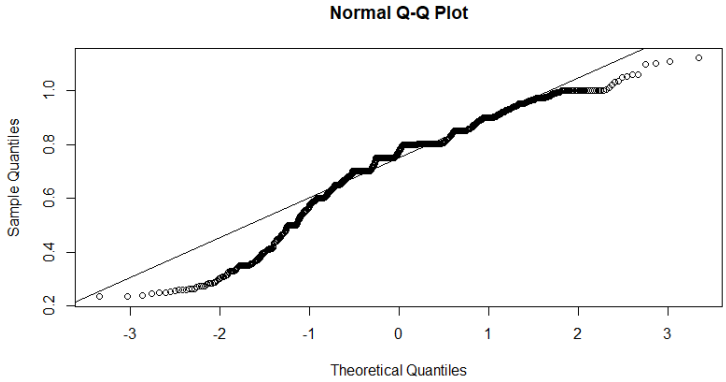


Figure 5: Box plots for response variable vs categorical predictor variables

All three boxplots show that 'actual_productivity' has slight skewness to the left in most of the categories. However, as there are no statistically significant reasons for the possible outliers present in the boxplots, they are not omitted from the dataset.

c. The mean of a Response variable grouped by a Categorical variable.



Q-Q plot drawn to check the normality shows a significant deviation from the straight-line at the two ends concluding that it does not follow a normal distribution.

Figure 6: Q-Q plot for normality assumption

```
> levene_test_result <- leveneTest(actual_productivity ~ department, data = GWP)
> print(levene_test_result)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   1  42.142 1.241e-10 ***
      1195
```

Figure 7: R output - Levene test for homoscedasticity

P value obtained for the test suggests that there's enough evidence conclude that the Homogeneity of Variance assumption is also violated under 5% significance level.

```
> kruskal_test_result <- kruskal.test(actual_productivity ~ day, data = GWP)
> print(kruskal_test_result)

        Kruskal-Wallis rank sum test

data:  actual_productivity by day
Kruskal-Wallis chi-squared = 4.3055, df = 5, p-value = 0.5063

>
> kruskal_test_result <- kruskal.test(actual_productivity ~ quarter, data = GWP)
> print(kruskal_test_result)

        Kruskal-Wallis rank sum test

data:  actual_productivity by quarter
Kruskal-Wallis chi-squared = 38.307, df = 4, p-value = 9.685e-08

>
> kruskal_test_result <- kruskal.test(actual_productivity ~ department, data = GWP)
> print(kruskal_test_result)

        Kruskal-Wallis rank sum test

data:  actual_productivity by department
Kruskal-Wallis chi-squared = 27.288, df = 1, p-value = 1.753e-07
```

Figure 8: R output - Kruskal Wallis test readings

Since assumptions for one way ANOVA are violated and hence the non-parametric Kruskal Wallis test is used to test the equality of means.

According to the Levene test output, there's enough evidence to say that the variables 'quarter' and 'department' have a significant effect on the response variable whereas there is no evidence to say that there's an effect of the variable 'day' at 5% significance level.

02. Descriptive analysis for qualitative data

a. Distribution of the quantitative variables

According to the above result we can see idle_time, idle_men, no_of_style_change have very

```
targeted_productivity      smv              wip            over_time      incentive
Min.   :0.0700        Min.   : 2.90   Min.   :    0.0   Min.   :    0   Min.   :   0.00
1st Qu.:0.7000        1st Qu.: 3.94   1st Qu.:    0.0   1st Qu.: 1440   1st Qu.:   0.00
Median :0.7500        Median :15.26   Median :  586.0   Median : 3960   Median :   0.00
Mean   :0.7296        Mean   :15.06   Mean   :  687.2   Mean   : 4567   Mean   :  38.21
3rd Qu.:0.8000        3rd Qu.:24.26   3rd Qu.: 1083.0   3rd Qu.: 6960   3rd Qu.:  50.00
Max.   :0.8000        Max.   :54.56   Max.   :23122.0   Max.   :25920   Max.   :3600.00
  idle_time          idle_men       no_of_style_change no_of_workers
Min.   :  0.0000   Min.   : 0.0000   Min.   :0.0000     Min.   : 2.00
1st Qu.:  0.0000   1st Qu.: 0.0000   1st Qu.:0.0000     1st Qu.: 9.00
Median :  0.0000   Median : 0.0000   Median :0.0000     Median :34.00
Mean   :  0.7302   Mean   : 0.3693   Mean   :0.1504     Mean   :34.61
3rd Qu.:  0.0000   3rd Qu.: 0.0000   3rd Qu.:0.0000     3rd Qu.:57.00
Max.   :300.0000   Max.   :45.0000   Max.   :2.0000     Max.   :89.00
```

Figure 9: Five number summary for quantitative variables

low mean values. Most of the observations are zero for these variables.

b. Correlation Analysis

In here we calculate correlation coefficients to understand the linear relationships between the response variable and numerical independent variables.

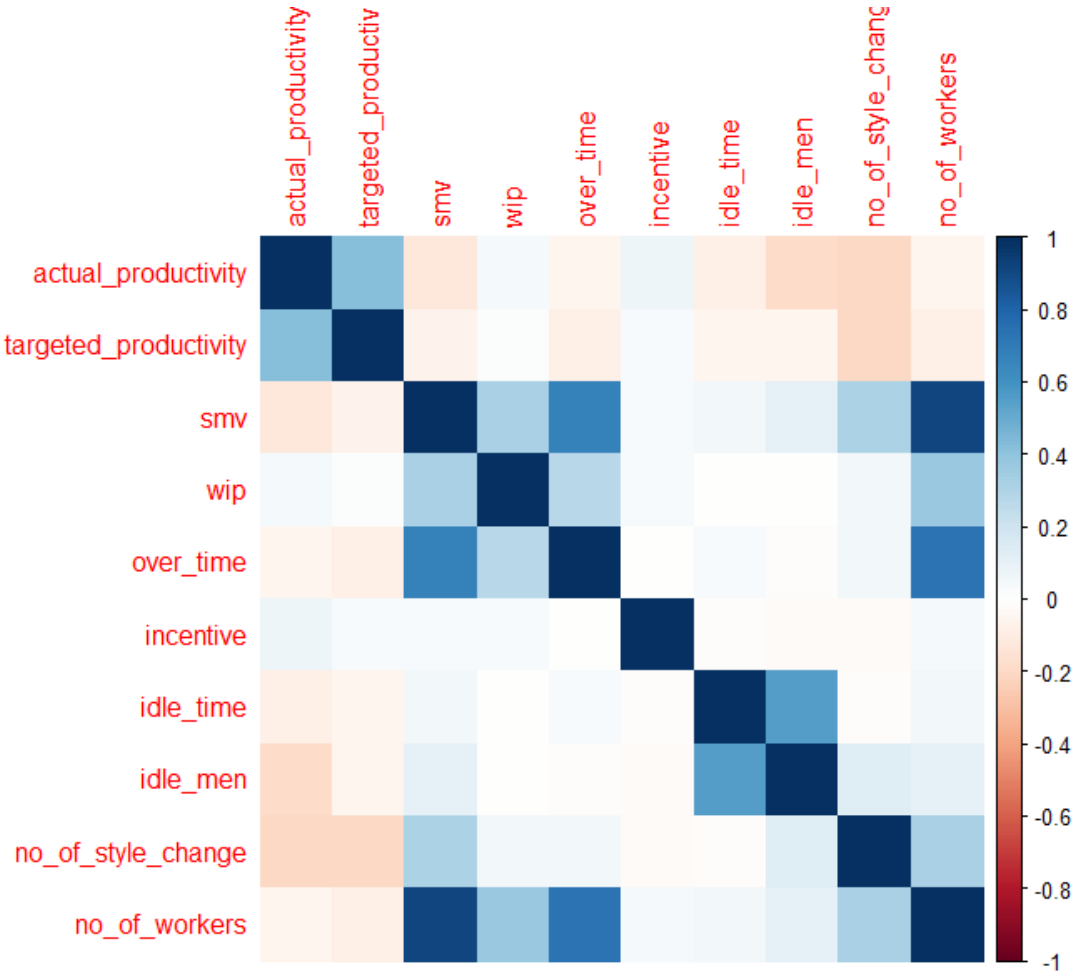


Figure 10: Correlation heatmap

- In the above heatmap it is quite evident productivity mainly depends on the target productivity as having a target will motivate and boost the employees.
- Most of the other variables have a weak correlation with the actual_productivity.
- We can observe that there is a strong correlation between no_of_workers and smv variables.
- We can observe that there is a strong correlation between over_time and no_of_workers variables.

Regression Analysis

01. Model fitting

a. Forward selection method – without considering interaction terms

- Null model

```
> lm_null=lm(actual_productivity~1)
> anova(lm_null)
Analysis of Variance Table

Response: actual_productivity
            Df Sum Sq  Mean Sq F value Pr(>F)
Residuals 1196 36.413 0.030446
> summary(lm_null)

Call:
lm(formula = actual_productivity ~ 1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.50139 -0.08478  0.03824  0.11516  0.38535

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.735091   0.005043   145.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1745 on 1196 degrees of freedom
```

Figure 11: R output - Summary of the forward selection null model

Predictor	d.f.	ESS	F- Statistic	P-Value
quarter	1192	35.565	7.1117	1.170E-05
smv	1195	35.871	18.082	2.281E-05
department	1195	36.134	9.2462	2.411E-03
targeted_productivity	**1195**	**29.9413**	**258.31**	**2.200E-16**
wip	1195	36.332	2.6896	1.013E-01
over-time	1195	36.306	3.5216	6.082E-02
incentive	1195	36.2	7.0416	8.070E-03
idle_time	1195	36.175	7.8629	5.128E-03
idle_men	1195	35.211	40.816	2.391E-10
no_of_style_change	1195	34.848	53.694	4.300E-13
no_of_workers	1195	36.291	4.0322	4.486E-02
day	1191	36.305	0.7121	6.144E-01

‘target_productivity’ has the lowest p value making it the most significant. Hence it is added to the model as the 1st predictor variable.

- Null model + target_productivity

Predictor	F-Value	P-Value
quarter	6.4905	3.630E-05
smv	12.698	3.806E-04
department	5.1279	0.02372
wip	2.6896	0.1013
over-time	0.4167	0.5187

incentive	5.7464	0.01667
idle_time	4.7793	0.029
idle_men	**38.011**	**9.61E-10**
no_of_style_change	21.948	3.12E-06
no_of_workers	0.7379	0.3905
day	0.7464	0.5888

- Null model + target_productivity + idle_men

Predictor	F-Value	P-Value
quarter	6.1238	7.062E-05
smv	8.9636	2.811E-03
department	2.9873	8.418E-02
wip	2.2116	1.372E-01
over-time	0.6341	4.260E-01
incentive	5.3553	2.083E-02
idle_time	2.1727	1.407E-01
no_of_style_change	**15.994**	**6.745E-05**
no_of_workers	0.0571	8.112E-01
day	0.796	5.526E-01

Since the lowest p value is for no_of_style_change, it is added to the model as the 3rd predictor variable.

Similarily, by considering the most significant p value, we can add variables to the model as follows,

- 4th predictor variable as quarter
- 5th predictor variable as incentive
- 6th predictor variable as smv
- 7th predictor variable as no_of_workers
- 8th predictor variable as department
- Null model + targeted_productivity + idle_men + no_of_style_change + quarter + incentive + smv + no_of_workers + department + 9th predictor variable

Predictor	**F-Value**	**P-Value**
wip	2.5502	0.1105497
over_time	2.5563	0.1101221
idle_time	0.8748	0.3498341
day	0.7828	0.562118

Since all the p values are greater than 0.05 which is the significance level, there's no significant 9th variable to be added. Also, as the interaction terms are not being considered no more variables are added to the model.

- Therefore, final model can be written as:

$$\boldsymbol{actual\ productivity} = \boldsymbol{target\ productivity} + \boldsymbol{idle\ men} + \boldsymbol{no\ of\ style\ change} + \boldsymbol{quarter} + \boldsymbol{incentive} + \boldsymbol{smv} + \boldsymbol{no\ of\ workers} + \boldsymbol{department}$$

```
> summary(lm_tmchqincsmvworkdept)

Call:
lm(formula = actual_productivity ~ targeted_productivity + idle_men +
    no_of_style_change + quarter + incentive + smv + no_of_workers +
    department)

Residuals:
     Min       1Q   Median       3Q      Max
-0.54901 -0.06252  0.02213  0.07756  0.50887

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            1.030e-01  4.689e-02   2.196 0.028287 *
targeted_productivity  7.116e-01  4.554e-02  15.627  < 2e-16 ***
idle_men              -7.633e-03  1.351e-03  -5.649 2.02e-08 ***
no_of_style_change    -3.661e-02  1.138e-02  -3.217 0.001330 **
quarterQuarter2        9.528e-05  1.158e-02   0.008 0.993437
quarterQuarter3       -2.052e-02  1.320e-02  -1.554 0.120506
quarterQuarter4       -1.456e-02  1.284e-02  -1.134 0.257027
quarterQuarter5        9.083e-02  2.402e-02   3.781 0.000164 ***
incentive              5.372e-05  2.738e-05   1.962 0.050041 .
smv                   -6.124e-03  9.773e-04  -6.266 5.17e-10 ***
no_of_workers          4.995e-03  6.849e-04   7.292 5.56e-13 ***
departmentfinishing    9.885e-02  2.582e-02   3.828 0.000136 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.15 on 1185 degrees of freedom
Multiple R-squared:  0.2674,    Adjusted R-squared:  0.2606
F-statistic: 39.33 on 11 and 1185 DF,  p-value: < 2.2e-16
```

Figure 12: R output - Summary of the forward selection final modal

$$\boldsymbol{actual\ productivity} = \mathbf{0.103} + \mathbf{0.7116}(\boldsymbol{target\ productivity}) + \mathbf{0.007633}(\boldsymbol{idle\ men}) - \mathbf{0.03661}(\boldsymbol{no\ of\ style\ change})\ \mathbf{9.528} \times \mathbf{10^{-5}}(\boldsymbol{quarter}: \boldsymbol{Quarter2}) - \mathbf{0.02052}(\boldsymbol{quarter}: \boldsymbol{Quarter3}) - \mathbf{0.01456}(\boldsymbol{quarter}: \boldsymbol{Quarter4}) + \mathbf{0.09083}(\boldsymbol{quarter}: \boldsymbol{Quarter5}) + \mathbf{5.372} \times \mathbf{10^{-5}}(\boldsymbol{incentive}) - \mathbf{0.006124}(\boldsymbol{smv}) + \mathbf{0.004995}(\boldsymbol{no\ of\ workers}) + \mathbf{0.09885}(\boldsymbol{department}: \boldsymbol{finishing})$$

Forward selection without considering any interaction term provides an adjusted R^2 value of 0.2606.

b. Backward elimination method – without considering interaction terms

Summary of the full model under the backward elimination below provides the p values for each variable considered in the model. From the p value approach, we can see that the largest p value is for Sunday category in the variable 'day'. Since all the categories under the 'day' variable has a value greater than 0.05, it is removed from the model.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.55817 -0.06104  0.02243  0.07930  0.50372

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             9.341e-02  4.808e-02   1.943 0.052281 .
quarterQuarter2         2.789e-03  1.167e-02   0.239 0.811213
quarterQuarter3        -1.533e-02  1.334e-02  -1.149 0.250782
quarterQuarter4        -1.021e-02  1.298e-02  -0.787 0.431549
quarterQuarter5         9.109e-02  2.502e-02   3.641 0.000284 ***
departmentfinishing     1.066e-01  2.606e-02   4.090 4.60e-05 ***
targeted_productivity   7.042e-01  4.585e-02  15.357  < 2e-16 ***
smv                    -5.984e-03  9.799e-04  -6.107 1.38e-09 ***
wip                     5.391e-06  3.196e-06   1.687 0.091928 .
over_time              -3.208e-06  2.082e-06  -1.541 0.123625
incentive               5.285e-05  2.783e-05   1.899 0.057832 .
idle_time               4.090e-04  4.183e-04   0.978 0.328280
idle_men               -8.750e-03  1.651e-03  -5.300 1.39e-07 ***
no_of_style_change     -4.069e-02  1.209e-02  -3.365 0.000789 ***
no_of_workers           5.340e-03  7.369e-04   7.247 7.71e-13 ***
daySaturday             1.402e-02  1.600e-02   0.876 0.381269
daySunday               8.749e-04  1.528e-02   0.057 0.954362
dayThursday            -2.046e-03  1.567e-02  -0.131 0.896095
dayTuesday              2.162e-02  1.529e-02   1.414 0.157758
dayWednesday            6.784e-03  1.518e-02   0.447 0.655015
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1499 on 1177 degrees of freedom
Multiple R-squared:  0.2736,    Adjusted R-squared:  0.2618
F-statistic: 23.33 on 19 and 1177 DF,  p-value: < 2.2e-16
```

Figure 14: Summary of the backward elimination full model

- Similar to the elimination of the ‘day’ variable, further analysis using p value approach with 5%significance level result in the elimination of the variables idle_time, wip, over_time and incentive.

```
> summary(lm_withoutinc)

Call:
lm(formula = actual_productivity ~ quarter + department + targeted_productivity +
    smv + idle_men + no_of_style_change + no_of_workers)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55231 -0.06168  0.02265  0.07865  0.50652

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             0.1015807  0.0469378   2.164 0.030652 *
quarterQuarter2         0.0025178  0.0115296   0.218 0.827176
quarterQuarter3        -0.0205543  0.0132204  -1.555 0.120275
quarterQuarter4        -0.0144016  0.0128587  -1.120 0.262943
quarterQuarter5         0.0912721  0.0240473   3.796 0.000155 ***
departmentfinishing     0.0988670  0.0258544   3.824 0.000138 ***
targeted_productivity   0.7142045  0.0455741  15.671  < 2e-16 ***
smv                    -0.0061842  0.0009780  -6.324 3.61e-10 ***
idle_men               -0.0076634  0.0013527  -5.665 1.84e-08 ***
no_of_style_change     -0.0373960  0.0113856  -3.284 0.001051 **
no_of_workers           0.0050489  0.0006852   7.369 3.22e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1502 on 1186 degrees of freedom
Multiple R-squared:  0.2651,    Adjusted R-squared:  0.2589
F-statistic: 42.78 on 10 and 1186 DF,  p-value: < 2.2e-16
```

Figure 13: R output - Summary of the backward elimination final model

- Since all the remaining variables have significant p value under 5% significance level, final model can be written as:

$$\begin{aligned}\boldsymbol{actual\ productivity} &= \mathbf{0.1016} + \mathbf{0.002518}\ (\boldsymbol{quarter}: \boldsymbol{Quarter2}) \\ &- \mathbf{0.02055}\ (\boldsymbol{quarter}: \boldsymbol{Quarter3}) - \mathbf{0.0144}\ (\boldsymbol{quarter}: \boldsymbol{Quarter4}) \\ &+ \mathbf{0.09127}\ (\boldsymbol{quarter}: \boldsymbol{Quarter5}) + \mathbf{0.09887}\ (\boldsymbol{department}: \boldsymbol{finishing}) \\ &+ \mathbf{0.7142}\ (\boldsymbol{target\ productivity}) - \mathbf{0.006184}\ (\boldsymbol{smv}) - \mathbf{0.00766}\ (\boldsymbol{idle\ men}) \\ &- \mathbf{0.037396}\ (\boldsymbol{no\ of\ style\ change}) + \mathbf{0.00509}\ (\boldsymbol{no\ of\ workers})\end{aligned}$$

Above model derived using the backward elimination method without considering any interaction terms yield an adjusted R^2 value of 0.2589**.**

Furthermore, when the model is fitted using the R^2 approach, the final model obtained provided an adjusted R^2 value of 0.2626

c. Stepwise method – without considering interaction terms

Stepwise selection method carried out for the model result in the following final model:

$$\begin{aligned}\boldsymbol{actual\ productivity} &= \mathbf{0.1024} + \mathbf{0.002474}\ (\boldsymbol{quarter}: \boldsymbol{Quarter2}) \\ &- \mathbf{0.01737}\ (\boldsymbol{quarter}: \boldsymbol{Quarter3}) - \mathbf{0.01066}\ (\boldsymbol{quarter}: \boldsymbol{Quarter4}) \\ &+ \mathbf{0.09024}\ (\boldsymbol{quarter}: \boldsymbol{Quarter5}) + \mathbf{0.1064}\ (\boldsymbol{department}: \boldsymbol{finishing}) \\ &+ \mathbf{0.7024}(\boldsymbol{target\ productivity}) - \mathbf{0.005964}\ (\boldsymbol{smv}) + \mathbf{4.994} \times \mathbf{10^{-6}}(\boldsymbol{wip}) \\ &- \mathbf{3.261} \times \mathbf{10^{-6}}(\boldsymbol{over\ time}) + \mathbf{4.956} \times \mathbf{10^{-5}}(\boldsymbol{incentive}) \\ &- \mathbf{0.007795}(\boldsymbol{idle\ men}) - \mathbf{0.04191}(\boldsymbol{no\ of\ style\ change}) \\ &+ \mathbf{0.005348}\ (\boldsymbol{no\ of\ workers})\end{aligned}$$

The model fitted using stepwise selection method yields an adjusted R^2 value of 0.2626.

It's clear that the adjusted R^2 values obtained by each method differs as they do not possess the same set of predictor variables in the final model. Thus, those R^2 values obtained can be compared as follows.

Goodness of fit testing for the fitted models – without considering interaction terms

	Forward Selection	*Backward Selection*		*Stepwise Selection*
		P- value approach	R^2 approach	
Adjusted R^2 value	0.2606	0.2589	0.2626	0.2626

All three methods provide models explain similar amount of variation in the response variable. However, as all the R^2 values are less than 0.7, none of them can be considered a good fit. A possible reason for this could be the absence of interaction terms in the model.

In order to detect the presence of multicollinearity, calculation of VIF values is carried out in Figure 15, and it shows that there is a presence of some multicollinearity among some of the predictor variables as there are some variables with VIF value greater than 5.

```
> vif_values <- vif(lm_full1)
> print(vif_values)
                          GVIF Df GVIF^(1/(2*Df))
day                   1.231360  5        1.021030
quarter               1.292959  4        1.032638
department            8.824723  1        2.970644
targeted_productivity 1.072268  1        1.035504
smv                   6.119847  1        2.473832
wip                   1.247161  1        1.116764
over_time             2.587107  1        1.608448
incentive             1.057637  1        1.028415
idle_time             1.503822  1        1.226304
idle_men              1.550433  1        1.245164
no_of_style_change    1.424214  1        1.193405
no_of_workers        14.238519  1        3.773396
```

Figure 15: R output - VIF values

To acknowledge the multicollinearity among the variables, the model is refitted considering the two-way interaction terms. Thus, the results obtained can be interpreted as follows.

Goodness of fit testing for the fitted models – considering two-way interaction terms

	Forward Selection	*Backward Selection*	*Stepwise Selection*
Adjusted R^2 value	0.4437 = 44.37%	0.4631=46.31%	0.4657=46.57%

From the above table, we can see that the stepwise selection carried out considering the two-way interactions yield in the highest adjusted R^2 value thus, it is the most fitting among the models we have fitted.

Final model obtained from the stepwise selection after considering the two-way interactions possess the composition of variables as follows:

$$
\begin{aligned}
&actual\ productivity \\
&\quad = quarter + department + targeted\ productivity + smv + wip + over\ time \\
&\quad + incentive + idle\ time + idle\ men + no\ of\ style\ change + no\ of\ workers \\
&\quad + day + quarter : department + department : targeted\ productivity \\
&\quad + department : smv + department : over\ time + department : incentive \\
&\quad + department : no\ of\ workers + targeted\ productivity : day + smv : day \\
&\quad + wip : day + targetedproductivity : wip \\
&\quad + targeted\ productivity : over\ time + targeted\ productivity : idle\ men \\
&\quad + targeted\ productivity : no\ of\ style\ change \\
&\quad + targeted\ productivity : no\ of\ workers + smv : over\ time \\
&\quad + over\ time : incentive + idle\ time : no\ of\ workers + incentive : idle\ men
\end{aligned}
$$

02. Checking the assumptions of the MLR model

a. Linearity between the response variable and predictors

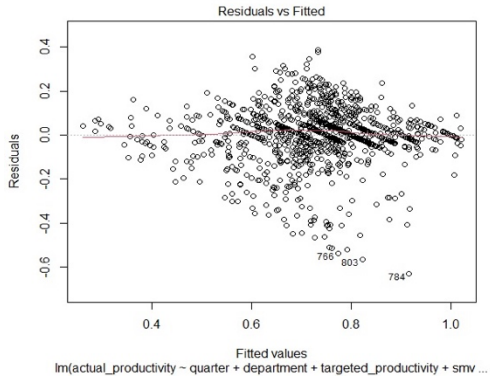


Figure 16: Scatterplot of residuals vs fitted - stepwise model with 2 way interactions

The scatterplot of residuals vs fitted values drawn for the final model selected shows a pattern around the horizontal line concluding that the linearity assumption is not satisfied.

b. Normally distributed residuals with mean 0

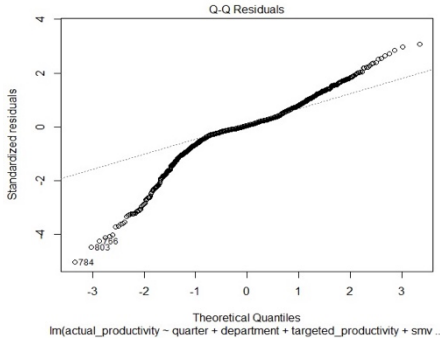


Figure 17: Plot between theoretical quantiles and standardized residuals to check the normality assumption

The Q-Q plot has significant deviations from the straight line drawn which proof that the normality assumption is being violated.

c. Independence of the residual error terms

Since the scatter plot in Figure 16 shows a pattern, we can conclude that the residual error terms are not independent.

d. Constant variance assumption

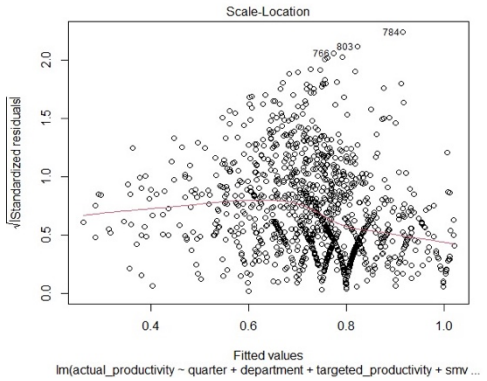


Figure 18: Plot between fitted values and standardized residuals

This plot shows that the residuals are not spread equally along the ranges of predictors as the range above the horizontal line is high. This implies that the residuals may not have constant variance

e. Absence of multicollinearity

VIF values calculated in Figure 15 concludes that this assumption is violated as there are some variables with a VIF value higher than 5.

Discussion and Conclusion

Descriptive analysis carried out concludes the following results about the productivity dataset:

- Qualitative variable 'quarter' has significantly fewer count for Quarter5 compared to other categories but no logical reason for that can be derived from the information available.
- Frequency of observations for sewing department is higher than the finishing department.
- Most of the quantitative variables present in this data set are skewed towards right. Reason for this is because several variables have significant number of observations scattered around 0 showing right skewness whereas the remaining variables such as 'target_productivity' possess slightly right skewed distribution.
- The correlation heat map drawn to identify the strengths of the linear relationship among variables shows that there are some associations between the predictor variables.

Thus, it is clear from the descriptive analysis that the categorical variables have an effect on the response variable but the strength of it or the effect of each category on the response cannot be predicted without fitting a regression model also there are some significant multicollinearities present within the predictor variables in the dataset.

Under the regression analysis carried out considering the aforesaid variables, several conclusions can be drawn as follows:

- Model fitting carried out using all three methods forward selection, backward elimination, and stepwise selection without considering the interaction terms result in a lower adjusted R^2 value. Which allows to explain about 26% of the variation in the response variable after accounting the number of predictor variables added to the model.
- There's no best method out of forward, backward, and stepwise when considering this fit as all of them gives more or less the same precision.
- Two-way interaction terms affect the model significantly as the refitted model considering the two- way interaction terms possess a significantly higher adjusted R^2 value compared to the previous model fit which allows to explain about 47% of the variation of the model, after accounting for the predictors added to the model.
- Refitted model by using the stepwise selection has the best adjusted R^2 value among the three methods used hence can be selected as the final model to predict the from the analysis conducted.

- Since the two-way interaction terms tend to increase the adjusted R^2 value of the model it is logical to say that the adjusted R^2 can be further increased by refitting the model with three-way interactions introduced to the model.

However, due to the complexity of adding three-way interaction terms to the model, it is not included in the scope of this project hence the final model to predict the productivity can be presented as below.

$$
\begin{aligned}
&\boldsymbol{actual\ productivity} \\
&\quad = \boldsymbol{quarter + department + targeted\ productivity + smv + wip} \\
&\quad + \boldsymbol{over\ time + incentive + idle\ time + idle\ men + no\ of\ style\ change} \\
&\quad + \boldsymbol{no\ of\ workers + day + quarter : department} \\
&\quad + \boldsymbol{department : targeted\ productivity + department : smv} \\
&\quad + \boldsymbol{department : over\ time + department : incentive} \\
&\quad + \boldsymbol{department : no\ of\ workers + targeted\ productivity : day} \\
&\quad + \boldsymbol{smv : day + wip : day + targetedproductivity : wip} \\
&\quad + \boldsymbol{targeted\ productivity : over\ time + targeted\ productivity : idle\ men} \\
&\quad + \boldsymbol{targeted\ productivity : no\ of\ style\ change} \\
&\quad + \boldsymbol{targeted\ productivity : no\ of\ workers + smv : over\ time} \\
&\quad + \boldsymbol{over\ time : incentive + idle\ time : no\ of\ workers} \\
&\quad + \boldsymbol{incentive : idle\ men}
\end{aligned}
$$

- Assumptions checking for the MLR model selected as the final model gives proof that the model does not satisfy any of the assumptions made, hence the statistically logical solution is using transformation techniques and/or use other regression modelling techniques to get a better fitting model which satisfy the modeling assumptions.

Dataset information

- Kaggle link to the data set: https://www.kaggle.com/datasets/ishadss/productivity-prediction-of-garment-employees
- All the R codes and the data set used to carry out the analysis are uploaded to the LMS with the report.

Contribution to the project

INDEX NO.	NAME	TASKS COMPLETED
15652	Inuka Chandipa Sathsara	Descriptive analysis
15577	Darshika Wijesena	Introduction, Discussion and Conclusion, Structuring the report
15666	Dulakshi Wilegoda	Regression analysis
15598	Lasani Balasuriya	Literature Review
15355	Sandali Gunathilaka	Regression analysis