Data_EDA

Team OK: Kara Wei, Ouyang Du

2023-10-17

```
data <- read.csv("/Users/karawei/Desktop/GRAD 3rd/ESE 527/garments_worker_productivity.csv")
#data <- read.csv("C:/Users/ThinkPad/Documents/GitHub/OK/worker_productivity.csv")


data_org <- read.csv("/Users/karawei/Desktop/GRAD 3rd/ESE 527/garments_worker_productivity.csv")
#data_org <- read.csv("C:/Users/ThinkPad/Documents/GitHub/OK/worker_productivity.csv")

data$department = replace(data$department, data$department=="sweing", "sewing")
data_org$department = replace(data_org$department, data_org$department=="finishing ", "finishing")
data_org$department = replace(data_org$department, data_org$department=="sweing", "sewing")
for (j in 1:length(data$department)){
  if(data$department[j]=="sewing"){data$department[j]<-1}
  else if(data$department[j]=="finishing "){data$department[j]<-2}
  else if(data$department[j]=="finishing"){data$department[j]<-2}
}

# change day of the week to number  Monday=1, Sunday=7
for (i in 1:length(data$day)){
  if(data$day[i]=="Monday"){data$day[i]<-as.integer(1)}
  else if(data$day[i]=="Tuesday"){data$day[i]<-as.integer(2)}
  else if(data$day[i]=="Wednesday"){data$day[i]<-as.integer(3)}
  else if(data$day[i]=="Thursday"){data$day[i]<-as.integer(4)}
  else if(data$day[i]=="Friday"){data$day[i]<-as.integer(5)}
  else if(data$day[i]=="Saturday"){data$day[i]<-as.integer(6)}
  else if(data$day[i]=="Sunday"){data$day[i]<-as.integer(7)}
}

# changing the date to date format
data$date <- as.Date(data$date,format="%m/%d/%Y" )

#Changing Quarter to numbers
for (k in 1:length(data$quarter)){
  if(data$quarter[k]=="Quarter1"){data$quarter[k]<-as.integer(1)}
  else if(data$quarter[k]=="Quarter2"){data$quarter[k]<-as.integer(2)}
  else if(data$quarter[k]=="Quarter3"){data$quarter[k]<-as.integer(3)}
  else if(data$quarter[k]=="Quarter4"){data$quarter[k]<-as.integer(4)}
  else if(data$quarter[k]=="Quarter5"){data$quarter[k]<-as.integer(5)}
}

sewing<-subset(data,data$department==1)
finishing<-subset(data,data$department==2)
sewingTeam01<-subset(sewing,sewing$team==1)
```

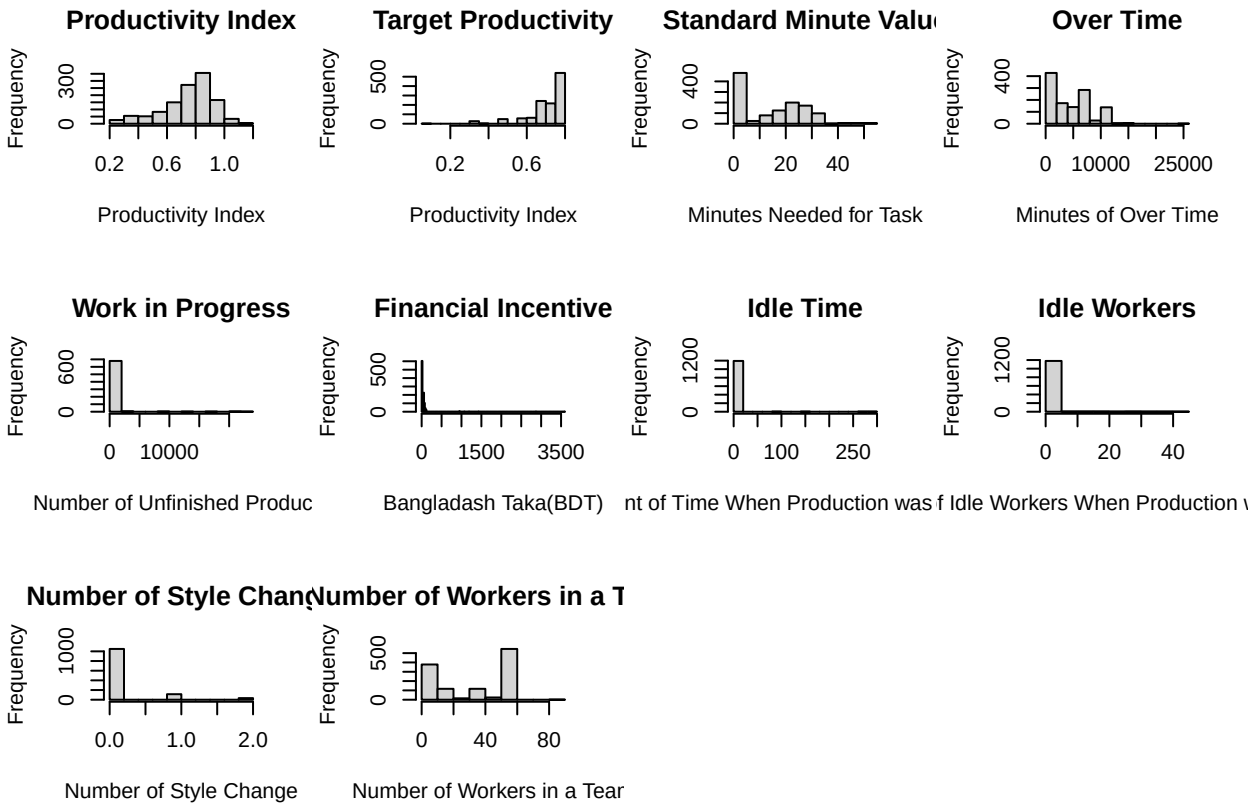
```
sewingTeam02<-subset(sewing,sewing$team==2)
sewingTeam03<-subset(sewing,sewing$team==3)
sewingTeam04<-subset(sewing,sewing$team==4)
sewingTeam05<-subset(sewing,sewing$team==5)
sewingTeam06<-subset(sewing,sewing$team==6)
sewingTeam07<-subset(sewing,sewing$team==7)
sewingTeam08<-subset(sewing,sewing$team==8)
sewingTeam09<-subset(sewing,sewing$team==9)
sewingTeam10<-subset(sewing,sewing$team==10)
sewingTeam11<-subset(sewing,sewing$team==11)
sewingTeam12<-subset(sewing,sewing$team==12)

finishingTeam01<-subset(finishing,finishing$team==1)
finishingTeam02<-subset(finishing,finishing$team==2)
finishingTeam03<-subset(finishing,finishing$team==3)
finishingTeam04<-subset(finishing,finishing$team==4)
finishingTeam05<-subset(finishing,finishing$team==5)
finishingTeam06<-subset(finishing,finishing$team==6)
finishingTeam07<-subset(finishing,finishing$team==7)
finishingTeam08<-subset(finishing,finishing$team==8)
finishingTeam09<-subset(finishing,finishing$team==9)
finishingTeam10<-subset(finishing,finishing$team==10)
finishingTeam11<-subset(finishing,finishing$team==11)
finishingTeam12<-subset(finishing,finishing$team==12)
```

We first convert all the categorical variables into dummy variables. Bause the dataset only has wip values for the sewing department, we decide to split the dataset into two subcategories - finishing department and sewing department.

Histograms of all numerical variables:

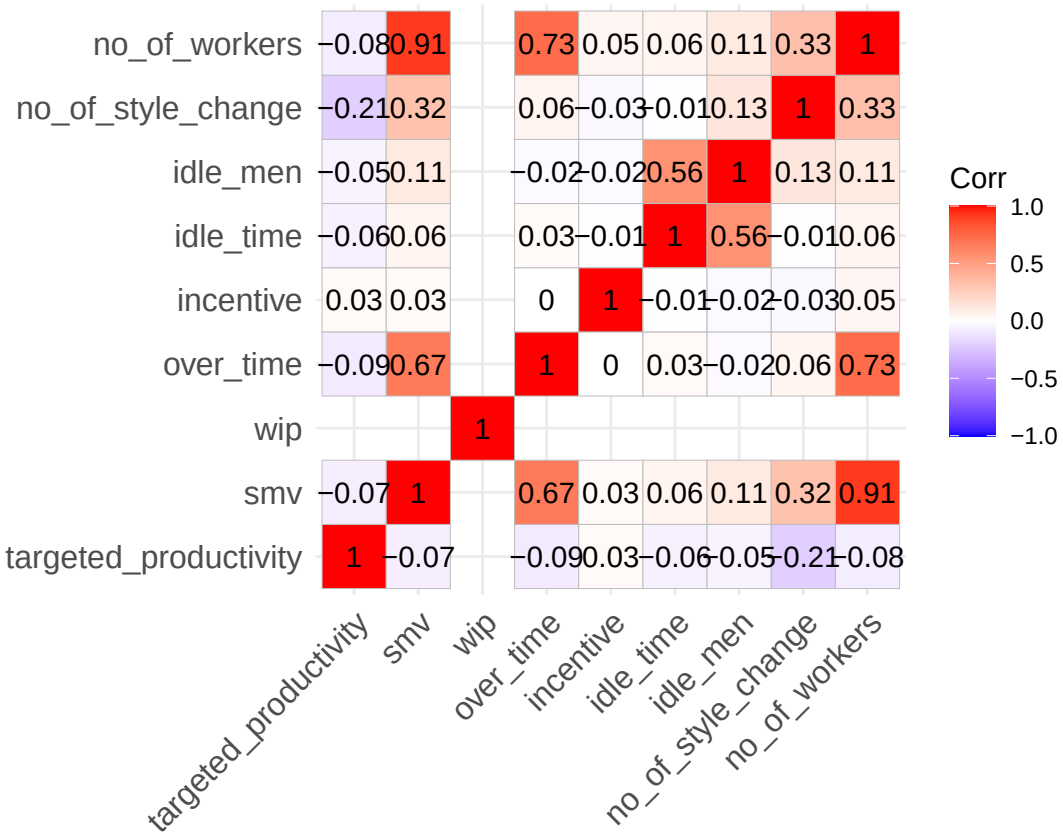
```
par(mfrow = c(3, 4))
hist(data[["actual_productivity"]],main="Productivity Index",xlab="Productivity Index")
hist(data[["targeted_productivity"]],main="Target Productivity",xlab="Productivity Index")
hist(data[["smv"]],main="Standard Minute Value",xlab="Minutes Needed for Task")
hist(data[["over_time"]],main="Over Time",xlab="Minutes of Over Time")
hist(data[["wip"]],main="Work in Progress",xlab="Number of Unfinished Products")
hist(data[["incentive"]],main="Financial Incentive",xlab="Bangladash Taka(BDT)",breaks=200)
hist(data[["idle_time"]],main="Idle Time",xlab="Amount of Time When Production was Interrupted")
hist(data[["idle_men"]],main="Idle Workers",xlab="Number of Idle Workers When Production was Interrupte
hist(data[["no_of_style_change"]],main="Number of Style Change",xlab="Number of Style Change")
hist(data[["no_of_workers"]],main="Number of Workers in a Team",xlab="Number of Workers in a Team")
```



We then draw he distributions of all numerical variables. From the numerical variables, we notice that there are tails for several variables:target productivity, over time, financial incentives, etc.

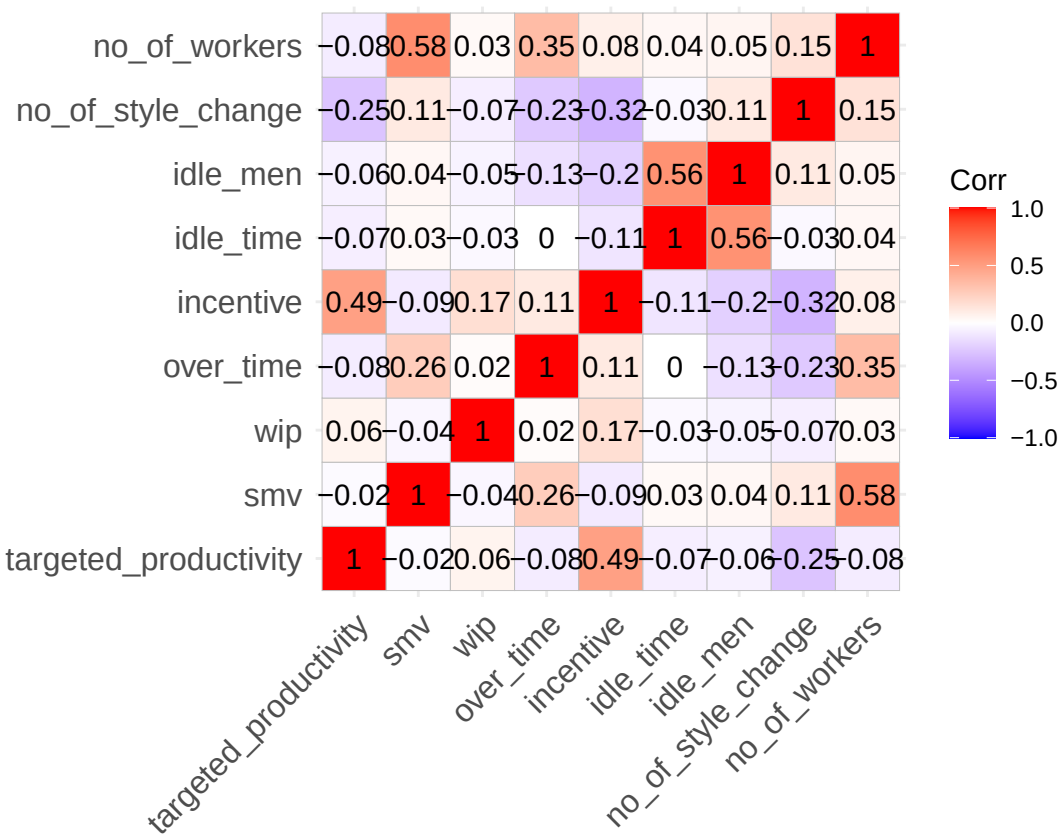
We create correlation map for numerical variables above:

```
correlation_matrix <- cor(data[6:14])
ggcorrplot(correlation_matrix,lab = TRUE)
```



For sewing department:

```
sewing_correlation_matrix <- cor(sewing[6:14])
ggcorrplot(sewing_correlation_matrix,lab = TRUE)
```

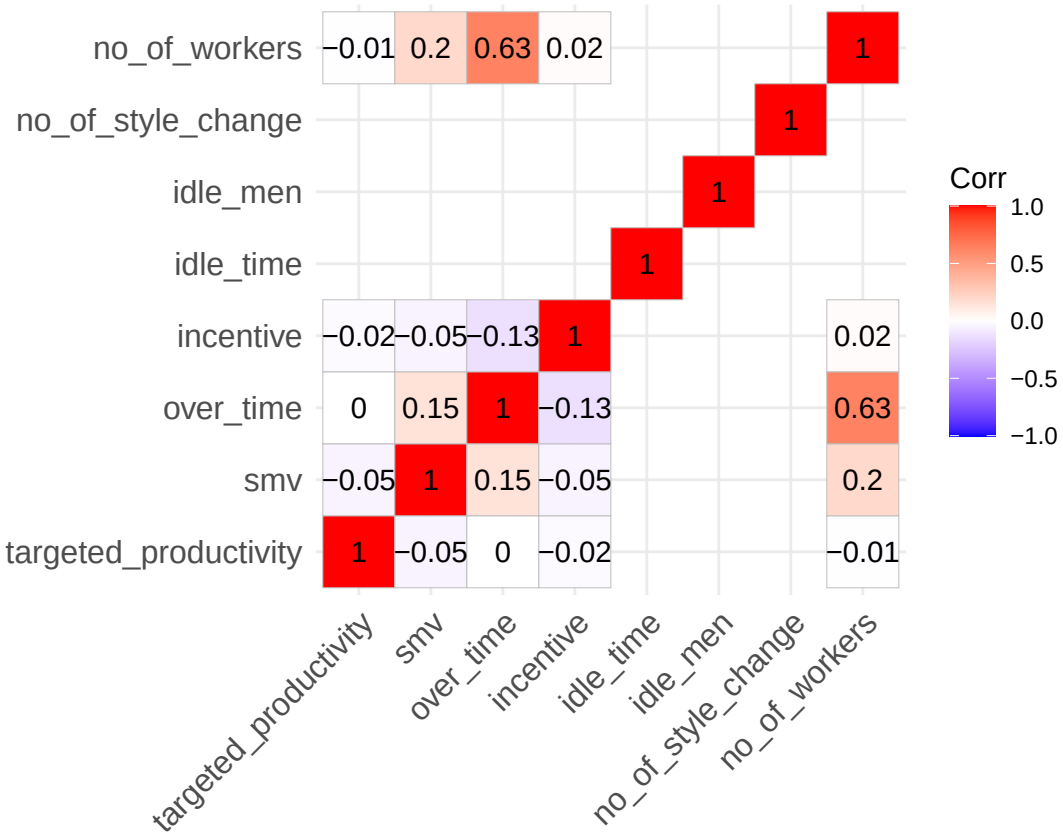


For finishing department:

```
finishing_correlation_matrix <- cor(finishing[,c(6:7,9:14)])
```

```
## Warning in cor(finishing[, c(6:7, 9:14)]): the standard deviation is zero
```

```
ggcorrplot(finishing_correlation_matrix,lab = TRUE)
```




```
head(finishing[,c(6:7,9:14)])
```

```
##    targeted_productivity  smv over_time incentive idle_time idle_men
## 2                   0.75 3.94       960         0         0        0
## 7                   0.75 3.94       960         0         0        0
## 14                  0.65 3.94       960         0         0        0
## 15                  0.75 2.90       960         0         0        0
## 16                  0.75 3.94      2160         0         0        0
## 17                  0.80 2.90       960         0         0        0
##    no_of_style_change no_of_workers
## 2                   0             8
## 7                   0             8
## 14                  0             8
## 15                  0             8
## 16                  0            18
## 17                  0             8
```

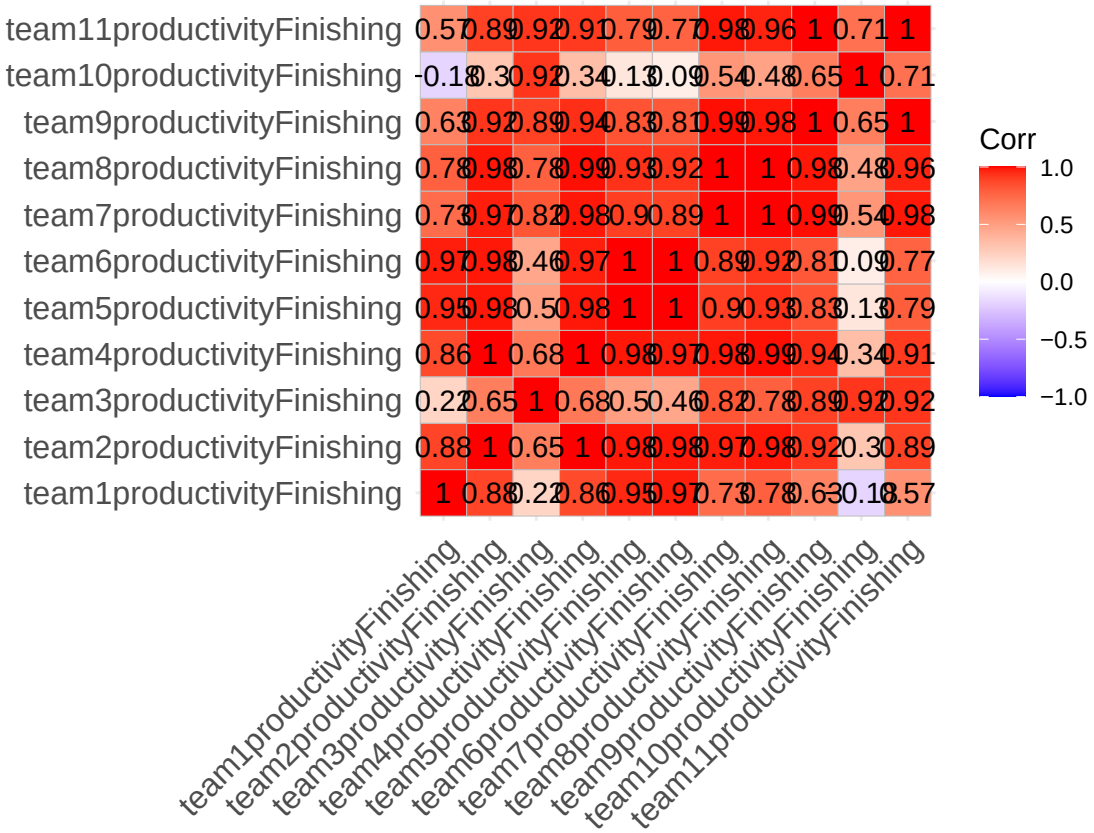
Notice that lots of the values are missing here. That's because they have mostly 0 values in the sub-datset per finishing department.

Take covariance between two every two variables in team finishing and team sewing:

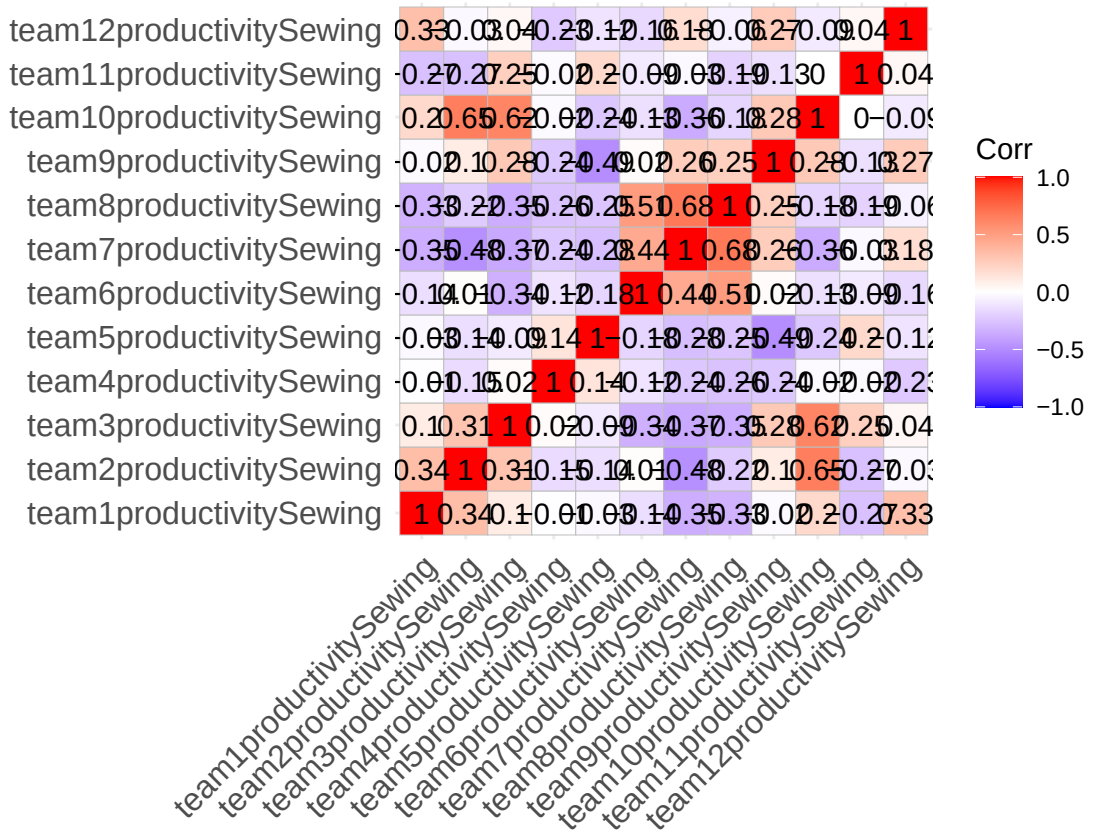
```
finishing_file_path <- "/Users/karawei/Downloads/combinedProductivityDataforFinishingTeams.csv"
sewing_file_path <- "/Users/karawei/Documents/GitHub/OK/combinedDataforProductivity.csv"

# Read the CSV file into a data frame
finishing_data <- read.csv(finishing_file_path)
sewing_data <- read.csv(sewing_file_path)

finishing_correlation_matrix <- cor(finishing_data[2:12], use = "complete.obs")
sewing_correlation_matrix <- cor(sewing_data[2:13], use = "complete.obs")
ggcorrplot(finishing_correlation_matrix,lab = TRUE)
```




```
ggcorrplot(sewing_correlation_matrix,lab = TRUE)
```




```
sewing_remove = na.omit((sewing_data[2:13]))
nrow(sewing_remove)
```

```
## [1] 44
```

```
finishing_remove = na.omit((finishing_data))
nrow(finishing_remove)
```

```
## [1] 1
```

```
head(finishing_remove)
```

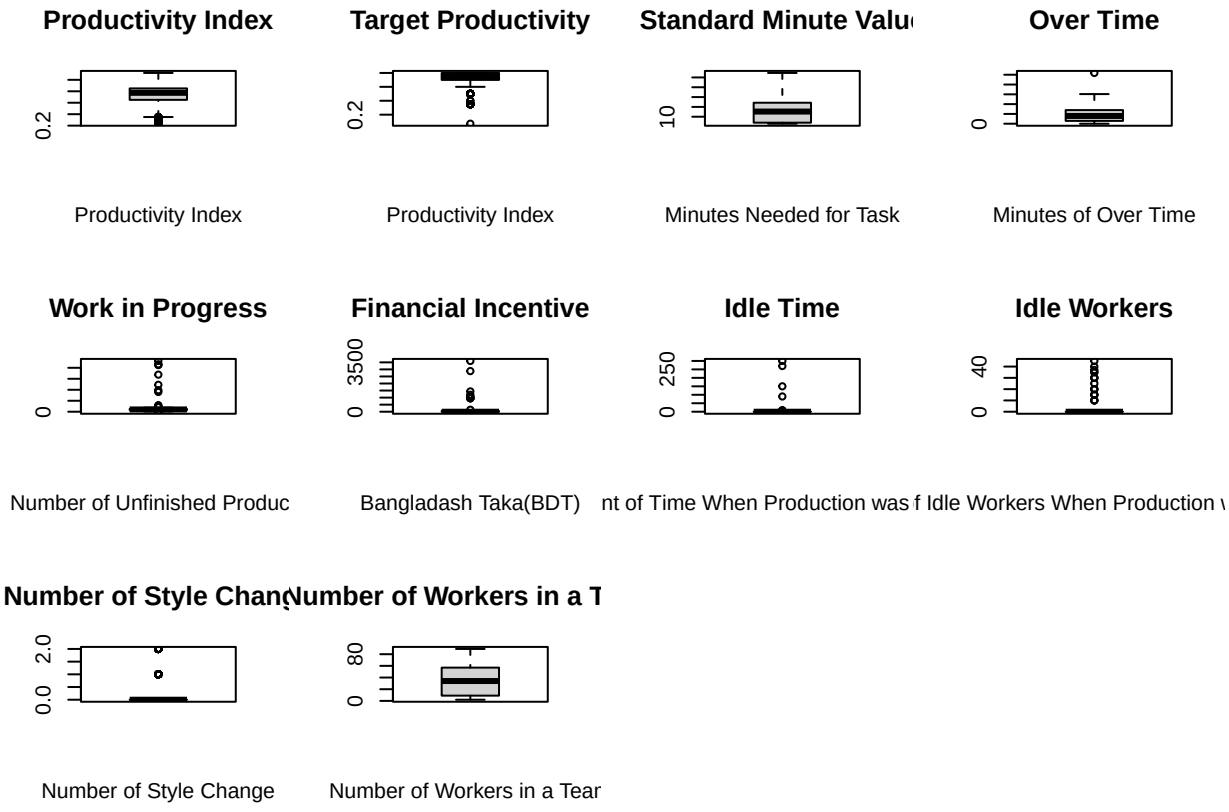
```
##          date team1productivityFinishing team2productivityFinishing
## 26 01/31/2015                  0.9718667                  0.9718667
##    team3productivityFinishing team4productivityFinishing
## 26                  0.9718667                  0.9718667
##    team5productivityFinishing team6productivityFinishing
## 26                  0.9718667                  0.9718667
##    team7productivityFinishing team8productivityFinishing
## 26                  0.9718667                  0.9718667
##    team9productivityFinishing team10productivityFinishing
## 26                  0.9718667                   0.9718667
##    team11productivityFinishing team12productivityFinishing
## 26                   0.9718667                   0.9718667
```

CONCERNS: By using the argument above. I remove all missing values and calculated the correlation matrix for producitivity between teams in both departments. However, I am worried that:

1) After we remove all missing values, the correlation might not be too reliable
2) I can;t run the solution for team12ProductivityFishing. Because there's too many missing values.
3) I am not sure if calculating the productivity would mean anything statistically. Because we have not yet established relationship between productivity and other variables. And teams collaborate during the manufacturing process instead of at the final productivity. -> sol: But it shows productivity between teams might not be independent.

We draw box plots for all variables in original dataset:

```
par(mfrow = c(3, 4))
boxplot(data[["actual_productivity"]],main="Productivity Index",xlab="Productivity Index")
boxplot(data[["targeted_productivity"]],main="Target Productivity",xlab="Productivity Index")
boxplot(data[["smv"]],main="Standard Minute Value",xlab="Minutes Needed for Task")
boxplot(data[["over_time"]],main="Over Time",xlab="Minutes of Over Time")
boxplot(data[["wip"]],main="Work in Progress",xlab="Number of Unfinished Products")
boxplot(data[["incentive"]],main="Financial Incentive",xlab="Bangladash Taka(BDT)",breaks=200)
boxplot(data[["idle_time"]],main="Idle Time",xlab="Amount of Time When Production was Interrupted")
boxplot(data[["idle_men"]],main="Idle Workers",xlab="Number of Idle Workers When Production was Interru
boxplot(data[["no_of_style_change"]],main="Number of Style Change",xlab="Number of Style Change")
boxplot(data[["no_of_workers"]],main="Number of Workers in a Team",xlab="Number of Workers in a Team")
```



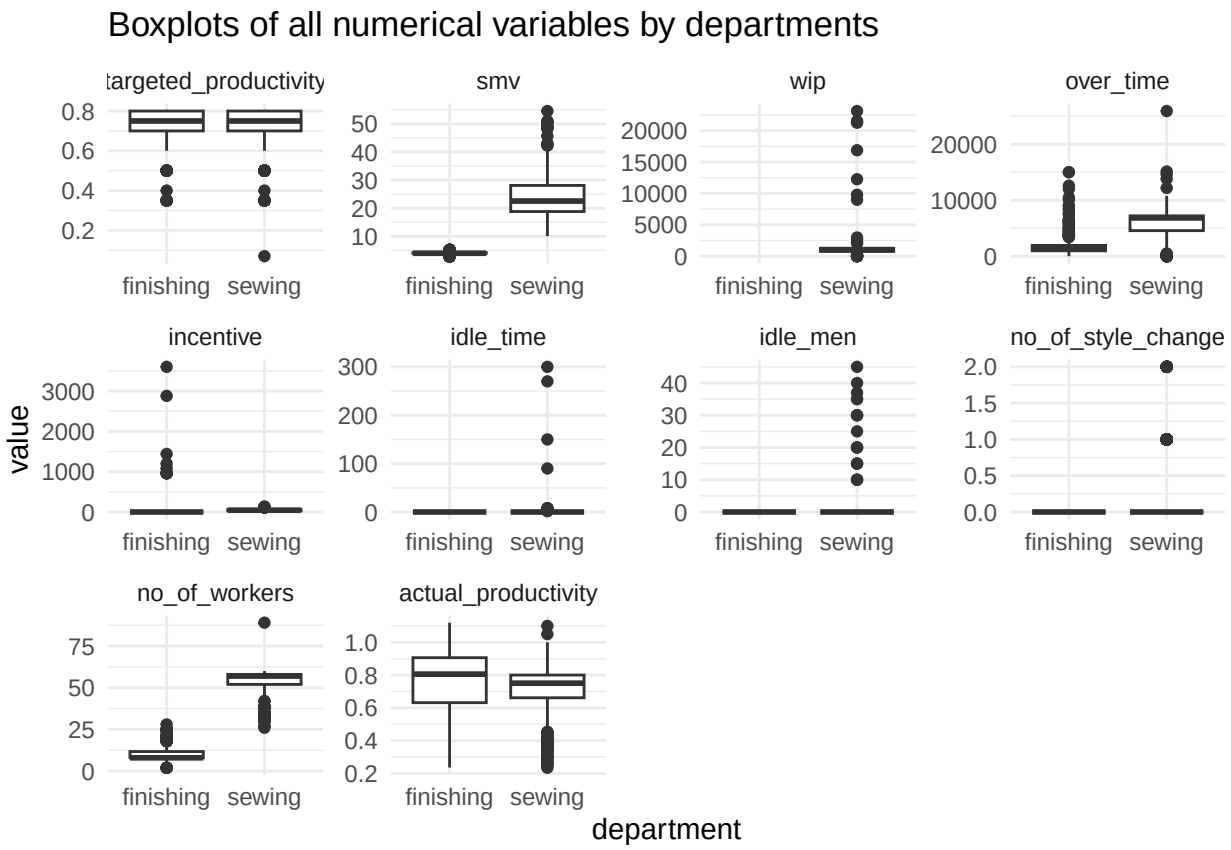
From the box plots, we see that most of the variables have outliers.

```
# ggplot(gather(data[,-c(1:5)]), aes(key,value)) +
# geom_boxplot() +
# facet_wrap(~key, scales = 'free') +
# labs(title = "Boxplots of all numerical variables")+
#   theme_minimal()
```

```
data_no_team = data_org[,c(1:4,6:15)]
ggplot(melt(data_no_team),aes(x=department,y=value)) +
facet_wrap(~variable, scales="free") +
geom_boxplot()+
  labs(title = "Boxplots of all numerical variables by departments")+
  theme_minimal()
```

```
## Using date, quarter, department, day as id variables
```

```
## Warning: Removed 506 rows containing non-finite values (`stat_boxplot()`).
```



If we zoom in the box plots per department, sewing department has higher mean for must of the numerical values. For most of the numerical variables - targeted_productivity, smv, wip, idle_time, idel_men, and no_of_style_change, no_of_workers, and actual_productivity - sewing department have higher outliers.

```
# ggplot(melt(data_org),aes(x=department)) +
# facet_wrap(~variable, scales="free") +
# geom_bar()+
#   labs(title = "Bar charts of all numerical variables by departments")+
#   theme_minimal()
```

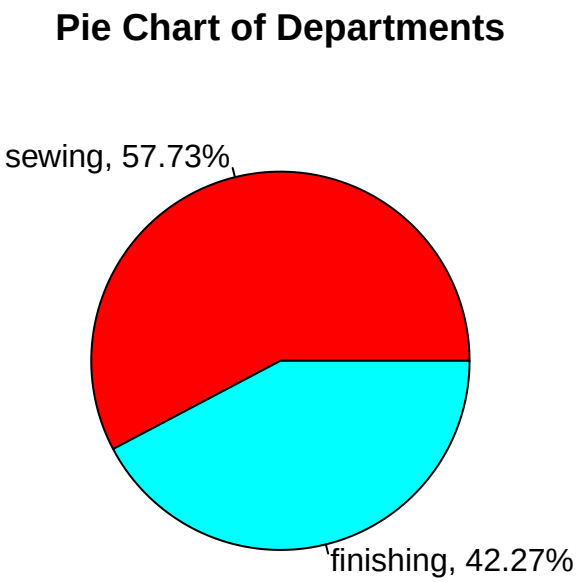
Create pie chart for department to see the distribution:

```
counts_of_departments = c(length(data_org$department[data_org$department=="sewing"]), length(data_org$d
labels_of_departments = c("sewing", "finishing")
departments_pie_labels <- paste0(labels_of_departments, ", ", round(100 * counts_of_departments/sum(cou

# counts_of_quarters = c(length(data_prod$quarter[data_prod$quarter=="Quarter1"]), length(data_prod$qua
# labels_of_quarters = c("Quarter1", "Quarter2","Quarter3","Quarter4","Quarter5")
# quarters_pie_labels <- paste0(labels_of_quarters, ", ", round(100 * counts_of_quarters/sum(counts_of_

# counts_of_days = c(length(data_prod$day[data_prod$day=="Monday"]),
# length(data_prod$day[data_prod$day=="Tuesday"]),
# length(data_prod$day[data_prod$day=="Wednesday"]),
# length(data_prod$day[data_prod$day=="Thursday"]),
# length(data_prod$day[data_prod$day=="Friday"]),
# length(data_prod$day[data_prod$day=="Saturday"]),
# length(data_prod$day[data_prod$day=="Sunday"]))
# labels_of_days = c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday")
# days_pie_labels <- paste0(labels_of_days, ", ", round(100 * counts_of_days/sum(counts_of_days), 2), "
```

```
pie(counts_of_departments, labels = departments_pie_labels,main = "Pie Chart of Departments",col = rainb
```

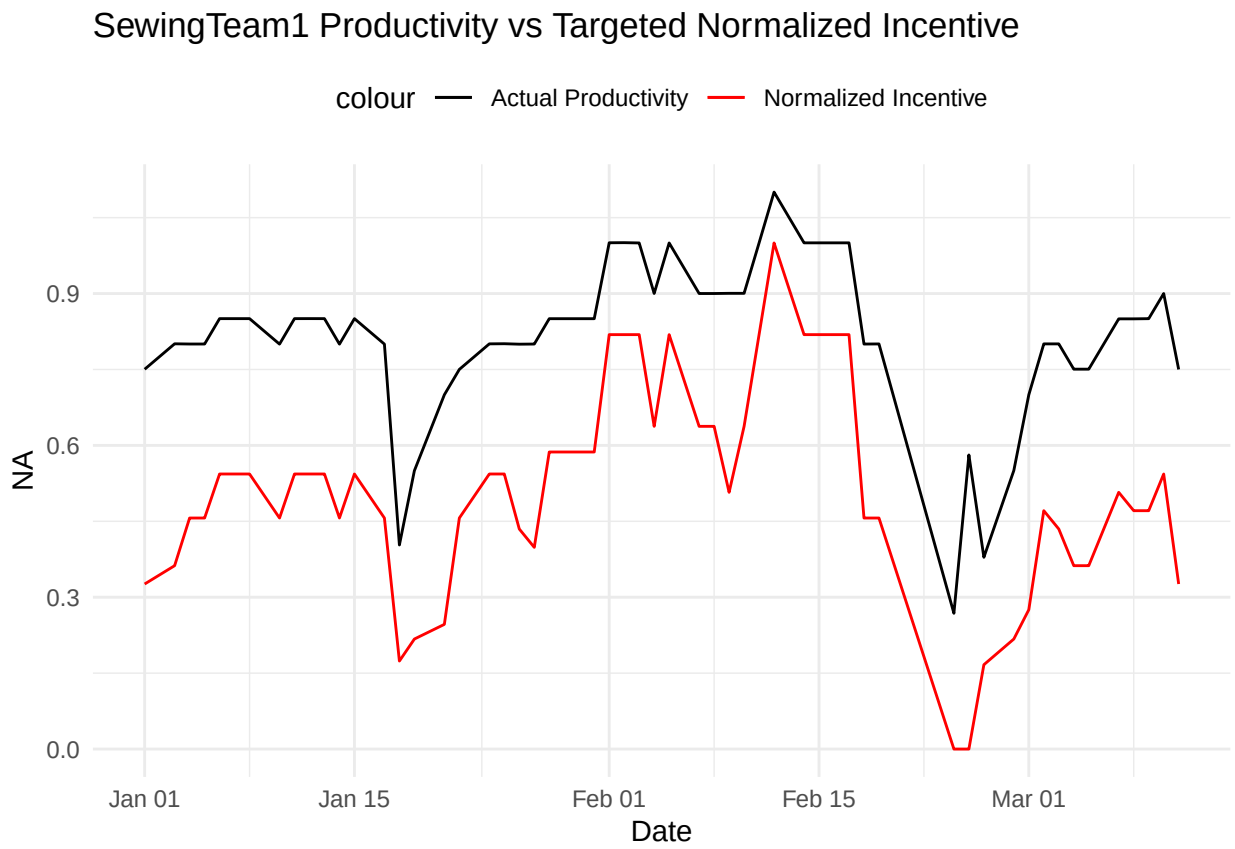



```
#pie(counts_of_quarters, labels = quarters_pie_labels,main = "Pie Chart of Quarters",col = rainbow(leng
#pie(counts_of_days, labels = days_pie_labels,main = "Pie Chart of Days",col = rainbow(length(counts_of
```

We see that sewing department has more corresponding rows than finishing.

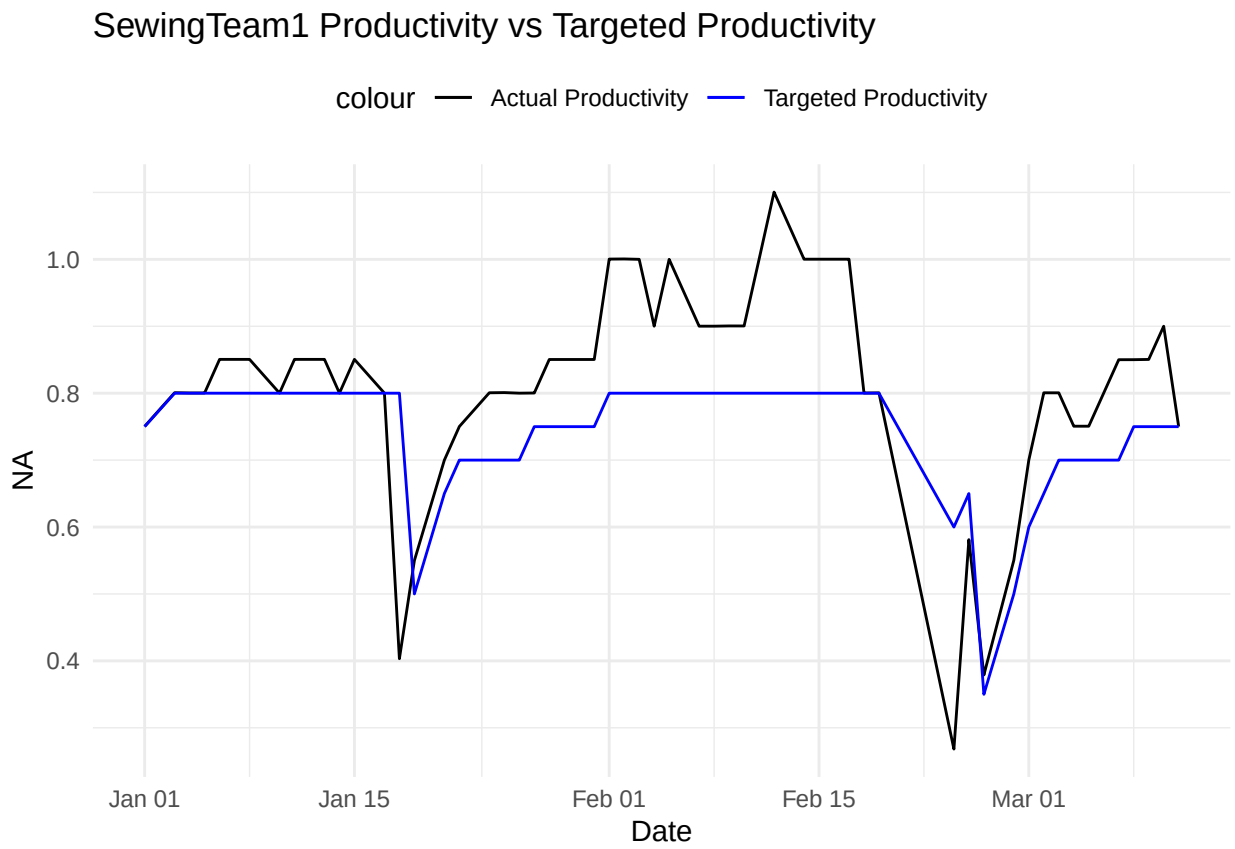
#Incentive ##

```
ggplot(data = sewingTeam01, aes(x = sewingTeam01$date)) +
  geom_line(aes(y= sewingTeam01$actual_productivity,color="Actual Productivity"))+
  geom_line(aes(y = sewingTeam01$incentive/max(sewingTeam01$incentive), color = "Normalized Incentive"))
  labs(x="Date", y = NA, title = "SewingTeam1 Productivity vs Targeted Normalized Incentive")+
  scale_color_manual(values = c("Actual Productivity" = "black", "Normalized Incentive" = "red")) +
  theme_minimal() +
  theme(legend.position = "top")
```



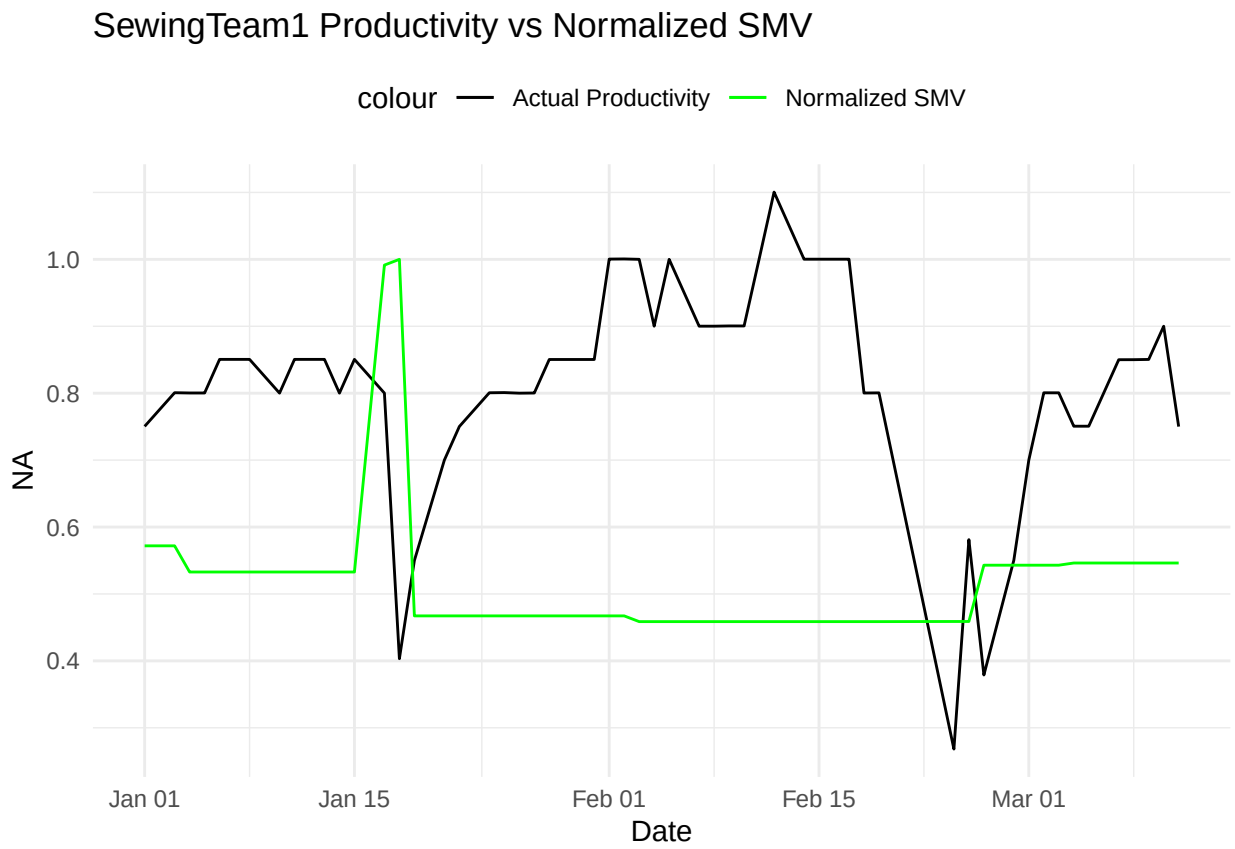
#Target Productivity ##

```
ggplot(data = sewingTeam01, aes(x = sewingTeam01$date)) +
  geom_line(aes(y= sewingTeam01$actual_productivity,color="Actual Productivity"))+
  geom_line(aes(y = sewingTeam01$targeted_productivity, color = "Targeted Productivity"))+
  labs(x="Date", y = NA, title = "SewingTeam1 Productivity vs Targeted Productivity")+
  scale_color_manual(values = c("Actual Productivity" = "black", "Targeted Productivity" = "blue")) +
  theme_minimal() +
  theme(legend.position = "top")
```



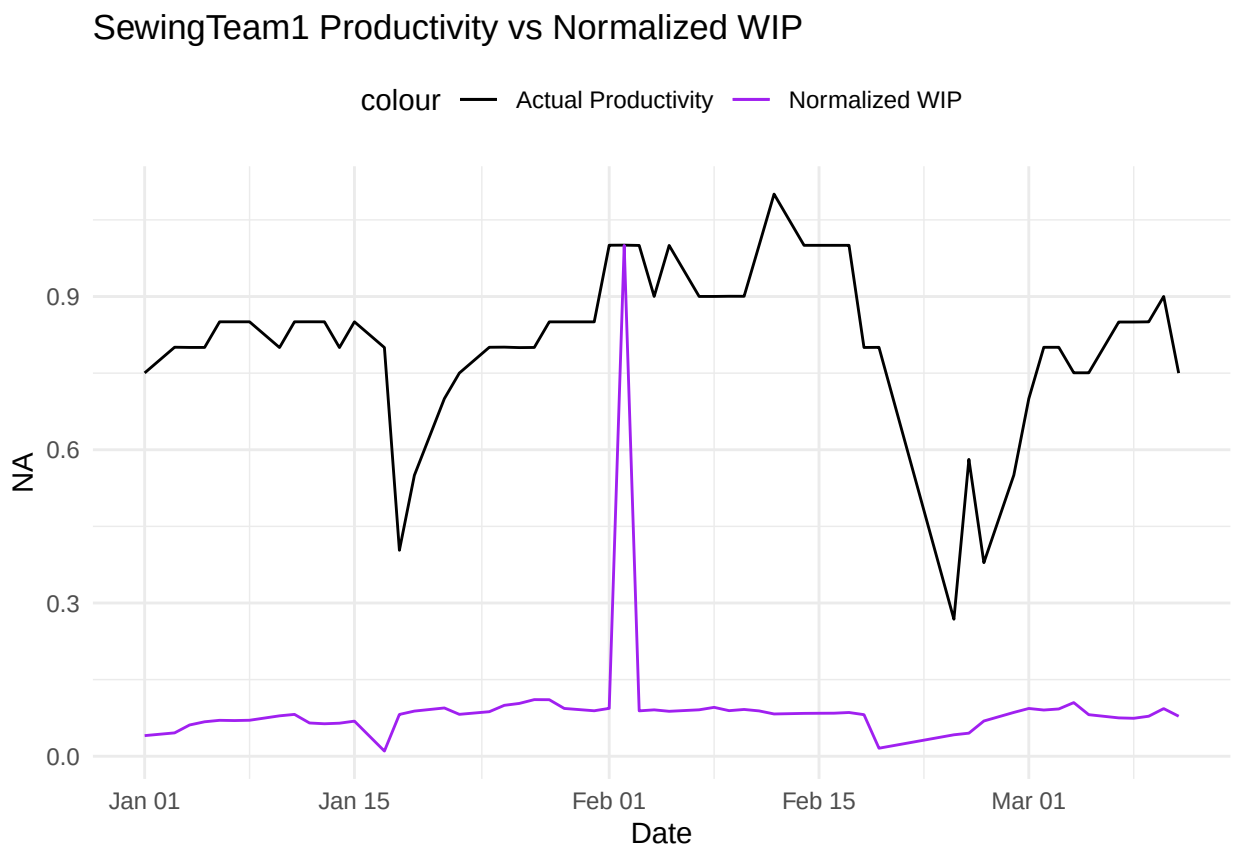
#SMV ##

```
ggplot(data = sewingTeam01, aes(x = sewingTeam01$date)) +
  geom_line(aes(y= sewingTeam01$actual_productivity,color="Actual Productivity"))+
  geom_line(aes(y = sewingTeam01$smv/max(sewingTeam01$smv), color = "Normalized SMV"))+
  labs(x="Date", y = NA, title = "SewingTeam1 Productivity vs Normalized SMV")+
  scale_color_manual(values = c("Actual Productivity" = "black", "Normalized SMV" = "green")) +
  theme_minimal() +
  theme(legend.position = "top")
```



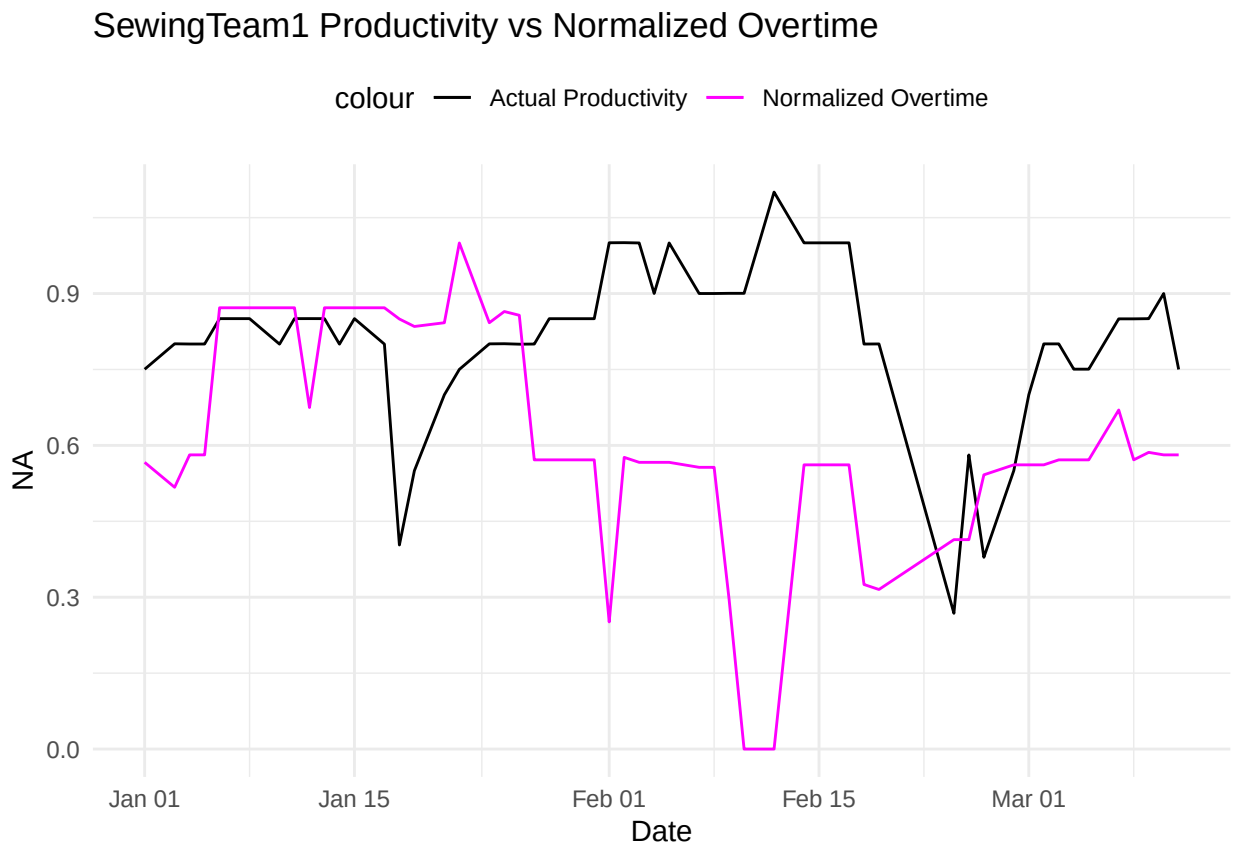
#WIP ##

```
ggplot(data = sewingTeam01, aes(x = sewingTeam01$date)) +
  geom_line(aes(y= sewingTeam01$actual_productivity,color="Actual Productivity"))+
  geom_line(aes(y = sewingTeam01$wip/max(sewingTeam01$wip), color = "Normalized WIP"))+
  labs(x="Date", y = NA, title = "SewingTeam1 Productivity vs Normalized WIP")+
  scale_color_manual(values = c("Actual Productivity" = "black", "Normalized WIP" = "purple")) +
  theme_minimal() +
  theme(legend.position = "top")
```



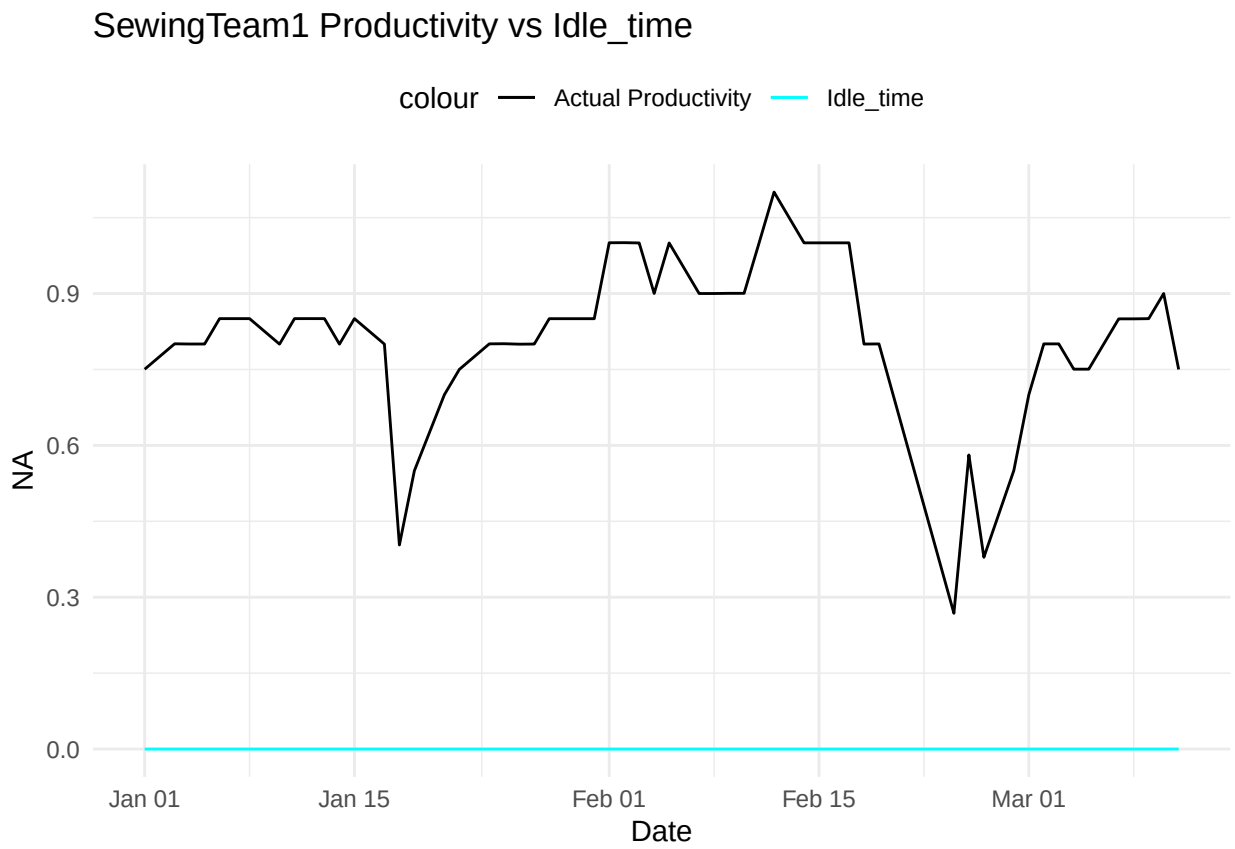
#overtime ##

```
ggplot(data = sewingTeam01, aes(x = sewingTeam01$date)) +
  geom_line(aes(y= sewingTeam01$actual_productivity,color="Actual Productivity"))+
  geom_line(aes(y = sewingTeam01$over_time/max(sewingTeam01$over_time), color = "Normalized Overtime"))+
  labs(x="Date", y = NA, title = "SewingTeam1 Productivity vs Normalized Overtime")+
  scale_color_manual(values = c("Actual Productivity" = "black", "Normalized Overtime" = "magenta")) +
    theme_minimal() +
    theme(legend.position = "top")
```



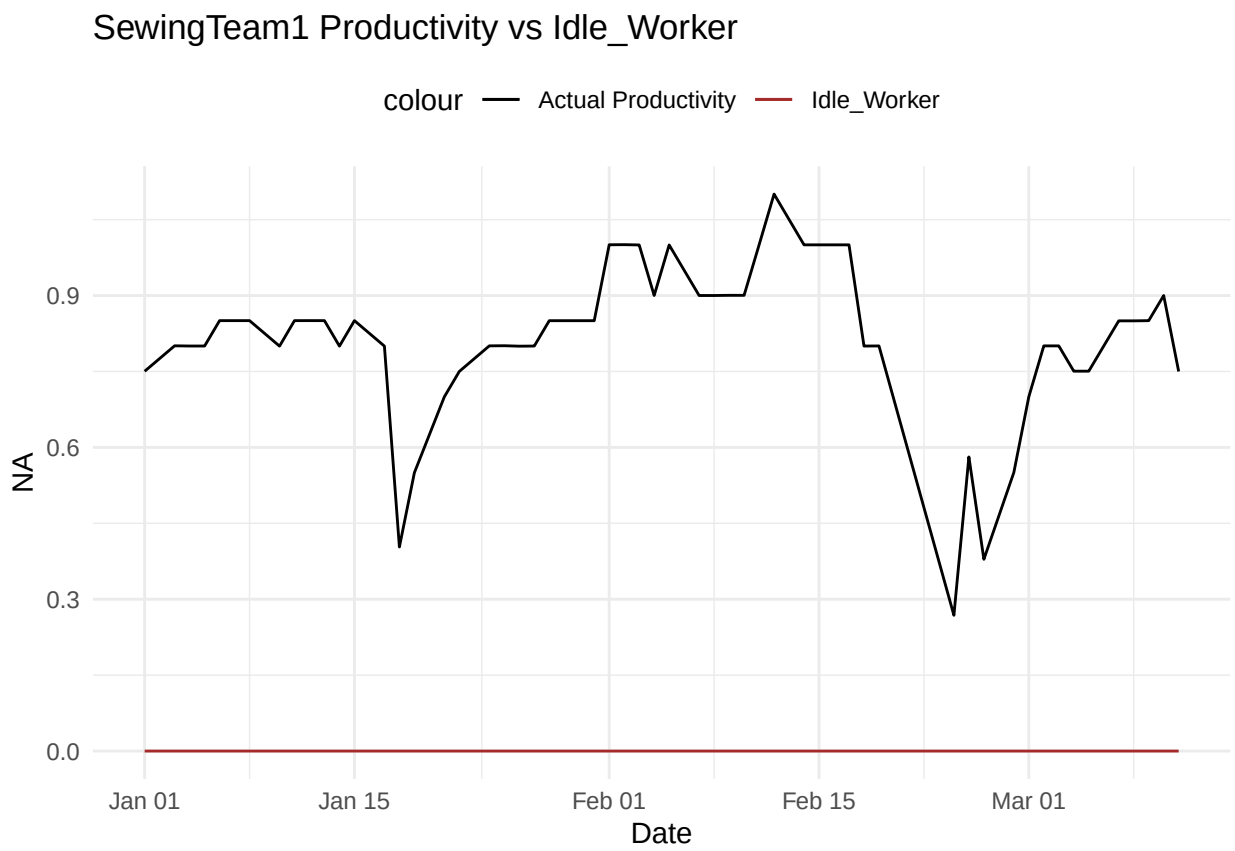
#Idle_time ##

```
ggplot(data = sewingTeam01, aes(x = sewingTeam01$date)) +
  geom_line(aes(y= sewingTeam01$actual_productivity,color="Actual Productivity"))+
  geom_line(aes(y = sewingTeam01$idle_time, color = "Idle_time"))+
  labs(x="Date", y = NA, title = "SewingTeam1 Productivity vs Idle_time")+
  scale_color_manual(values = c("Actual Productivity" = "black", "Idle_time" = "cyan")) +
  theme_minimal() +
  theme(legend.position = "top")
```



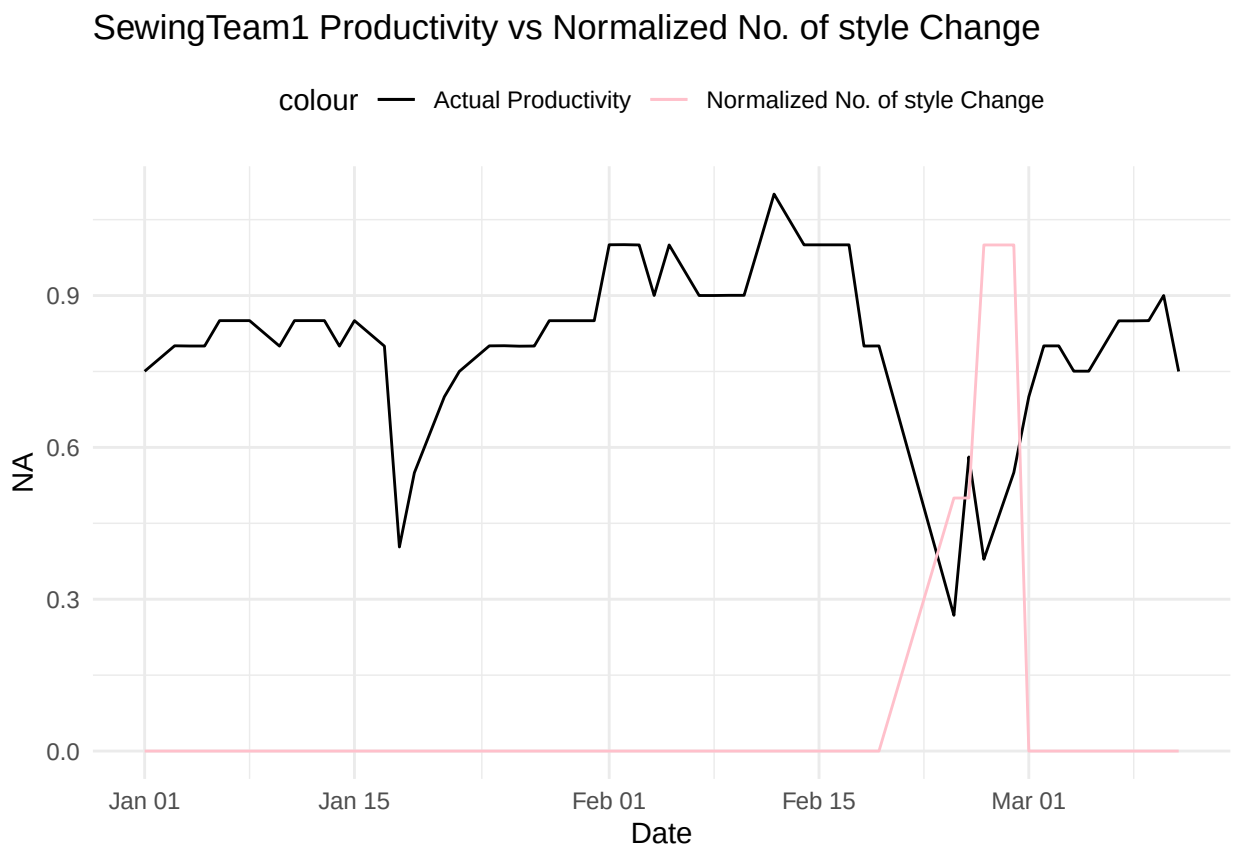
#Idle_men ##

```
ggplot(data = sewingTeam01, aes(x = sewingTeam01$date)) +
  geom_line(aes(y= sewingTeam01$actual_productivity,color="Actual Productivity"))+
  geom_line(aes(y = sewingTeam01$idle_men, color = "Idle_Worker"))+
  labs(x="Date", y = NA, title = "SewingTeam1 Productivity vs Idle_Worker")+
  scale_color_manual(values = c("Actual Productivity" = "black", "Idle_Worker" = "brown")) +
  theme_minimal() +
  theme(legend.position = "top")
```



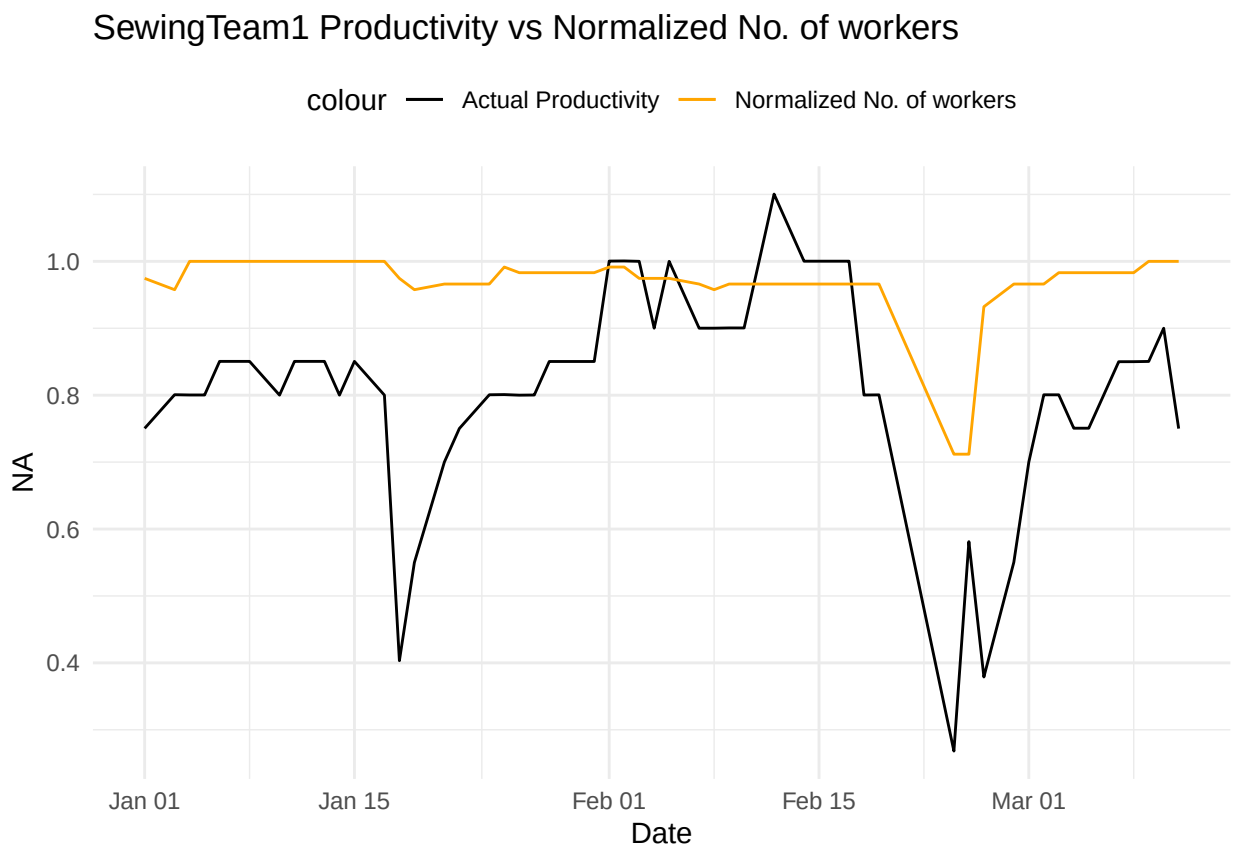
#No. of style Change ##

```
ggplot(data = sewingTeam01, aes(x = sewingTeam01$date)) +
  geom_line(aes(y= sewingTeam01$actual_productivity,color="Actual Productivity"))+
  geom_line(aes(y = sewingTeam01$no_of_style_change/max(sewingTeam01$no_of_style_change), color = "Norma
  labs(x="Date", y = NA, title = "SewingTeam1 Productivity vs Normalized No. of style Change")+
  scale_color_manual(values = c("Actual Productivity" = "black", "Normalized No. of style Change" = "pi
  theme_minimal() +
  theme(legend.position = "top")
```



#Normalized No. of workers

```
ggplot(data = sewingTeam01, aes(x = sewingTeam01$date)) +
  geom_line(aes(y= sewingTeam01$actual_productivity,color="Actual Productivity"))+
  geom_line(aes(y = sewingTeam01$no_of_workers/max(sewingTeam01$no_of_workers), color = "Normalized No.
  labs(x="Date", y = NA, title = "SewingTeam1 Productivity vs Normalized No. of workers")+
  scale_color_manual(values = c("Actual Productivity" = "black", "Normalized No. of workers" = "orange"
  theme_minimal() +
  theme(legend.position = "top")
```



Learning Objects:

We first zoom in the sewing team since there's more data available and no missing value contained. We want to explore the data from following perspectives:

1) How's productivity affected by each variables? (simply looking at the time series plots)
2) Is there a team/teams that outperformed the rest? What set them apart?
3) Is it possible if teams collaborate?
4) How does the workflow look like?