Analysis of absenteeism at work

Clarison James Dsilva

INTRODUCTION

Absenteeism in the workplace is a common phenomenon. Absenteeism, also referred to as a “bottom-line killer”, impacts the availability of the workforce and the profitability of organizations. The high competitiveness in the market, professional development combined with the development of organizations and the pressure to reach increasingly audacious goals, create increasingly overburdened employees and end up acquiring some disturbance in the state of health related to the type of work activity, including depression considered the evil of the 21st century. Taking employees to absenteeism. Absenteeism is defined as absence to work as expected, represents for the company the loss of productivity and quality of work.

The data set has been taken from UCI - Absenteeism at work. The database used has 21 attributes and 740 records from documents that prove that they are absent from work and was collected from January 2008 to December 2016 at a courier company in Brazil.

PROBLEM STATEMENT

To determine how various factors are affecting the absenteeism at work and plot different graphs to analyse the same.

```
#install.packages(scales)
#install.packages(readxl)
#install.packages(gridExtra)
#install.packages(plyr)

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readxl)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.4     v purrr   0.3.4
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidyr)
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following object is masked from 'package:purrr':
##
##     compact
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

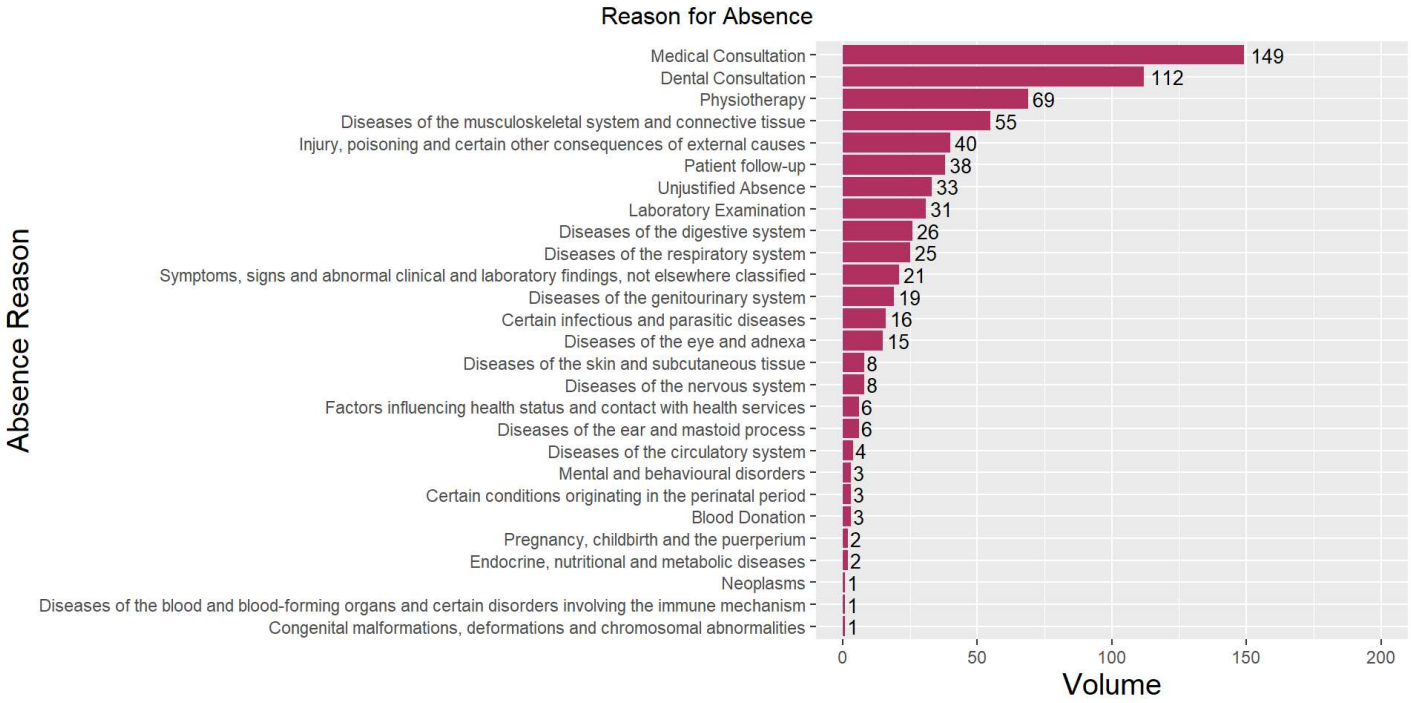
```
library(stringr)
```

```
absent <- read_excel("Absenteeism_at_work.xls")

reason <- read_excel("Absenteeism_at_work.xls","Reason for Absense")

seasons <- read_excel("Absenteeism_at_work.xls","Seasons")

dayofweek <- read_excel("Absenteeism_at_work.xls","Day of Week")

education <- read_excel("Absenteeism_at_work.xls","Education")

month <- read_excel("Absenteeism_at_work.xls","Month")

finaldf <- left_join(absent,reason, by = c("Reason for absence"="Reason for Absense ID")) %>%
  left_join(seasons, by = c("Seasons"="Season ID")) %>%
  left_join(dayofweek, by = c("Day of the week"="Day of Week ID"))  %>%
  left_join(education, by = c("Education"="Education ID")) %>%
  left_join(month, by = c("Month of absence"="Month ID"))

finaldf
```

```
## # A tibble: 740 x 26
##       ID `Reason for absence` `Month of absence` `Day of the week` Seasons
##    <dbl>                <dbl>              <dbl>             <dbl>   <dbl>
##  1    11                   26                  7                 3       1
##  2    36                    0                  7                 3       1
##  3     3                   23                  7                 4       1
##  4     7                    7                  7                 5       1
##  5    11                   23                  7                 5       1
##  6     3                   23                  7                 6       1
##  7    10                   22                  7                 6       1
##  8    20                   23                  7                 6       1
##  9    14                   19                  7                 2       1
## 10     1                   22                  7                 2       1
## # ... with 730 more rows, and 21 more variables: Transportation expense <dbl>,
## #   Distance from Residence to Work <dbl>, Service time <dbl>, Age <dbl>,
## #   Work load Average/day <dbl>, Hit target <dbl>, Disciplinary failure <dbl>,
## #   Education <dbl>, Son <dbl>, Social drinker <dbl>, Social smoker <dbl>,
## #   Pet <dbl>, Weight <dbl>, Height <dbl>, Body mass index <dbl>,
## #   Absenteeism time in hours <dbl>, Absense Reason <chr>,
## #   Season Details <chr>, Day of Week Name <chr>, Education Details <chr>, ...
```

```
##### **Code Output :**
# Separate columns are created to define the numerical categories as characters and are stored i
n finaldf.
```

```
# ##### **Visual Analysis :**
# This Visual represents number of absentees by Absence Reasons. Data is grouped by absence reas
on and a horizontal bar graph is plot between reason of absence and number of absentees (Volum
e).

by_reason <- finaldf %>%
  group_by(`Absense Reason`)  %>%
  dplyr :: summarise(count = n()) %>%
  drop_na()

plot_absent <- ggplot(by_reason ,aes(x = reorder(`Absense Reason`,count),y = count)) + geom_bar
(stat = "identity", fill= 'maroon') + coord_flip() +
  ylim(0,200) +
  geom_text(aes(label = count), hjust = -0.2,size=3.5) +
  labs(x = "Absence Reason",y = "Volume") +
  theme(axis.title.x = element_text(size=15),axis.title.y = element_text(size = 15))

grid.arrange(plot_absent,top = "Reason for Absence")
```

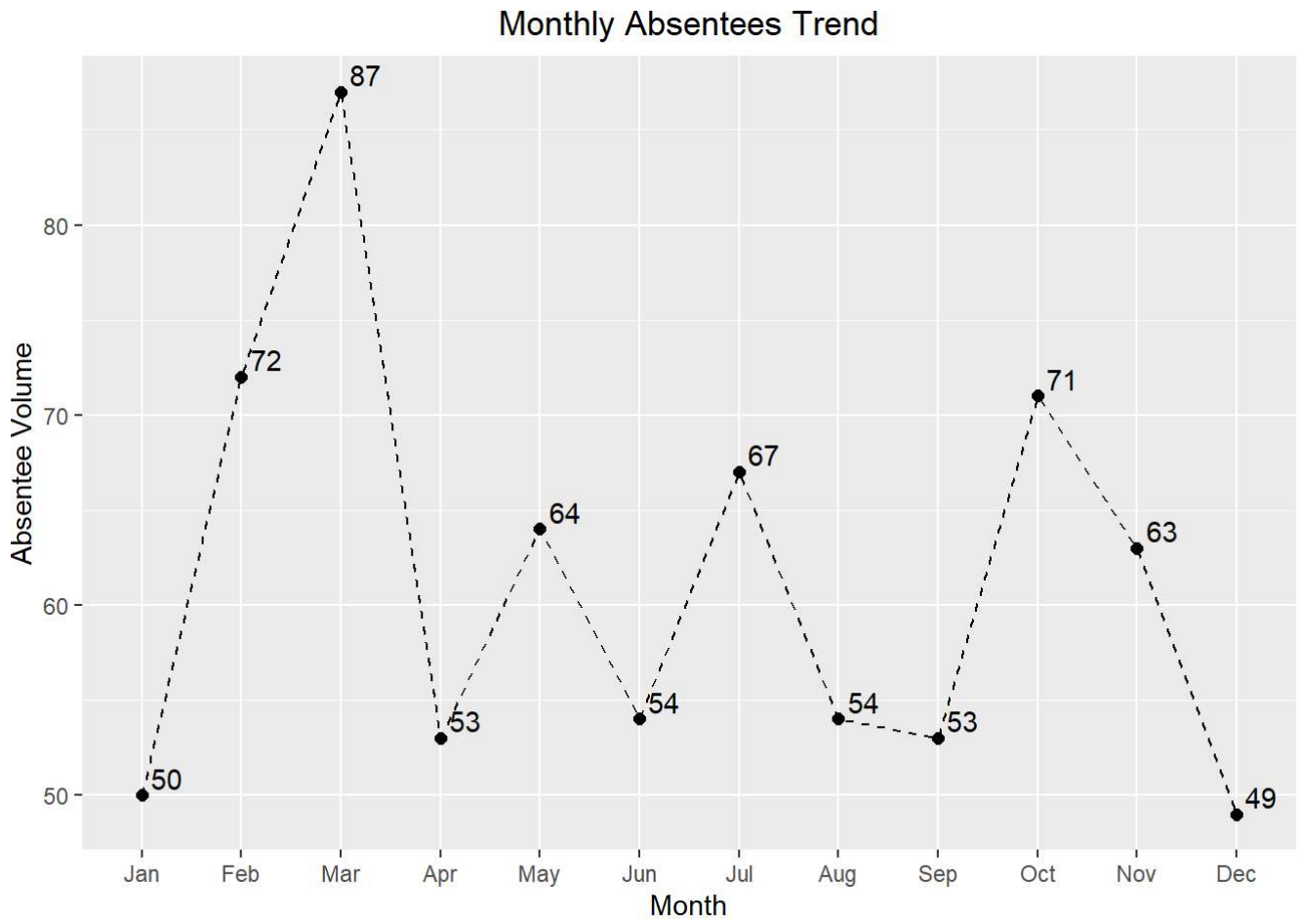



```
# ##### **Insights :**
# From the plot, we can see that the major reasons for absenteeism are Medical, Dental consultat
ion, Physiotherapy. Most of the reasons for absentees are disease of some kind.
```

```
# ##### **Visual Analysis :**
# This Visual represents  absenteeism trend with respect to month.


by_month <- finaldf %>%
  group_by(Month,`Month of absence`)  %>%
  dplyr:: summarise(count = n())  %>%
  drop_na()
```

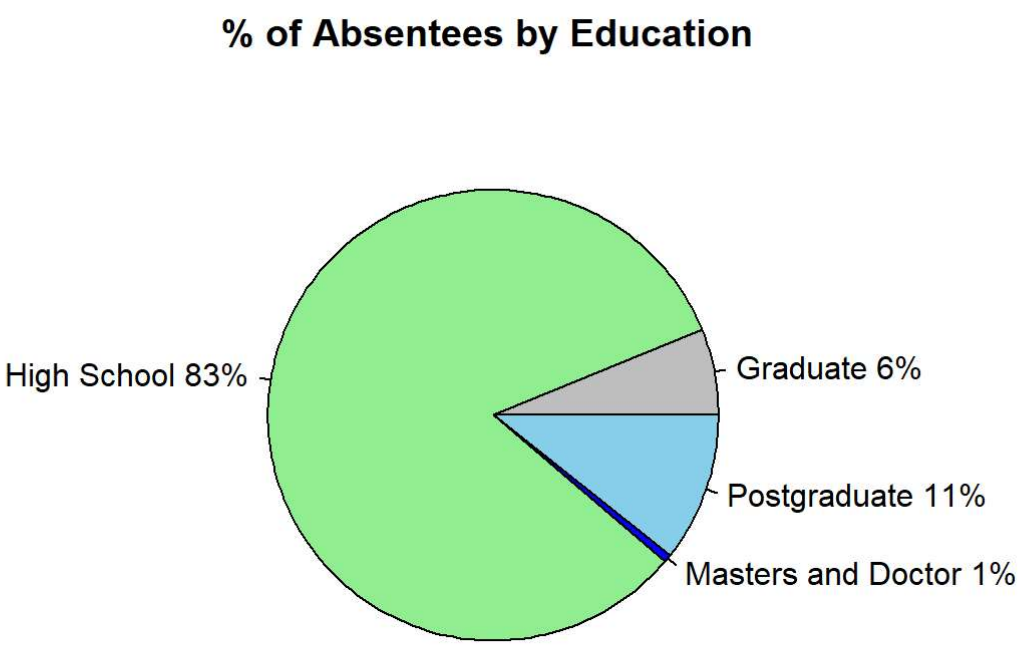
```
## `summarise()` has grouped output by 'Month'. You can override using the `.groups` argument.
```

```
ggplot(by_month,aes(x = reorder(`Month`, `Month of absence`), y = count, group = 1)) +
  geom_line(linetype = "dashed") +
  ggtitle("Monthly Absentees Trend") +
  labs(x="Month", y="Absentee Volume") +
  geom_point(size = 2) +
  geom_text(aes(label = count), vjust = -0.3, hjust = -0.3) +
  theme(plot.title = element_text(hjust = 0.5))
```



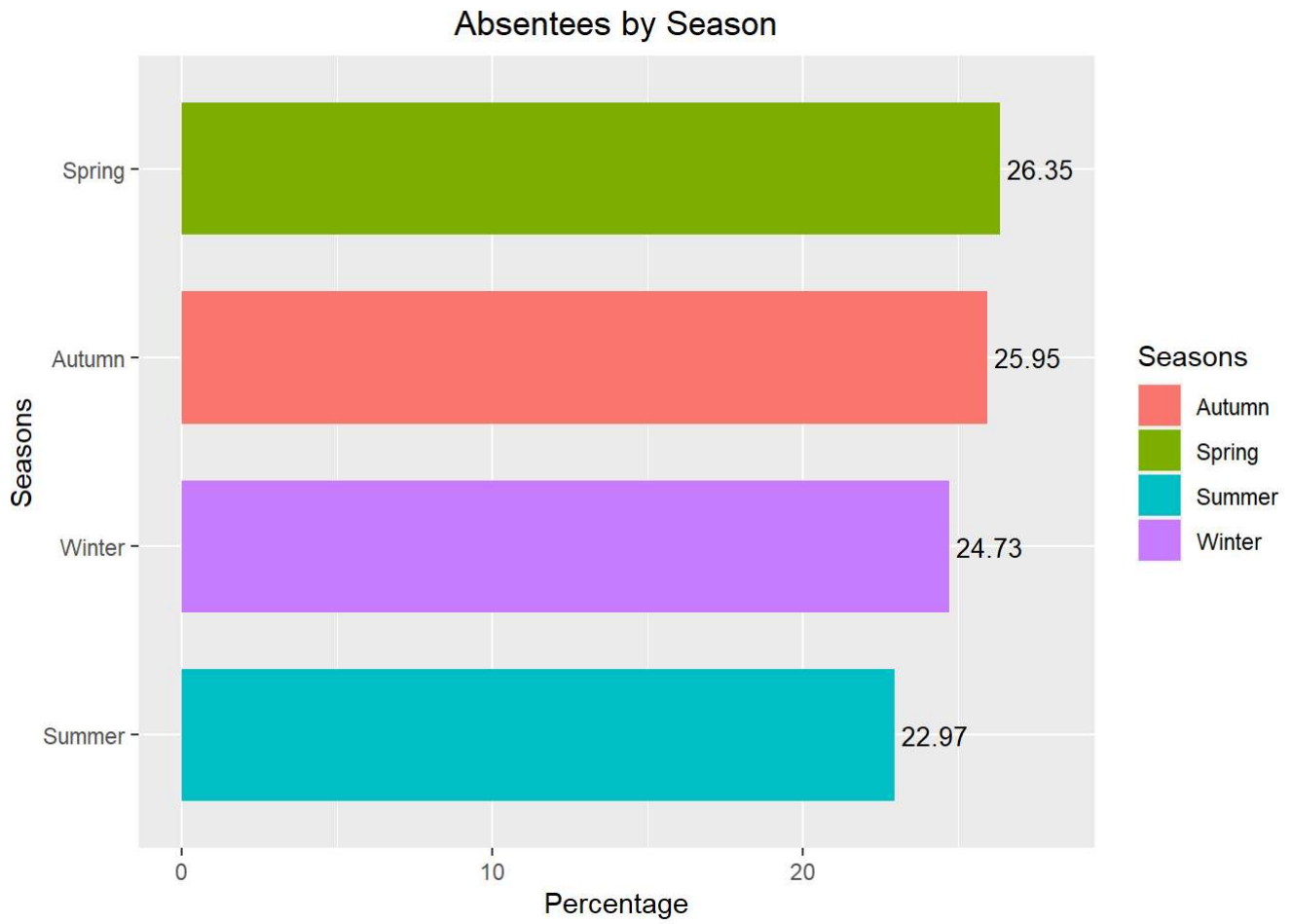

```
# ##### **Insights :**
# The plot clearly depicts that the highest number of leaves have been taken in March (87) and t
he least in December (49). The reason for least leaves in December might be due to holidays in D
ecember.
```

```
# ##### **Visual Analysis :**
# This Visual represents Percentage of absentees with respect to educational details.


by_edu<- finaldf %>%
  group_by(`Education Details`)  %>%
  dplyr:: summarise(count = n())

pctbyedu  <- round(by_edu$count/sum(by_edu$count)*100)
lable <- paste(by_edu$`Education Details`, pctbyedu)
lable <- paste(lable, "%" ,sep = "")
pie(by_edu$count, labels = lable, main = "% of Absentees by Education ",  col = c("grey", "light
green", "blue", "sky blue"))
```




```
# ##### **Insights :**
# Most absentees are workers who's highest education is high school(83%), as the level of educat
ion increases, we can see that the absenteeism percentage decreases.
```

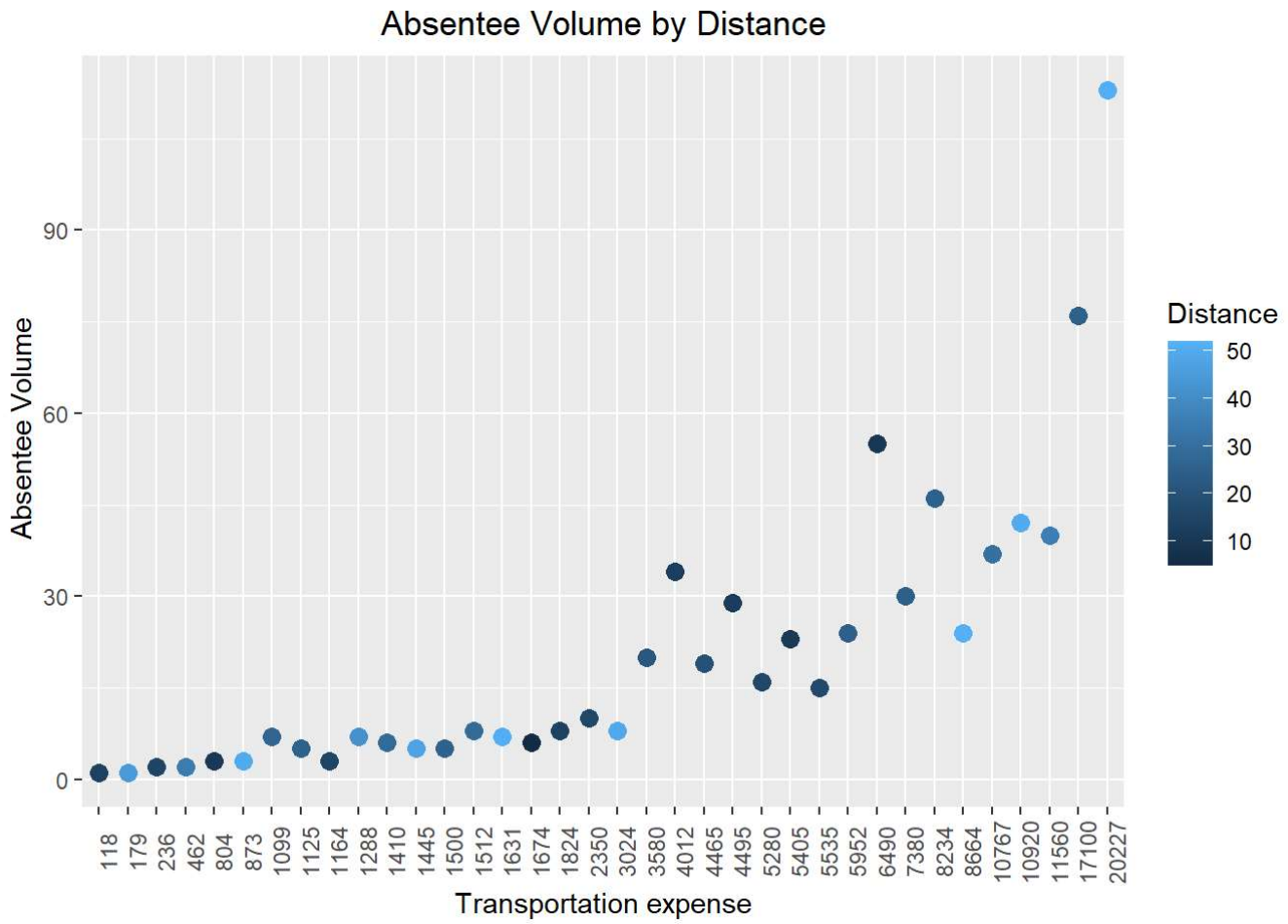
```
# ##### **Visual Analysis :**
# This Visual represents percentage of absentees by season.

by_season <- finaldf %>%
  group_by(`Season Details`)  %>%
  dplyr:: summarise(count = n(), percentage=(count/nrow(finaldf))*100)

ggplot(by_season, aes(x = reorder(`Season Details`,percentage), y = round(percentage,2), fill =
 `Season Details`)) +
  geom_bar(stat = "identity", width = 0.7, position = position_dodge(width = 0.5)) + coord_flip
() +
  ggtitle("Absentees by Season") +
  labs(x = "Seasons", y = "Percentage", fill = "Seasons") +
  geom_text(aes(label = round(percentage,2)), hjust = -0.1, size = 3.5) +
  theme(plot.title = element_text(hjust = 0.5) ) +
  ylim(0,28)
```




```
# ##### **Insights :**
# The season with the most number of absentees is Spring (26.35%) and least is Summer (22.97%).T
here is a gradual increase between the seasons.
```

```
###### **Visual Analysis :**
# This Visual represents impact of transportation expense and distance between Residence to Work
on absenteeism.

by_ID <- finaldf %>%
  group_by(ID)  %>%
  dplyr:: summarise(expense = sum(`Transportation expense`, na.rm = TRUE),
                    Distance = max(`Distance from Residence to Work`,na.rm = TRUE), Absent = n
())


ggplot(by_ID, aes( x = factor(expense), y = Absent, color = Distance)) + geom_point(size = 3) +
  scale_fill_brewer(palette = "blue") +
  theme(axis.text.x = element_text(angle = 90)) +
  ggtitle("Absentee Volume by Distance") +
  labs(x = "Transportation expense", y = " Absentee Volume") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning in pal_name(palette, type): Unknown palette blue
```



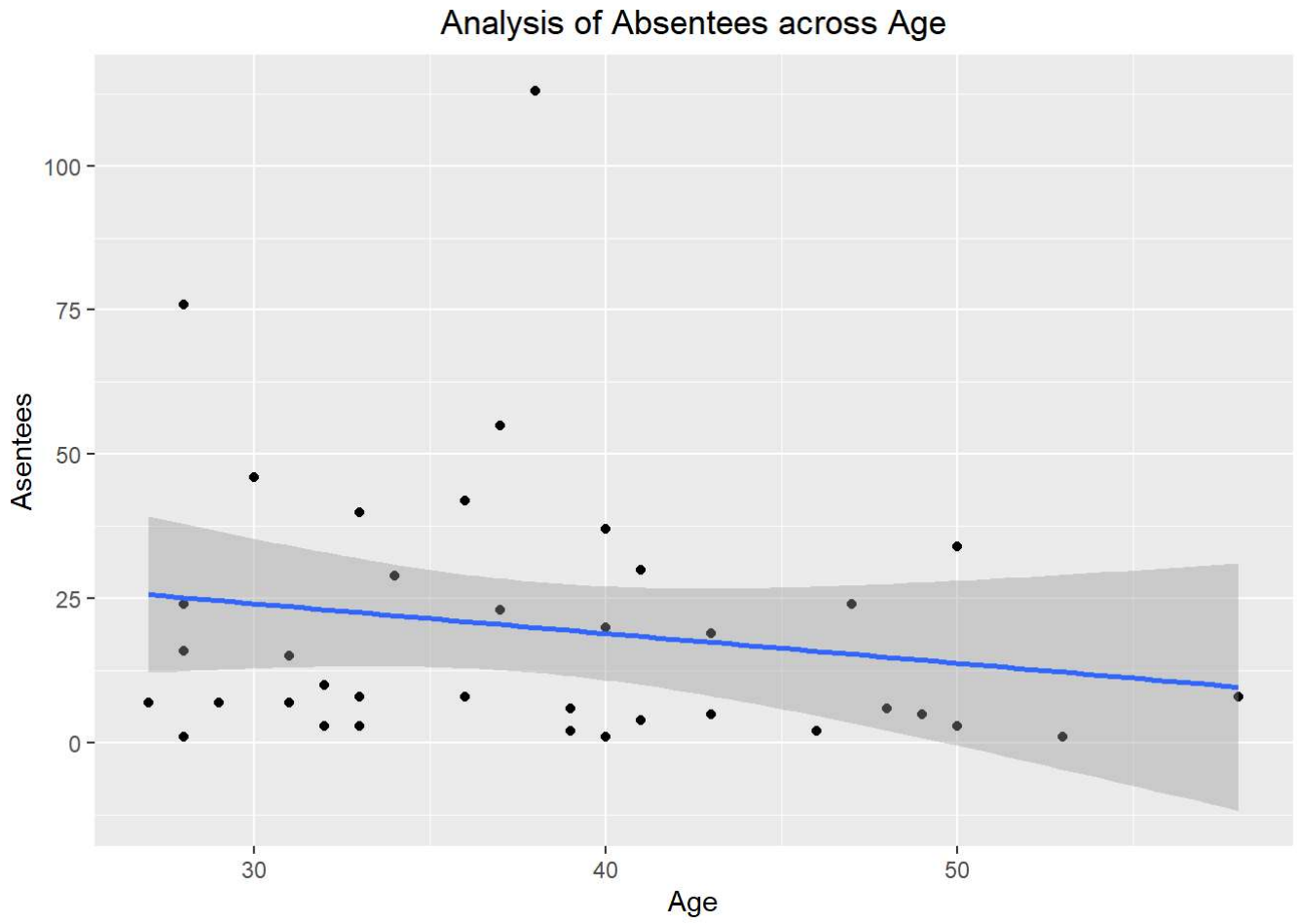

```
###### **Insights :**
# When the transportation expense is high, absentees are also the highest in most of the cases.W
e can also see that distance and transportation expense are not directly proportional,there are
 cases where distance is more and expense is also more but absentees are less.
```

```
# ### *Visual Analysis* :
# This visual represents absentees according to age of the employees.

age_for_leaves <- subset(absent,select = c('ID','Age'))
Count_for_ID <- age_for_leaves %>% group_by(ID,Age) %>%
  dplyr::summarise(c=n())
```

```
## `summarise()` has grouped output by 'ID'. You can override using the `.groups` argument.
```

```
Count_for_ID %>%
ggplot() +
  aes(x= Age,y=c) +
  geom_point() +
  geom_smooth(method = 'lm') +
  ggtitle('Analysis of Absentees across Age') +
  labs( x='Age',
       y='Asentees')+

 theme(plot.title = element_text(hjust = 0.5))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

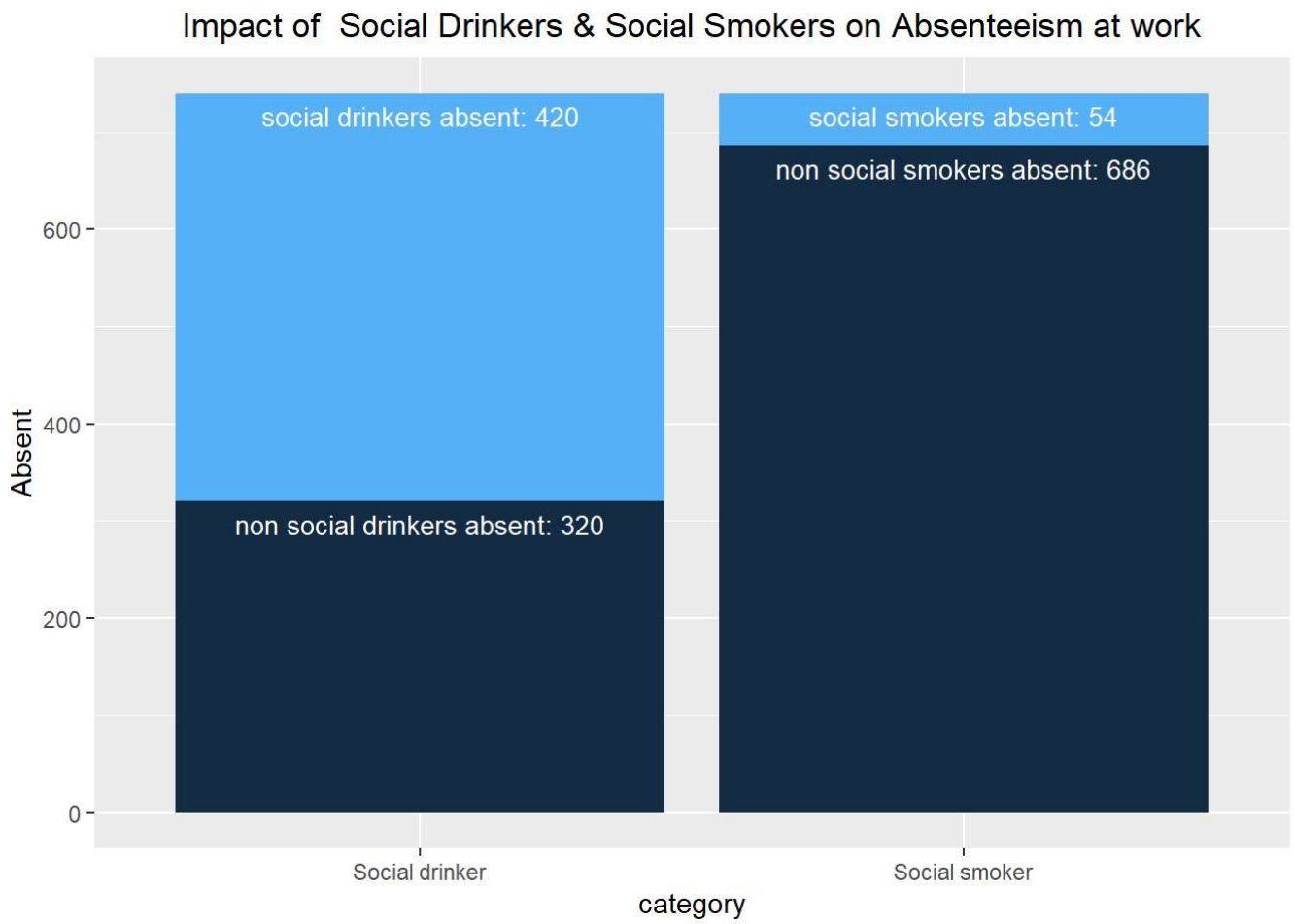



```
# ### *Insights* :
# Most of the employees who have taken leave are young people of age group 30-40. Older employee
s tend to take fewer leaves according to the analysis.
```

```
# ##### **Visual Analysis :**
# This Visual represents Number of absentees with respect to day of the week.


by_dayofweek <- finaldf %>%
  group_by(`Day of Week Name`)  %>%
  dplyr::summarise(count = n())



ggplot(by_dayofweek,aes(x=reorder(`Day of Week Name`,count),y=count)) +
  geom_segment( aes( xend=`Day of Week Name`, yend=0),color="skyblue") +
  geom_point( color="blue", size=4, alpha=0.6) +
  theme_light() +
  coord_flip() +
  ggtitle("Absentees by Days of the week") +
  labs(x = "Day of the week", y = "Number of Absentees")+
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank(),
    plot.title = element_text(hjust = 0.5)
  ) +
  theme(panel.grid.major = element_blank(),
    panel.grid.minor = element_blank())
```

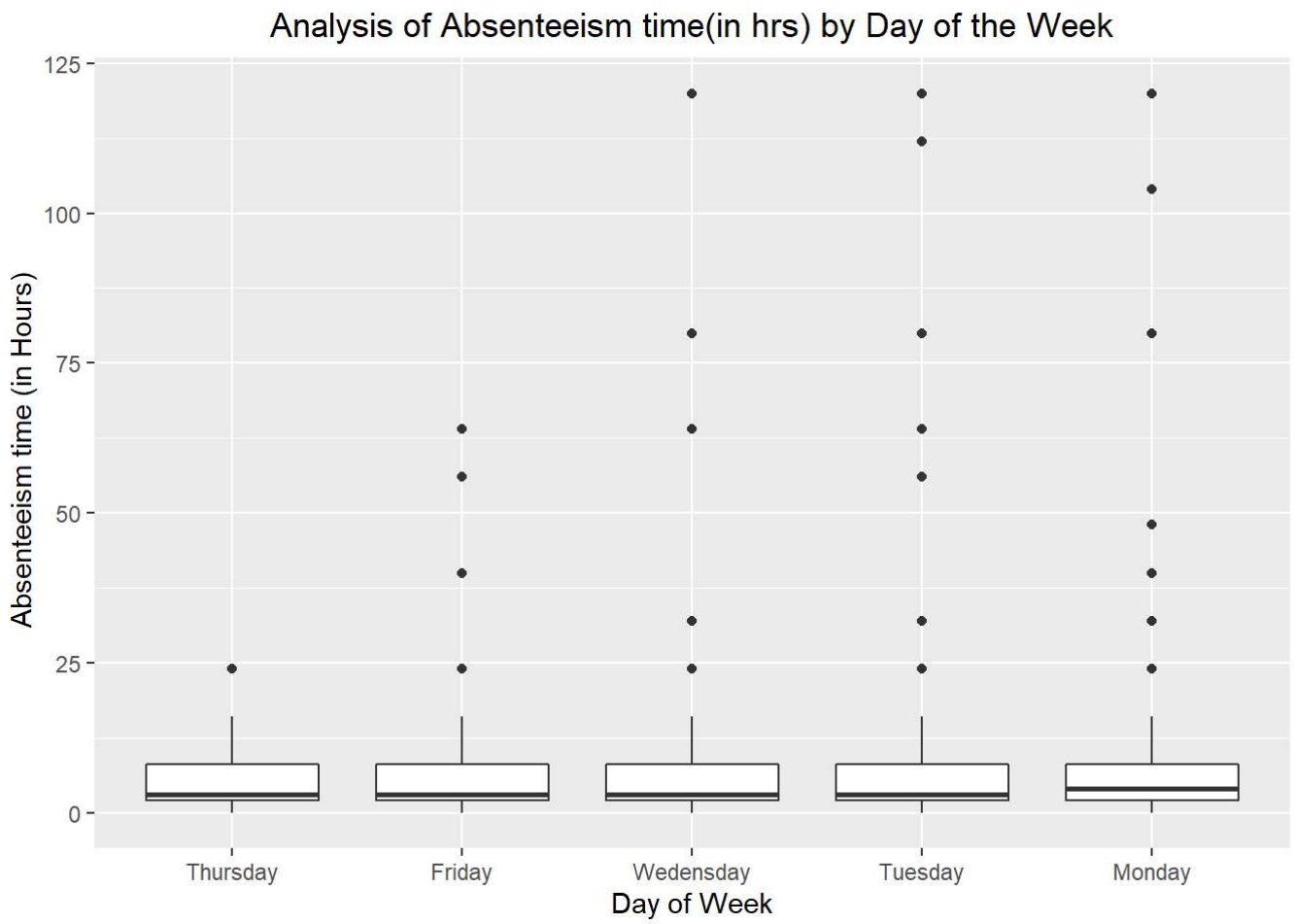
```
# ##### **Insights :**
# Monday's of every week has the most number of absentees (>150), as it is the immediate working
day after the weekend and people usually tend to make it a long weekend.
```

```
# ##### **Visual Analysis :**
# This Visual represents impact of  social smokers and social drinkers on absenteeism at work.



by_socialdrinker <- finaldf %>%
  group_by("y/n"=as.integer(`Social drinker`))  %>%
  dplyr::summarise(Absent=n()) %>%
  mutate(category ="Social drinker")

by_socialsmoker <- finaldf %>%
  group_by("y/n"=as.integer(`Social smoker`)) %>%
  dplyr::summarise(Absent=n()) %>%
  mutate(category ="Social smoker")

by_social <- rbind(by_socialdrinker,by_socialsmoker)


by_social_cumsum <- ddply(by_social, "category", transform, Absent_cumsum=cumsum(Absent))

ggplot(by_social_cumsum, aes(x=category,y=Absent,fill=`y.n`)) +
  geom_bar(stat="identity")  +
  geom_text(aes(y=Absent_cumsum,label = ifelse(category== "Social drinker" & `y.n`==1 ,
                                               paste0("social drinkers absent: ", Absent),
                                        ifelse(category== "Social drinker" & `y.n`==0,
                                               paste0("non social drinkers absent: ", Absent),
                                        ifelse(category== "Social smoker" & `y.n`==1,
                                               paste0("social smokers absent: ", Absent),
                                        paste0("non social smokers absent: ", Absent))))), vjust
=1.6, color="white", size=3.5) +
  ggtitle("Impact of  Social Drinkers & Social Smokers on Absenteeism at work") +
  theme(plot.title = element_text(hjust = 0.5),legend.position = "none")
```



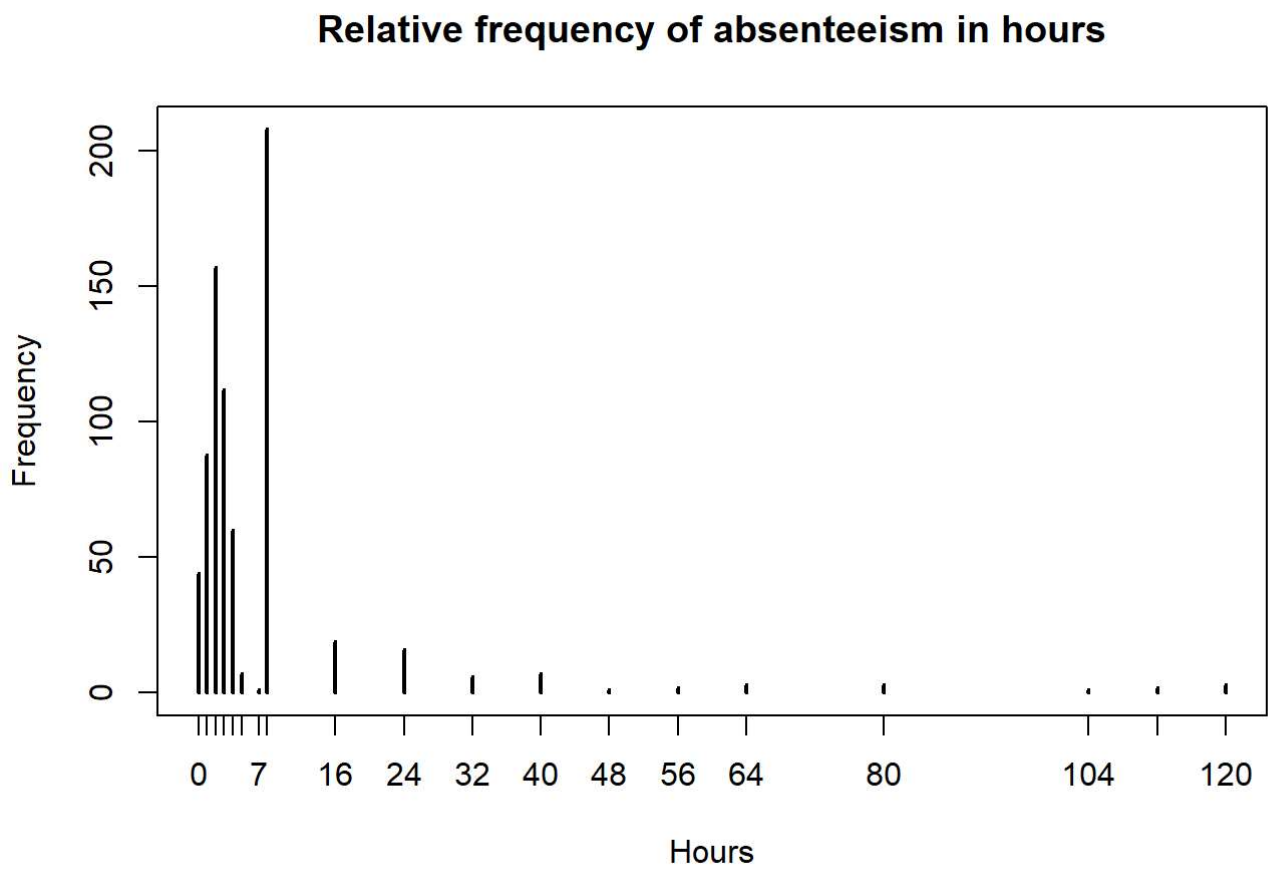

```
# ##### **Insights :**
# We can analyze from the plot that most of the absentees are social drinkers whereas social smo
kers contribute less to the absentees.
```

```
# ##### **Visual Analysis :**
# This Visual represents Analysis of Absenteeism time(in hours) by Day of the Week


box_plot <- finaldf %>% select(`Day of Week Name`, `Season Details`, `Absenteeism time in hours`
)


ggplot(box_plot,aes(x = reorder(`Day of Week Name`,`Absenteeism time in hours`) , y = `Absenteei
sm time in hours`)) + geom_boxplot() +
  ggtitle("Analysis of Absenteeism time(in hrs) by Day of the Week") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Day of Week", y = " Absenteeism time (in Hours)")
```




```
##### **Insights :**
#We can analyze from the plot that the spread of absenteeism (time in hours) is more on Monday,
 Tuesday and Wednesday.
```

```
# ### *Visual Analysis* :
# The Visual represents relative frequency of absenteeism time in hours.

frequency_table =table(finaldf$`Absenteeism time in hours`)
plot(frequency_table  ,xlab = "Hours" ,ylab = "Frequency",main="Relative frequency of absenteeis
m in hours")
```




```
# ### *Insights* :
# Most of the workers have been absent for 8 hours in total(more than 200 employees)
```