school data analysis

Vamshi Krishna Korutla

`02/02/2025`

Introduction:

In this analysis, we explore a dataset provided by Gloria containing student absenteeism data, which will help us understand the relationship between student absences and their academic performance, represented by GPA. By analyzing this dataset, we aim to uncover patterns in absenteeism based on factors like the type of absence (excused vs. unexcused), the presence of special education plans (IEP or 504 plans), and the overall impact on academic performance. The findings from this analysis could be valuable for making data-driven decisions to improve attendance and student success.

#load necessary libraries

```
library(ggplot2)
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(tidyr)
library(lubridate)
```

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

```
library(ggthemes)
library(forcats)
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

Load the dataset

```
school_data <- read_excel("Schooldata.xlsx", sheet = "Sheet1")
```

Cleaning and transforming the data

```
colnames(school_data) <- make.names(colnames(school_data)) # Ensure proper column names
```

Extracting relevant columns

```
school_data <- school_data %>%
  rename(Total_Absences = Total.Absences, GPA = GPA, Plan = Plan, CA_2023_24 = CA.2023.24, Excus
ed = Excused, Unexcused = Unexcused)
```

Cleaning the absence data

```
school_data$Total_Absences <- as.numeric(school_data$Total_Absences)
school_data$Excused <- as.numeric(school_data$Excused)
school_data$Unexcused <- as.numeric(school_data$Unexcused)
```

Extracting numeric absence values from CA_2023_24 column

```
school_data <- school_data %>%
  mutate(CA_2023_24 = as.numeric(gsub("[^0-9]", "", CA_2023_24)))

school_data <- school_data %>% filter(!is.na(Total_Absences))
```

Categorizing students based on IEP and 504 Plan

```
school_data <- school_data %>%
  mutate(IEP_504 = case_when(
    grepl("IEP", Plan, ignore.case = TRUE) ~ "IEP",
    grepl("504", Plan, ignore.case = TRUE) ~ "504",
    TRUE ~ "None"
  ))
```

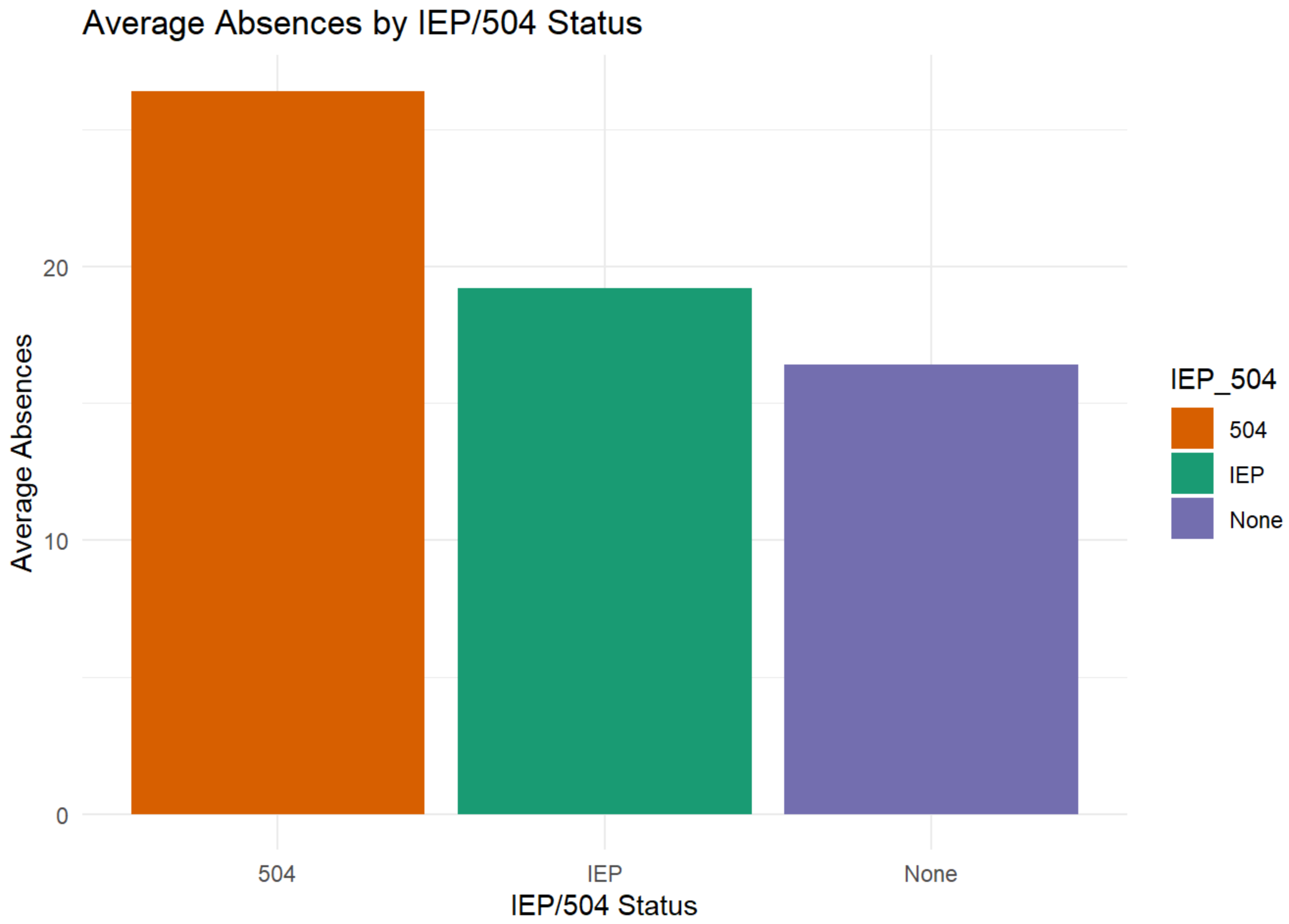
Summary statistics of absenteeism

```
absence_summary <- school_data %>%
  group_by(IEP_504) %>%
  summarise(Average_Absences = mean(Total_Absences, na.rm = TRUE),
            Average_Excused = mean(Excused, na.rm = TRUE),
            Average_Unexcused = mean(Unexcused, na.rm = TRUE))
print(absence_summary)
```

```
## # A tibble: 3 × 4
##   IEP_504 Average_Absences Average_Excused Average_Unexcused
##   <chr>              <dbl>           <dbl>             <dbl>
## 1 504                 26.4            19.2             18   
## 2 IEP                 19.2            14.7              5.54
## 3 None                16.4            14.8              4.88
```

Bar plot for absenteeism by IEP/504 status

```
ggplot(absence_summary, aes(x = IEP_504, y = Average_Absences, fill = IEP_504)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Average Absences by IEP/504 Status", x = "IEP/504 Status", y = "Average Absence
s") +
  scale_fill_manual(values = c("IEP" = "#1b9e77", "504" = "#d95f02", "None" = "#7570b3"))
```



The bar plot visualizes the average number of absences for students based on their IEP/504 status. The plot helps us understand if students with an IEP or 504 plan tend to have more or fewer absences compared to those without such plans. The colors represent different groups: IEP, 504, and None. Correlation between absences and GPA

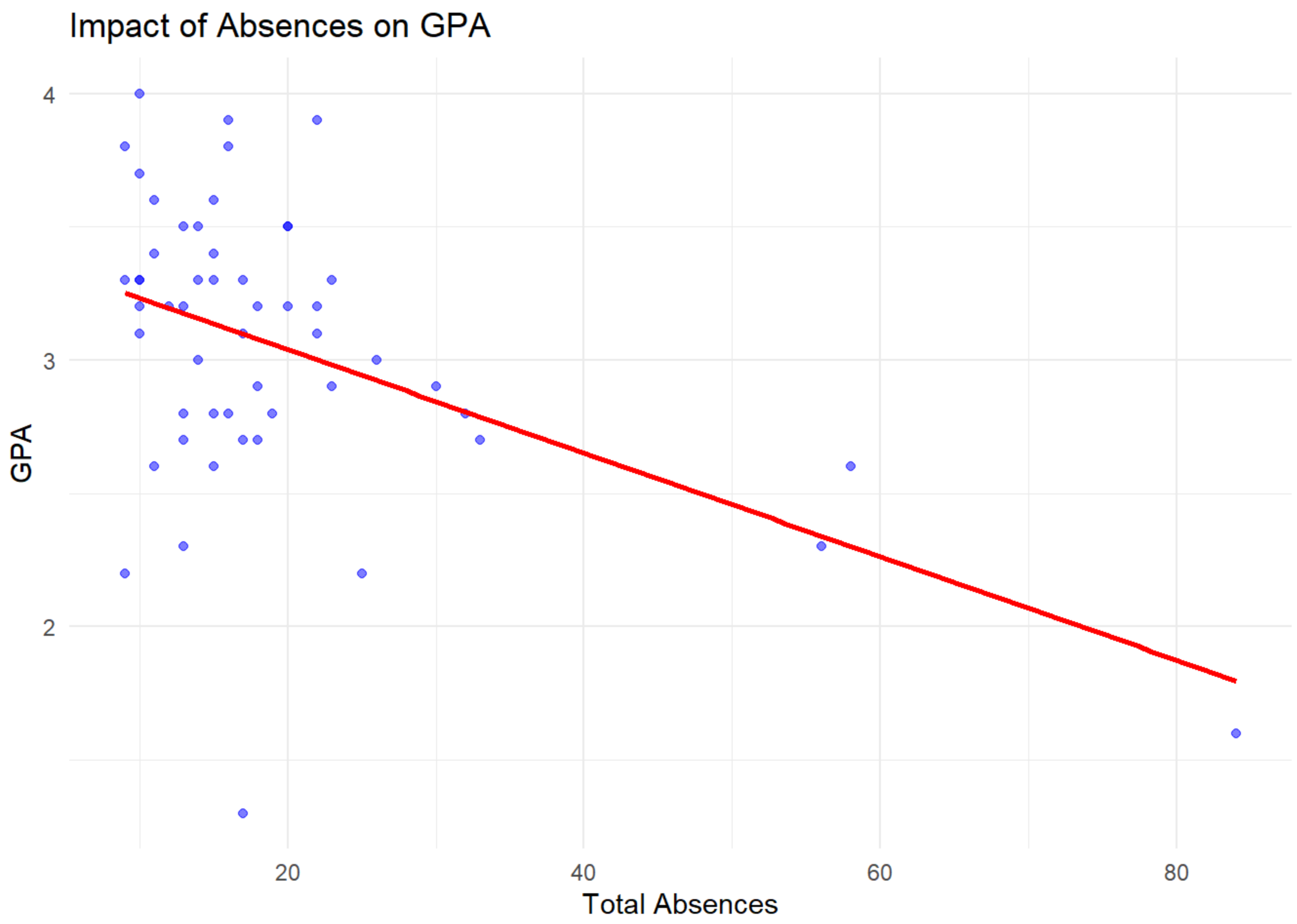
```
correlation_data <- school_data %>% select(Total_Absences, GPA) %>% na.omit()
cor_value <- cor(correlation_data$Total_Absences, correlation_data$GPA, method = "pearson")
print(paste("Correlation between absences and GPA:", round(cor_value, 2)))
```

```
## [1] "Correlation between absences and GPA: -0.47"
```

Scatter plot for absences vs GPA

```
ggplot(correlation_data, aes(x = Total_Absences, y = GPA)) +
  geom_point(alpha = 0.5, color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  theme_minimal() +
  labs(title = "Impact of Absences on GPA", x = "Total Absences", y = "GPA")
```

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



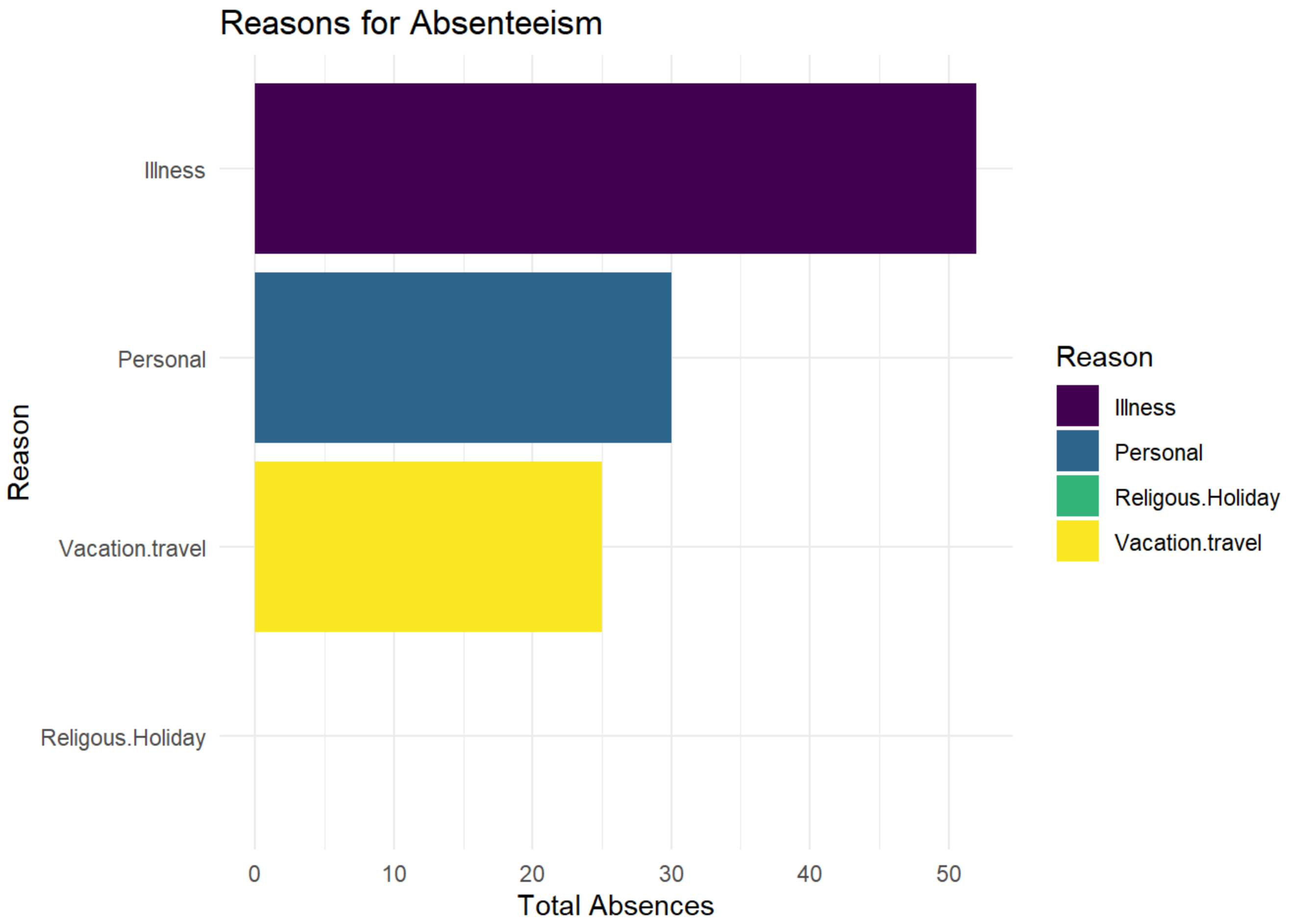
###

The scatter plot shows the relationship between student absences and GPA.A negative correlation can be observed, which implies that increased absences tend to lower GPA.The red line represents the linear trend that helps to identify the general direction of the relationship.

Absences by reason category

```
if("Illness" %in% colnames(school_data)) {
  reason_summary <- school_data %>%
    select(Illness, `Vacation.travel`, `Religous.Holiday`, `Personal`) %>%
    summarise_all(~ sum(. > 0, na.rm = TRUE)) %>%
    pivot_longer(cols = everything(), names_to = "Reason", values_to = "Count")

  ggplot(reason_summary, aes(x = fct_reorder(Reason, Count), y = Count, fill = Reason)) +
    geom_bar(stat = "identity") +
    coord_flip() +
    theme_minimal() +
    labs(title = "Reasons for Absenteeism", x = "Reason", y = "Total Absences") +
    scale_fill_viridis_d()
}
```



###

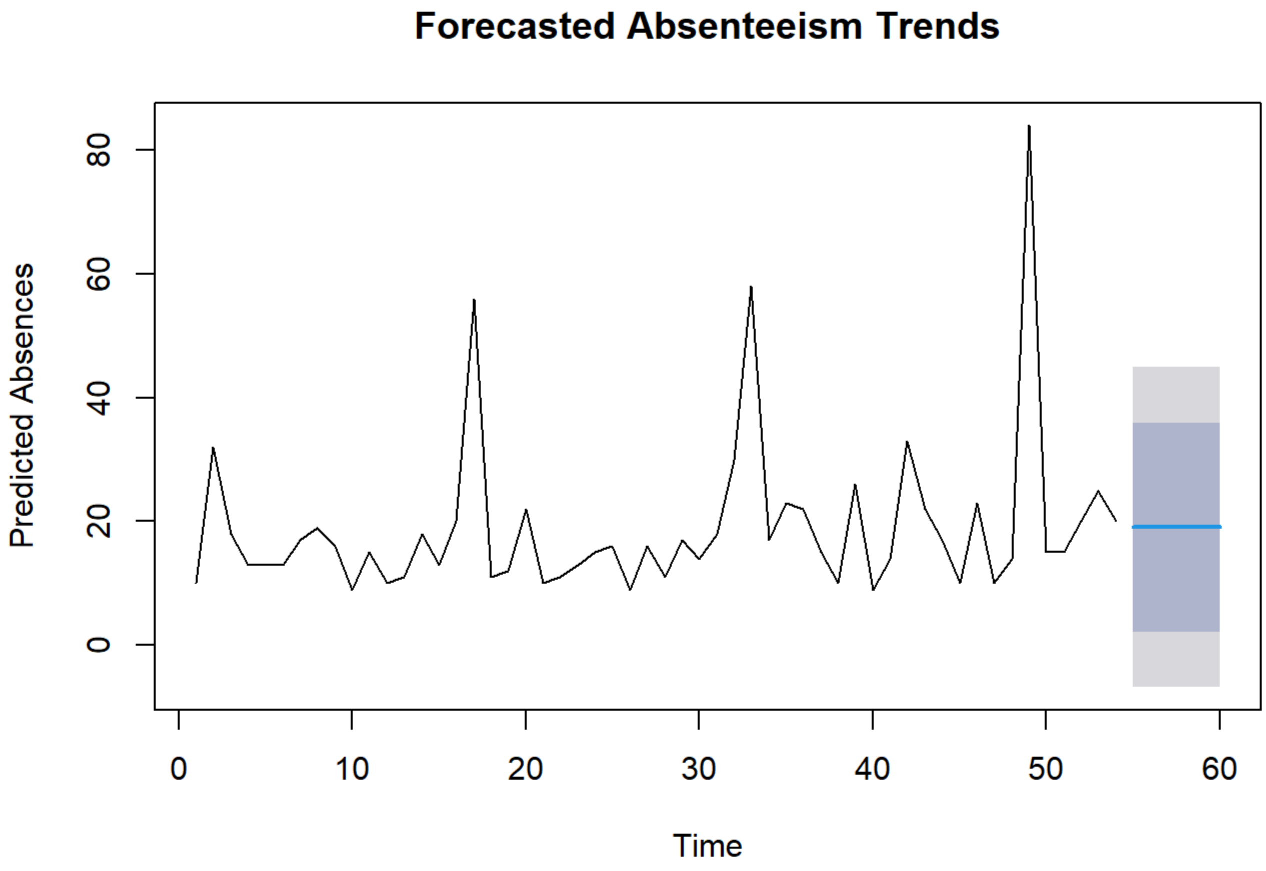
The horizontal bar chart displays the reasons for absenteeism, such as illness, vacation, religious holidays, and personal reasons.It helps to identify which reasons contribute the most to student absences, providing insights into patterns.The chart is ordered from the most frequent reason to the least frequent.

Predicting future absenteeism trends

```
absences_ts <- ts(school_data$Total_Absences, frequency = 1)
ab_model <- auto.arima(absences_ts)
predictions <- forecast(ab_model, h = 6)
```

Plot future absenteeism trend

```
plot(predictions, main = "Forecasted Absenteeism Trends", ylab = "Predicted Absences", xlab = "T
ime")
```



###

The time series forecast plot predicts absenteeism trends for the next 6 periods based on historical data.This helps in identifying potential future patterns and preparing interventions if absenteeism is expected to increase. The forecasted trend line shows the predicted rise or fall in absenteeism, helping administrators make data-driven decisions.

Recommendations for Stakeholders

1. Implement attendance monitoring systems and early interventions for high absenteeism.

2. Encourage students to participate in extracurricular activities to increase engagement.

3. Work with healthcare providers to support students with frequent illness absences.

4. Schedule important school activities at times when absenteeism is lower.

5. Educate parents and students on the importance of consistent attendance.

Conclusion:

The insights from this analysis highlight the significant impact absenteeism has on student performance, particularly for students with special educational plans. It’s clear that addressing absenteeism can be a key factor in improving GPA outcomes. By identifying the most common reasons for absenteeism and predicting future trends, this analysis offers actionable steps for schools to consider in managing and mitigating absenteeism. The data-driven approach provided by this study, using the dataset from Gloria, can help administrators better understand attendance patterns and take proactive measures to support students.

Reference:

Dataset of school absenteeism provided by Gloria.