ADS-506 Final Project - Baggage Complaints

Team 4: Sowmiya Kanmani Maruthavanan, Ben Ogle, Vicky van der Wagt

2023-11-20

Import libraries

```
knitr::opts_chunk$set(echo = TRUE)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

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

```
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(forecast)
```

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

```
library(zoo)
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(plotly)
```

```
##
## Attaching package: ’plotly’
```

```
## The following object is masked from ’package:ggplot2’:
##
##     last_plot
```

```
## The following object is masked from ’package:stats’:
##
##     filter
```

```
## The following object is masked from ’package:graphics’:
##
##     layout
```

```
library(gridExtra)
```

```
##
## Attaching package: ’gridExtra’
```

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

Pre-Processing

Import data

```
airlines_df <- read.csv("baggagecomplaints.csv")
head(airlines_df)
```

```
##          Airline   Date Month Year Baggage Scheduled Cancelled Enplaned
## 1 American Eagle Jan-04     1 2004   12502     38276      2481   992360
## 2 American Eagle Feb-04     2 2004    8977     35762       886  1060618
## 3 American Eagle Mar-04     3 2004   10289     39445      1346  1227469
## 4 American Eagle Apr-04     4 2004    8095     38982       755  1234451
## 5 American Eagle May-04     5 2004   10618     40422      2206  1267581
## 6 American Eagle Jun-04     6 2004   13684     39879      1580  1347303
```

```r
summary(airlines_df)
```

```
##    Airline              Date               Month           Year     
##  Length:252         Length:252         Min.   : 1.00   Min.   :2004  
##  Class :character   Class :character   1st Qu.: 3.75   1st Qu.:2005  
##  Mode  :character   Mode  :character   Median : 6.50   Median :2007  
##                                        Mean   : 6.50   Mean   :2007  
##                                        3rd Qu.: 9.25   3rd Qu.:2009  
##                                        Max.   :12.00   Max.   :2010  
##     Baggage        Scheduled       Cancelled          Enplaned      
##  Min.   : 1033   Min.   : 3553   Min.   :   0.00   Min.   : 423446  
##  1st Qu.: 1910   1st Qu.: 5566   1st Qu.:  25.75   1st Qu.: 686520  
##  Median :12224   Median :36696   Median : 533.00   Median :1391112  
##  Mean   :12614   Mean   :28128   Mean   : 703.76   Mean   :2203871  
##  3rd Qu.:19359   3rd Qu.:42162   3rd Qu.:1078.50   3rd Qu.:4111049  
##  Max.   :41787   Max.   :50837   Max.   :3712.00   Max.   :6137271  
```

```r
dim(airlines_df)
```

```
## [1] 252   8
```

The dataset contains 252 rows and 8 columns

Check for null values

```r
missing <- colSums(is.na(airlines_df))
print(missing)
```

```
##   Airline      Date     Month      Year   Baggage Scheduled Cancelled  Enplaned 
##         0         0         0         0         0         0         0         0 
```

There are no missing values in this dataset.

Check distribution of airlines

```r
airline_counts <- table(airlines_df$Airline)

# Display the counts
print(airline_counts)
```

```
## 
## American Eagle       Hawaiian         United 
##             84             84             84 
```

Convert categorical columns into factors

```r
#converting month and year to categorical
airlines_df$Month <- factor(airlines_df$Month)
airlines_df$Year <- factor(airlines_df$Year)
airlines_df$Airline <- factor(airlines_df$Airline)
```

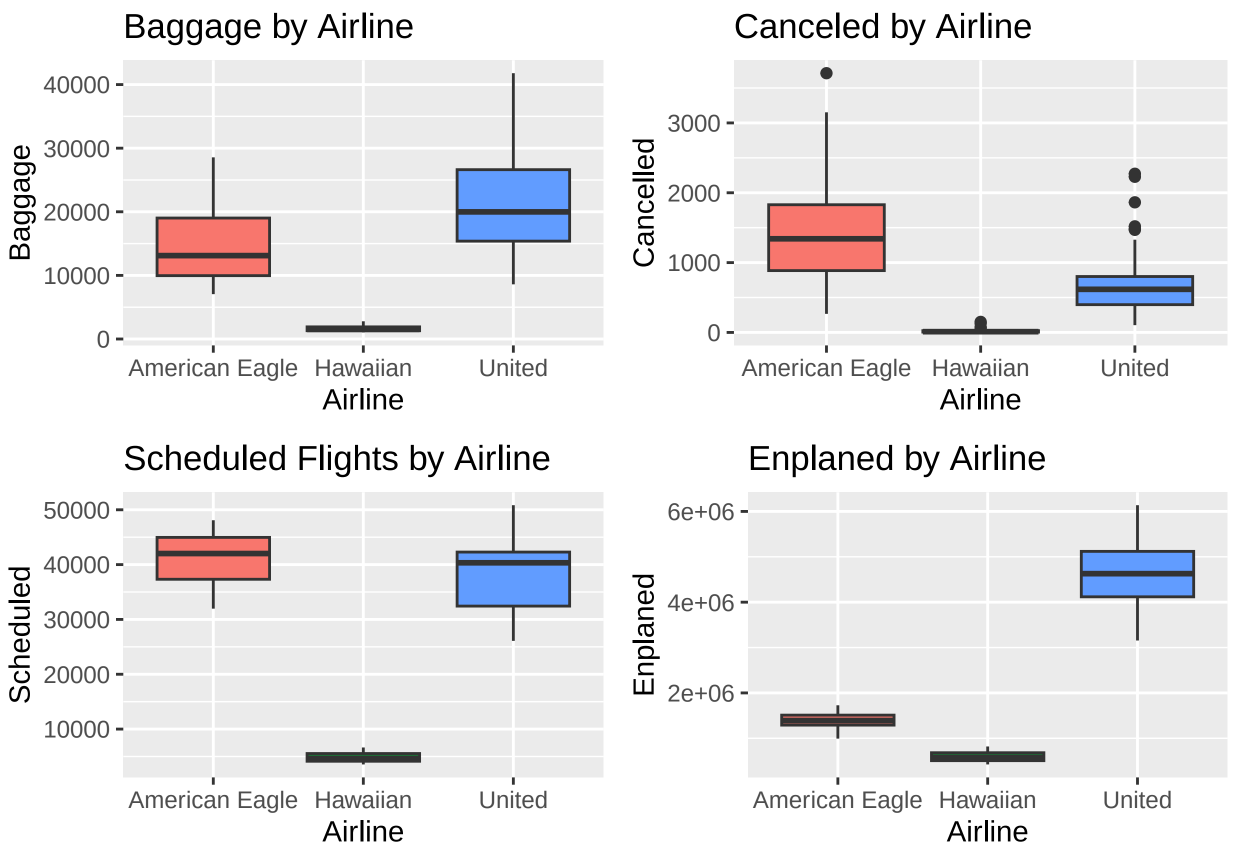
Transform data into time series

The time series starts in 2004 and ends in 2010. The frequency is set to 12 because the data is monthly and there are 12 months in a year.

```r
# Create time series for 3 airlines
american_eagle.ts <- ts(airlines_df[airlines_df$Airline == "American Eagle", "Baggage"],
                        start = c(2004,1), end = c(2010,12), freq = 12)

hawaiian.ts <- ts(airlines_df[airlines_df$Airline == "Hawaiian", "Baggage"],
                        start = c(2004,1), end = c(2010,12), freq = 12)

united.ts <- ts(airlines_df[airlines_df$Airline == "United", "Baggage"],
                        start = c(2004,1), end = c(2010,12), freq = 12)
```

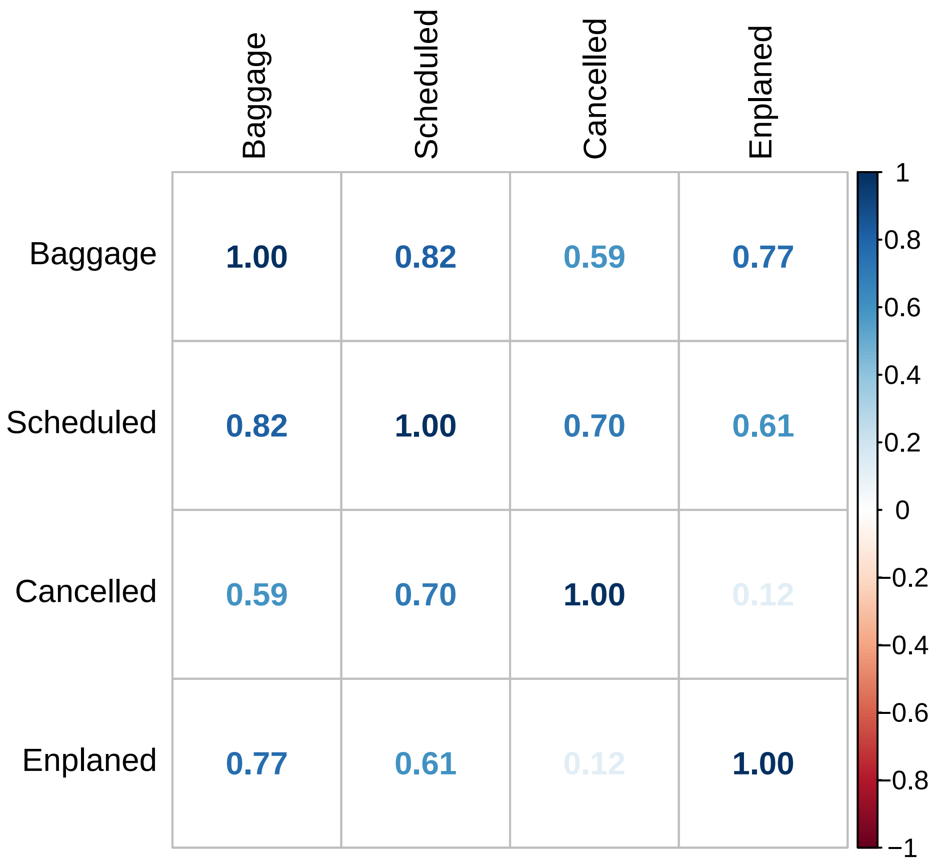
Check for outliers

```r
baggage_box <- ggplot(airlines_df, aes(x = Airline, y = Baggage, fill=Airline)) +
  geom_boxplot() +
  labs(title = "Baggage by Airline")+
  theme(legend.position = "none")

cancelled_box <- ggplot(airlines_df, aes(x = Airline, y = Cancelled, fill=Airline)) +
  geom_boxplot() +
  labs(title = "Canceled by Airline")+
  theme(legend.position = "none")

scheduled_box <- ggplot(airlines_df, aes(x = Airline, y = Scheduled, fill=Airline)) +
  geom_boxplot() +
  labs(title = "Scheduled Flights by Airline") +
  theme(legend.position = "none")


enplaned_box <- ggplot(airlines_df, aes(x = Airline, y = Enplaned, fill=Airline)) +
  geom_boxplot() +
  labs(title = "Enplaned by Airline")+
  theme(legend.position = "none")

grid.arrange(baggage_box, cancelled_box, scheduled_box, enplaned_box, ncol = 2)
```



Correlation plot

```
numerical <- airlines_df[sapply(airlines_df, is.numeric)]
cor_matrix <- cor(numerical)
corrplot(cor_matrix,
         method = "number",
         tl.col = "black")
```



	Baggage	Scheduled	Cancelled	Enplaned
Baggage	1.00	0.82	0.59	0.77
Scheduled	0.82	1.00	0.70	0.61
Cancelled	0.59	0.70	1.00	0.12
Enplaned	0.77	0.61	0.12	1.00

There are positive correlations between all the numerical variables ‘Baggage’, ‘Scheduled’, ‘Canceleled’, and ‘Enplaned’. High correlations exist between ‘Baggage’ and ‘Scheduled’, as well as ‘Baggage’ and ‘Enplaned.’ Moderate correlations exist between ‘Cancelled’ and ‘Baggage’, as well as ‘Bagged’ and ‘Enplaned.’ There is a very weak relationship between ‘Cancelled’ and ‘Emplaned.’

Feature Creation

```
airlines_df$Cancelled_prop <- (airlines_df$Cancelled / airlines_df$Scheduled) * 100
airlines_df$Flights <- airlines_df$Scheduled - airlines_df$Cancelled
airlines_df$Enplaned_per_flight <- airlines_df$Enplaned / airlines_df$Flights
airlines_df$Complaints_per_enplaned <- (airlines_df$Baggage / airlines_df$Enplaned) * 100
```

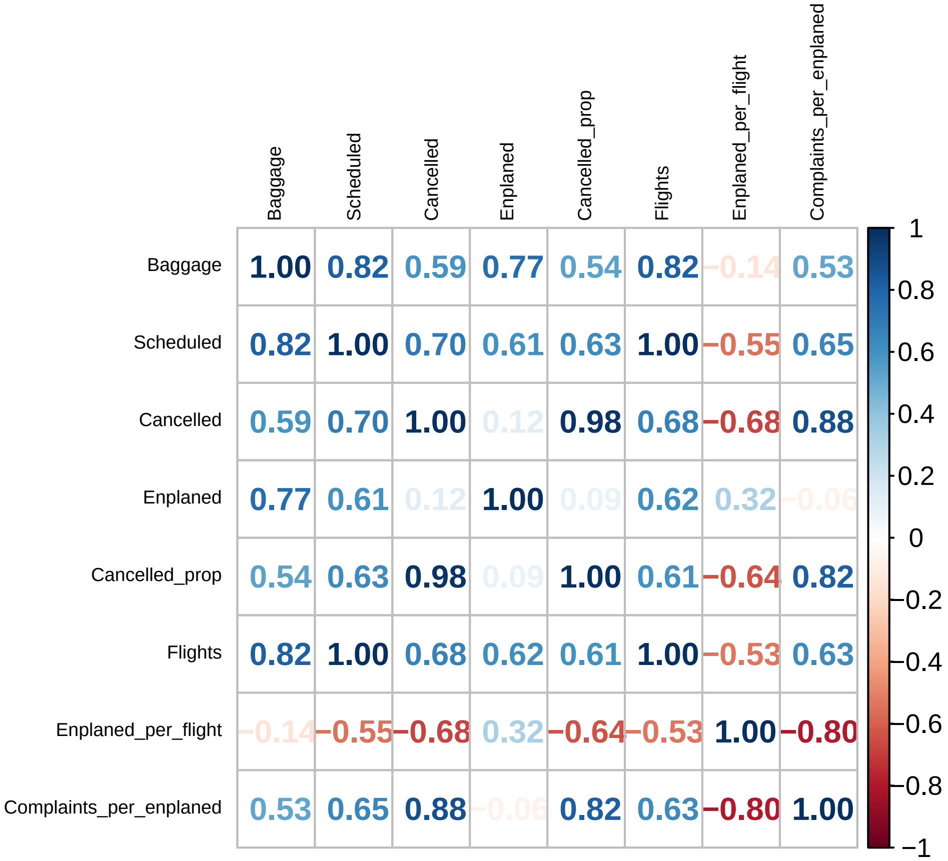
Created 4 new calculated features:

- Cancelled_prop: the proportion of cancelled flights
- Flights: the number of flights that were not cancelled
- Enplaned_per_flight: the number of enplaned passengers per flight
- Complaints per enplaned: the proportion of complaints over the total number of passengers

Correlation matrix after feature creation

View the correlation matrix after new features are created

```
numerical <- airlines_df[sapply(airlines_df, is.numeric)]
cor_matrix <- cor(numerical)
corrplot(cor_matrix,
         method = "number",
         tl.col = "black",
         tl.cex = 0.6)
```



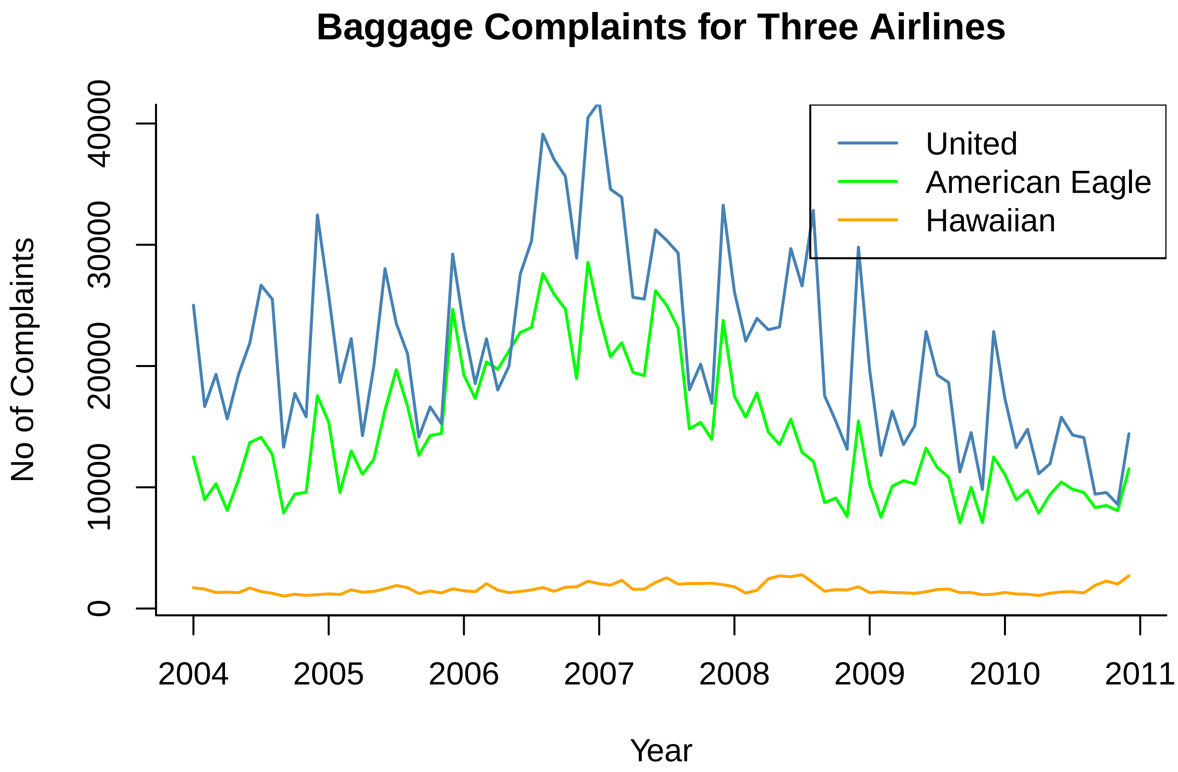
	Baggage	Scheduled	Cancelled	Enplaned	Cancelled_prop	Flights	Enplaned_per_flight	Complaints_per_enplaned
Baggage	1.00	0.82	0.59	0.77	0.54	0.82	−0.14	0.53
Scheduled	0.82	1.00	0.70	0.61	0.63	1.00	−0.55	0.65
Cancelled	0.59	0.70	1.00	0.12	0.98	0.68	−0.68	0.88
Enplaned	0.77	0.61	0.12	1.00	0.09	0.62	0.32	−0.06
Cancelled_prop	0.54	0.63	0.98	0.09	1.00	0.61	−0.64	0.82
Flights	0.82	1.00	0.68	0.62	0.61	1.00	−0.53	0.63
Enplaned_per_flight	−0.14	−0.55	−0.68	0.32	−0.64	−0.53	1.00	−0.80
Complaints_per_enplaned	0.53	0.65	0.88	−0.06	0.82	0.63	−0.80	1.00

The correlation plot after feature creation elicits some new insights:

- Complaints per enplaned indiviudal have a strong positive correlation with the proportion of cancelled flights
- There is a strong negative relationship between the number of enplaned individuals per flight, and the number of complaints per enplaned individual
- There is a moderately strong relationship between the proportion of cancelled flights and the number of scheduled flights
- There is a strong positive relationship between the number of flights and the number of baggage complaints

Exploratory Data Analysis

```
# create time series plot
plot(american_eagle.ts, xlab = "Year", ylab = "No of Complaints", ylim = c(1000, 40000), col="green",lw
     bty="l", main = "Baggage Complaints for Three Airlines")
lines(hawaiian.ts, col="orange", lwd = 1.5, bty="l")
lines(united.ts, col="steelblue", lwd = 1.5, bty="l")

# Add a legend
legend("topright", legend = c("United", "American Eagle", "Hawaiian"), col = c("steelblue", "green", "o
```

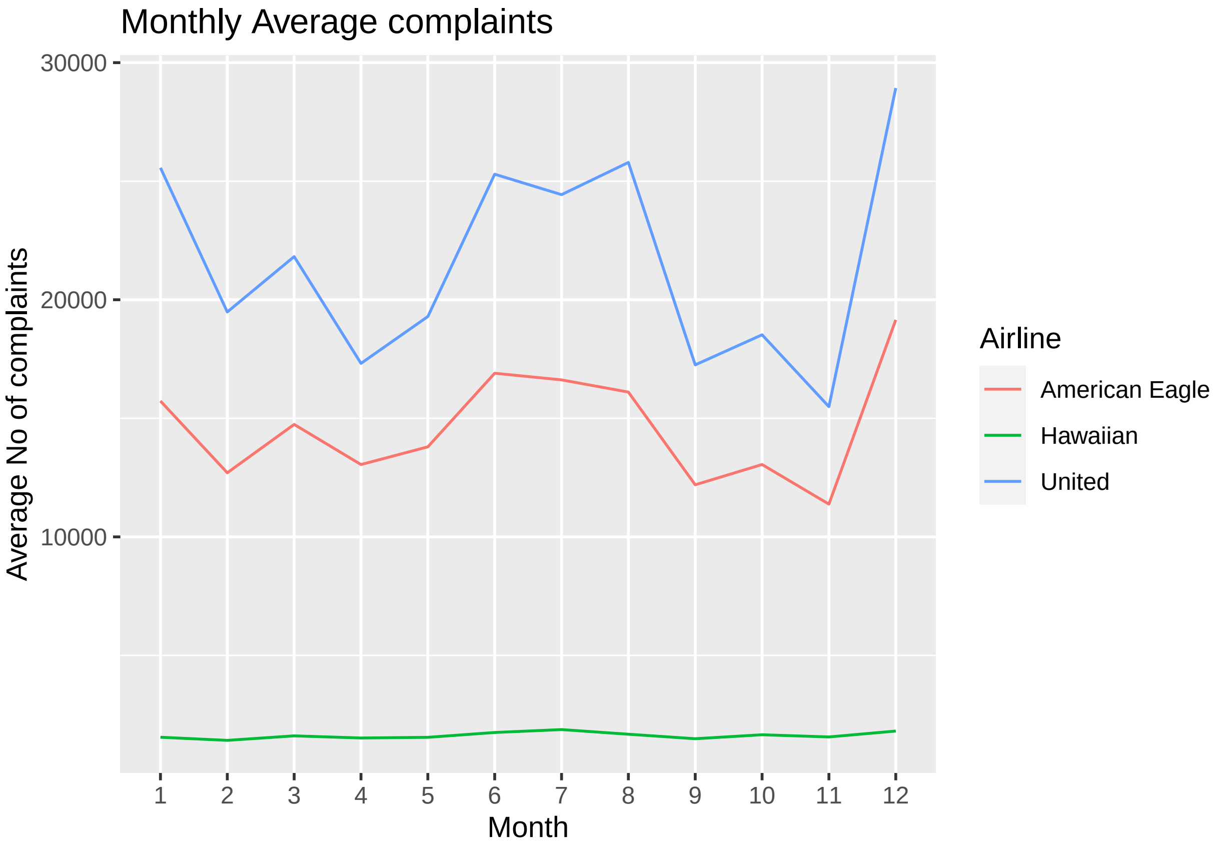


Throughout the dataset, United Airlines consistently receives the highest number of complaints regarding mishandled baggage each month, while Hawaiian Airlines consistently records the lowest number of complaints in every month.

The above plot shows a mild seasonality for American Eagle and United Airlines as there is a gradual increase in the baggage complaints at the beginning of each year. Additionally, it does not exhibit any trend as there is an increase and decrease in the baggage complaints for the three airlines.

Monthly Average Complaints for each airline

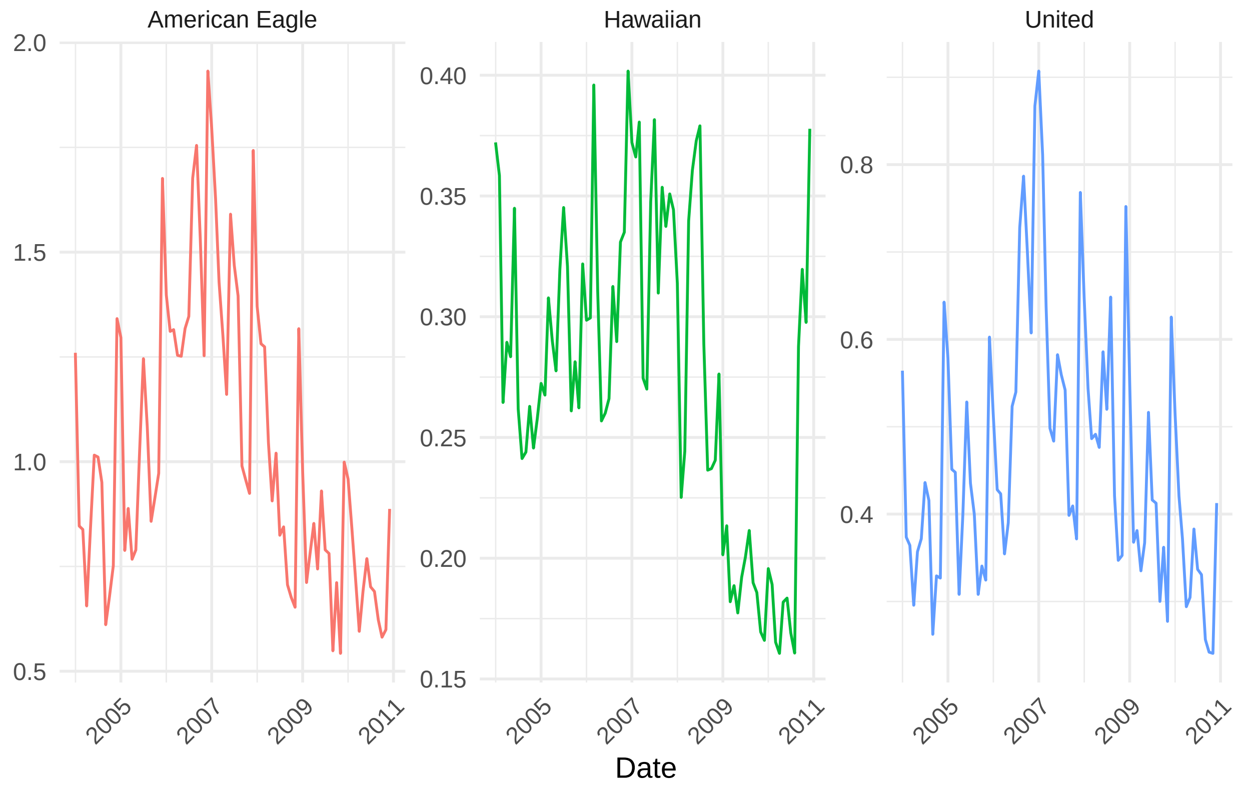
```
# Monthly average complaints for each airline
avg_complaints <- airlines_df %>%
  group_by(Airline, Month) %>% summarise_at(vars(Baggage), list(Avg_Complaints = mean))

ggplot(avg_complaints, aes(x = Month, y = Avg_Complaints)) +
  geom_line(aes(color = Airline, group = Airline)) +
  labs(title = 'Monthly Average complaints',
       x = 'Month',
       y = 'Average No of complaints')
```



Based on the above chart, on average, United consistently receives a higher average number of complaints in comparison to other airlines. Additionally, it is observed that the number of complaints shows an increase from May to August and from November to December. Furthermore, the trend for Hawaiian Airlines appears comparatively stable, with less fluctuation compared to the other airlines.

Proportion of complaints for each airline

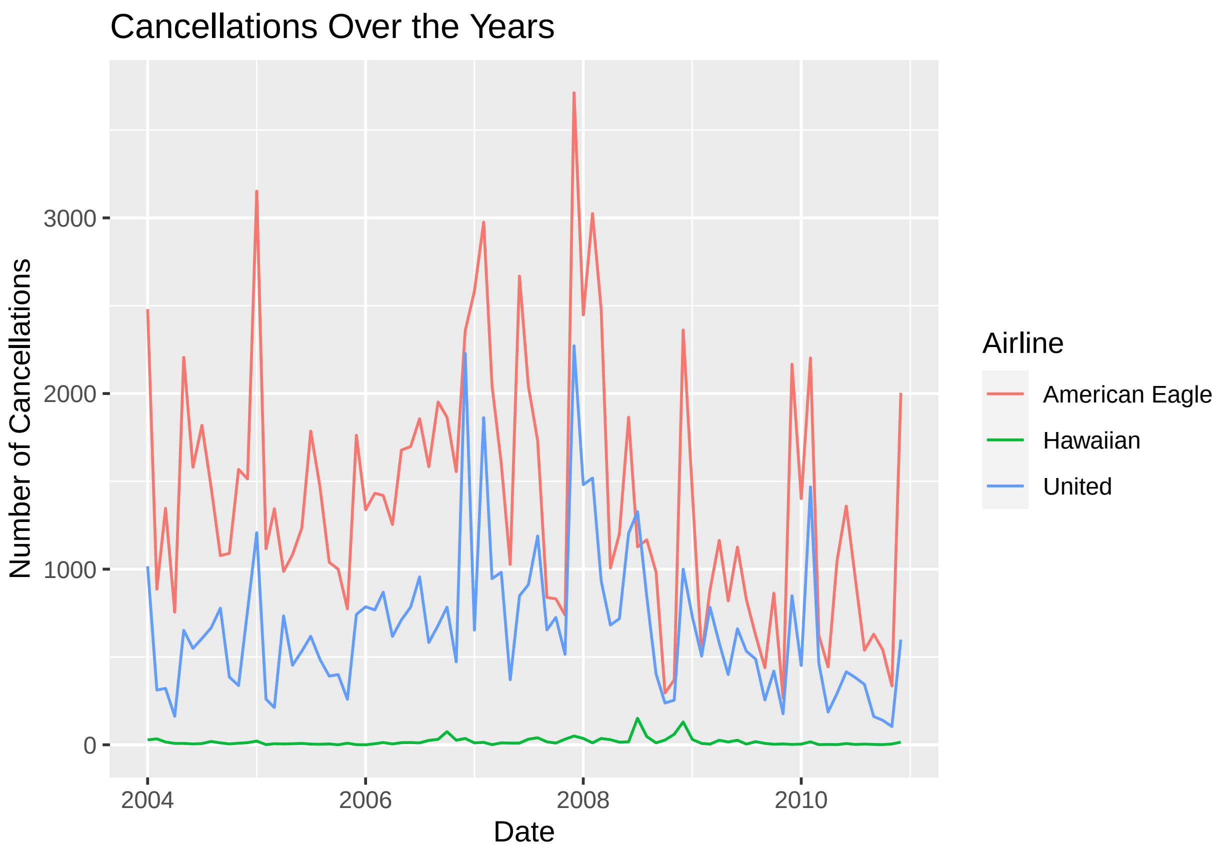
```
# Create a Date variable from the 'Date' column
airlines_df$Date <- as.Date(paste(airlines_df$Year, airlines_df$Month, "01", sep = "-"))

# Line plot for complaints for every month for each airline
ggplot(airlines_df, aes(x = Date, y = Complaints_per_enplaned, group = Airline, color = Airline)) +
  geom_line() +
  labs(title = "",
       x = "Date",
       y = "") +
  scale_x_date(date_labels = "%Y", date_breaks = "2 years") +
  theme_minimal() +
  facet_wrap(~Airline, scales = "free_y") +  # Separate plots for each airline
  theme(
    axis.title.y = element_blank(),
    legend.position = "none",
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```



Upon examining the individual charts for each airline, it becomes evident that there was a rise in trends from 2004 to 2006, followed by a subsequent decline. Moreover, the plot highlights that Hawaiian Airlines had a big increase in complaints in 2010, while other airlines had a more steady rise in the number of complaints.

Explore cancellations over the years

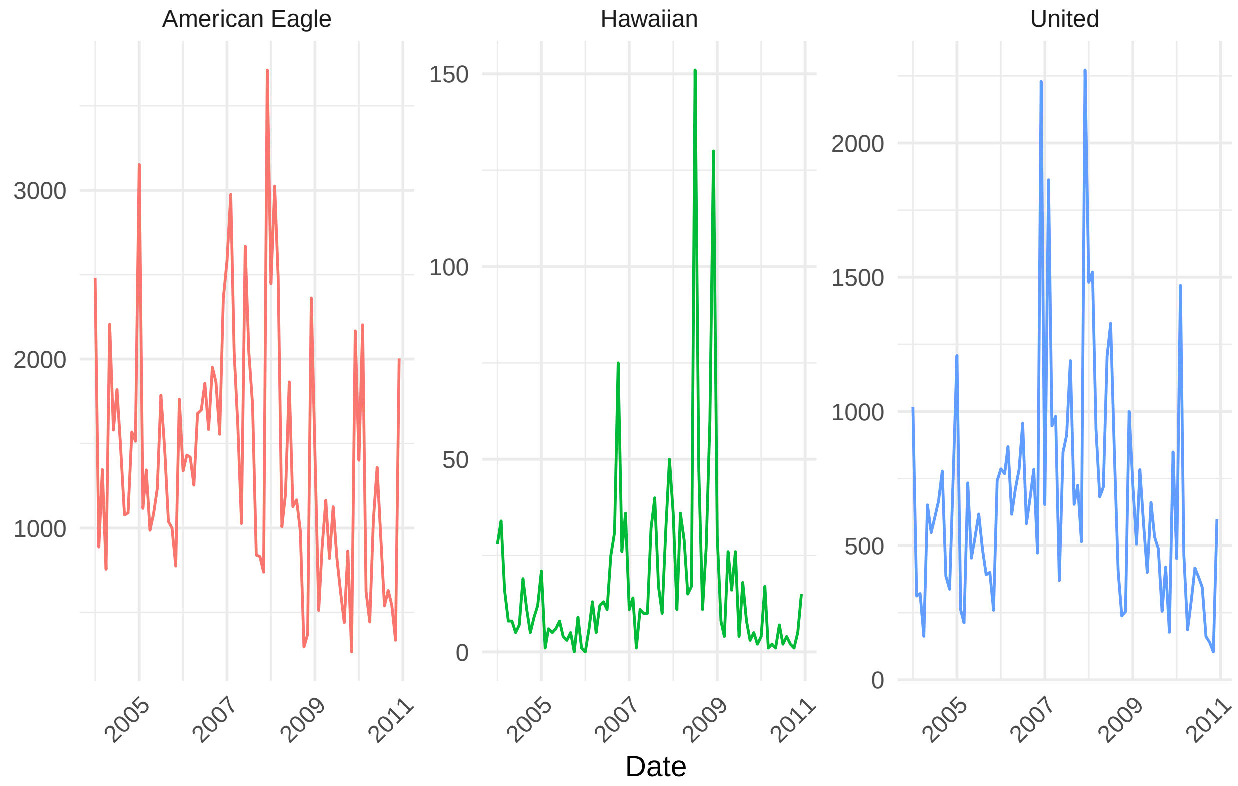
```
# Explore cancellations over the years
ggplot(airlines_df, aes(x = as.Date(paste(Year, Month, "01", sep = "-")), y = Cancelled, group = Airlin
  geom_line() +
  labs(title = "Cancellations Over the Years",
       x = "Date",
       y = "Number of Cancellations")
```



The above plot illustrates the annual trend of cancelled flights. It is observed that American Eagle has the highest count of cancelled flights and highest frequency of cancellations occur in January.

Explore the trend in the Cancelled flights for each airline

```r
# Create a Date variable from the 'Date' column
airlines_df$Date <- as.Date(paste(airlines_df$Year, airlines_df$Month, "01", sep = "-"))

# Line plot of cancelled flights for each airline
ggplot(airlines_df, aes(x = Date, y = Cancelled, group = Airline, color = Airline)) +
  geom_line() +
  labs(title = "",
       x = "Date",
       y = "") +
  scale_x_date(date_labels = "%Y", date_breaks = "2 years") +
  theme_minimal() +
  facet_wrap(~Airline, scales = "free_y") +  # Separate plots for each airline
  theme(
    axis.title.y = element_blank(),
    legend.position = "none",
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```



In the provided graph, American Eagle and United share a similar pattern in the number of cancelled flights over the years, while Hawaiian Airlines demonstrates relatively fewer ups and downs. Hawaiian flight cancellations reached their highest point in 2008 and started to decrease afterward.

Explore monthly Cancelled proportions

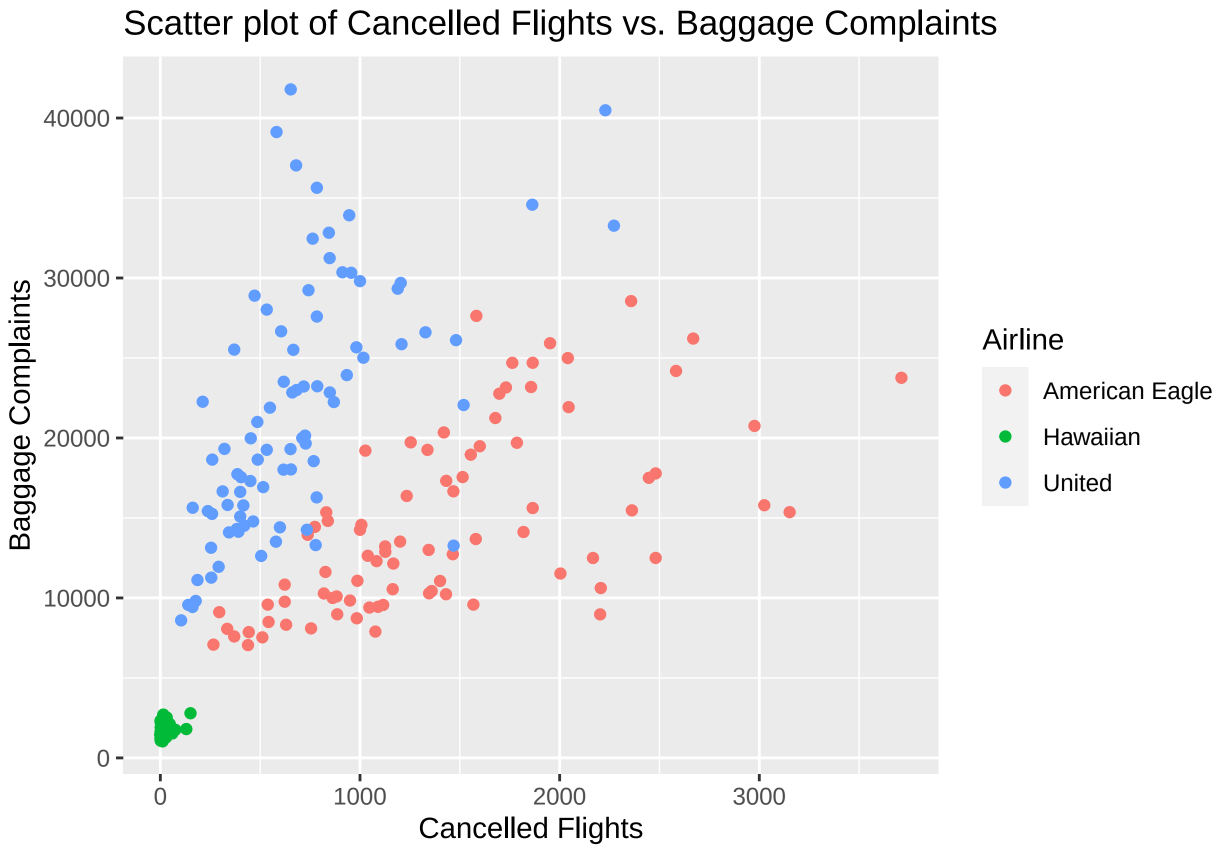
```
# Working on this

# ggplot(airlines_df, aes(x = Month, y = Cancelled_prop, color = Airline)) +
#   geom_line() +
#   theme_minimal() +
#   labs(title = "Cancelled Proportions of Each Airline Over Months",
#        x = "Month",
#        y = "") +
#   scale_color_discrete(name = "Airline") +
#   theme(legend.position = "top")
```

Relationship between Baggage Complaints and Cancelled Flights

```
# Calculate correlation coefficient
corr_coeff <- cor(airlines_df$Cancelled, airlines_df$Baggage)
cat("Correlation coefficient between Cancelled Flights and Baggage Complaints:", corr_coeff, "\n")
```

```
## Correlation coefficient between Cancelled Flights and Baggage Complaints: 0.5944247
```

```
# Scatter plot of Baggage vs. Cancelled
ggplot(airlines_df, aes(x = Cancelled, y = Baggage, color=Airline)) +
  geom_point() +
  labs(title = "Scatter plot of Cancelled Flights vs. Baggage Complaints",
       x = "Cancelled Flights",
       y = "Baggage Complaints")
```



A correlation coefficient of 0.5944 indicates a moderate positive correlation between the number of cancelled flights and the number of baggage complaints.

Based on the plot, it is observed that there is a general upward trend indicating that there is a tendency for higher baggage complaints when there are more cancelled flights. Moreover, it is important to note that correlation does not imply causation. While there is a statistical association between cancelled flights and baggage complaints, it doesn't necessarily mean that one causes the other. There could be other factors influencing both variables.

Modeling