

# DSO 545: HW 1

*Bradley Rava, Patrick Vossler, Simeng Shao*

*1/27/2019*

Load the data:

```
baggage = read.csv(here("HW1", "Baggage.csv"), header=T, stringsAsFactors = F)
indus_med = read.csv(here("HW1", "IndustryMedians.csv"), header=T)
head(baggage)
```

##	Airline	Date	Month	Year	Baggage	Scheduled	Cancelled	Enplaned
## 1	American Eagle	01/2004	1	2004	12502	38276	2481	992360
## 2	American Eagle	02/2004	2	2004	8977	35762	886	1060618
## 3	American Eagle	03/2004	3	2004	10289	39445	1346	1227469
## 4	American Eagle	04/2004	4	2004	8095	38982	755	1234451
## 5	American Eagle	05/2004	5	2004	10618	40422	2206	1267581
## 6	American Eagle	06/2004	6	2004	13684	39879	1580	1347303

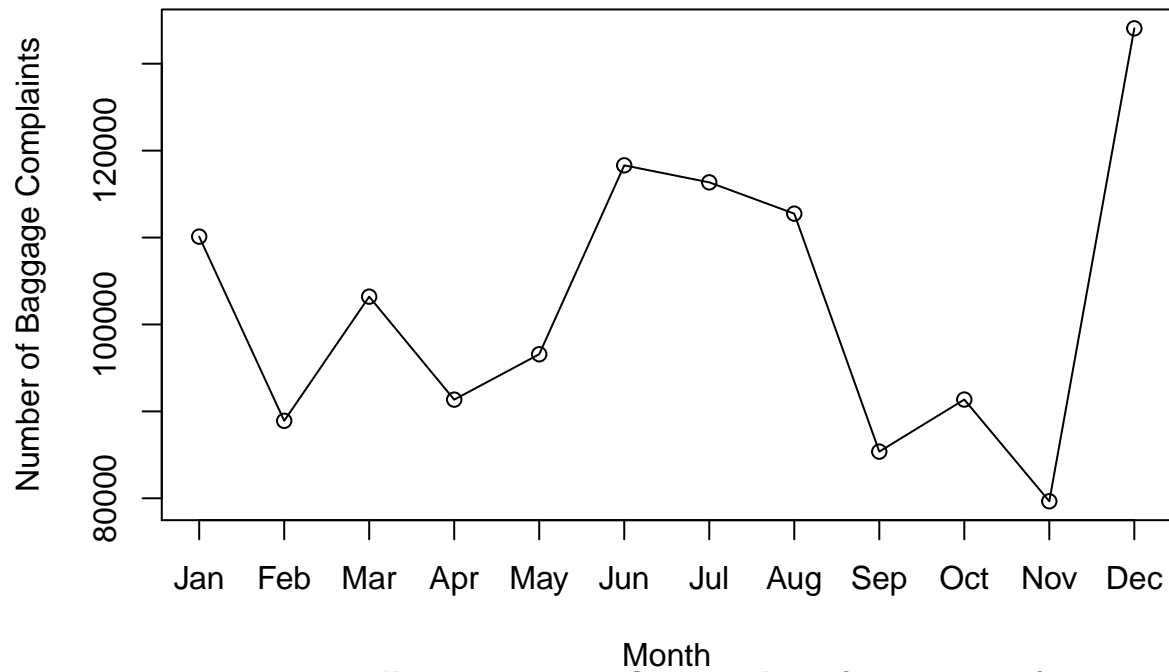
Process data:

```
baggage$Date = as.Date(paste0("02/", baggage$Date), "%d/%m/%Y")
baggage$Month = factor(baggage$Month, labels=c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))
baggage$Airline = as.character(baggage$Airline)
```

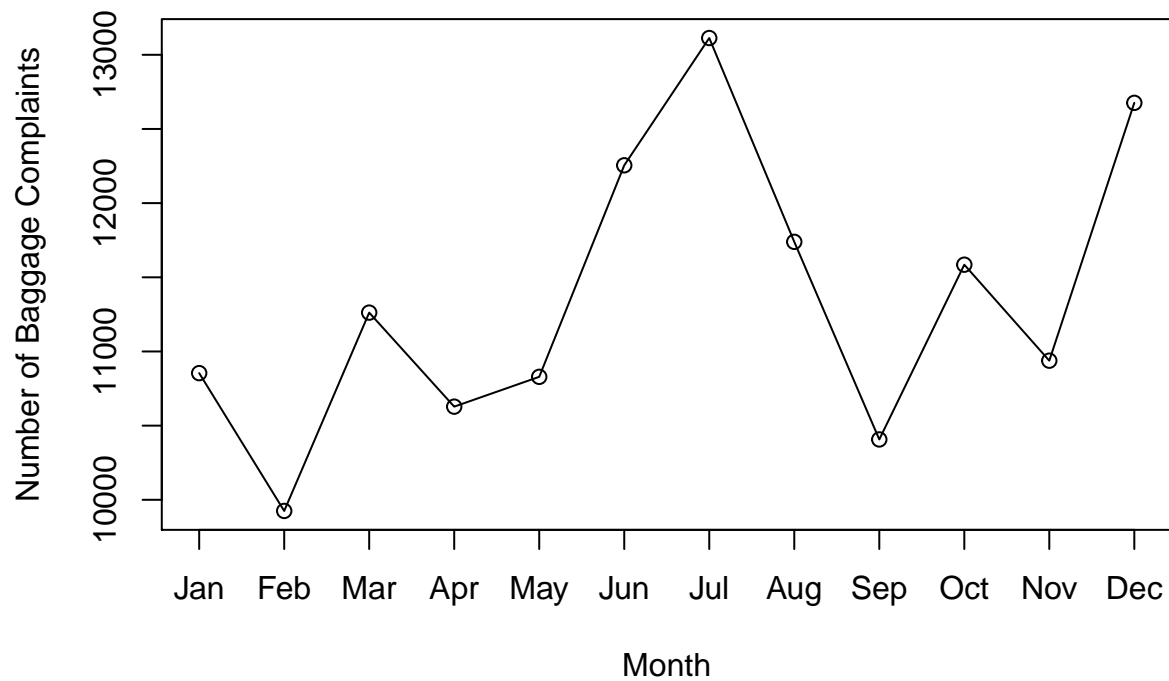
1. Explore baggage complaints over time: create 3 time series plots for the variable *Baggage* by Date for each of the airlines separately.

```
airlines = unique(baggage$Airline)
for(i in 1:length(airlines)){
  airline = airlines[i]
  data = baggage[baggage$Airline == airline,]
  res = aggregate(data["Baggage"], by=list(Month = data$Month), sum)
  plot(x=as.integer(res$Month), y=res$Baggage, type="o", xaxt="n", xlab="Month", ylab="Number of Baggage Complaints",
       axis(1, at = seq(1,12), labels = levels(res$Month)))
  title(paste(airline, "Baggage Complaints (2004-2010)"))
}
```

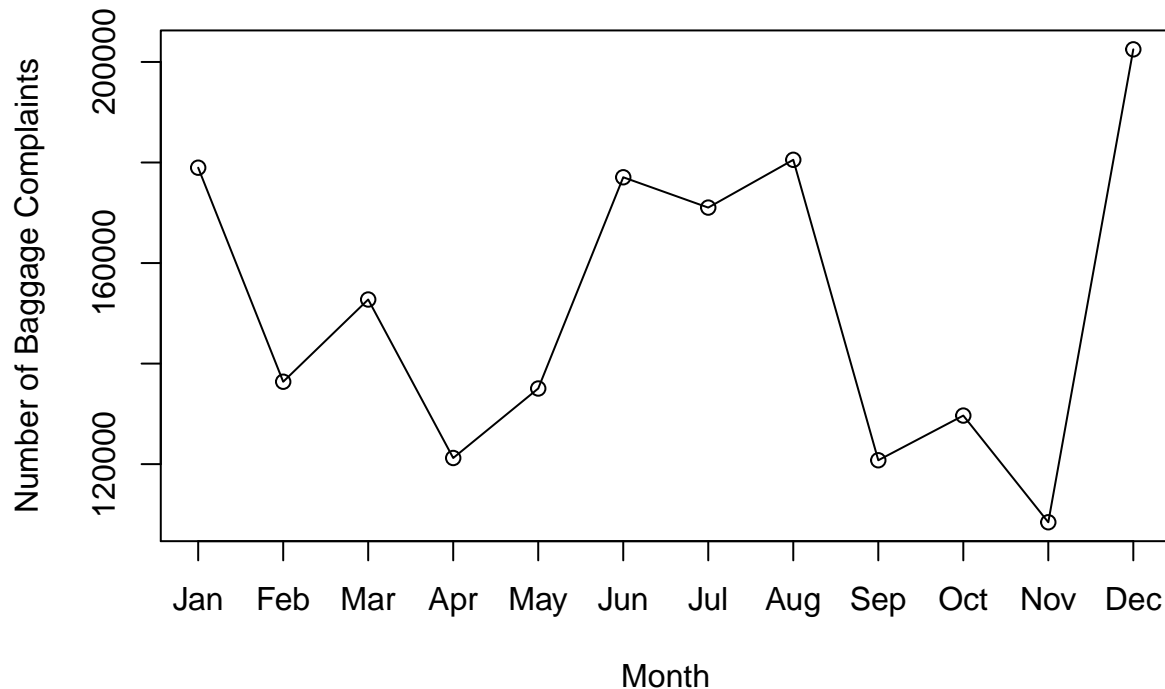
### American Eagle Baggage Complaints (2004–2010)



### Hawaiian Baggage Complaints (2004–2010)



## United Baggage Complaints (2004–2010)



### 2. Briefly describe what patterns you see in the plots

In some of the plots we see a cyclical pattern with the number of baggage complaints increasing during the winter holiday travel season (November-January). There is often another spike in baggage complaints in the summer likely when families are going on summer vacations.

- American Eagle
  - We see that the cyclical yearly trend described above holds for American Eagle. Furthermore we see that there is an increase in the total number of complaints in 2006-2008 and then the number of complaints drops back down from 2009 onward.
- Hawaiian Airlines
  - Compared to American Eagle, Hawaiian Airlines has a smaller number of complaints each month. This is expected because Hawaiian Airlines is a smaller airline compared to American Eagle. Whereas American Eagle had a spike in baggage complaints during the winter holiday travel season, Hawaiian Airlines seems to have spikes in baggage complaints during the Spring and Summer. This perhaps could be because they see an influx of passengers wishing to travel to Hawaii during the Spring and Summer months.
  - The most concerning trend for Hawaiian Airlines is the trend of larger spikes in each of the successive years, culminating with a large spike in baggage complaints during the 2010 holiday season.
- United Airlines
  - Unsurprisingly United Airlines has a larger number of baggage complaints overall which can be explained by its much larger size compared to the other two companies.
  - Like American Eagle we see that United Airlines also experiences a surge in baggage claims during the holiday season. Additionally, it is interesting that both American Eagle and United Airlines have a spike in baggage complaints during 2006. Perhaps there was some external event that caused this for both airlines?
  - Since both American Eagle and United Airlines provide a variety of flights to domestic destinations it is not surprising to see that they have similar baggage complaint patterns in the summer and

winter months.

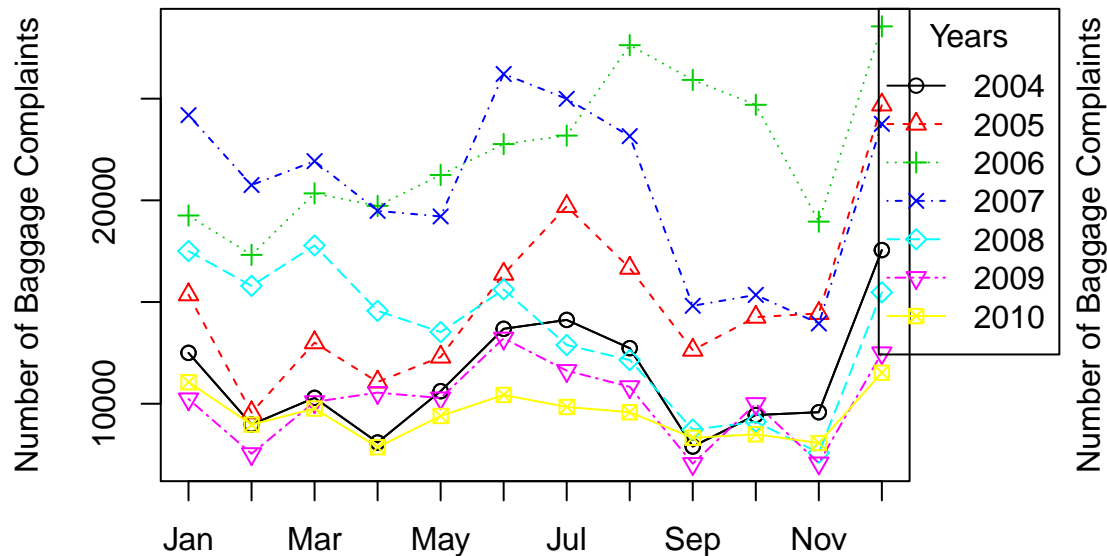
3.

```
airlines = unique(baggage$Airline)
for(i in 1:length(airlines)){
  airline = airlines[i]
  data = baggage[baggage$Airline == airline,]
  res = aggregate(data["Baggage"], by=list(Month = data$Month, Year = data$Year), sum)
  years = unique(data$Year)
  plot_dat = res[res$Year == years[1],]

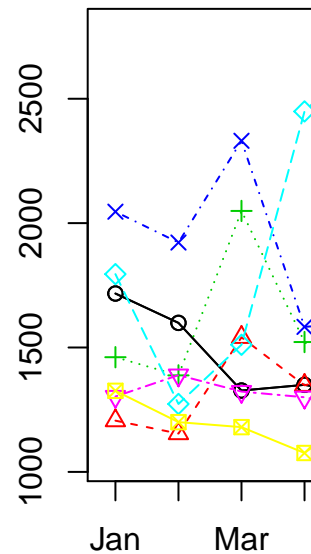
  #bottom, left, top, right margin
  par(mar=c(7.1, 4.1, 3.1, 8.9), xpd=TRUE)

  plot(x=as.integer(plot_dat$Month), y=plot_dat$Baggage, type="o", xaxt="n", xlab="", ylab="Number of Baggage Complaints",
       axis(1, at = seq(1,12), labels = levels(res$Month))
       title(paste(airline, "Baggage Complaints"))
       for(j in 1:length(years)){
         plot_dat = res[res$Year == years[j],]
         lines(x=as.integer(plot_dat$Month), y=plot_dat$Baggage, type="o", lty=j, col=j, pch=j)
       }
       legend("topright", inset=c(-0.2,0), legend=years, pch=1:length(years), lty=1:length(years), col=1:length(years))
}
}
```

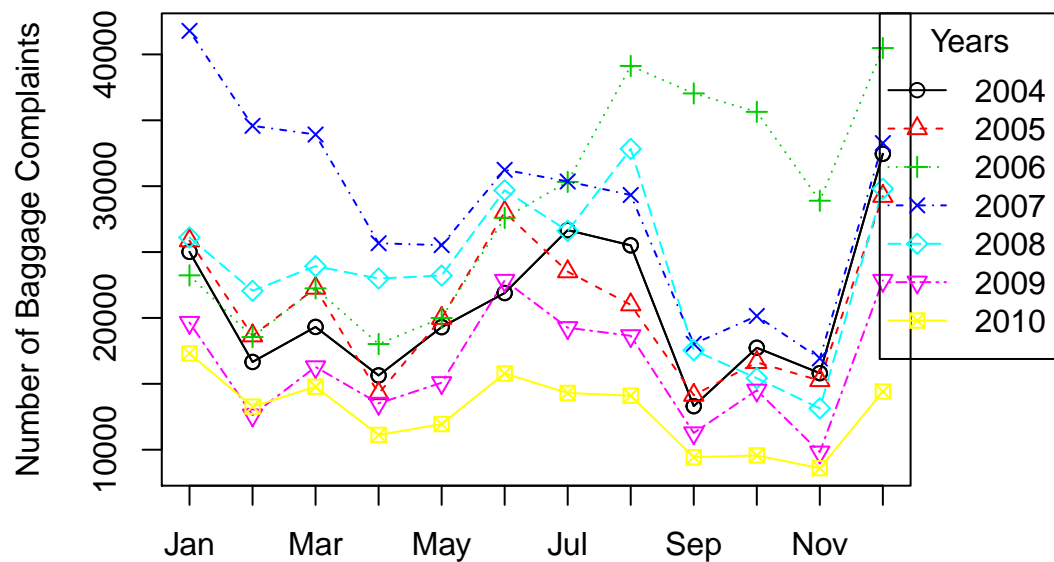
**American Eagle Baggage Complaints**



**Hawaiian Airlines Baggage Complaints**



## United Baggage Complaints



4. Describe the patterns in the plot

5. Plot all three airline Baggage data by Date on one graph.

```
# Maybe do this on the log scale?
airlines = unique(baggage$Airline)
airline = airlines[1]

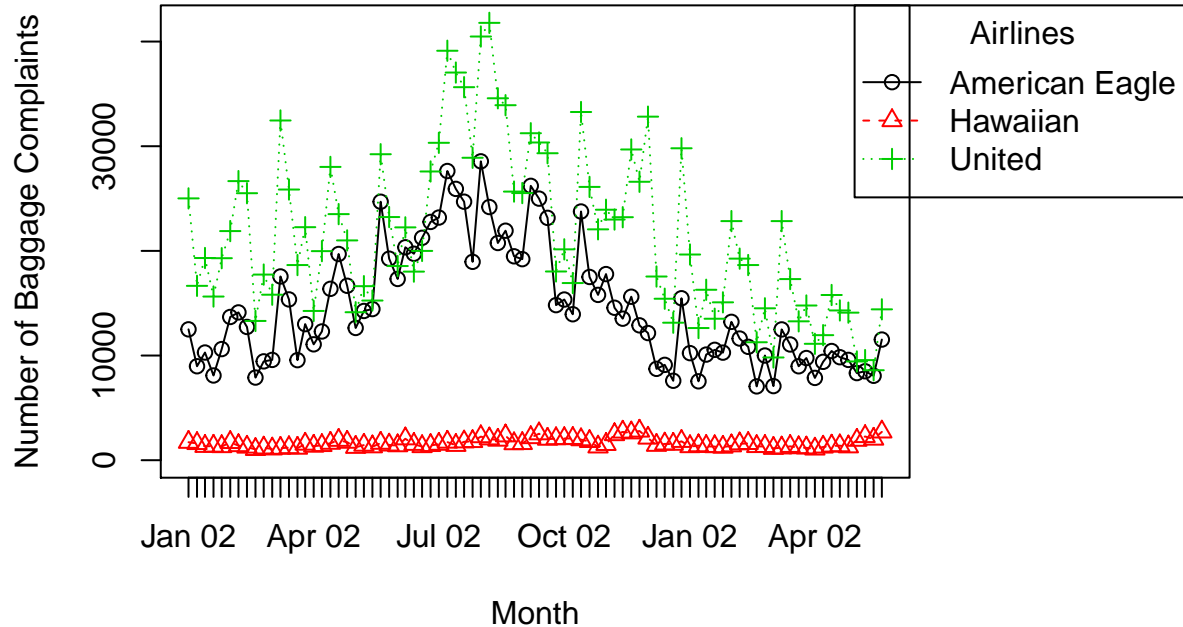
total_aggregated = aggregate(baggage["Baggage"], by=list(Date = baggage$Date,Airline=baggage$Airline),
res = total_aggregated[baggage$Airline == airline,]

par(mar=c(7.1, 4.1, 3.1, 8.9), xpd=TRUE)

plot(x=res$Date,y=res$Baggage,type="o",xaxt="n",xlab="Month", ylab="Number of Baggage Complaints",lty=1)

axis(1, res$Date, format(res$Date, "%b %d"))
title("Baggage Complaints for all 3 Airlines (2004-2010)")
for(i in 2:length(airlines)){
  airline = airlines[i]
  res = total_aggregated[baggage$Airline == airline,]
  lines(x=res$Date,y=res$Baggage,type="o",lty=i, col=i,pch=i)
}
legend("topright", inset=c(-0.375,0), legend=airlines, pch=1:length(airlines),lty=1:length(airlines),col=1:length(airlines))
```

## Baggage Complaints for all 3 Airlines (2004–2010)



6. Based on the graph in question 5., do some airlines have better baggage handling practices?

According to the plot, Hawaiian line seems to have much smaller baggage complaints throughout 2004-2010, which has been below 50000. Other two lines, however, are above 50000.

7. Based on the graph in question 5., which airline has the best record? The worst?

Based on the graph, Hawaiian has the best record, United has the worst record.

8. Based on the graph in question 5., are complaints getting better or worse over time?

There is no clear pattern that the curves are going up or down, in fact they all once increase and fluctuate back to the level where they started with. So based on the graph the complaints are not getting better nor worse.

9. Are the conclusions, you have drawn based on the graphs of the raw data you created, accurate? Are there any potential factors that may distort your conclusions and should be taken into consideration?

The conclusions are not necessarily accurate since we only looked at the number of baggage complaints of the three airlines. Chances are that Hawaiian is a smaller airline and have way fewer passengers than United or American Eagle. So we look at the ratio of ( $\#$  of complaints)/( $\#$  of boarded passengers), i.e., “baggage”/“enplaned” in our dataset.

10. Report the average of scheduled flights and the average of enplaned passengers by airline.

```
mean_scheduled = rep(0, length(unique(airlines)))
names(mean_scheduled) = unique(airlines)
for(i in 1:length(unique(airlines)))
{
  mean_scheduled[i] = mean(baggage[baggage$Airline == airlines[i],6])
}

mean_enplaned = rep(0, length(unique(airlines)))
names(mean_enplaned) = unique(airlines)
for(i in 1:length(unique(airlines)))
{
  mean_enplaned[i] = mean(baggage[baggage$Airline == airlines[i],8])
}
```

The average of scheduled flights are:

```
mean_scheduled
```

## American Eagle	Hawaiian	United
## 41314.048	4844.679	38225.298

The average of enplaned passengers are:

```
mean_enplaned
```

## American Eagle	Hawaiian	United
## 1396725.5	594174.2	4620712.3

11. What insights, ideas, and concerns does the data in the table in 10. provide you with?

The number of scheduled planes and enplaned passengers of United and Hawaiian are not on the same scale. Again this confirms our concern in problem 9 that simply looking at the number of complains is not fair for assessing the baggage handling practices of these companies.

12. Create Baggage % KPI that adjusts the total number of passenger complaints for size

```
baggage$Baggage_perc = baggage$Baggage / baggage$Enplaned * 100

mean_kpi = rep(0, length(unique(airlines)))
names(mean_kpi) = unique(airlines)
for(i in 1:length(unique(airlines)))
{
  mean_kpi[i] = mean(baggage[baggage$Airline == airlines[i],9])
}
```

The average Baggage % for each airline are:

```
for(i in 1: length(unique(airlines)))
  print(paste(unique(airlines), round(mean_kpi*100,2),"%")[i])
```

```
## [1] "American Eagle 103.3 %"
## [1] "Hawaiian 27.71 %"
## [1] "United 46.41 %"
```

13. Do the results in question 12 support your previous conclusions? Briefly explain.

The results in question 12 show that Hawaiian has the lowest **Baggage %**, United is the second; while American Eagle has the highest **Baggage %**. This result contradicts with our previous conclusions in that the worst baggage handling records belongs to American Eagle instead of United.

14. Superimpose all three time series on one graph to display Baggage % by Date.

```
airlines = unique(baggage$Airline)
airline = airlines[1]

perc_aggregated = aggregate(baggage["Baggage_perc"], by=list(Date = baggage$Date,Airline=baggage$Airline),
                             FUN=mean)

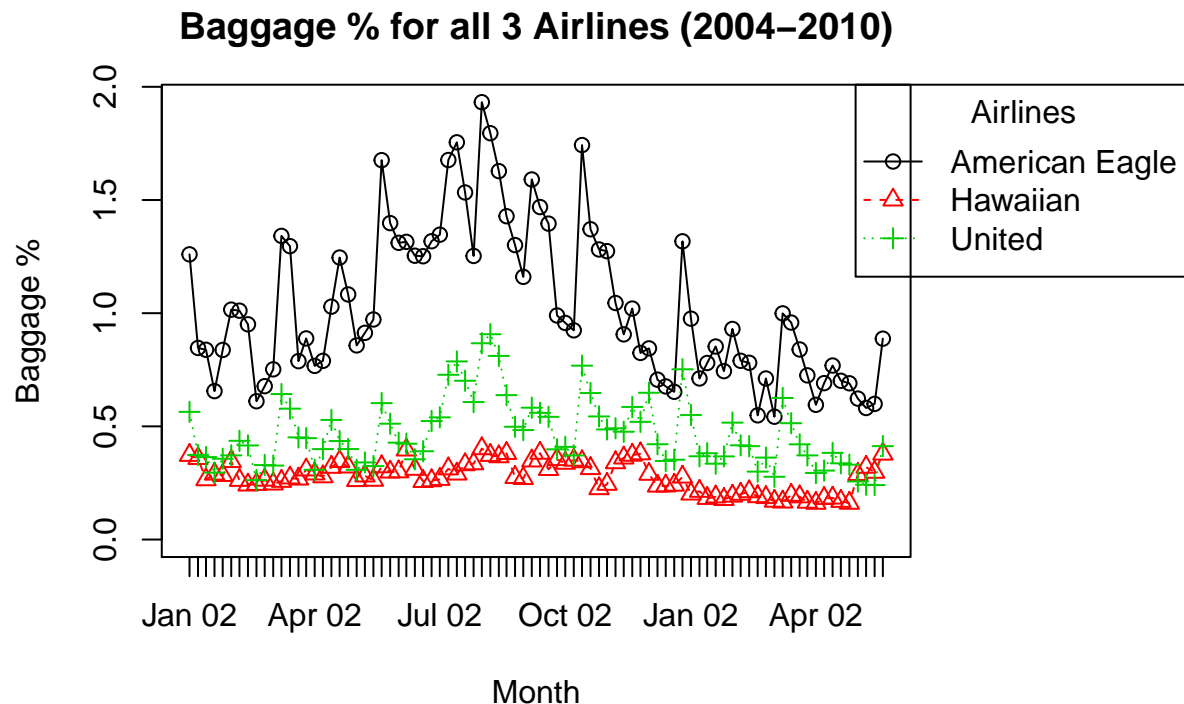
res1 = perc_aggregated[baggage$Airline == airline,]

par(mar=c(7.1, 4.1, 3.1, 8.9), xpd=TRUE)

plot(x=res1$Date,y=res1$Baggage_perc,type="o",xaxt="n",xlab="Month", ylab="Baggage %",lty=1, col=1, pch=1)

axis(1, res1$Date, format(res1$Date, "%b %d"))
title("Baggage % for all 3 Airlines (2004-2010)")
for(i in 2:length(airlines)){
  airline = airlines[i]
  res1 = perc_aggregated[baggage$Airline == airline,]
  lines(x=res1$Date,y=res1$Baggage_perc,type="o",lty=i, col=i,pch=i)
}
legend("topright", inset=c(-0.375,0), legend=airlines, pch=1:length(airlines),lty=1:length(airlines),col=1:length(airlines))
```





15. In addition to the graph in question 14., would plotting each series on a separate graph be beneficial and why? Create a graph to support your answer.

16. Based on the analysis of KPI Baggage %, have any of your conclusions drawn in questions 6. - 8. changed? Briefly discuss.

17. Superimpose time series plots of monthly averages of Baggage % by time for the three airlines

Plot the mean of each month for all of the years

18. Discuss common patterns all three time series exhibit in question 17.

19. Create a timeplot of Baggage %, add average line for Baggage % and a trendline of monthly average Baggage % for each airline.

Create monthly averages over total data and plot lm

20. Prepare a brief (one paragraph) executive summary of your findings.

## Case 2: CEO Compensation

Load CEO compensation data

```
#ceo_comp = read.table(here("HW1", "CEOcompensation.txt"), header=T, sep = "\t")
```

1. What is the number of female CEOs?