

# DSO 545: HW 1

*Bradley Rava, Patrick Vossler, Simeng Shao*

*1/27/2019*

## Case 1: Baggage Data

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

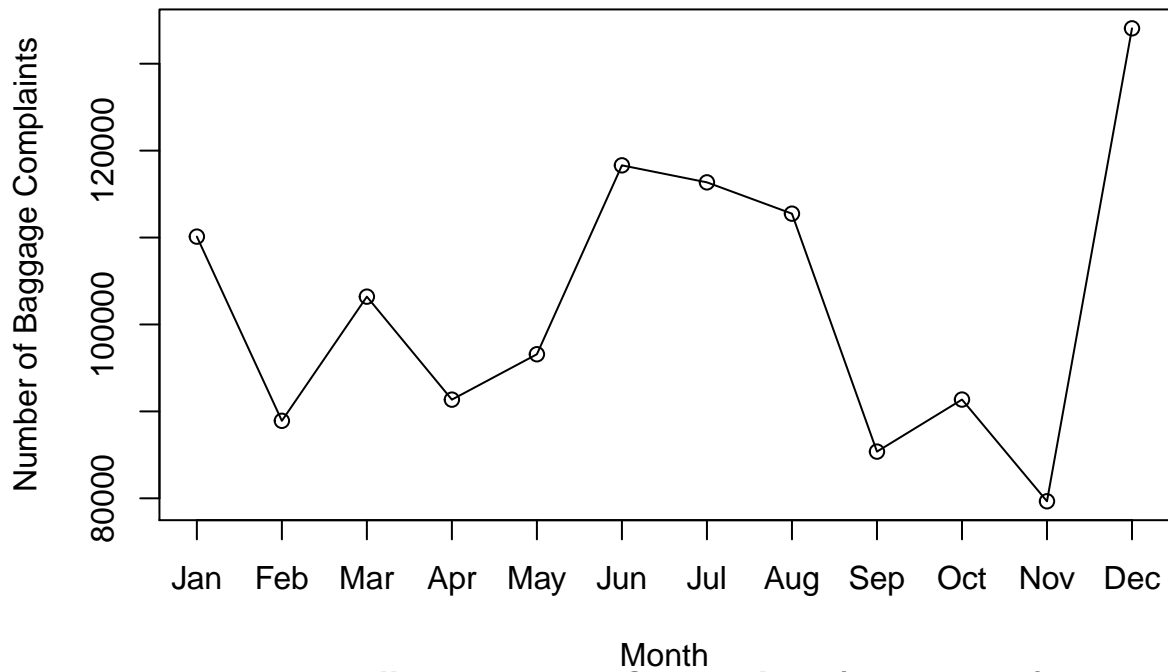
Process data:

```
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)
```

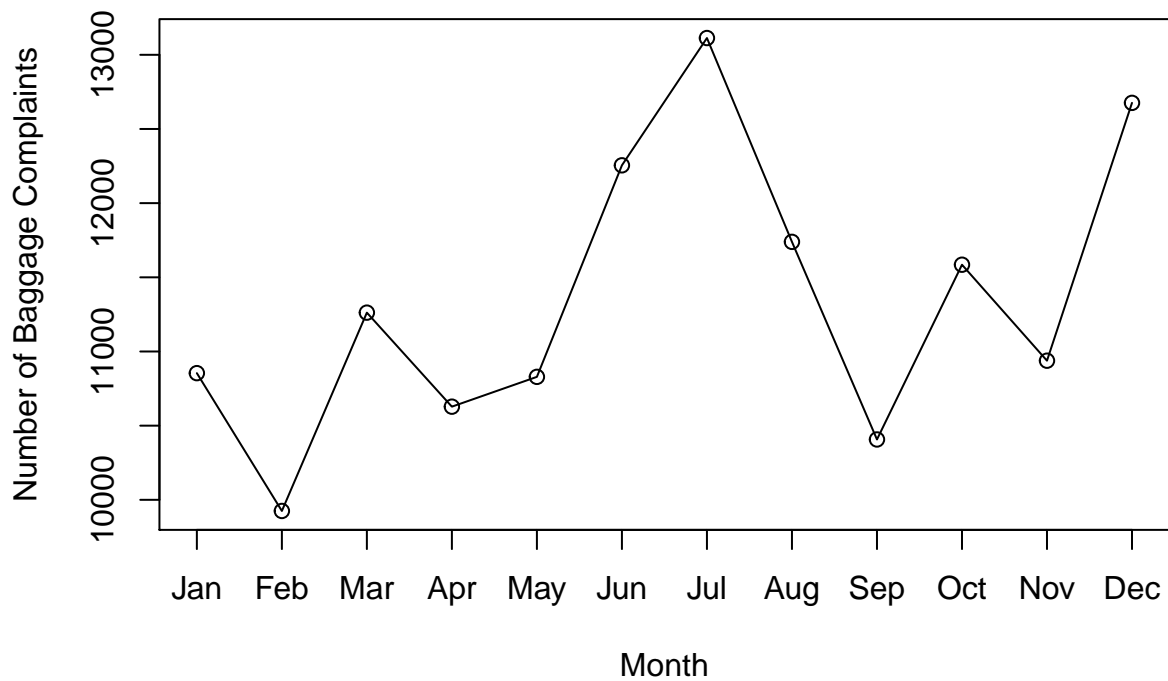
1. Explore baggage complaints over time: create 3 time series plots for the variable *Baggage* by Date for each of the airlines separately.

```
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)"))
}
```

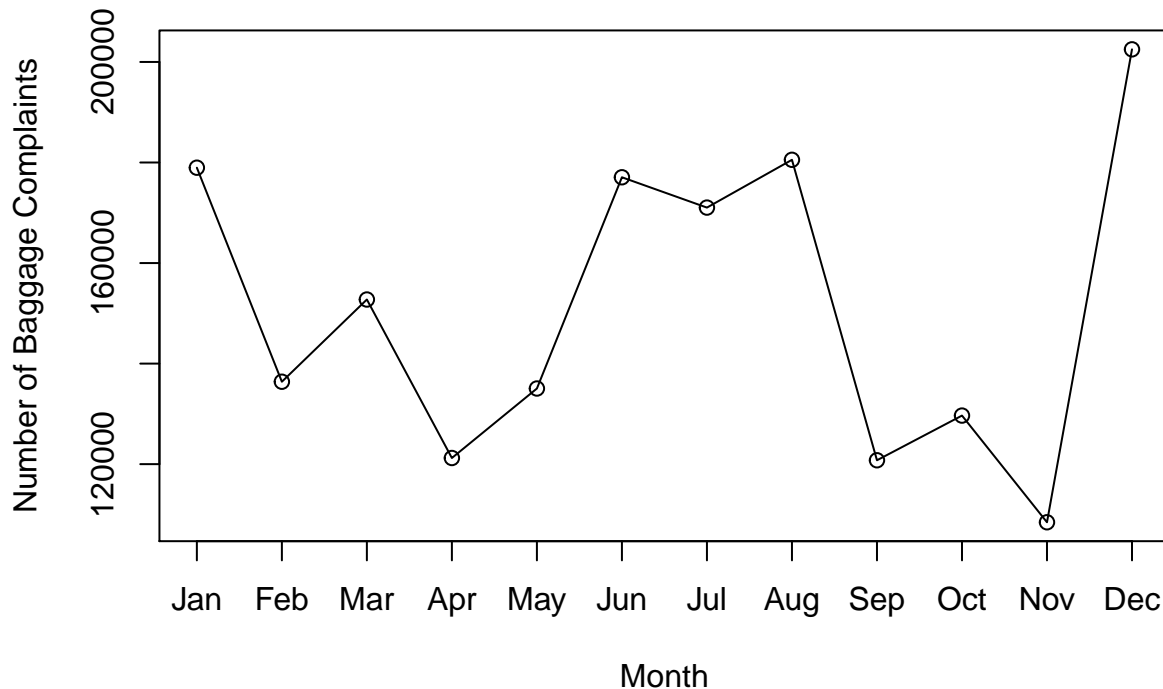
### American Eagle Baggage Complaints (2004–2010)



### 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.

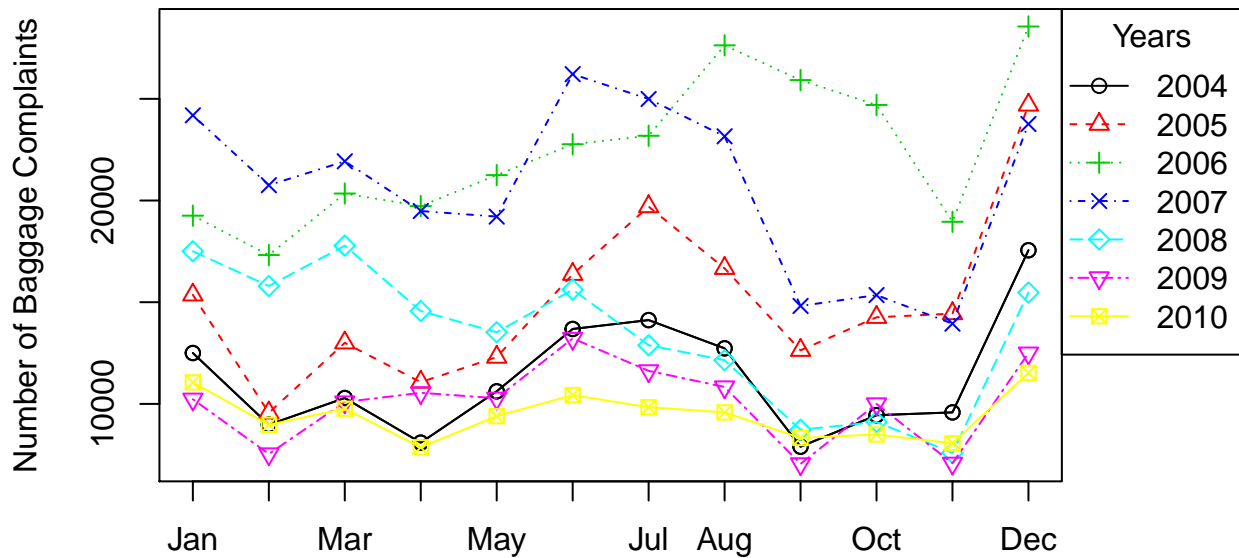
3.

```
airlines = unique(baggage$Airline)
airline = airlines[1]
data = baggage[baggage$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, 4.9), xpd=TRUE)

plot(x=as.integer(plot_dat$Month), y=plot_dat$Baggage, type="o", xaxt="n", xlab="", ylab="Number of Baggage
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



```
airline = airlines[2]
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],]

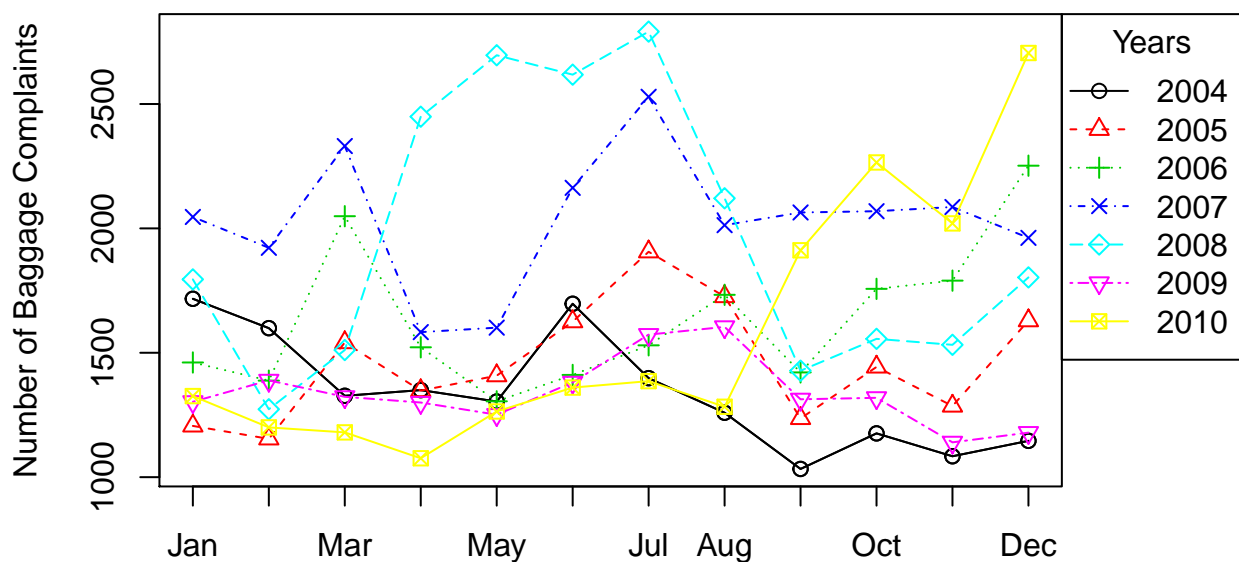
#bottom, left, top, right margin
par(mar=c(7.1, 4.1, 3.1, 4.9), xpd=TRUE)
```

```

plot(x=as.integer(plot_dat$Month),y=plot_dat$Baggage,type="o",xaxt="n",xlab="",
     ylab="Number of Baggage Complaints",lty=1, col=1, pch = 1,ylim = c(min(res$Baggage),max(res$Baggage)),
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), title="Years")

```

## Hawaiian Baggage Complaints



```

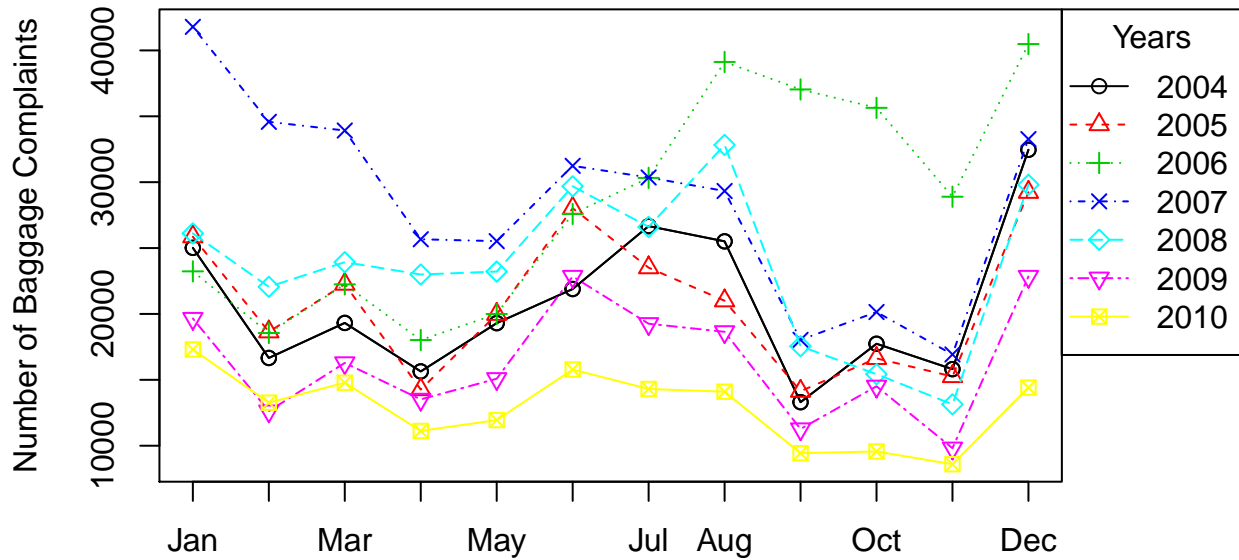
airline = airlines[3]
data = baggage[baggage$Airline == airline,]
res = aggregate(data["Baggage"], by=list(Month = data$Month, Year = data$Year), sum)
years = unique(data$Year)
plot_dat = res[res$Year == years[1],]

#bottom, left, top, right margin
par(mar=c(7.1, 4.1, 3.1, 4.9), xpd=TRUE)

plot(x=as.integer(plot_dat$Month),y=plot_dat$Baggage,type="o",xaxt="n",xlab="",
     ylab="Number of Baggage Complaints",lty=1, col=1, pch = 1,ylim = c(min(res$Baggage),max(res$Baggage)),
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), title="Years")

```

## United Baggage Complaints



### 4. Describe the patterns in the plot

For American Airlines we see that there is an increase in the number of overall complaints for the years 2005-2007 but that the number of overall baggage complaints decreases later in the time period from 2009-2010.

For Hawaiian Airlines we see a large spike in baggage complaints in 2008 compared to the other years. Furthermore we see that the number of baggage complaints noticeably increases towards the end of 2010 in comparison to the other years.

Of the three airlines, United Airlines has the most consistent number of baggage complaints year over year compared to the other airlines, except for 2006 where there is a larger number of total complaints. This mirrors American Airlines which saw an increase in total number of baggage complaints for the period between 2005-2007.

### 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),
                              FUN=mean)

res = total_aggregated[airline == airline,]
res_ts = ts(res$Baggage, frequency = 12, start = 2004)
tsp = attributes(res_ts)$tsp
dates = seq(as.Date("2004-01-02"), by = "month", along = res_ts)

#bottom, left, top, right margin
par(mar=c(7.1, 4.1, 3.1, 4.9), xpd=TRUE)

plot(res_ts,type="o",xaxt="n",xlab="Month",
      ylab="Number of Baggage Complaints",lty=1, col=1, pch = 1,
```

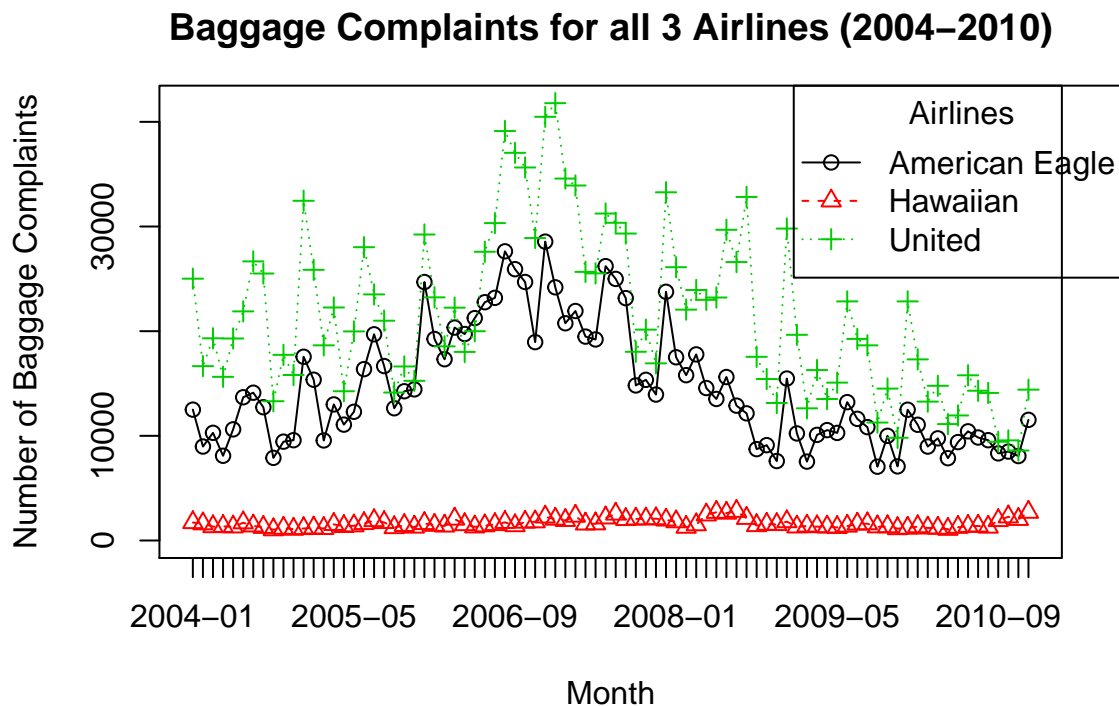
```

ylim= c(0,max(total_aggregated$Baggage)))+ axis(1, at = seq(tsp[1], tsp[2], along = res_ts),
labels = format(dates, "%Y-%m"))

## numeric(0)

title("Baggage Complaints for all 3 Airlines (2004-2010)")
for(i in 2:length(airlines)){
  airline = airlines[i]
  res = total_aggregated[airline == airline,]
  res_ts = ts(res$Baggage, frequency = 12, start = 2004)
  lines(res_ts,type="o",lty=i, col=i,pch=i)
}
legend("topright", inset=c(-0.075,0), legend=airlines,
pch=1:length(airlines),lty=1:length(airlines),col=1:length(airlines), title="Airlines")

```



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 5000. Other two lines, however, are above 5000.

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 complains 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)

res = perc_aggregated[perc_aggregated$Airline == airline,]
res_ts = ts(res$Baggage_perc, frequency = 12, start = 2004)
tsp = attributes(res_ts)$tsp
dates = seq(as.Date("2004-01-02"), by = "month", along = res_ts)

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

plot(res_ts,type="o",xaxt="n",xlab="Month", ylab="Number of Baggage %",lty=1, col=1, pch = 1,ylim= c(0,100))

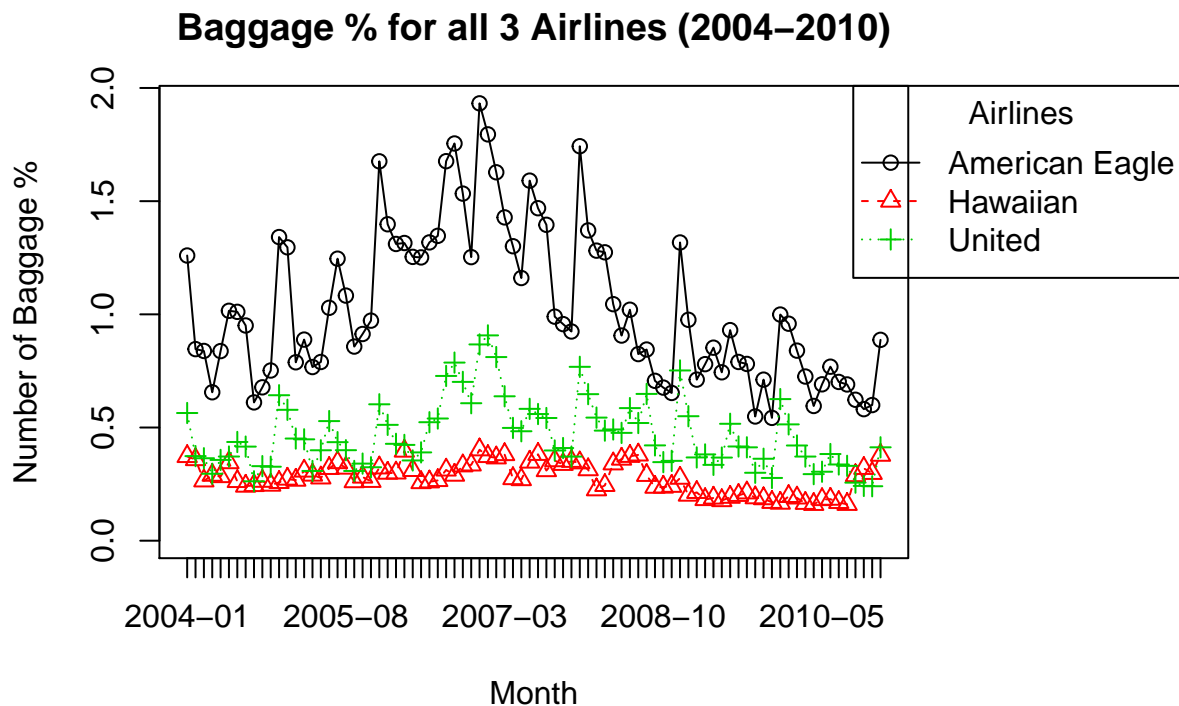
## numeric(0)

title("Baggage % for all 3 Airlines (2004-2010)")
for(i in 2:length(airlines)){
  airline = airlines[i]
  res = perc_aggregated[perc_aggregated$Airline == airline,]
```

```

res_ts = ts(res$Baggage_perc, frequency = 12, start = 2004)
lines(res_ts,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.

```

airlines = unique(baggage$Airline)
airline = airlines[1]
perc_aggregated = aggregate(baggage["Baggage_perc"], by=list(Date = baggage$Date,Airline=baggage$Airline), FUN=mean)
res = perc_aggregated[perc_aggregated$Airline == airline,]

res = perc_aggregated[baggage$Airline == airline,]
res_ts = ts(res$Baggage_perc, frequency = 12, start = 2004)
tsp = attributes(res_ts)$tsp
dates = seq(as.Date("2004-01-02"), by = "month", along = res_ts)

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

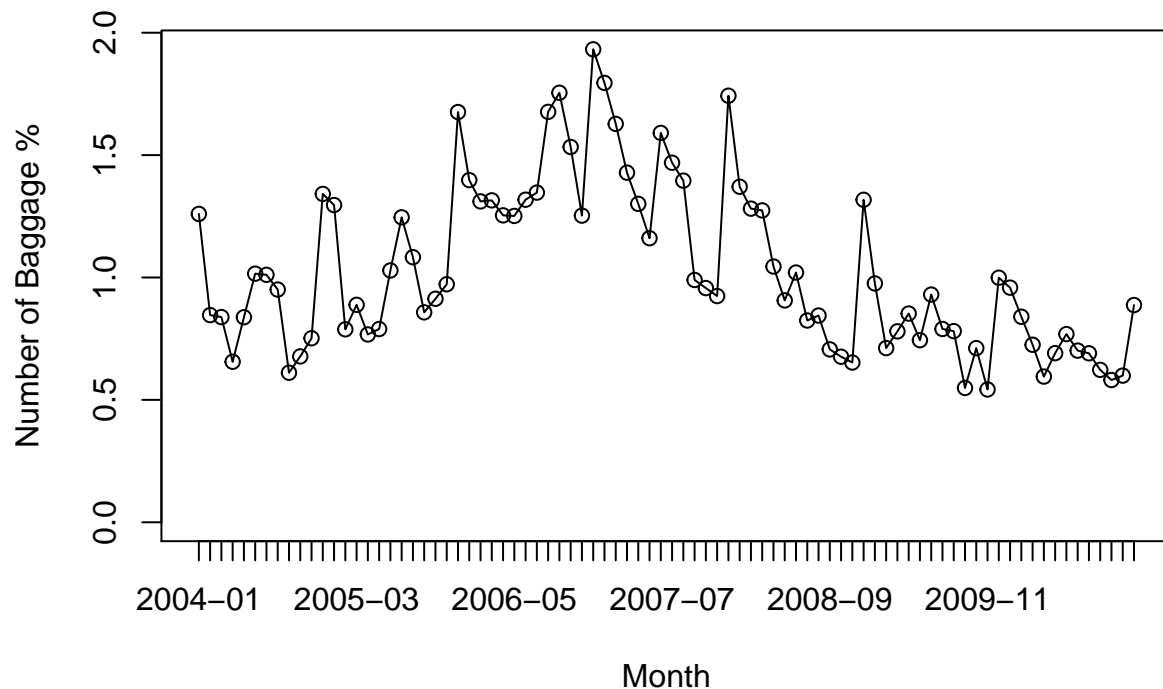
plot(res_ts,type="o",xaxt="n",xlab="Month", ylab="Number of Baggage %",lty=1, col=1, pch = 1,ylim=c(0,max(res_ts)))

## numeric(0)

title(paste0("Baggage % for ", airline, " (2004-2010)"))

```

## Baggage % for American Eagle (2004–2010)



```

airline = airlines[2]
perc_aggregated = aggregate(baggage["Baggage_perc"], by=list(Date = baggage$Date,Airline=baggage$Airline),
res = perc_aggregated[perc_aggregated$Airline == airline,]

res = perc_aggregated[baggage$Airline == airline,]
res_ts = ts(res$Baggage_perc, frequency = 12, start = 2004)
tsp = attributes(res_ts)$tsp
dates = seq(as.Date("2004-01-02"), by = "month", along = res_ts)

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

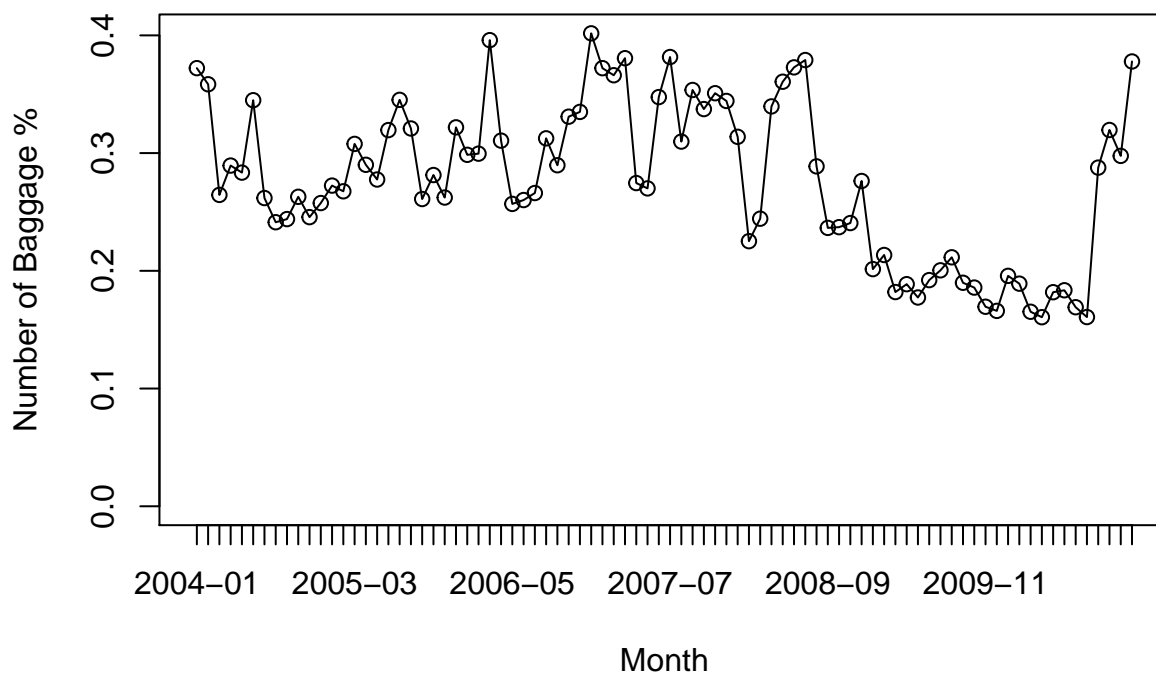
plot(res_ts,type="o",xaxt="n",xlab="Month", ylab="Number of Baggage %",lty=1, col=1, pch = 1,ylim=c(0,m

## numeric(0)

title(paste0("Baggage % for ", airline, " (2004-2010)"))

```

## Baggage % for Hawaiian (2004–2010)



```
airline = airlines[3]
perc_aggregated = aggregate(baggage["Baggage_perc"], by=list(Date = baggage$Date,Airline=baggage$Airline),
res = perc_aggregated[perc_aggregated$Airline == airline,]

res = perc_aggregated[baggage$Airline == airline,]
res_ts = ts(res$Baggage_perc, frequency = 12, start = 2004)
tsp = attributes(res_ts)$tsp
dates = seq(as.Date("2004-01-02"), by = "month", along = res_ts)

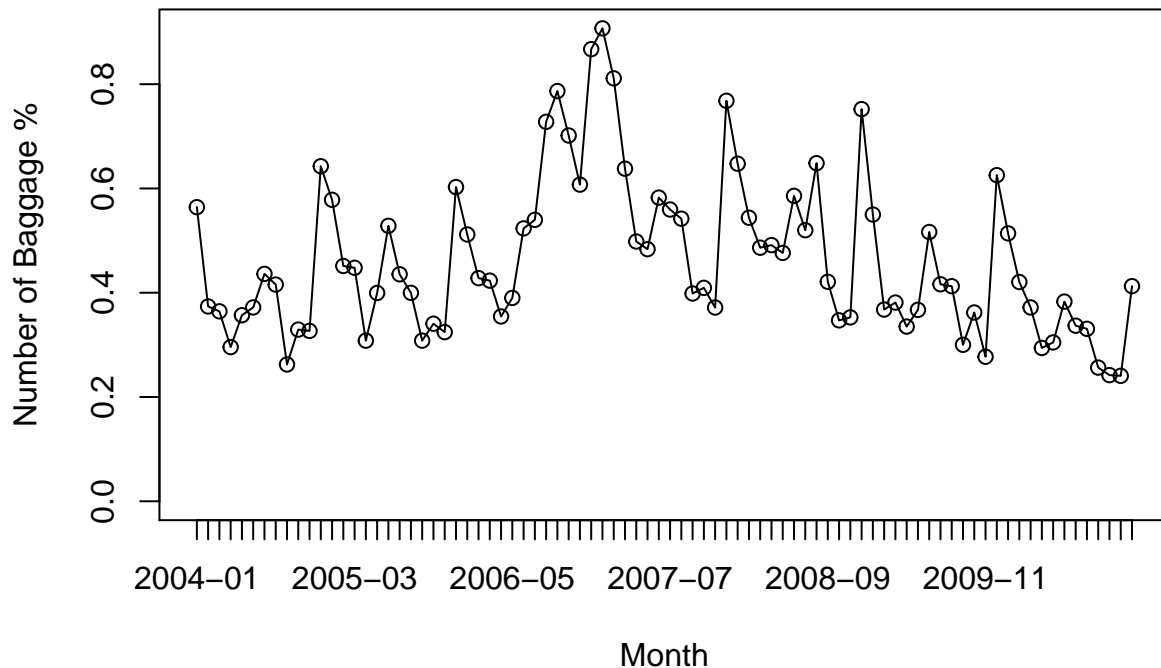
# par(mar=c(7.1, 4.1, 3.1, 8.9), xpd=TRUE)

plot(res_ts,type="o",xaxt="n",xlab="Month", ylab="Number of Baggage %",lty=1, col=1, pch = 1,ylim=c(0,m

## numeric(0)

title(paste0("Baggage % for ", airline, " (2004-2010)"))
```

## Baggage % for United (2004–2010)



Plotting each series on a separate graph is beneficial because this way we can pay a closer look to how every curve fluctuated. In the previous plot, since the range for American Eagle is too big, it is hard to tell how the curve of Hawaiian changed over time.

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

The conclusion for the best service and worst service has been changed. If we look at the KPI Baggage %, we would find that Hawaiian still has the best service, whilst American Eagle has the worst service.

How complaints are changing over time remains non-significant. For United the Baggage % level seems pretty stable; For American Eagle Baggage % seems to rise up and fall back to the beginning level; for Hawaiian it seems that the Baggage % once seems to drop but at the end of 2010 it increases rapid to the highest level. Therefore, by the current data we cannot tell whether the complaints level are becoming better or worse.

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

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

perc_averaged= aggregate(baggage["Baggage_perc"], by=list(Date = baggage$Month,Airline=baggage$Airline)
res = perc_averaged[perc_averaged$Airline == airline,]
plot(x=1:12, y=res$Baggage_perc, type="o",xlab="Month", ylab="Month Average of Baggage %",lty=1, col=1,

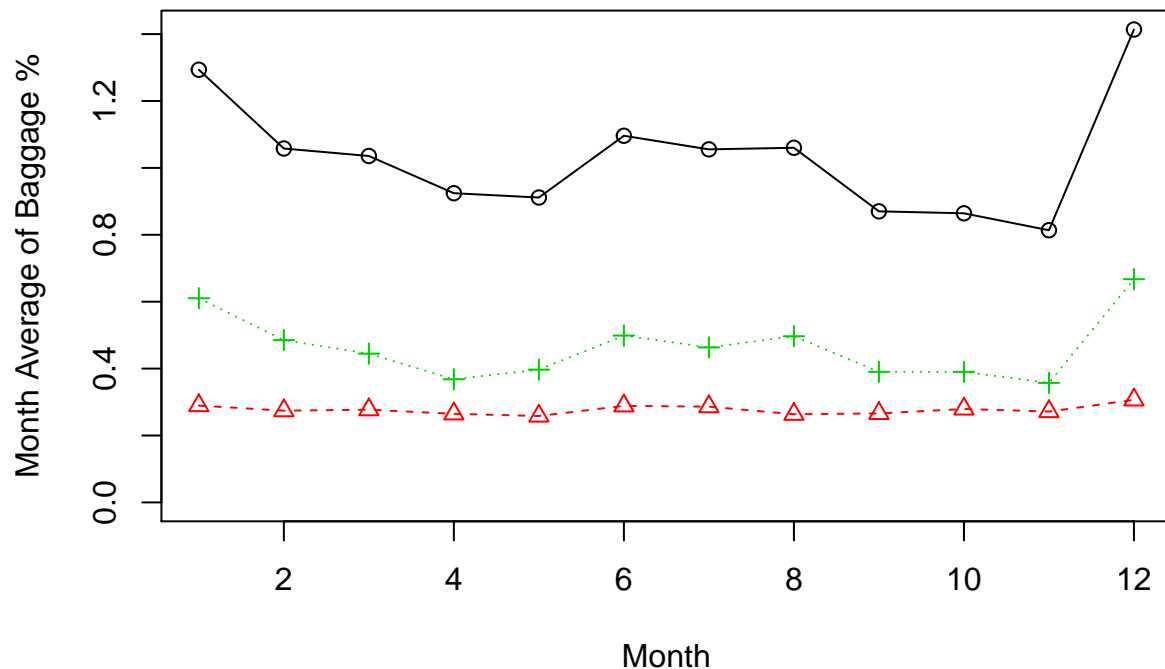
title("Monthly Averages of Baggage % for all 3 Airlines (2004-2010)")
for(i in 2:length(airlines)){
  airline = airlines[i]
```

```

res = perc_averaged[perc_averaged$Airline == airline,]
lines(res$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))

```

## Monthly Averages of Baggage % for all 3 Airlines (2004–2010)



```

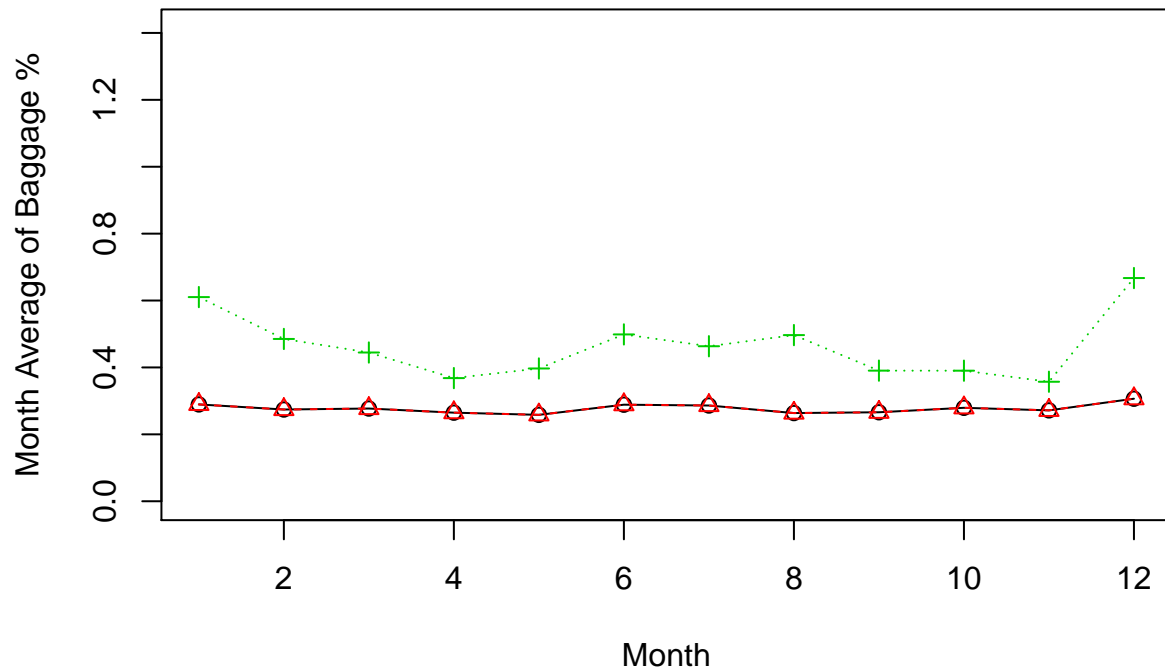
airline = airlines[2]

perc_averaged= aggregate(baggage["Baggage_perc"], by=list(Date = baggage$Month,Airline=baggage$Airline),
res = perc_averaged[perc_averaged$Airline == airline,]
plot(x=1:12, y=res$Baggage_perc, type="o",xlab="Month", ylab="Month Average of Baggage %",lty=1, col=1,

title("Monthly Averages of Baggage % for all 3 Airlines (2004-2010)")
for(i in 2:length(airlines)){
  airline = airlines[i]
  res = perc_averaged[perc_averaged$Airline == airline,]
  lines(res$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))

```

## Monthly Averages of Baggage % for all 3 Airlines (2004–2010)

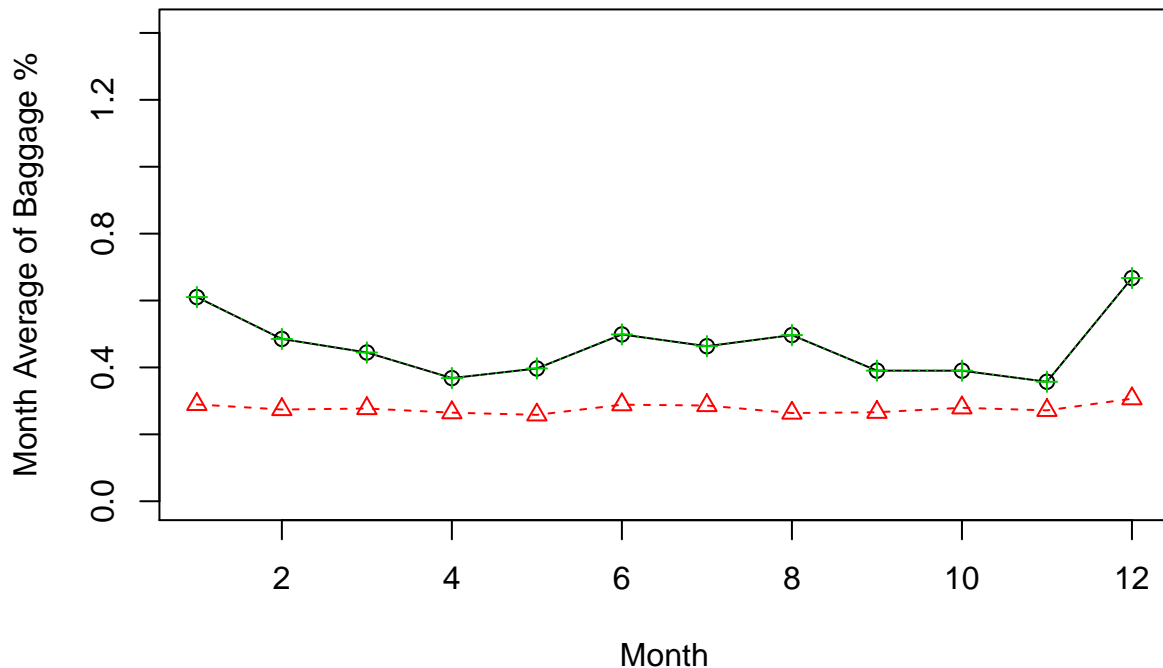


```
airline = airlines[3]

perc_averaged= aggregate(baggage["Baggage_perc"], by=list(Date = baggage$Month,Airline=baggage$Airline),
res = perc_averaged[perc_averaged$Airline == airline,]
plot(x=1:12, y=res$Baggage_perc, type="o",xlab="Month", ylab="Month Average of Baggage %",lty=1, col=1,

title("Monthly Averages of Baggage % for all 3 Airlines (2004-2010)")
for(i in 2:length(airlines)){
  airline = airlines[i]
  res = perc_averaged[perc_averaged$Airline == airline,]
  lines(res$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))
```

## Monthly Averages of Baggage % for all 3 Airlines (2004–2010)



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

The common patterns that all three series share are as follows: The Baggage % begins to drop during the first 4-5 months, then it will hit the highest point in June, and stay at a high level till August, then it will continue dropping before it soars in Nov.-Dec.

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

```
airlines = unique(baggage$Airline)
airline = airlines[1]
data = baggage[baggage$Airline == airline,]
res = aggregate(data["Baggage_perc"], 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_perc, type="o",
     xaxt="n", xlab="", ylab="Baggage % ", lty=2, col=2, pch = 2,
     ylim = c(min(res$Baggage_perc), max(res$Baggage_perc)))
axis(1, at = seq(1, 12), labels = levels(res$Month))
title(paste(airline, "Baggage %"))
for(j in 1:length(years)){
  plot_dat = res[res$Year == years[j],]
  lines(x=as.integer(plot_dat$Month),
```



```

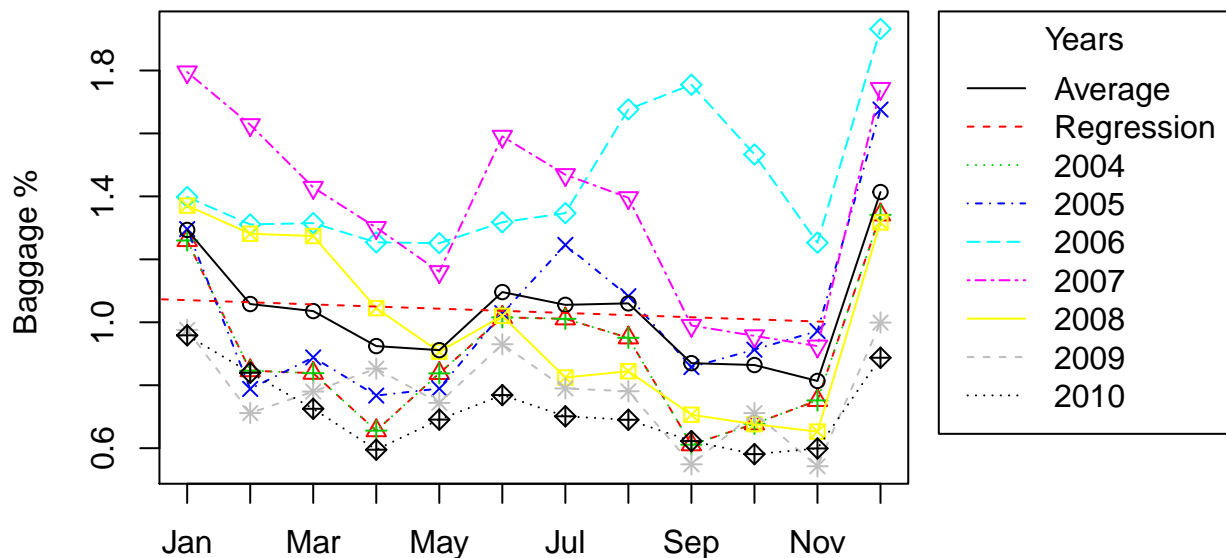
y=plot_dat$Baggage_perc,type="o",lty=j+2, col=j+2,pch=j+2)
}
# add average line
perc_averaged= aggregate(res["Baggage_perc"], by=list(Date = res$Month), mean)
lines(x=as.integer(plot_dat$Month),y=perc_averaged$Baggage_perc, type="o",lty=1, col=1,pch=1)

legend("topright", inset=c(-0.43,0), legend=c("Average", "Regression", years), lty=1:(length(years)+2),

# add regression line
lm_perc = lm(perc_averaged$Baggage_perc~as.integer(plot_dat$Month))
clip(min(as.integer(plot_dat$Month))-0.48,
     max(as.integer(plot_dat$Month))-0.9,
     min(res$Baggage),max(res$Baggage_perc))
abline(lm_perc, lty = 2, col=2)

```

### American Eagle Baggage %



```

airline = airlines[2]
data = baggage[baggage$Airline == airline,]
res = aggregate(data["Baggage_perc"], by=list(Month = data$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_perc,type="o",
     xaxt="n",xlab="", ylab="Baggage % ",lty=2, col=2, pch = 2,
     ylim = c(min(res$Baggage_perc),max(res$Baggage_perc)))
axis(1,at = seq(1,12),labels = levels(res$Month))
title(paste(airline,"Baggage %"))
for(j in 1:length(years)){
  plot_dat = res[res$Year == years[j],]

```

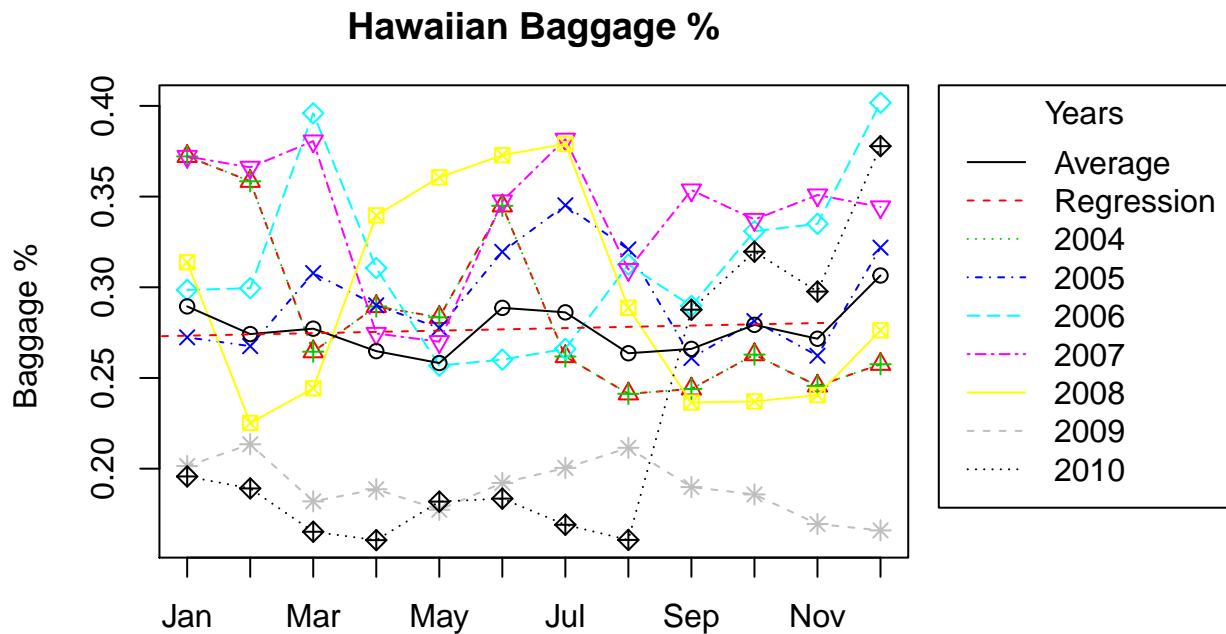
```

lines(x=as.integer(plot_dat$Month),
      y=plot_dat$Baggage_perc,type="o",lty=j+2, col=j+2,pch=j+2)
}
# add average line
perc_averaged= aggregate(res["Baggage_perc"], by=list(Date = res$Month), mean)
lines(x=as.integer(plot_dat$Month),y=perc_averaged$Baggage_perc, type="o",lty=1, col=1,pch=1)

legend("topright", inset=c(-0.43,0), legend=c("Average", "Regression", years), lty=1:(length(years)+2),

# add regression line
lm_perc = lm(perc_averaged$Baggage_perc~as.integer(plot_dat$Month))
clip(min(as.integer(plot_dat$Month))-0.48,
     max(as.integer(plot_dat$Month))-0.9,
     min(res$Baggage),max(res$Baggage_perc))
abline(lm_perc, lty = 2, col=2)

```



```

airline = airlines[3]
data = baggage[baggage$Airline == airline,]
res = aggregate(data["Baggage_perc"], by=list(Month = data$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_perc,type="o",
     xaxt="n",xlab="", ylab="Baggage % ",lty=2, col=2, pch = 2,
     ylim = c(min(res$Baggage_perc),max(res$Baggage_perc)))
axis(1,at = seq(1,12),labels = levels(res$Month))
title(paste(airline,"Baggage %"))
for(j in 1:length(years)){

```

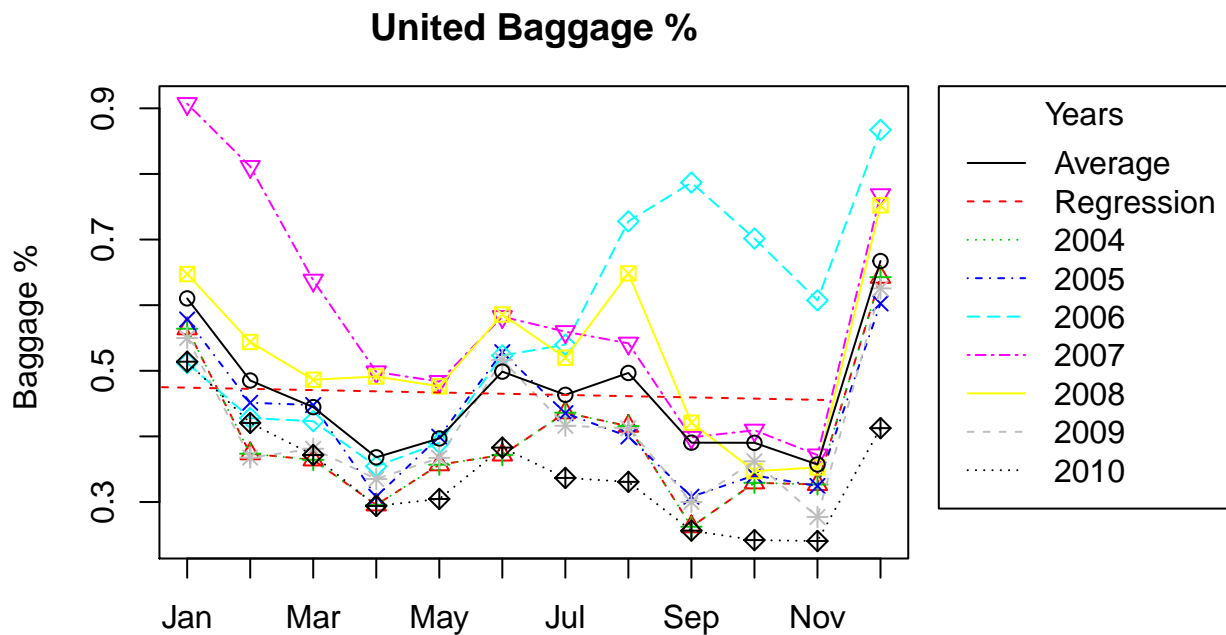
```

plot_dat = res[res$Year == years[j],]
lines(x=as.integer(plot_dat$Month),
      y=plot_dat$Baggage_perc,type="o",lty=j+2, col=j+2,pch=j+2)
}
# add average line
perc_averaged= aggregate(res["Baggage_perc"], by=list(Date = res$Month), mean)
lines(x=as.integer(plot_dat$Month),y=perc_averaged$Baggage_perc, type="o",lty=1, col=1,pch=1)

legend("topright", inset=c(-0.43,0), legend=c("Average", "Regression", years), lty=1:(length(years)+2),

# add regression line
lm_perc = lm(perc_averaged$Baggage_perc~as.integer(plot_dat$Month))
clip(min(as.integer(plot_dat$Month))-0.48,
     max(as.integer(plot_dat$Month))-0.9,
     min(res$Baggage),max(res$Baggage_perc))
abline(lm_perc, lty = 2, col=2)

```



For each airline, I superimposed the follow curves: \* “Average”: The average monthly Baggage % among the 7 years (black solid curve) \* “Regression”: The linear regression of the average monthly Baggage % (red curve), using *lm()* function \* 7 years of monthly Baggage %

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

There is an increase in the number of baggage complaints during the summer and winter holiday holiday seasons for all three airline carriers. These holiday season spikes in complaints is relatively consistent across the different years. When we looked at the time series plots using the KPI of Baggage % we saw that the Baggage % begins to drop during the first 4-5 months, then it will hit the highest point in June, and stay at a high level till August, then it will continue dropping before it soars in Nov.-Dec.

## Case 2: CEO Compensation

```
x <- read.table(here("HW1", "CEOcompensation.txt"), header = T, sep = "\t", quote = "\"", row.names = 1)
```

Question 1: What is the number of female CEO's?

```
num.ceo.f <- length(x$CEO[x$Gender == "F"])
print(paste("The number of female CEO's is", num.ceo.f))

## [1] "The number of female CEO's is 2"
```

Question 2: What is the age of the youngest CEO?

```
age.ceo.min <- min(x$Age)
print(paste("The age of the youngest CEO is", age.ceo.min))

## [1] "The age of the youngest CEO is 45"
```

Question 3: What is the age of the oldest CEO?

```
age.ceo.max <- max(x$Age)
print(paste("The age of the oldest CEO is", age.ceo.max))

## [1] "The age of the oldest CEO is 81"
```

Question 4: What is the average age of a CEO?

```
age.ceo.avg <- round(mean(x$Age), 2)
print(paste("The average age of a CEO is", age.ceo.avg))

## [1] "The average age of a CEO is 58.38"
```

Question 5: What is the total CEO 2008 salary?

```
tot.2008.sal <- sum(x$X2008.Salary)
print(paste("The total CEO 2008 salary is", paste0(tot.2008.sal,"0"), "million"))

## [1] "The total CEO 2008 salary is 201.80 million"
```

Question 6: How many CEOs have joined a company as a CEO? (Hint: CEOs can always be founders. Founders can't always be CEOs)

```
## Here we claim that a CEO joined the company as a CEO if the number of years she was at the company equals the number of years she was with the company

yearCheck <- sum(x$Years.as.company.CEO == x$Years.with.company)
print(paste(yearCheck, "CEO's joined a company as a CEO"))
```

```
## [1] "40 CEO's joined a company as a CEO"
```

**Question 7: What is the average amount of time a CEO worked for a company before becoming a CEO? (Use two decimal digit precision)**

```
beforeCEO <- round(mean(x$Years.with.company - x$Years.as.company.CEO), 2)
print(paste("The average amount of time a CEO worked for a company before becoming a CEO is", beforeCEO))

## [1] "The average amount of time a CEO worked for a company before becoming a CEO is 11.51 years"
```

**Question 8: Which industry in the data set has largest number CEO's?**

```
numCEO <- aggregate(x$CEO, list(Industry = x$Industry), FUN = length)
industry.max.ceo <- numCEO[[1]][which(numCEO[[2]] == max(numCEO[[2]]))]
print(paste("The industry with the largest number of CEO's is", industry.max.ceo))

## [1] "The industry with the largest number of CEO's is Oil & Gas Operations"
```

**Question 9: What is the average CEO 2008 Compensation? Note that 2008 compensation for a CEO consists of a total four components: Salary, Bonus, other (including vested restricted stock grants, LTIP (long-term incentive plan) payouts, and perks), and stock gains. (Use two decimal digit precision)**

```
totCompensation <- round(mean(x$X2008.Salary +
                             x$X2008.Bonus + x$X2008.Other +
                             x$X2008.Stock.gains), 2)
print(paste("The average CEO 2008 Compensation is", totCompensation, "million"))

## [1] "The average CEO 2008 Compensation is 18.68 million"
```

**Question 10: Which CEO did get paid the largest compensation amount in 2008?**

```
ceoCompensation <- x$X2008.Salary + x$X2008.Bonus + x$X2008.Other + x$X2008.Stock.gains
maxCompensation <- which(ceoCompensation == max(ceoCompensation))
print(paste("The CEO with the largest compensation amount in 2008 is", x$CEO[maxCompensation]))

## [1] "The CEO with the largest compensation amount in 2008 is Lawrence J Ellison"
```

**Question 11: What is the corresponding amount? (Use two decimal digit precision)**

```
ceoCompensation <- x$X2008.Salary + x$X2008.Bonus + x$X2008.Other + x$X2008.Stock.gains
print(paste("The corresponding amount is", max(ceoCompensation), "million"))

## [1] "The corresponding amount is 556.98 million"
```

**Question 12:** Which industry does correspond to the second largest total CEO compensation in 2008? (Hint: check `sort()`, `order()` functions).

```
ceoCompensation <- x$X2008.Salary + x$X2008.Bonus + x$X2008.Other + x$X2008.Stock.gains
secondMaxCompensation <- which(ceoCompensation == tail(sort(ceoCompensation), 2)[1])
print(paste("The industry with the second largest total CEO compensation is",
            x$Industry[secondMaxCompensation]))
```

```
## [1] "The industry with the second largest total CEO compensation is Oil & Gas Operations"
```

**Question 13:** Consider the following age groups: [45 – 50), [50 – 55), [55 – 60), [60 – 70), and [70 or more). Analyze age groups by industry and determine which age group corresponds to largest CEO average salary in 2008? Hint: 1. left end point is included; 2. nested if helps assign age category

```
ageGroups <- vector(mode="numeric", length=nrow(x))
ageGroups[which(x$Age >= 45 & x$Age < 50)] <- 1
ageGroups[which(x$Age >= 50 & x$Age < 55)] <- 2
ageGroups[which(x$Age >= 55 & x$Age < 60)] <- 3
ageGroups[which(x$Age >= 60 & x$Age < 70)] <- 4
ageGroups[which(x$Age >= 70)] <- 5
x$ageGroups <- ageGroups

groupByIndustry <- aggregate(x$X2008.Salary, list(ageGroups = x$ageGroups,
                                                industry = x$Industry), mean)
groupByAge <- as.data.frame(as.matrix(aggregate(x$X2008.Salary, list(ageGroups = x$ageGroups), mean)))

print(paste("The age group that has the highest average salary is group",
            groupByAge$ageGroups[which(groupByAge$x == max(groupByAge$x))],
            "which corresponds to [70 or more)"))
```

```
## [1] "The age group that has the highest average salary is group 5 which corresponds to [70 or more)"
```

```
print(paste("This corresponded to the industry of",
            groupByIndustry$industry[which(groupByIndustry$x == max(groupByIndustry$x))],
            "where the salary was", max(groupByIndustry$x), "million"))
```

```
## [1] "This corresponded to the industry of Media where the salary was 8.1 million"
```

```
## Import median data
y <- read.csv("IndustryMedians.csv")

## Calculate percent difference for each CEO
z <- cbind.data.frame(ceoCompensation, industry=x$Industry)

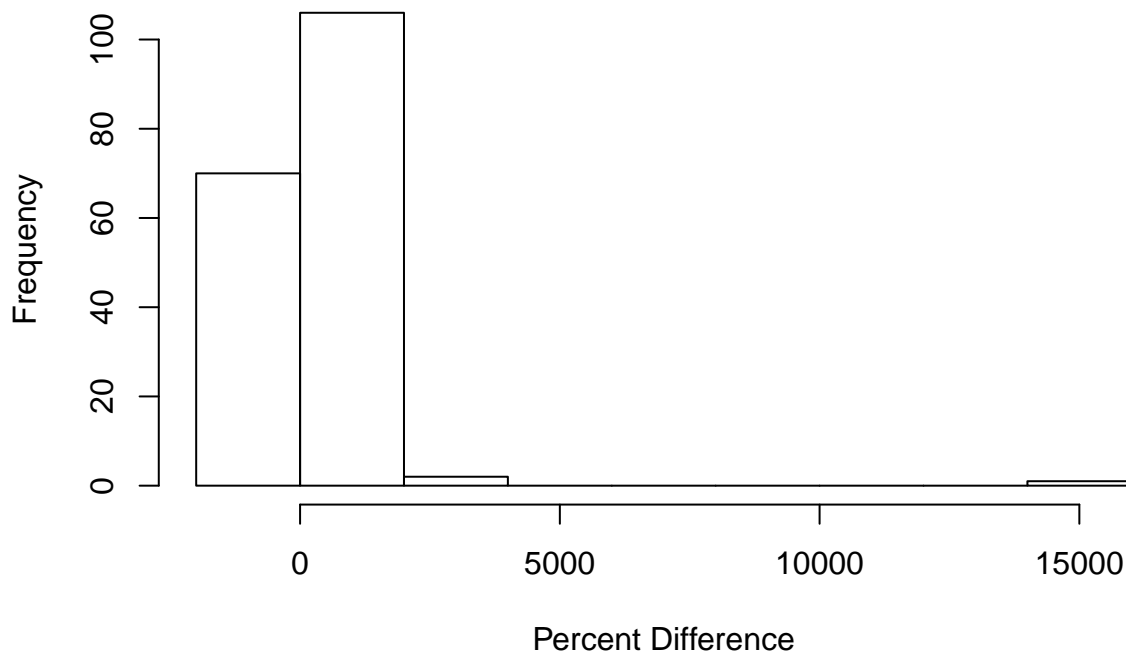
x$percentDiff <- sapply(1:nrow(z), function(i) {
  compensation <- z$ceoCompensation[i]
  indMed <- y$Total.compensation[which(y$Industry == z$industry[i])]
  return((compensation - indMed) / indMed*100)
})

## Look at the percent difference for each CEO
print(round(x$percentDiff,3))
```

```
## [1] 115.432 -70.843 -19.433 -66.287 -53.654 -6.595 346.538
## [8] -32.759 87.264 15.385 -61.048 0.000 1745.852 344.423
## [15] 40.000 167.249 -45.577 -98.077 -62.830 -3.111 -35.769
## [22] 34.061 -22.311 -26.538 -100.000 350.272 154.585 1020.019
## [29] 54.865 0.000 34.403 276.983 146.978 69.697 -79.189
## [36] -26.724 -5.677 146.736 413.833 873.333 100.296 131.731
## [43] -24.865 324.054 -23.351 -6.154 312.756 152.620 277.760
## [50] 51.592 94.894 -69.209 194.043 -91.238 -52.100 177.689
## [57] 30.809 22.432 57.308 65.836 3.222 325.101 -40.868
## [64] 34.263 96.070 595.879 2264.324 1022.096 17.254 0.000
## [71] -0.199 736.219 -62.421 -43.774 357.171 66.397 158.088
## [78] 33.904 0.199 -4.360 -43.869 477.692 681.275 5.132
## [85] 16.534 -54.670 17.078 78.065 504.035 336.217 599.801
## [92] 766.571 336.446 11.957 87.472 134.987 -3.079 -25.862
## [99] 29.262 78.927 6.595 99.746 15201.648 -20.120 107.721
## [106] 2032.567 -33.429 -69.248 581.222 -7.838 -50.136 61.753
## [113] 0.437 -71.912 1380.651 15.139 -59.778 174.641 355.460
## [120] -30.809 13.783 70.811 -23.748 598.638 103.125 244.923
## [127] 53.352 72.581 59.542 368.934 269.755 77.667 143.360
## [134] -6.773 -62.912 -55.022 254.743 93.548 -13.740 -36.842
## [141] -12.931 -60.577 60.187 -5.677 0.000 -53.275 -22.635
## [148] -31.893 733.242 59.565 -9.401 139.101 -25.073 145.623
## [155] -20.957 -38.865 -33.901 104.183 95.179 31.608 386.827
## [162] 26.879 -58.378 274.089 531.064 -69.975 -73.842 -38.889
## [169] -79.316 137.346 205.577 -15.385 231.322 271.346 91.092
## [176] -30.482 151.081 -5.405 446.725
```

```
hist(x$percentDiff, main = "Percent Difference for each CEO", xlab = "Percent Difference")
```

## Percent Difference for each CEO



**Question 14:** How many CEO's have received 100% or larger compensation relative to their respective median compensation?

```
numLarger <- length(which(x$percentDiff > 100))
cat(paste("The number of CEO's that recieved 100% or larger
          compensation relative to their respective median compensation is", numLarger))
```

```
## The number of CEO's that recieved 100% or larger
##          compensation relative to their respective median compensation is 63
```

**Question 15:** Is the following formula always true?

```
y$totalMedianCompensation <- y$Salary + y$Bonus + y$Other + y$Stock.Gains

cat(paste("There are a total of",
          length(which(y$totalMedianCompensation != y$Total.compensation)),
          "where the total median compensation formula
          does not match our given total median compensation. This means the formula is not always true."))
```

```
## There are a total of 26 where the total median compensation formula
## does not match our given total median compensation. This means the formula is not always true.
```