WW2 US Aircraft Costs

Mark Biegert

February 19, 2017

Objective

This worksheet illustrates how I cleaned and graphed some untidy data grabbed from the least tidy data that I know of- WW2 data stored on the hyperwar web site. This is great data, but it looks like it was OCR'ed from scanned WW2 documents. It doesn't get much worse than that.

Source Material

To illustrate how to use **R** to cleanup an untidy file. The file is labeled ".csv", but it is actually tab-separated.

Untidy Table

Load in some libraries

These are the standard libraries that I normally load.

```
require(dplyr)
require(tidyr)
require(magrittr)
require(stringr)
require(zoo)
require(ggplot2)
require(ggthemes)
require(extrafont)
```

Read in the Source File

I simply copied this file from the web site linked below. This was an ugly file:

- poor format
- much data missing
- inconsistant: NA's represented by "-", "$-", and empty cells. Currency symbol sometimes use, mostly not.

```
c = read.csv(file="RaWCost.csv",header=TRUE,sep="\t",na.strings=c("","NA"))
head(c)
```

```
                Type Model    X1941    X1942 X1943    X1944   X1945
1 Very Heavy Bombers                                          <NA>
2               <NA> B-29      <NA> 897,730        605,360  509,465
3     Heavy Bombers                                           <NA>
4               <NA> B-17  301,221  258,949        204,370  187,742
5               <NA> B-24  379,162  304,391        215,516        -
6               <NA> B-32        -  790,433     -  790,433        -
```

Basic Cleanup

I need to cleanup this data and reorganize it. I will:

- Standard R
- Tidyr
- Dplyr

```
c$Type=na.locf(c$Type)
c$Type = str_trim(c$Type)
c$Model = str_trim(c$Model)
c$Type=as.factor(c$Type)
c$Model= as.factor(c$Model)
c = gather(c,Year,Cost,X1941:X1945)
c$Model[c$Model==""] <- NA
c$Model= as.factor(c$Model)
c$Year=str_replace(c$Year,"X","")
c$Year=as.factor(c$Year)
c$Cost=str_replace(c$Cost,",","")
c$Cost=as.numeric(c$Cost)
c = c[complete.cases(c$Model),]
write.csv(c,"AircraftCostData.csv", row.names=FALSE)
head(c,6)
```

```
                Type Model Year   Cost
2 Very Heavy Bombers  B-29 1941     NA
4      Heavy Bombers  B-17 1941 301221
5      Heavy Bombers  B-24 1941 379162
6      Heavy Bombers  B-32 1941     NA
8     Medium Bombers  B-25 1941 180031
9     Medium Bombers  B-26 1941 261062
```

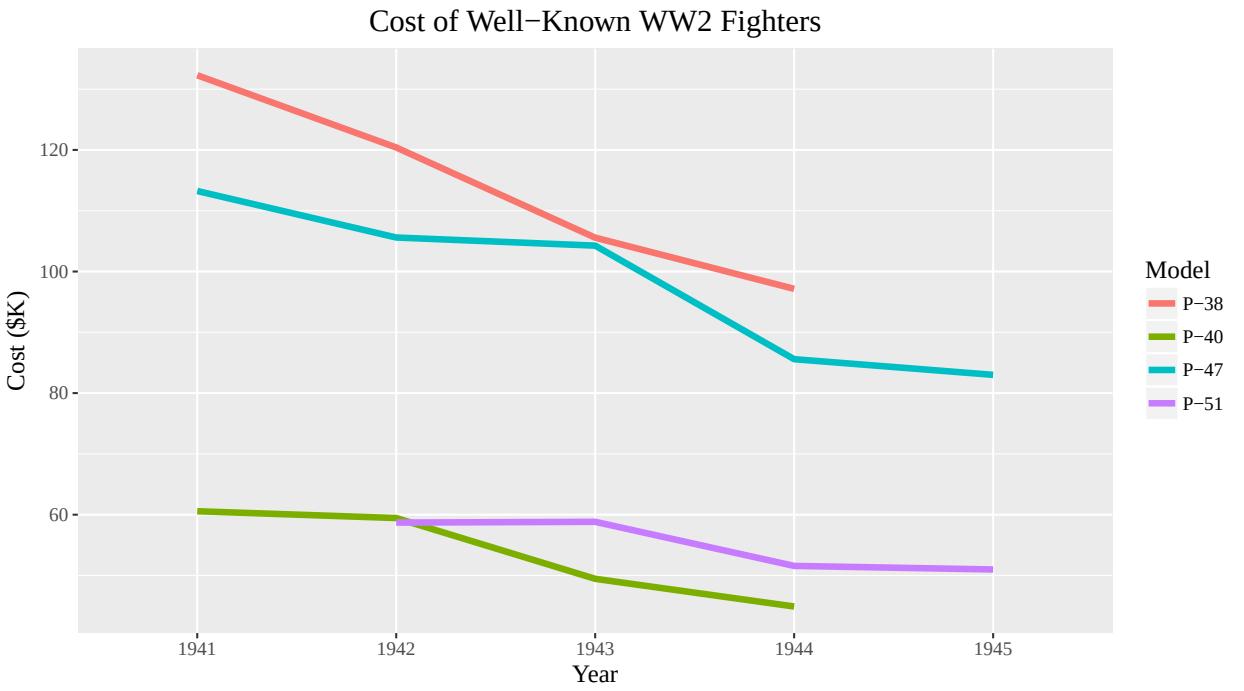

Figure 1: WW2 Fighter Cost Plot.

##Plot the Cost Data for Fighter Planes This worksheet is intended to illustrate how to clean files, so I will simply illustrate how to make a simple plot of the data.

```
target=c("P-38","P-40","P-47","P-51")
fighter=filter(c,Model %in%target)   #I am mainly interested in fighter planes
g=ggplot(data=fighter,aes(x=Year,y=Cost*0.001,group = Model, color = Model))
g=g+geom_line(stat="identity",size=1.6)
g=g+ggtitle("Cost of Well-Known WW2 Fighters")
g=g+ylab("Cost ($K)")+xlab("Year")
g=g+theme_get()
g=g+theme(plot.title = element_text(hjust = 0.5),text=element_text(size=13, family="Times"))
g
```

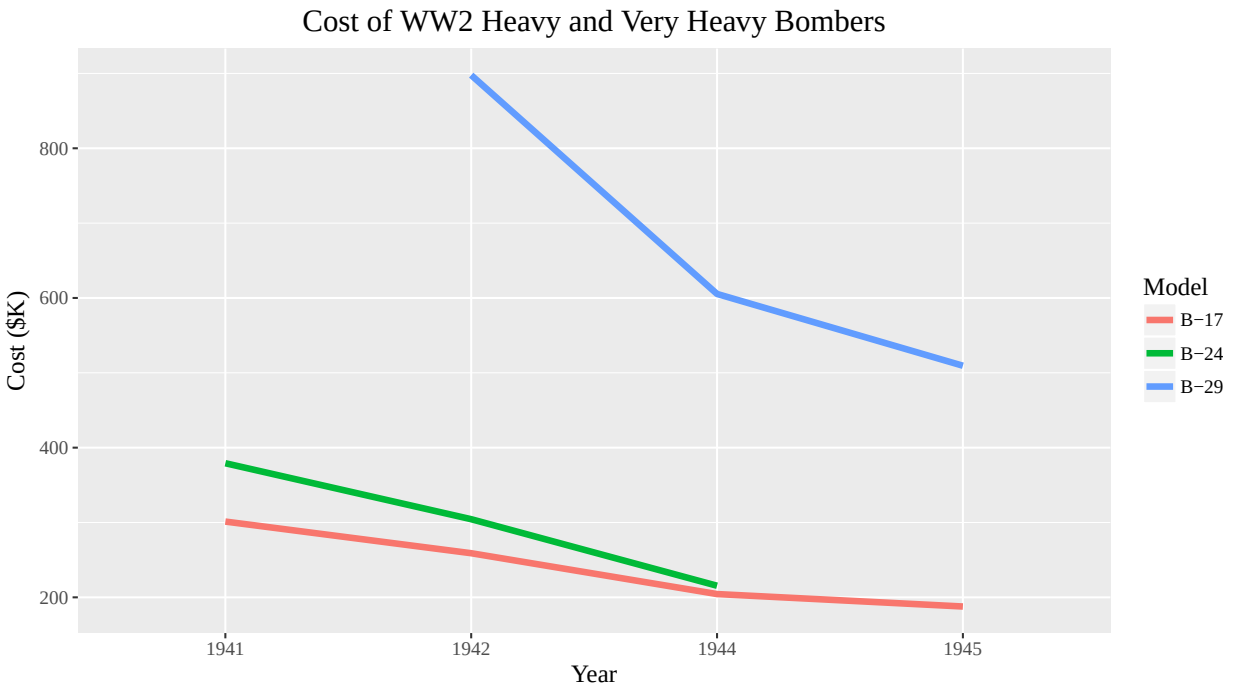

Figure 2: WW2 US Bomber Plot

##Plot the Bomber Cost Data Now that I am into it, I will take a quick peak at the bomber data.

```
target=c("B-17","B-24","B-29")
bomber=filter(c,Model %in%target)   #I am mainly interested in fighter planes
g=ggplot(data=na.omit(bomber),aes(x=Year,y=Cost*0.001,group = Model, color = Model))
g=g+geom_line(stat="identity",size=1.6)
g=g+ggtitle("Cost of WW2 Heavy and Very Heavy Bombers")
g=g+ylab("Cost ($K)")+xlab("Year")
g=g+theme_get()
g=g+theme(plot.title = element_text(hjust = 0.5),text=element_text(size=13, family="Times"))
g
```