

Stats 519: Homework 2 - R Graphics

Due on February 4, 2009

Dr. Stephen Lee 1:30

Monte J. Shaffer

Program 1 Preparing Data Frame

```
> lostDays = read.table('clipboard');
> colnames(lostDays) = c("case","daysLost","crewSize","foremanAge",
                        "foremanExp","crewExp","powerTools");

> lostDays;
  case daysLost crewSize foremanAge foremanExp crewExp powerTools
1     1       110      16         54         26        8          1
2     2         0       8         38         11        8          0
3     3        77       9         31          3        6          1
4     4        37       7         44          9        6          1
5     5        15       9         44         11        9          0
6     6        89      10         53         13        3          0
.     .         .       .          .          .        .          .
.     .         .       .          .          .        .          .
.     .         .       .          .          .        .          .
```

Problem 1

EDA (Exploratory Data Analysis) of daysLost

- Histogram Analysis

Program 2 Histograms

```
> par(mfrow=c(2,2));  
> hist(lostDays$daysLost,main="Default as 10 bins (max)");  
> hist(lostDays$daysLost,breaks=20,main="20 bins (max)");  
> hist(lostDays$daysLost,breaks=5,main="5 bins (max)");  
> hist(lostDays$daysLost,breaks=30,main="30 bins (max)");  
... OR ...  
> hist(lostDays$daysLost,breaks="Sturges");  
> hist(lostDays$daysLost,breaks="Scott");  
> hist(lostDays$daysLost,breaks="FD");
```

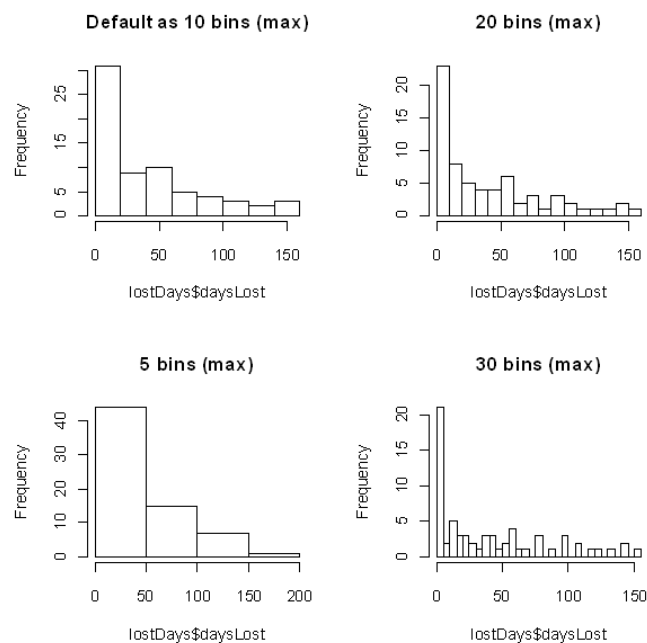


Figure 1: Varying Bins of Histogram

- Kernel Density Analysis

Program 3 Kernel Densities

```
> par(mfrow=c(2,2));
> plot(density(lastDays$daysLost),main="default kd");
> plot(density(lastDays$daysLost,bw=8),main="kd (bw=8)");
> plot(density(lastDays$daysLost,bw=24),main="kd (bw=24)");
> plot(density(lastDays$daysLost,bw=32),main="kd (bw=32)");
... OR ...
> plot(density(lastDays$daysLost,adjust=1/4));
```

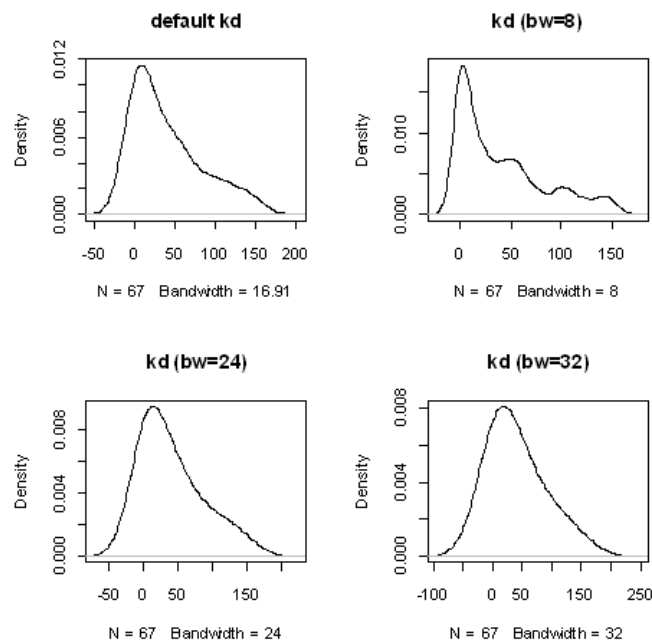


Figure 2: Varying Bandwidths of Kernel Density Estimates

This is a distribution with lots of zeros (clearly right), meaning there are censoring issues. Extreme value distributions would probably best fit the data; for example, an exponential decay from $\frac{1}{\theta}$; e.g.,

```
> library(MASS);
> F=(fitdistr(lastDays$daysLost,"exponential"));
```

Problem 2

Power Tools Influence on work days lost (daysLost)

- Strip and Box Charts

Program 4 Strip Charts / Bar Charts

```
> par(mfrow=c(2,2));  
> stripchart(lostDays$daysLost,"stack");  
> stripchart(lostDays$daysLost~lostDays$powerTools,method="stack");  
> boxplot(lostDays$daysLost);  
> boxplot(lostDays$daysLost~lostDays$powerTools);
```

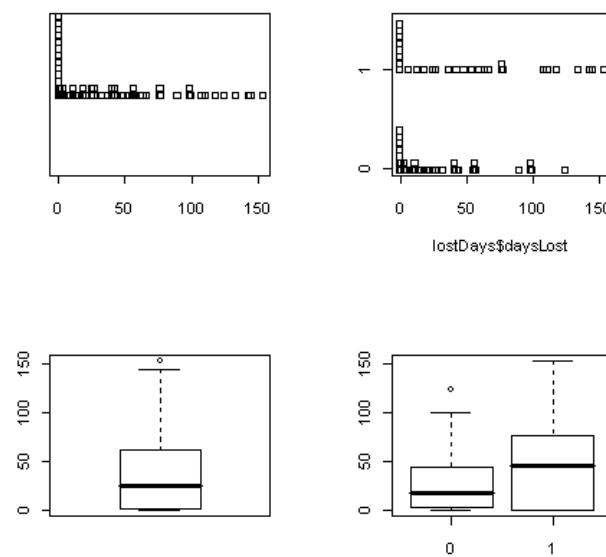


Figure 3: Power Tools Influence on **daysLost**

Apparently, if I cut off my hand with the power tool, I will miss more days of work. Seriously, the presence of power tools increases the variability of work potentially missed.

Problem 3

Other Influences on work days lost (`daysLost`)

- Scatterplots

Program 5 Scatterplots

```
> attach(lostDays); ## so lines are shorter
> par(mfrow=c(2,2));
> plot(crewSize,daysLost,main=round(cor(crewSize,daysLost),digits=3));
> plot(foremanAge,daysLost,main=round(cor(foremanAge,daysLost),digits=3));
> plot(foremanExp,daysLost,main=round(cor(foremanExp,daysLost),digits=3));
> plot(crewExp,daysLost,main=round(cor(crewExp,daysLost),digits=3));
```

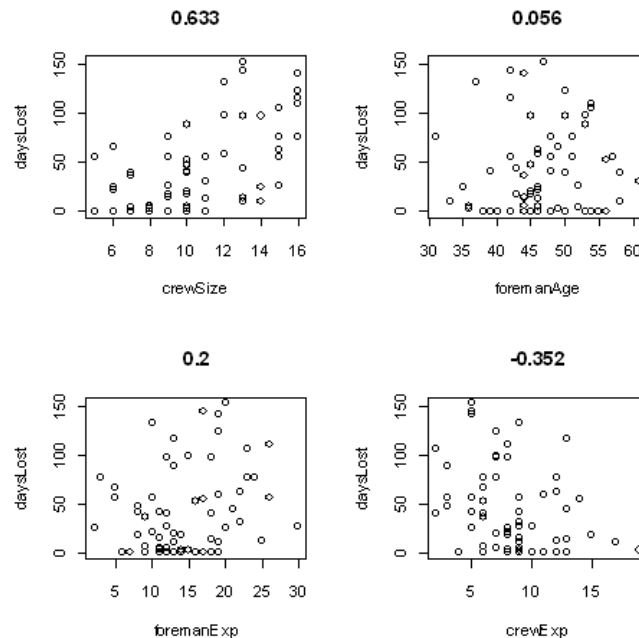


Figure 4: Experience Influence on **daysLost**

The correlations of each comparison are reported as the title (main) of each graph. In order of importance (without considering interactions or controlling for other elements): `crewSize` (more people means more `daysLost`); `crewExp` (more experienced crew means less `daysLost`); `foremanExp` (more experienced foreman means more `daysLost` - he is cautious before allowing them to return to work?); `foremanAge` (looks pretty irrelevant). A simple regression also can give preliminary relationships: e.g.,

```
> attach(lostDays);
> L=lm(daysLost~crewSize+crewExp+powerTools+foremanExp+foremanAge,lostDays);
> summary(L);
```

Problem 4

Interaction Effects

- 3D plots

Program 6 3D EDA

```
> attach(lostDays); ## so lines are shorter
> library(scatterplot3d);
> par(mfrow=c(1,1));
> scatterplot3d(cbind(crewSize,crewExp,daysLost), type='h', highlight.3d=T);
> library(rgl);
> par(mfrow=c(1,1));
> plot3d(foremanAge,crewExp,daysLost, type='s');
> plot3d(foremanExp,crewExp,daysLost, type='s');
> plot3d(powerTools,crewExp,daysLost, type='s');
> plot3d(crewSize,crewExp,daysLost, type='s');
> getwd();
[1] "C:/Documents and Settings/Monte/My Documents"
> rgl.snapshot("rgl.png");
```

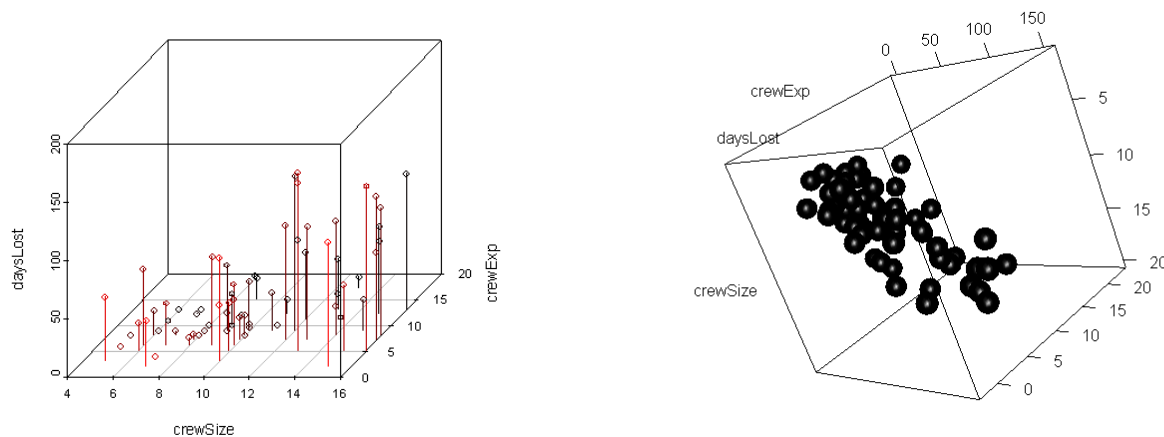


Figure 5: Crew Size/Experience on **daysLost**

It appears that Crew Size and Experience are good predictors of daysLost; a young (inexperienced), large crew.

Problem 5

Power Tools (Advanced)

- Scatterplot Matrix

Program 7 Scatterplot Matrix

```
> par(mfrow=c(1,1));
> attach(lostDays); ## so lines are shorter
> pairs(lostDays[,2:6], pch=16, col=2+powerTools);
> ## color = 2 is red and represents no powerTools
> ## color = 3 is green and represents presence of powerTools
```

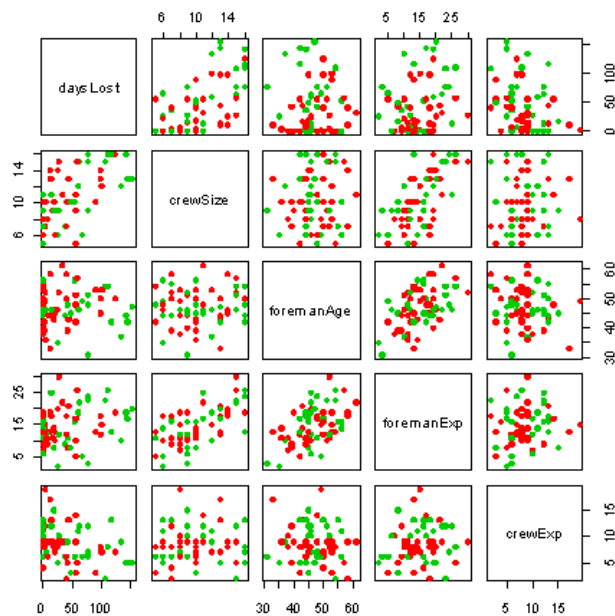


Figure 6: Paired Relationships by Power

- Stars Plot

Program 8 Stars Plot

```

> par(mfrow=c(1,1));
> attach(lostDays); ## so lines are shorter
> stars(lostDays[,2:6],key.loc=c(15,1.5), col.stars=2+powerTools);
> ## color = 2 is red and represents no powerTools
> ## color = 3 is green and represents presence of powerTools

```

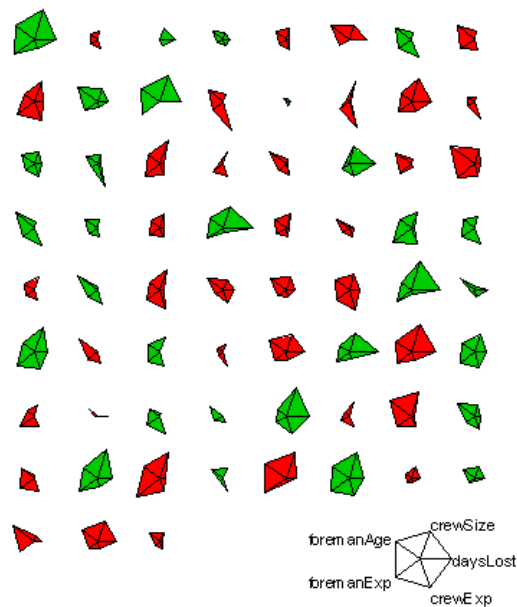


Figure 7: Stars of Each Case

- Parallel Coordinate Plots

Program 9 Parallel Coordinate Plots

```
> attach(lostDays); ## so lines are shorter  
> library(MASS);  
> parcoord(lostDays[,2:6],pch=16,col=2+powerTools);  
> ## color = 2 is red and represents no powerTools  
> ## color = 3 is green and represents presence of powerTools
```

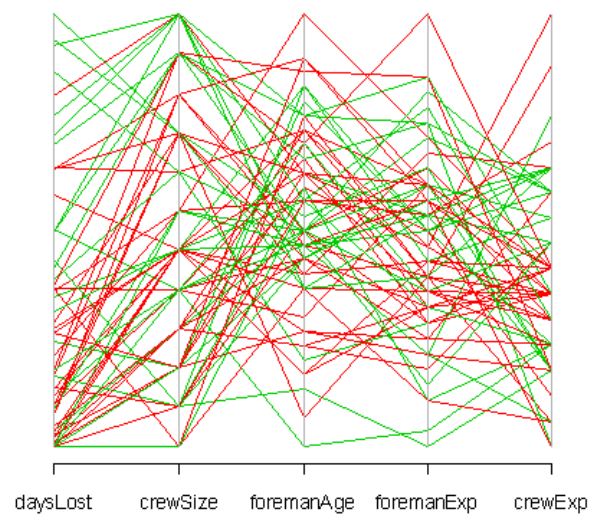


Figure 8: Parallel Coordinates by Power Tools

It appears there may be a relationship between the disconnect between the wise old man, and the inexperienced large crews. An older, experienced foreman loses lots of days regardless of power tools. On the other hand, the same older, experienced foreman minimizes days lost in other events. This may suggest a personal/likeability interaction (moderator) between the crew and the foreman (not captured in our data).

Problem 6

Conditioning Plots

- Conditioning Plots

Program 10 Conditioning Plots

```
> attach(lostDays); ## so lines are shorter
> par(mfrow=c(1,1));
> coplot(daysLost~crewExp|crewSize,data=lostDays,col="red");
> coplot(daysLost~crewExp|foremanExp,data=lostDays,col="red");
> coplot(daysLost~crewExp*crewSize|foremanExp,data=lostDays,col="red");
> coplot(daysLost~crewExp*powerTools|foremanExp,data=lostDays,col="red");
```

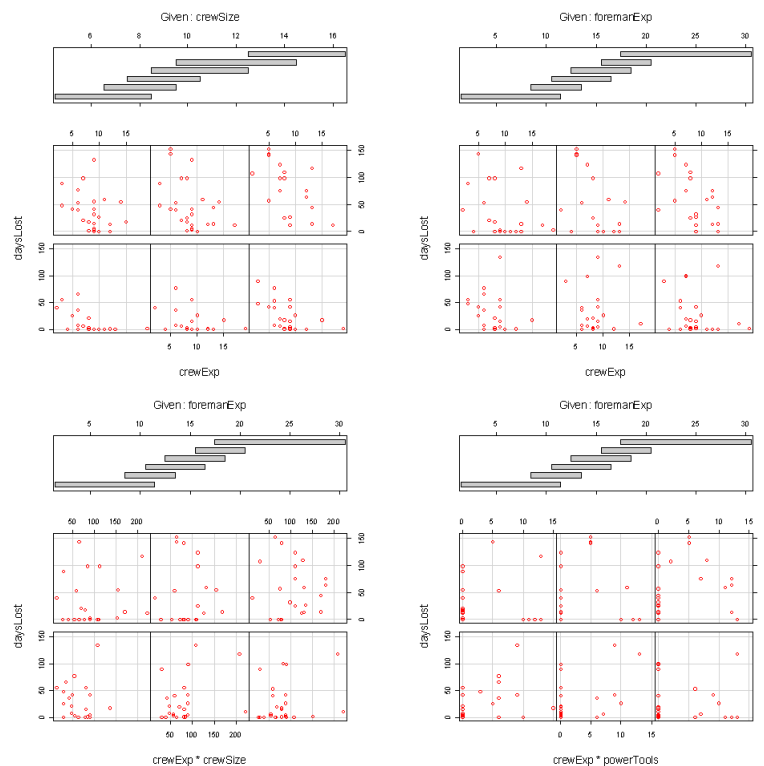


Figure 9: Conditional Effects on **daysLost**

As described on the previous problem, it appears that the interaction between the foreman and crew is interesting. "Love/Hate" or "Tough Love & Respect / Jerk" may be influencing the data. The first graph (crewSize/crewExp) is not surprising; especially if the crew is new, and the men are determining who is the Alpha Male in the group, unnecessary risks may surface (Happy Gilmore movie and construction days). The second graph suggests that crewExp is more important than foremanExp OR that the interpersonal dynamics between crew and foreman might exist. The last two graphs try to flush out interactions to explain the importance of foremanExp to daysLost considering crewExp and crewSize, powerTools respectively.