Initial Data Summaries

Lee Panter

```
#### Set Working Directory
WD="/Users/lee/Documents/GitHub/MSproject_RBC/MSproject_RBC/Scripts/Modeling/fbln-cd34"
setwd(WD)


### flowFilter
load("/Users/lee/Documents/Lee/School/CU Denver/MS_Project/Data:Scripts/FinalData/FilteredMergedData/

#### mdataFilter
load("/Users/lee/Documents/Lee/School/CU Denver/MS_Project/Data:Scripts/FinalData/FilteredMergedData/

#### seqFilter
load("/Users/lee/Documents/Lee/School/CU Denver/MS_Project/Data:Scripts/FinalData/FilteredMergedData/
```

Description

This script will produce numerical and graphical summaries of relevant models considered in each of the model developed as described in the ReadMe.

Begin Script

```
index.fbln=which(rownames(seqFilter)=="FBLN1")
index.cd34=which(rownames(seqFilter)=="CD34")
fbln=seqFilter[index.fbln,]
cd34=seqFilter[index.cd34,]
nFeature=mdataFilter$nFeature
nCount=mdataFilter$nCount
Perc.Mt=mdataFilter$Perc.Mt
subject.no=mdataFilter$subject.no
measurement=mdataFilter$measurement.name
dat=data.frame(subject.no, measurement, Perc.Mt, nCount, nFeature, cd34, fbln)
```

Exploratory Data Analysis

Following:

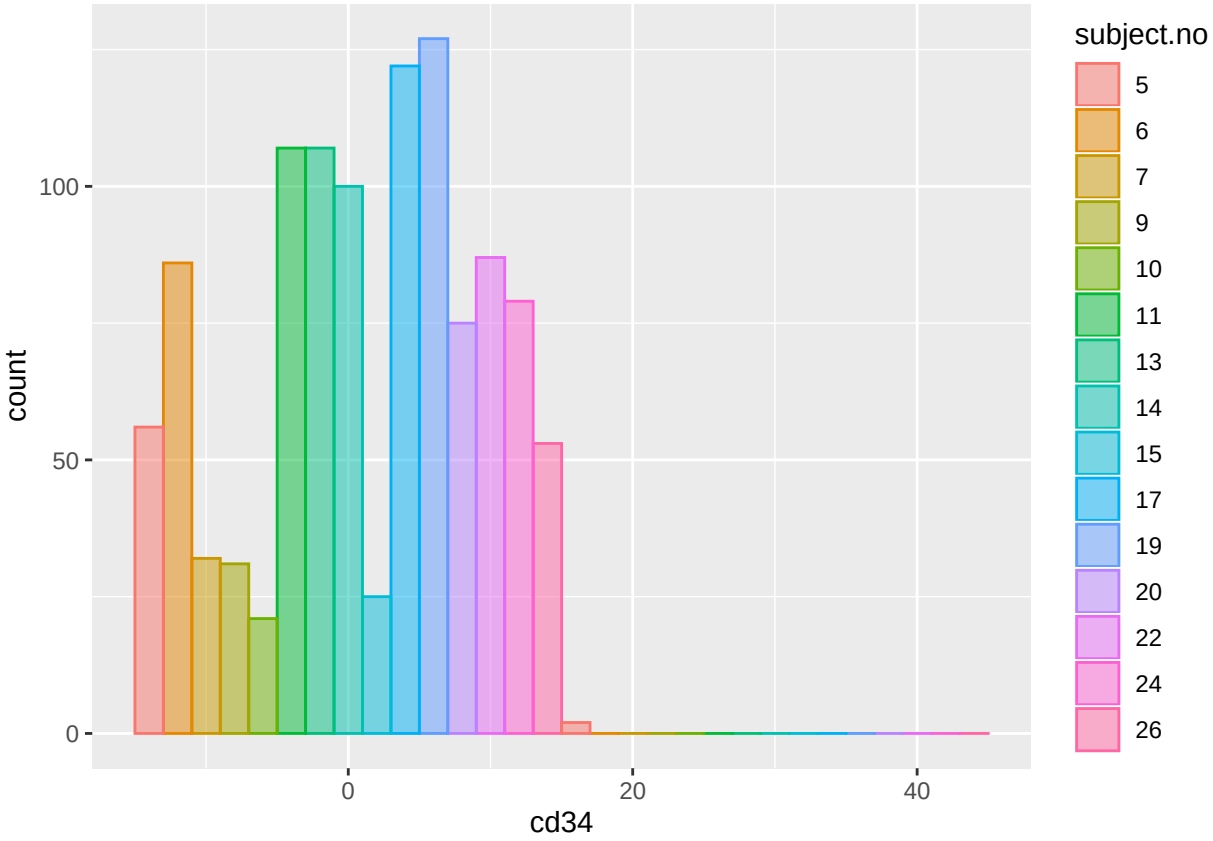
- Histogram plots of Predictor and Outcome
- Scatter Plot of Predictor v Outcome
- Numerical five number summaries of Predictor and Outcome

Predictor Summaries

```
# FIVE NUMBER SUMMARY
summary(dat$cd34)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4234  0.0000 19.0000
```

```
# HISTOGRAM
p1=ggplot(dat, aes(x=cd34,fill=subject.no, color=subject.no))+
  geom_histogram(alpha=0.5, position = "dodge", binwidth = 30)+
  theme(legend.position = "right")
p1
```

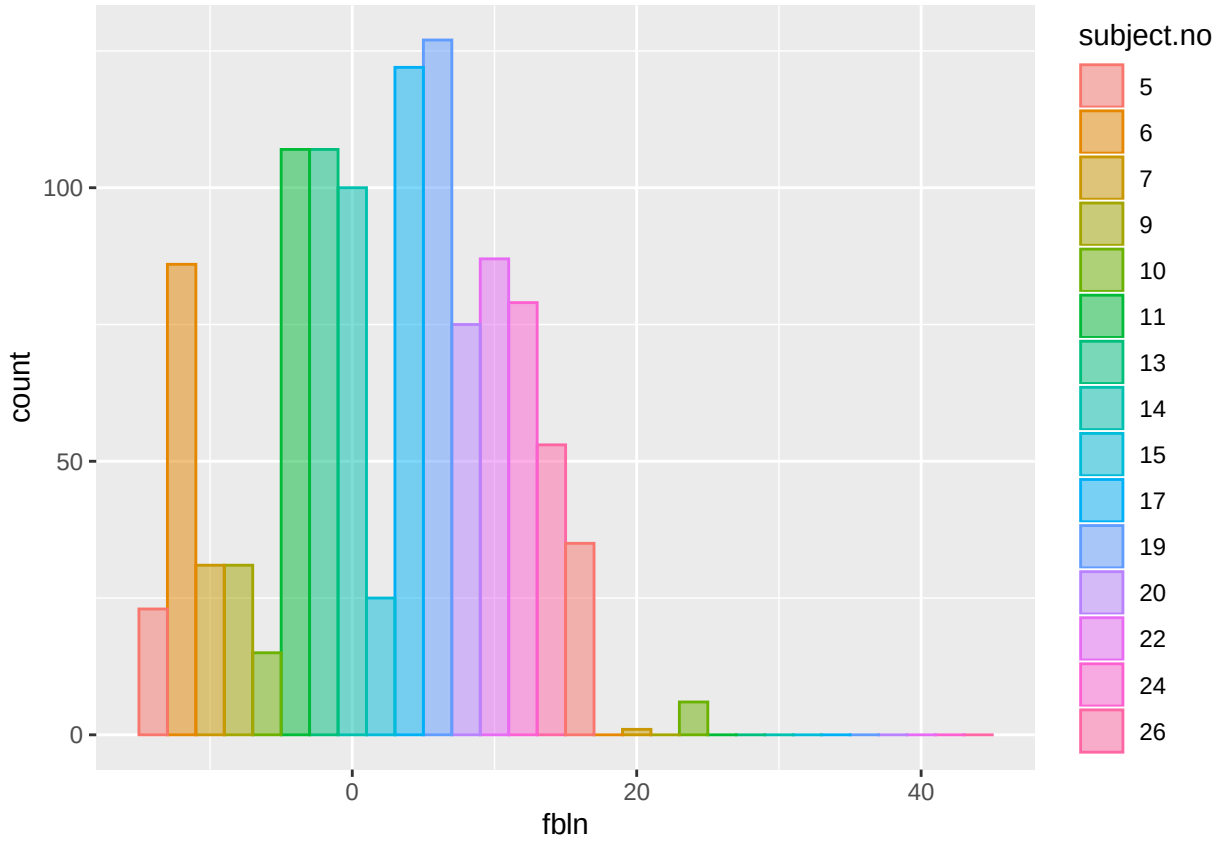


Outcome Summaries

```
summary(dat$fbln)
```

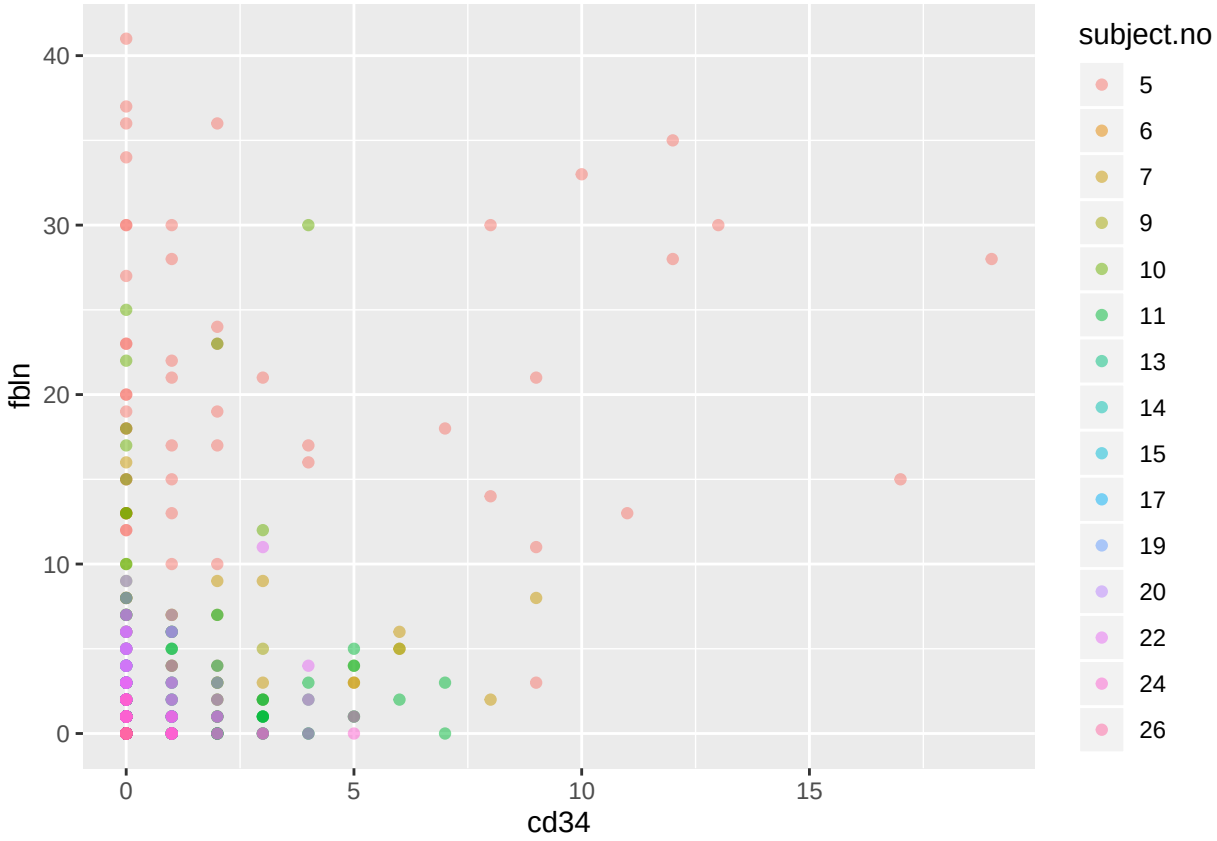
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.0     1.9     1.0    41.0
```

```
# HISTOGRAM
p2=ggplot(dat, aes(x=fbln,fill=subject.no, color=subject.no))+
  geom_histogram(alpha=0.5, position = "dodge", binwidth = 30)+
  theme(legend.position = "right")
p2
```



Scatter Plot Outcome ~ Predictor

```
p3=ggplot(dat, aes(x=cd34, y=fbln, color=subject.no))+
  geom_point(alpha=0.5)
p3
```



Transformed Variables (log transformations)

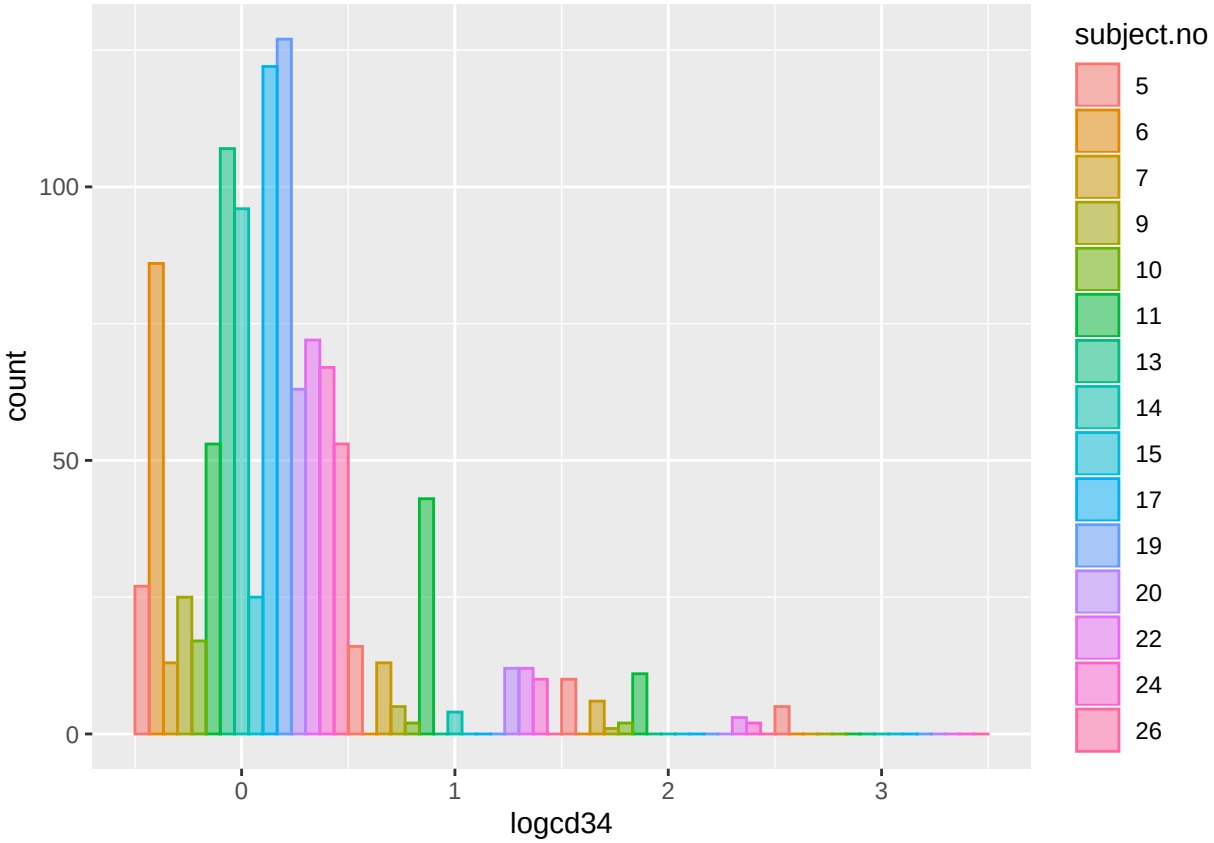
We will apply the transformation $Y = log(x + 1)$ to the outcome and response variables to create new-transformed variables.

```
dat$logcd34=log(cd34+1, base = exp(1))
dat$logfbln=log(fbln+1, base = exp(1))


# FIVE NUMBER SUMMARY
summary(dat$logcd34)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.169   0.000   2.996
```

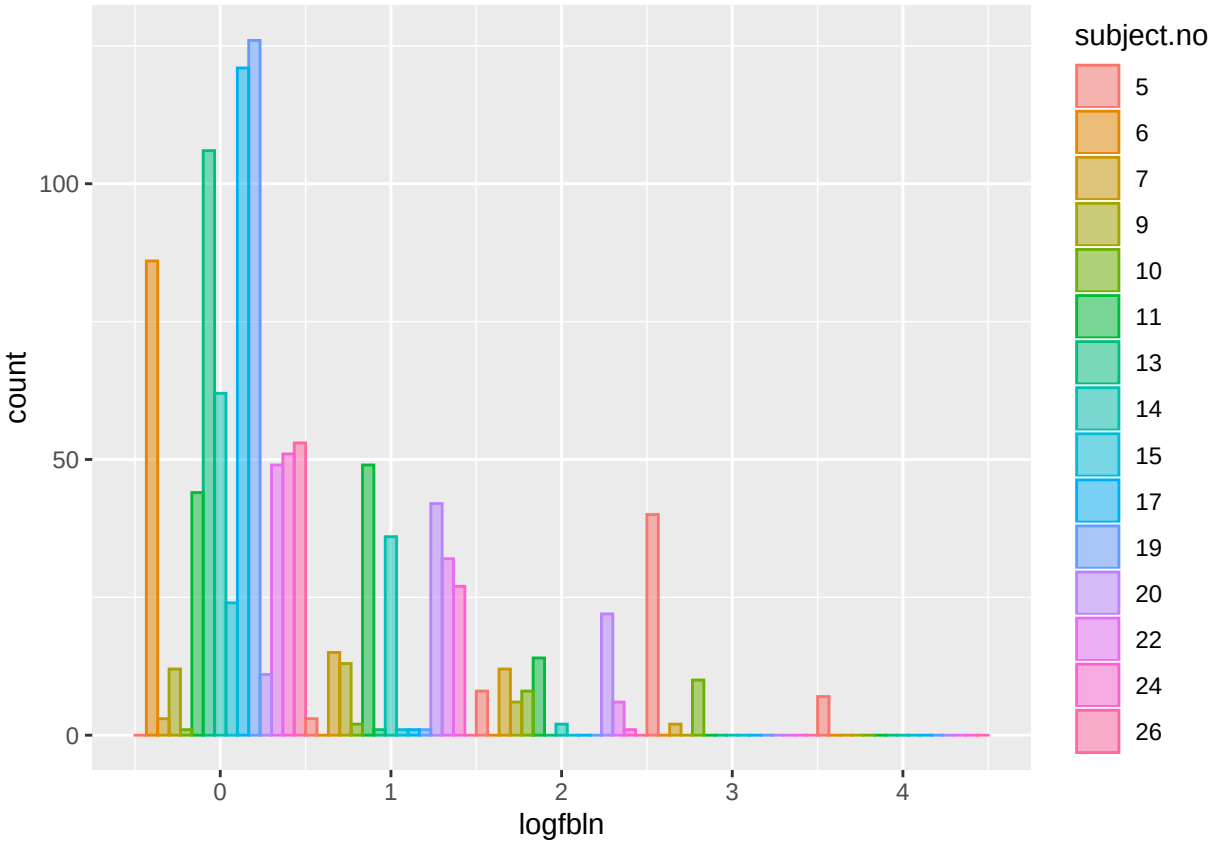
```
# HISTOGRAM
p1=ggplot(dat, aes(x=logcd34,fill=subject.no, color=subject.no))+
  geom_histogram(alpha=0.5, position = "dodge", binwidth = 1)+
  theme(legend.position = "right")
p1
```



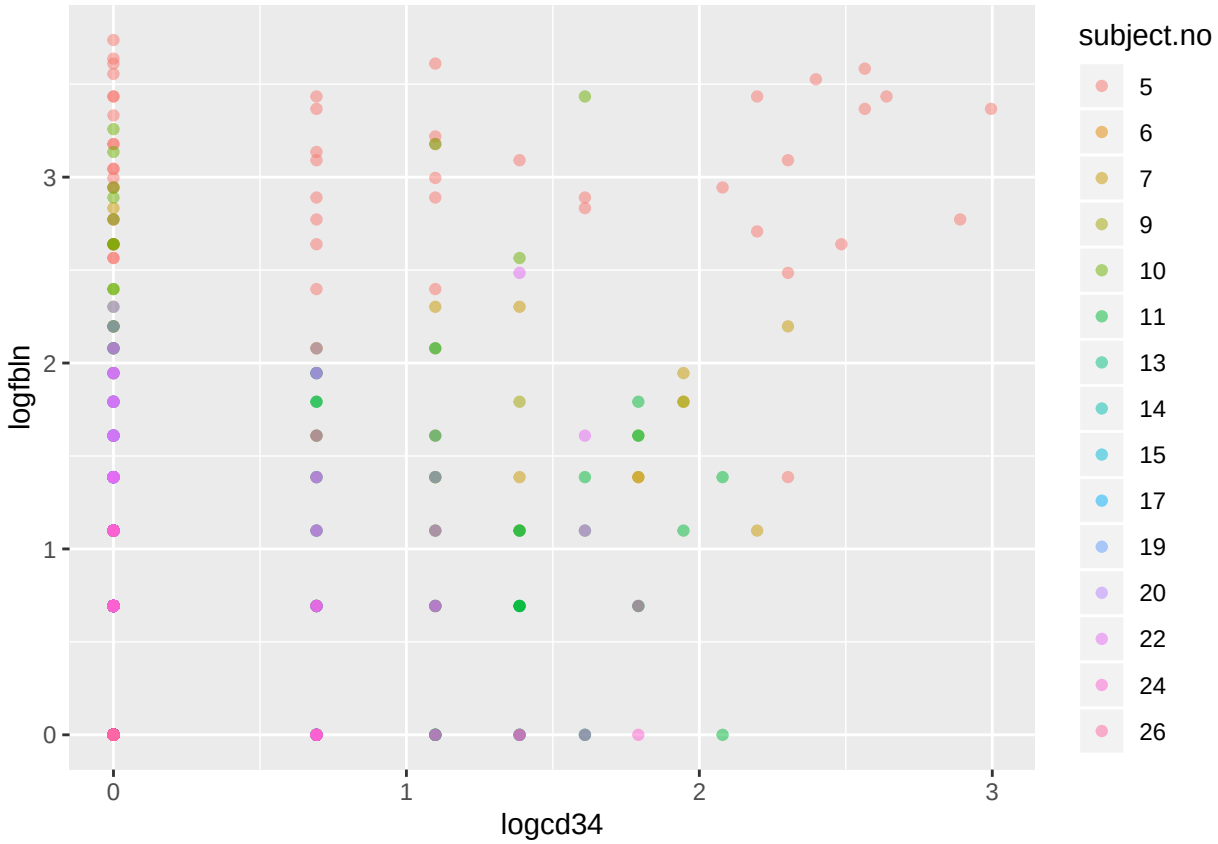

```
# FIVE NUMBER SUMMARY
summary(dat$logfbln)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4843  0.6931  3.7377
```

```
# HISTOGRAM
p1=ggplot(dat, aes(x=logfbln,fill=subject.no, color=subject.no))+
  geom_histogram(alpha=0.5, position = "dodge", binwidth = 1)+
  theme(legend.position = "right")
p1
```




```
p3=ggplot(dat, aes(x=logcd34, y=logfbln, color=subject.no))+
  geom_point(alpha=0.5)
p3
```


logfbln
logcd34
0
1
2
3
subject.no
5
6
7
9
10
11
13
14
15
17
19
20
22
24
26

End Script