Question_5

Mashhood Syed

October 2, 2014

Dataset 1 is based on US Census data and was part of a study done by some researchers at UCI. The dataset was used to in a ML algorithm to determine whether or not someone made >= $50k/year.
The link to the dataset is: http://archive.ics.uci.edu/ml/machine-learning -databases/adult/.

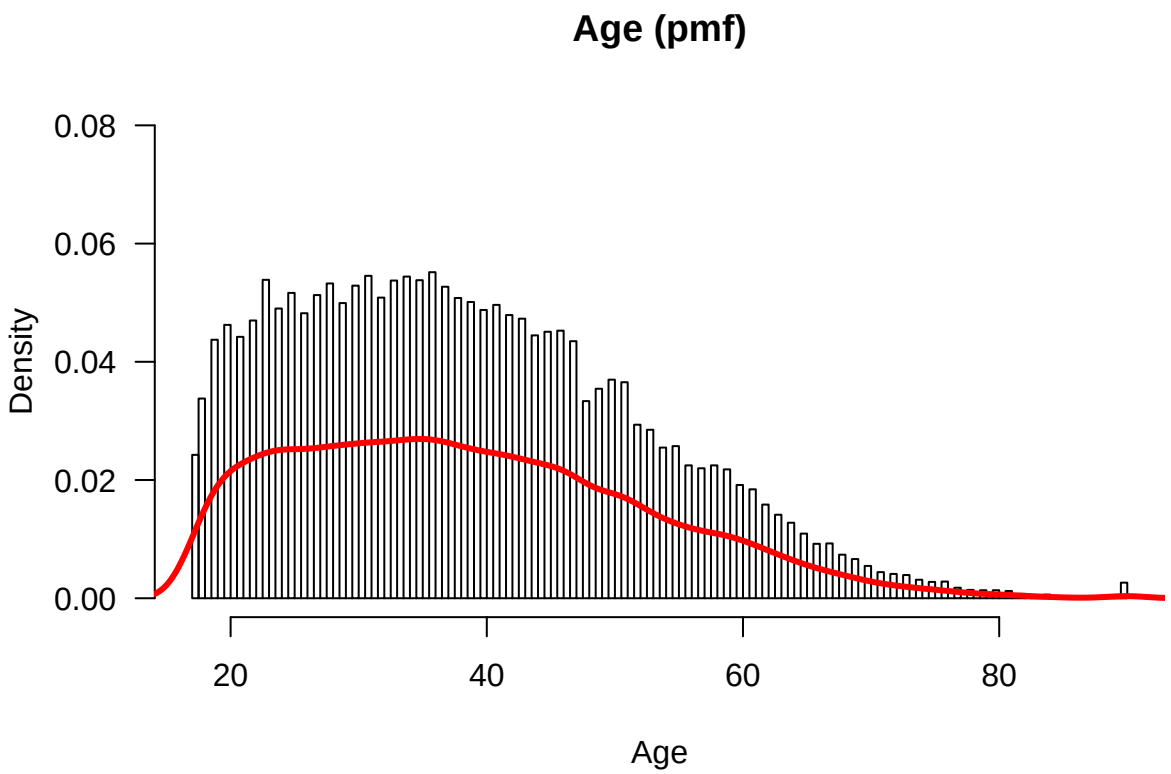
Dataset 2 is the count of people that came through 1 of three Express checkout cafe registers at the Wegman's in Hunt Valley on Friday October 3rd, 2014 between
3:07pm and 5:07pm. The data has 3 variables: Count of people, Hour Interval (3:07-4:07 = 1, 4:07-5:07 = 2), and register number.

Dataset 3 is the count of people that walked into Starbucks Coffee on Queen St in York, PA on Sunday, October 5, 2014 between 12pm and 2pm. The data has 3
variables: Count of people, 5 minute interval, and hour interval.

Question 5 Part 1 Dataset 1: We first set up the working directory, grab a subset of the original dataset and rename columns.

```
setwd("~/Desktop/Problem Set 1")
ds1 = "~/Desktop/Problem Set 1/random_variable1.tsv"
data1 = read.csv(ds1, header = F, sep = ",")
df1 = subset(data1, select = c("V1","V4","V6","V7","V9","V10","V15"))
colnames(df1)[1] = "Age"
colnames(df1)[2] = "Education"
colnames(df1)[3] = "Marital-status"
colnames(df1)[4] = "Occupation"
colnames(df1)[5] = "Race"
colnames(df1)[6] = "Gender"
colnames(df1)[7] = "IncomeLevel"
```

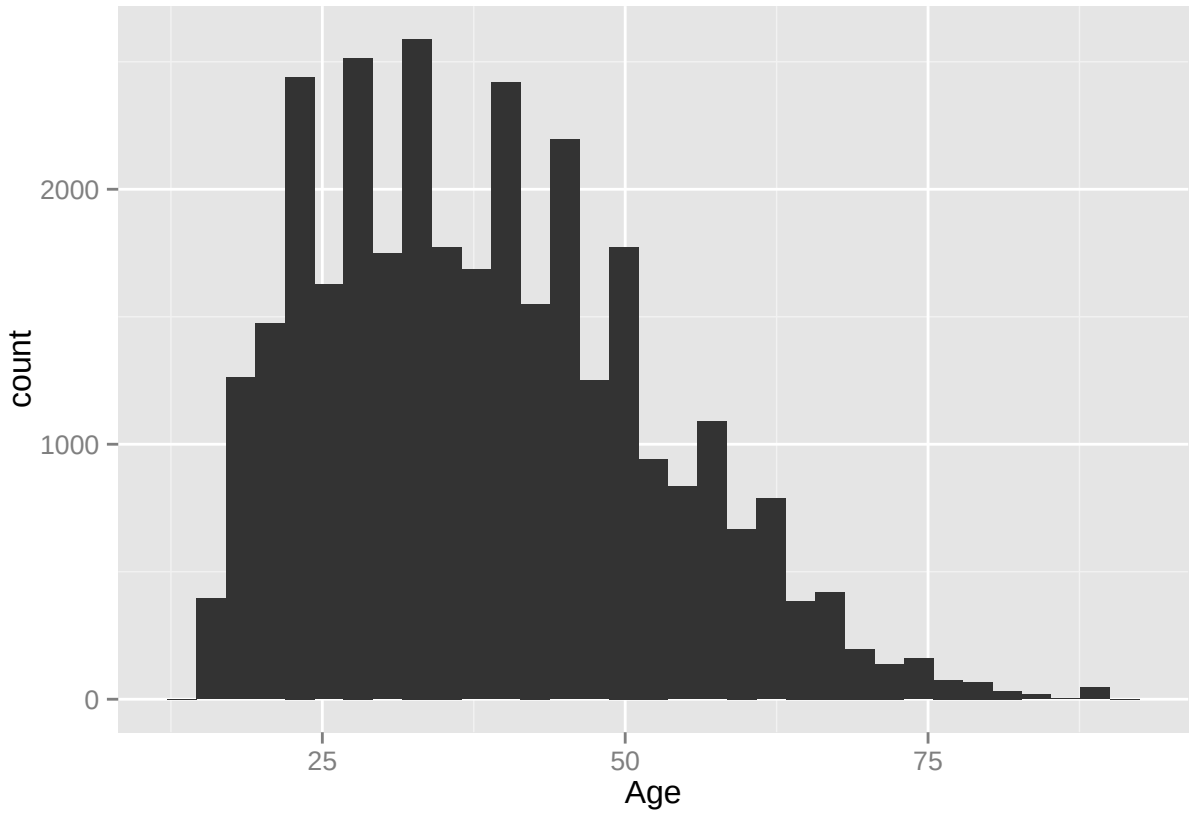
```
Here's a histogram representing the PMF for the Age Variable
```



Heres another histogram representing the frequency distribution for the Age Variable (using ggplot2)

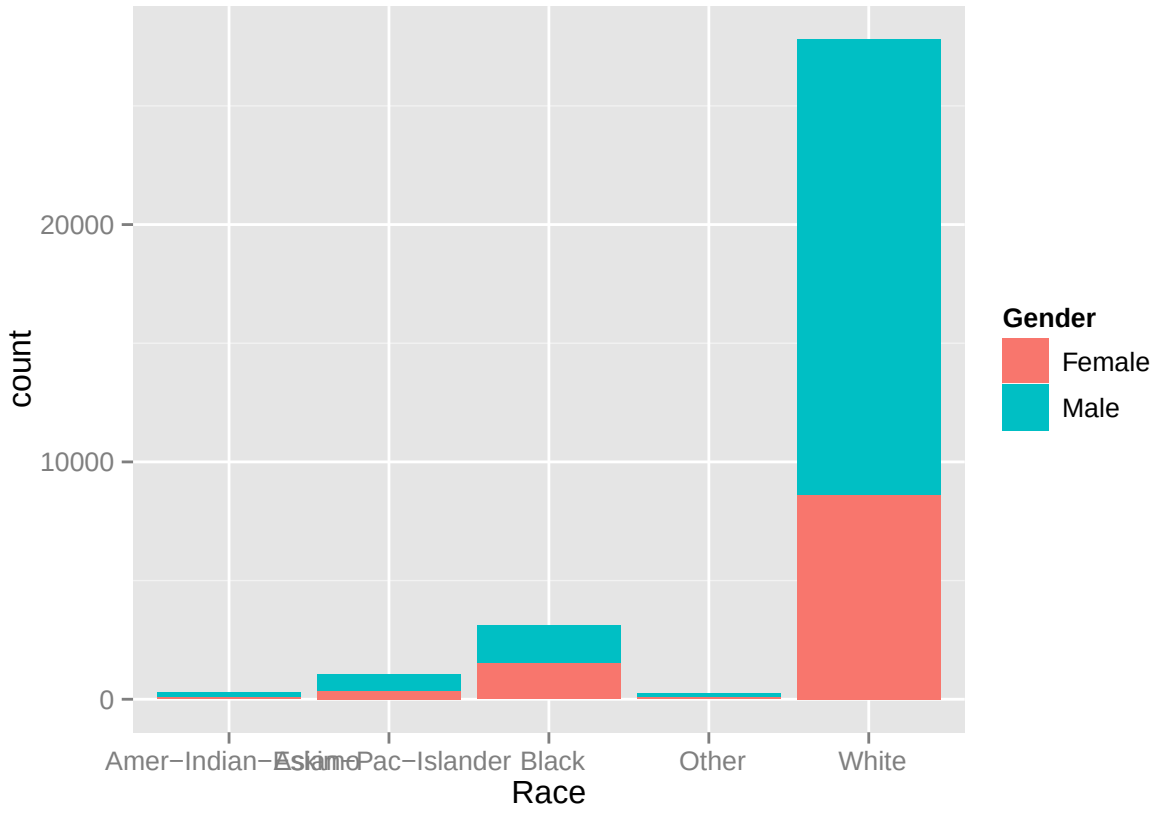
```
library(ggplot2)
x = ggplot(df1, aes(Age))
x + geom_histogram()
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



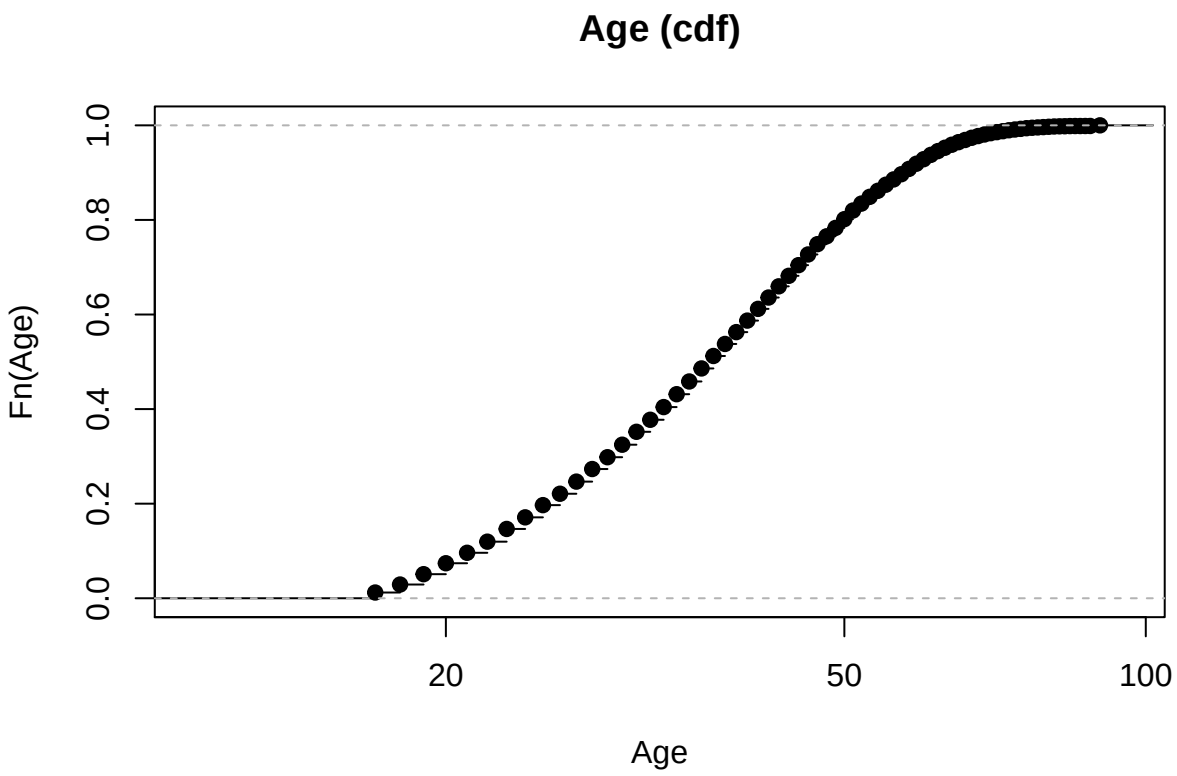
Here is a stacked histogram (using ggplot2) showing the distrbution of Sex for each Race

```
x = ggplot(df1, aes(Race, fill = Gender) )
x + geom_histogram()
```



Here is a simple CDF for the Age Variable

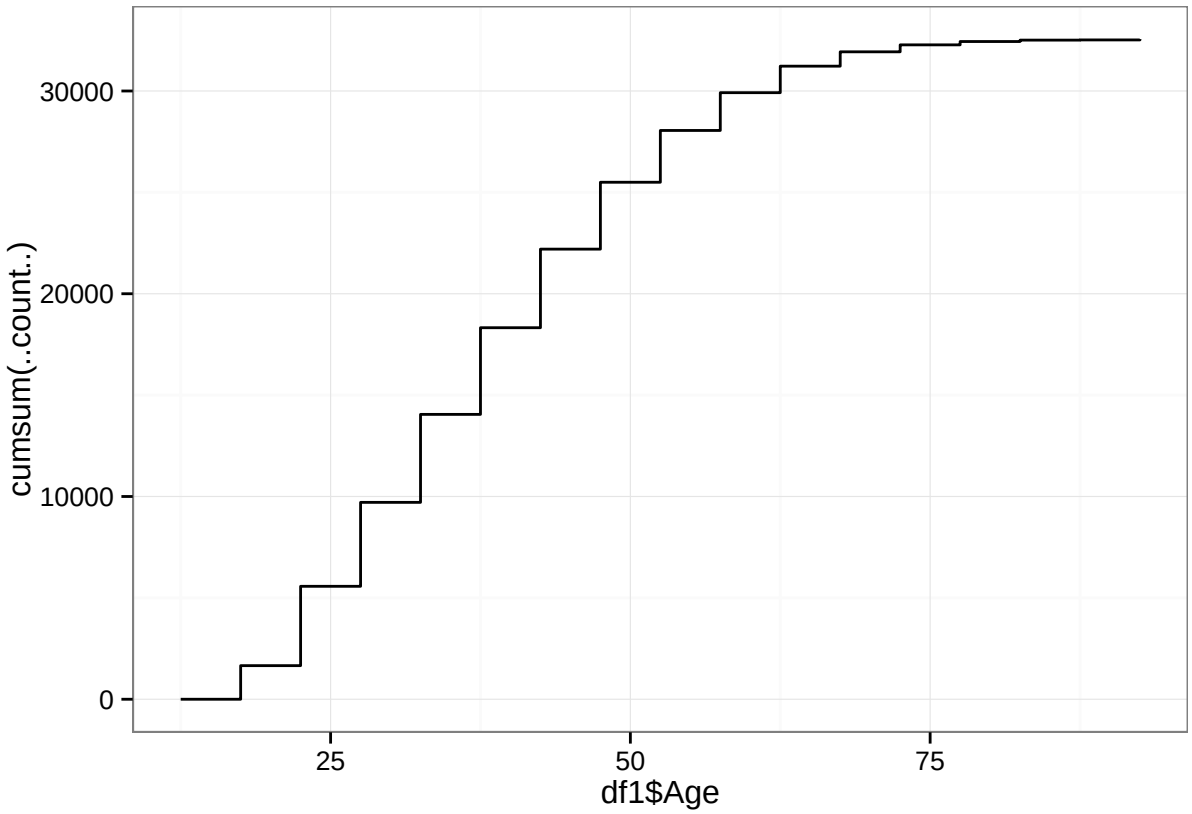
```
p = ecdf(df1$Age)
cdfAge = plot(p, ylab="Fn(Age)", xlab="Age", log="x", main="Age (cdf)")
```



Here is a more sophisticated CDF for the Age Variable

```
library(ggplot2)
ggplot(df1, aes(x = df1$Age),) + stat_bin(aes(y = cumsum(..count..)), geom = "step", colour = "black", i
```

```
## ymax not defined: adjusting position using y instead
```



Question 5 Part 2. Dataset 1:

Here is a summary of the data set

```
summary(df1)
```

```
##       Age                Education                    Marital-status
##  Min.   :17.0    HS-grad     :10501    Divorced              : 4443
##  1st Qu.:28.0    Some-college: 7291    Married-AF-spouse     :   23
##  Median :37.0    Bachelors   : 5355    Married-civ-spouse    :14976
##  Mean   :38.6    Masters     : 1723    Married-spouse-absent:  418
##  3rd Qu.:48.0    Assoc-voc   : 1382    Never-married         :10683
##  Max.   :90.0    11th        : 1175    Separated             : 1025
##                 (Other)      : 5134    Widowed               :  993
##             Occupation                    Race           Gender
##   Prof-specialty :4140    Amer-Indian-Eskimo:  311    Female:10771
##   Craft-repair   :4099    Asian-Pac-Islander: 1039    Male  :21790
##   Exec-managerial:4066    Black             : 3124
##   Adm-clerical   :3770    Other             :  271
##   Sales          :3650    White             :27816
##   Other-service  :3295
##  (Other)         :9541
##  IncomeLevel
##   <=50K:24720
##   >50K : 7841
##
##
##
```

```
##
##
```

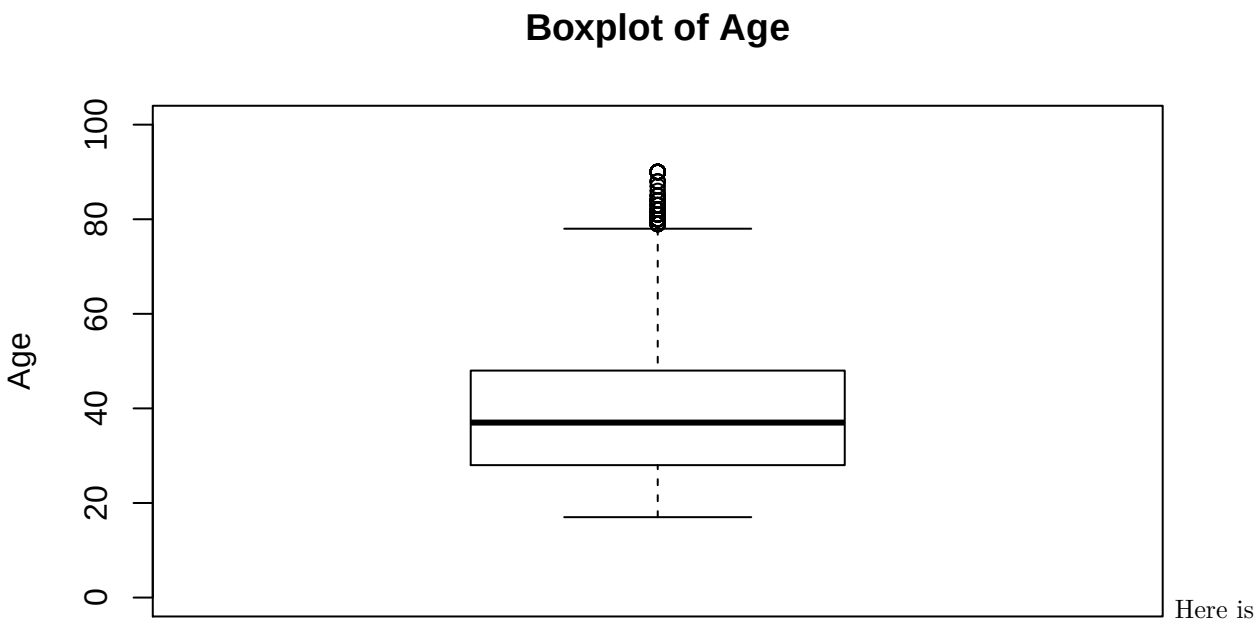
Here is a summary of the Age variable

```
summary(df1$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    17.0    28.0    37.0    38.6    48.0    90.0
```

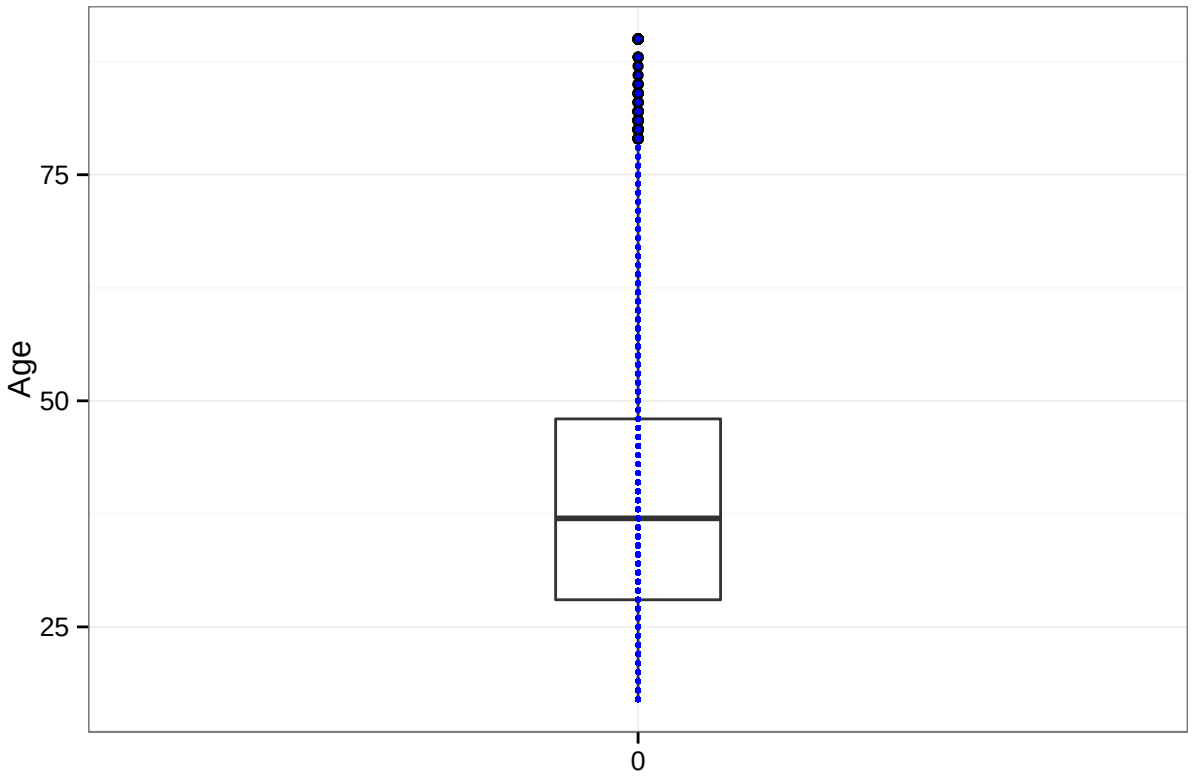
Here is a boxplot of the Age variable

```
boxplot(df1$Age, main="Boxplot of Age", ylab="Age", ylim=c(0,100))
```



Here is a boxplot using ggplot2 with Data Points for the Age variable

```
ggplot(df1, aes(x = factor(0), y = Age)) + geom_boxplot(width = .2) + ylab("Age") + xlab("") + geom_poi
```



Question 5 Part 3 Dataset 1:

Here is the MLE for the Age Variable

```
theta = c(0,1)
fn = function(theta) { sum( 0.5*(df1$Age-theta[1])^2/theta[2] + 0.5* log(theta[2]) ) }
nlm(fn, c(0,2), hessian = TRUE)
```

```
## Warning: NaNs produced
## Warning: NA/Inf replaced by maximum positive value
```

```
## $minimum
## [1] 101363
##
## $estimate
## [1]  38.58 186.06
##
## $gradient
## [1] -1.197e-03 -7.039e-07
##
## $hessian
##           [,1]     [,2]
## [1,] 175.00683 -0.00179
## [2,]  -0.00179  0.47012
##
## $code
## [1] 1
##
```

```
## $iterations
## [1] 27
```

I am not that sure on this part but from what I read online it looks as if the estimate of 38.58162 is very close to my actual mean of:

```
mean(df1$Age)
```

```
## [1] 38.58
```

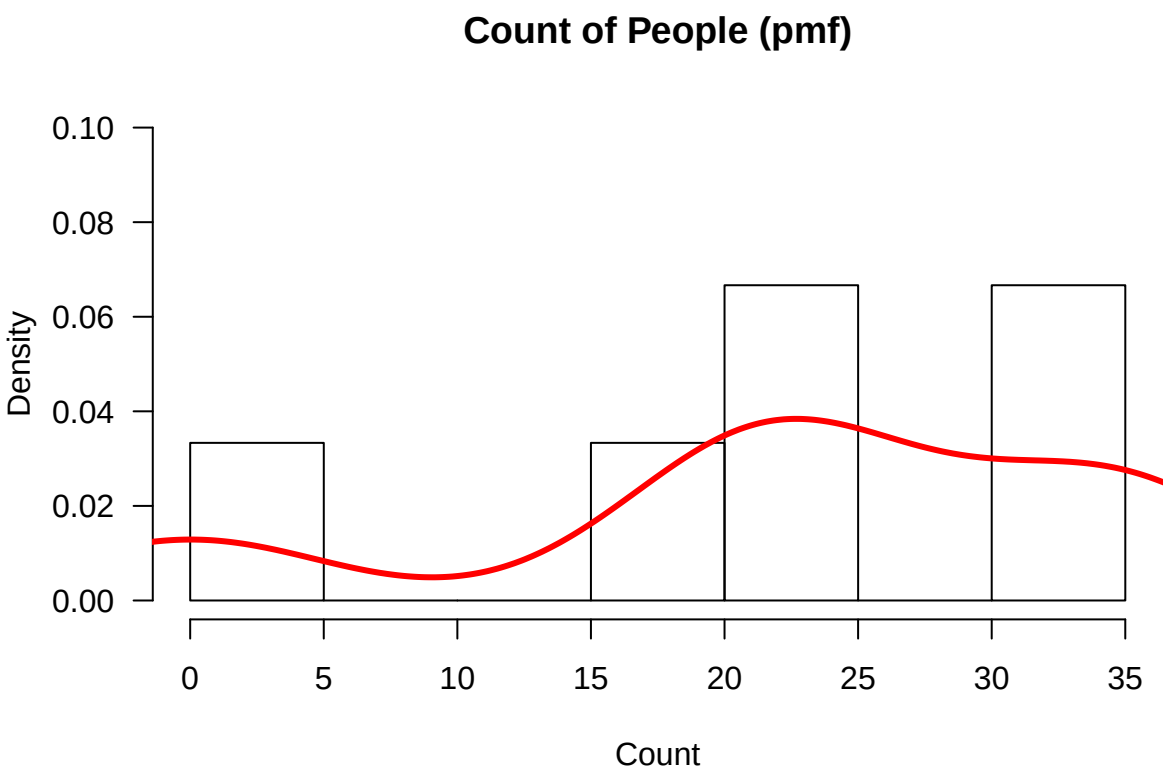
Question 5 Part 4 Dataset 1:

I expected to see a normal distribution. What I found is a normal distribution
with a "long right tail" in the 3rd and 4th quartiles. I think the reason for this is because there just are not that many people that still work for an income once you get into the age range of 65+.

Question 5 Part 1 Dataset 2: We first set up the working directory and place the data in a dataframe.

```
setwd("~/Desktop/Problem Set 1")
ds2 = "~/Desktop/Problem Set 1/random_variable2.tsv"
df2 = read.csv(ds2, header = T, sep = ",")
```

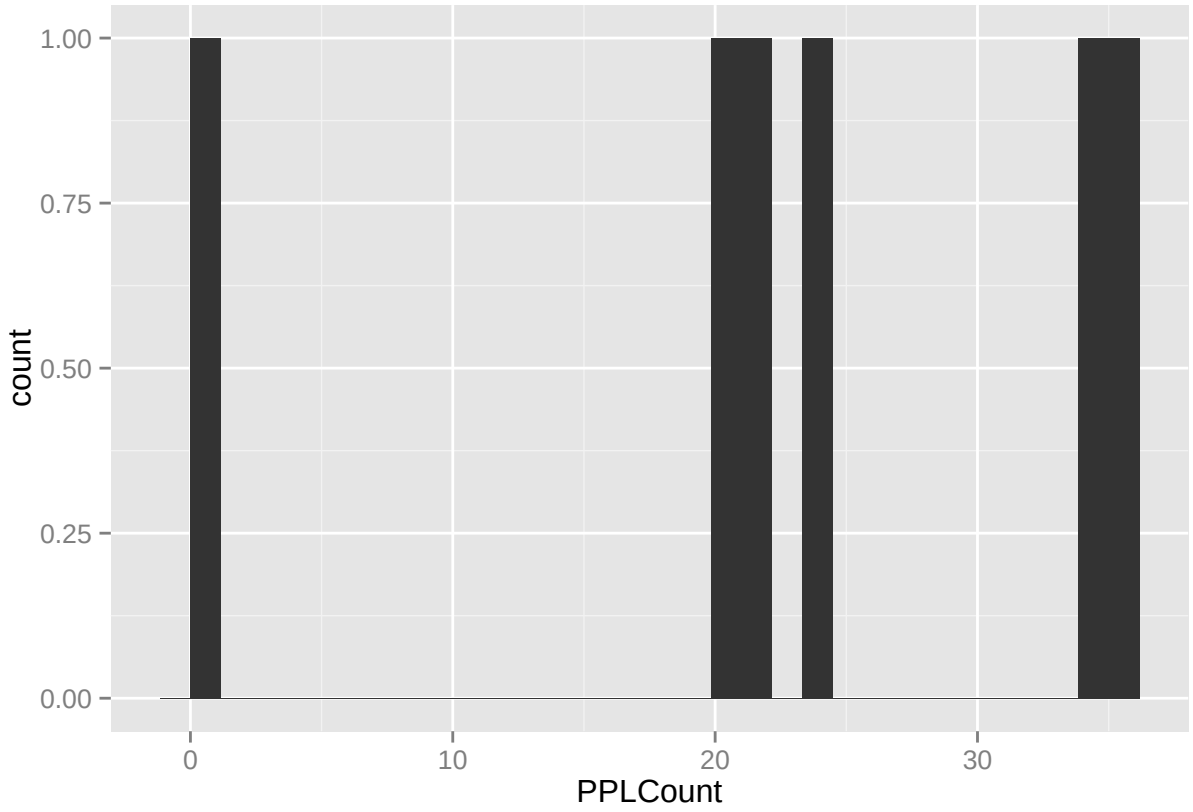
Here's a histogram representing the PMF for the Count of People:



Heres another histogram representing the frequency distribution for the PPLCount Variable (using ggplot2)

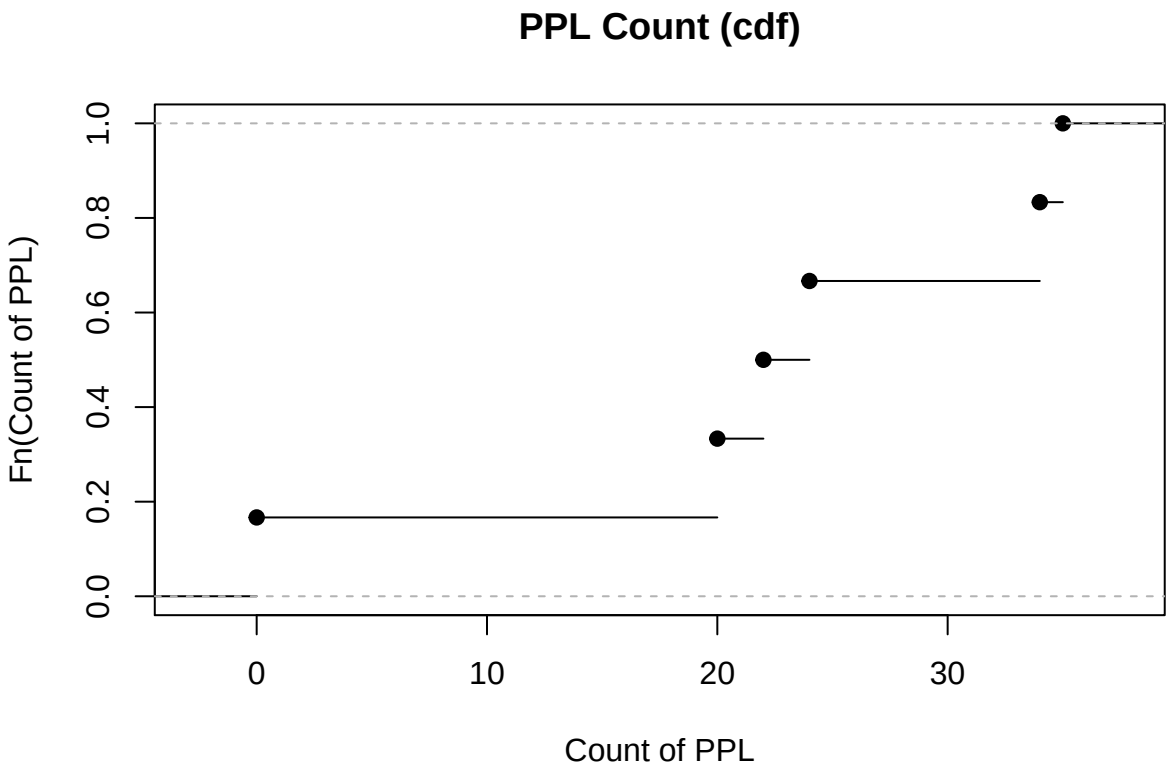
```
library(ggplot2)
x = ggplot(df2, aes(PPLCount))
x + geom_histogram()
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



Here is a simple CDF for the PPLCount Variable

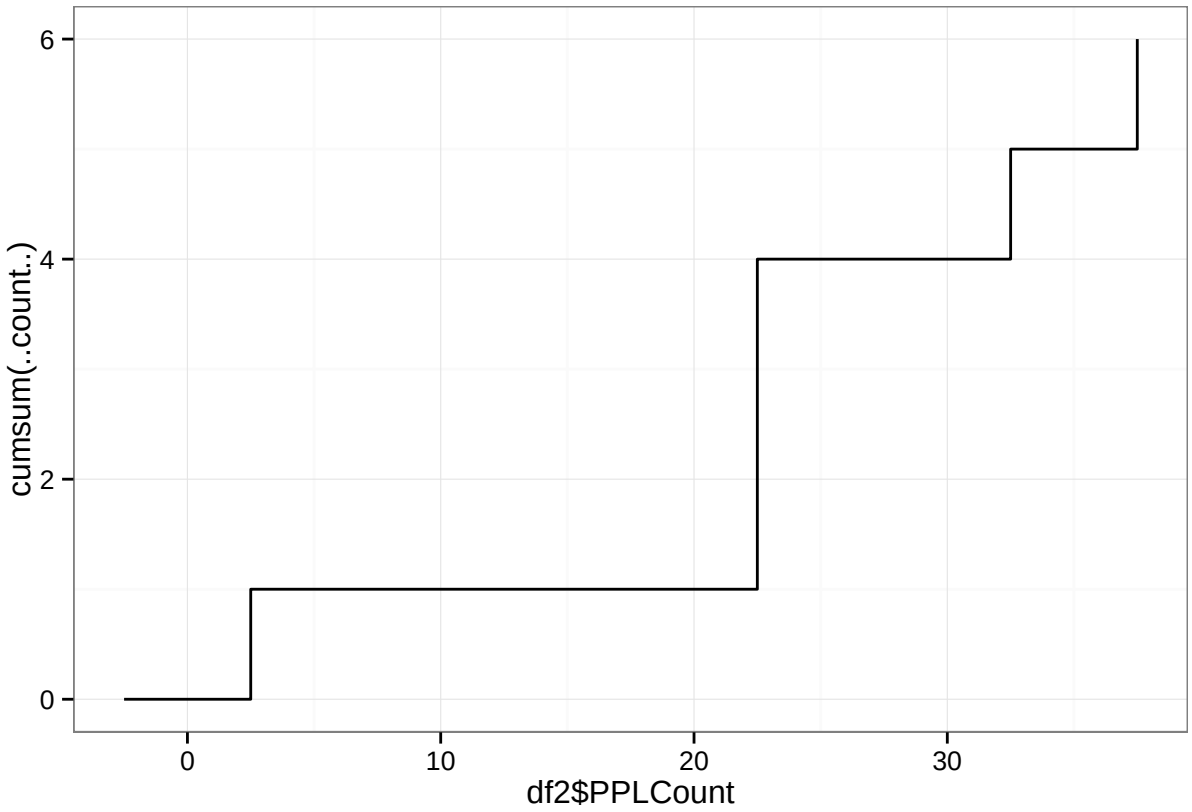
```
p = ecdf(df2$PPLCount)
cdfAge = plot(p, ylab="Fn(Count of PPL)", xlab="Count of PPL", main="PPL Count (cdf)")
```



Here is a more sophisticated CDF for the PPLCount Variable

```
library(ggplot2)
ggplot(df2, aes(x = df2$PPLCount),) + stat_bin(aes(y = cumsum(..count..)), geom = "step", colour = "bla
```

```
## ymax not defined: adjusting position using y instead
```



Question 5 Part 2 Dataset 2:

Here is a summary of the data set

```
summary(df2)
```

```
##     PPLCount       Register     TimeInterval
##  Min.   : 0.0   Min.   :1.00   Min.   :1.0
##  1st Qu.:20.5   1st Qu.:1.25   1st Qu.:1.0
##  Median :23.0   Median :2.00   Median :1.5
##  Mean   :22.5   Mean   :2.00   Mean   :1.5
##  3rd Qu.:31.5   3rd Qu.:2.75   3rd Qu.:2.0
##  Max.   :35.0   Max.   :3.00   Max.   :2.0
```

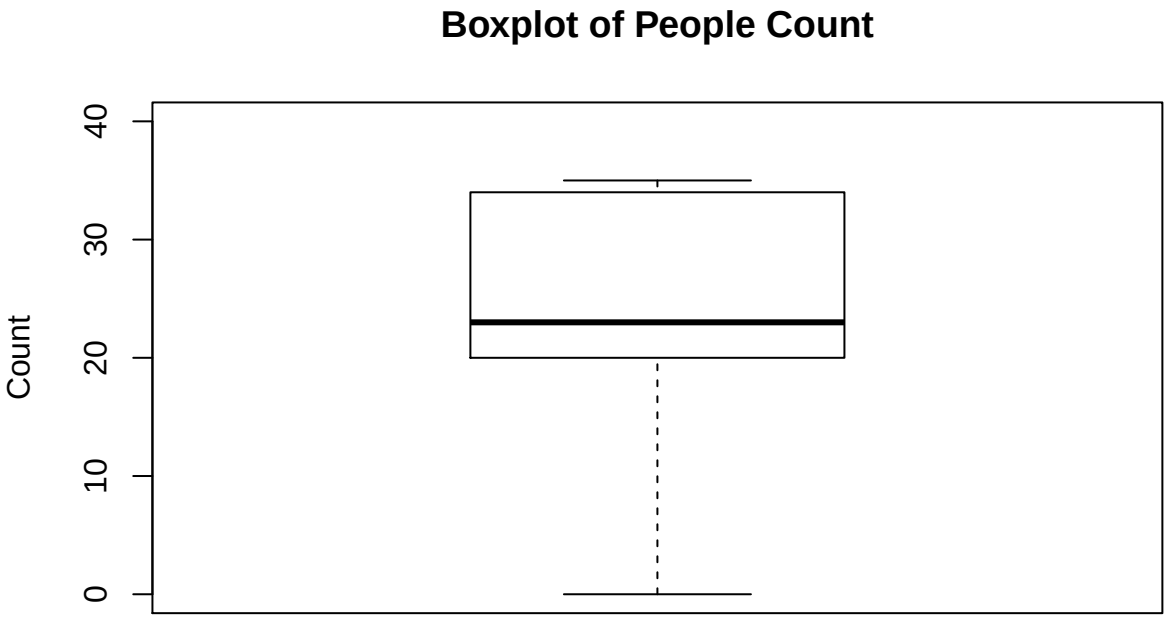
Here is a summary of the PPLCount variable

```
summary(df2$PPLCount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    20.5    23.0    22.5    31.5    35.0
```

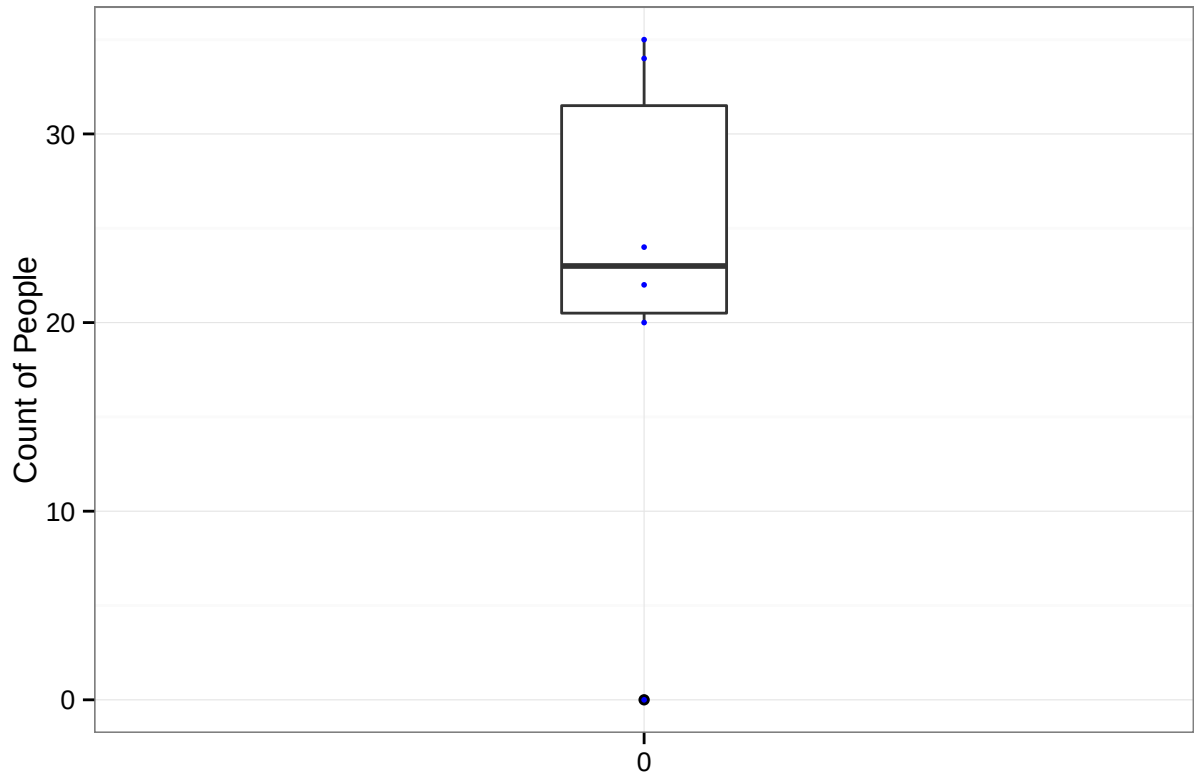
Here is a boxplot of the PPLCount variable

```
boxplot(df2$PPLCount, main="Boxplot of People Count", ylab="Count", ylim=c(0,40))
```



Here is a boxplot using ggplot2 with Data Points for the PPLCount variable

```
ggplot(df2, aes(x = factor(0), y = PPLCount)) + geom_boxplot(width = .2) + ylab("Count of People") + xla
```



Question 5 Part 3 Dataset 2:

```
Here is the MLE for the PPLCount Variable
```

```r
theta = c(0,1)
fn = function(theta) { sum( 0.5*(df2$PPLCount-theta[1])^2/theta[2] + 0.5* log(theta[2]) ) }
nlm(fn, c(0,2), hessian = TRUE)
```

```
## Warning: NaNs produced
## Warning: NA/Inf replaced by maximum positive value
```

```
## $minimum
## [1] 17.69
##
## $estimate
## [1]  22.5 133.9
##
## $gradient
## [1]  2.010e-07 -2.414e-09
##
## $hessian
##            [,1]       [,2]
## [1,]  4.480e-02 -3.741e-07
## [2,] -3.741e-07  1.672e-04
##
## $code
## [1] 1
##
## $iterations
## [1] 22
```

I am not that sure on this part but from what I read online it looks as if the estimate of 22.49 is very close to my actual mean of:

```r
mean(df2$PPLCount)
```

```
## [1] 22.5
```

Question 5 Part 4 Dataset 2:

After conducting the experiment and looking at my data, I realized that it would have been more valuable to have broken up each hour into smaller intervals. Its important for me to mention that there were a total of 4 registers. 3 registers were in my line of sight from where I was sitting. Therefore only three of the 4 total registers were included in the observations.

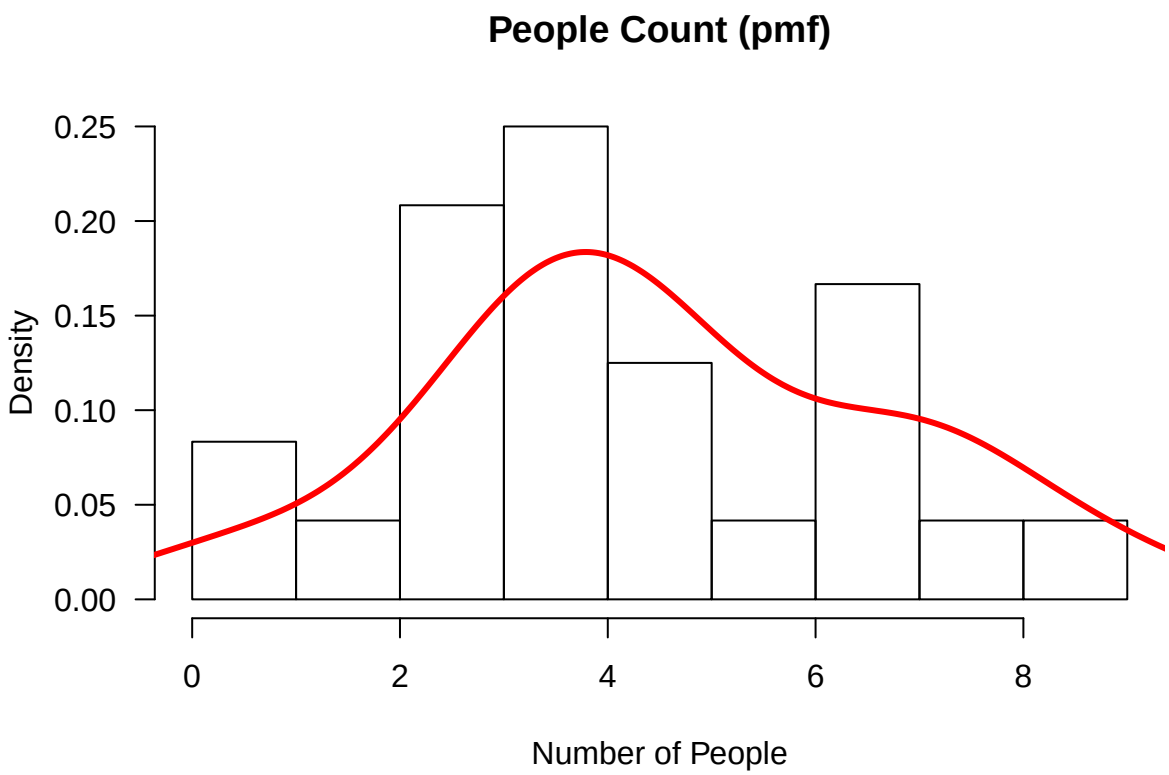
At the time of data collection, I was interested in seeing if there would be a great difference between the number of people going to one particular register over another. And for this reason, I didnt pay much attention to the time interval part. From the data you can see that the number of people that checked out of registers 1 and 2 are about equal when you take the total per hour count. My hypothesis was that the 2nd register would have more people checked out because the view of the 1st register was partialy blocked from the customers line of sight by a 1/2 foot wide column. What I ended up observing was that as the number of people in one checkout line grew past 3, the next person in the line would stop to see if any other registers were open. This behavior is what allowed each register to have roughly the same number of people checkout over the course of the 2 hour observation.

—

Question 5 Part 1 Dataset 3: We first set up the working directory, grab a subset of the original dataset and rename columns.

```
setwd("~/Desktop/Problem Set 1")
ds3 = "~/Desktop/Problem Set 1/random_variable3.tsv"
df3 = read.csv(ds3, header = T, sep = ",")
```

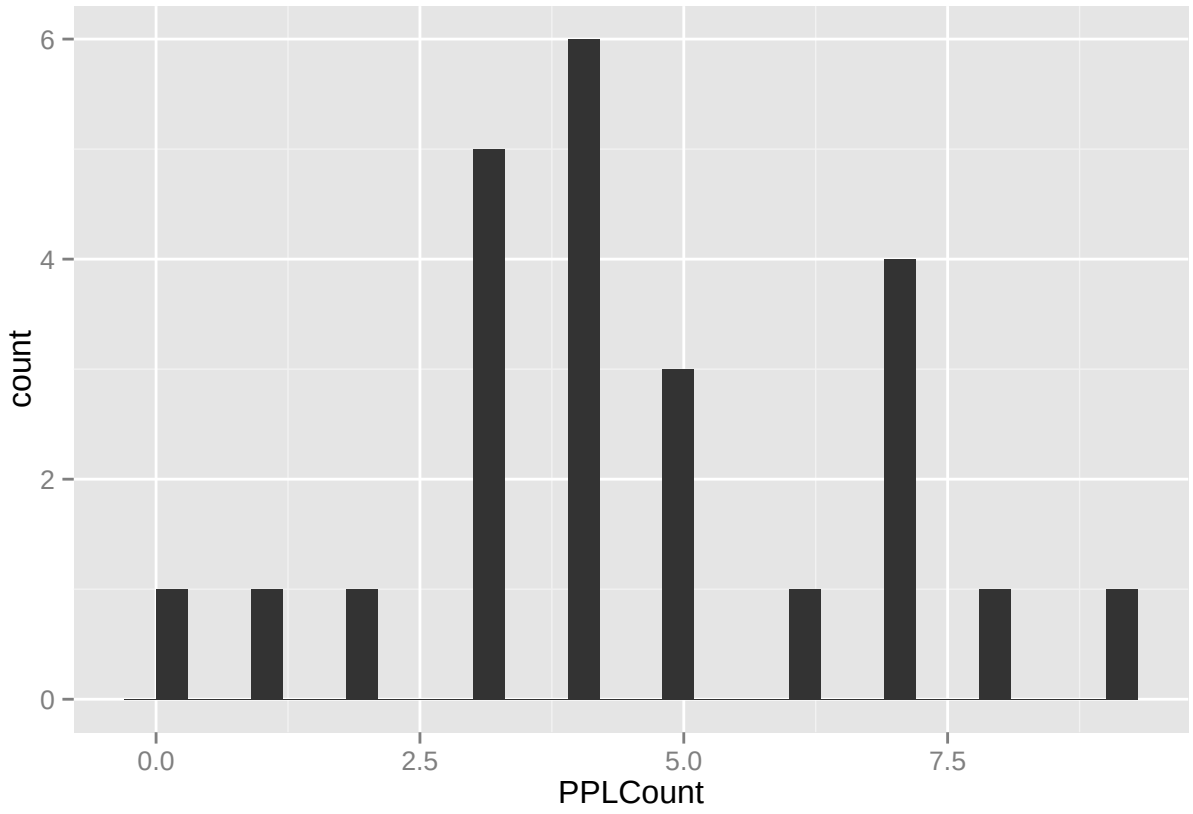
Here's a histogram representing the PMF for the PPLCount Variable



Heres another histogram representing the frequency distribution for the PPLCount Variable (using ggplot2)

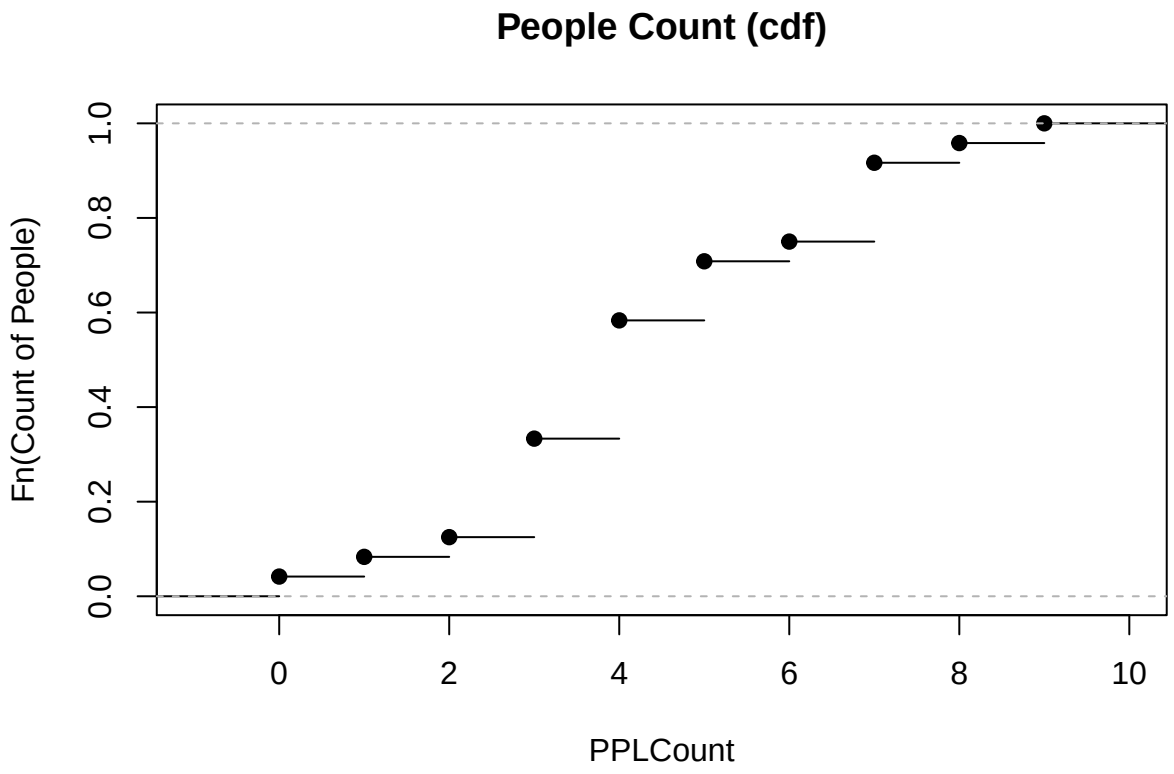
```
library(ggplot2)
x = ggplot(df3, aes(PPLCount))
x + geom_histogram()
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



Here is a simple CDF for the PPLCount Variable

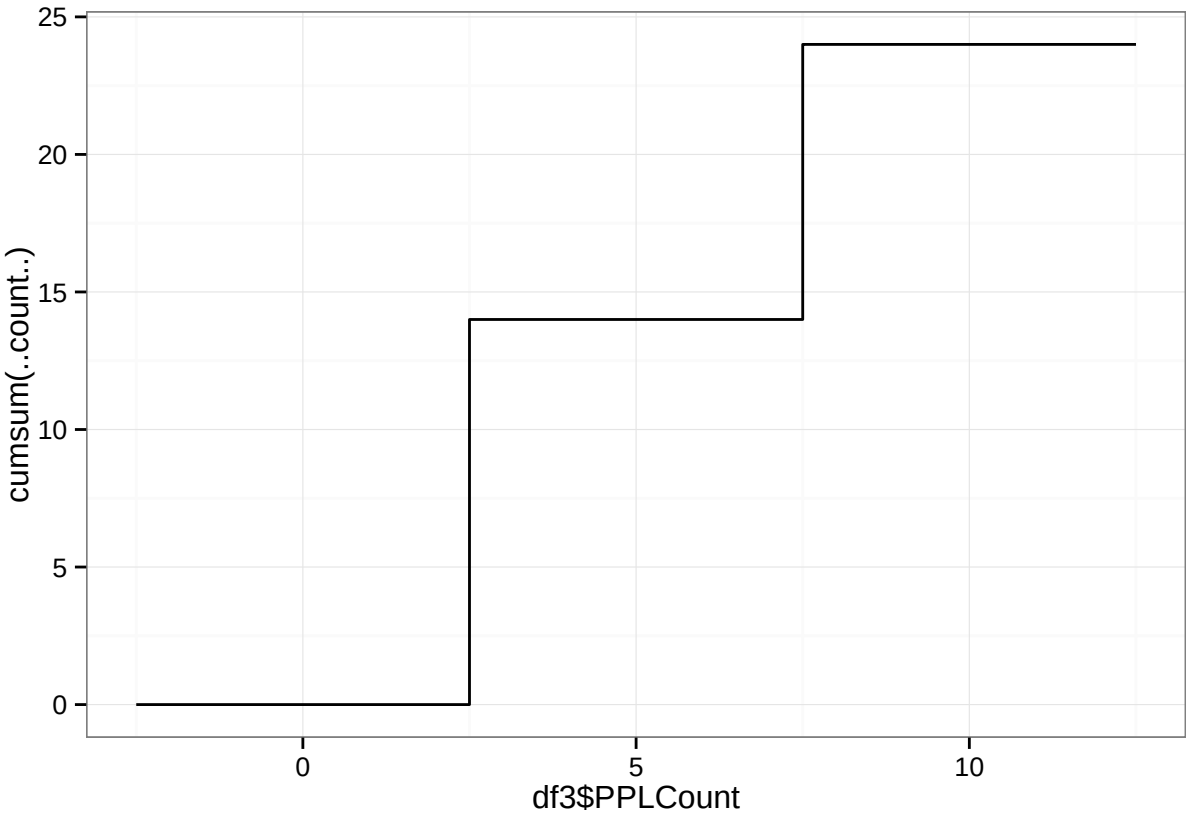
```
p = ecdf(df3$PPLCount)
cdfAge = plot(p, ylab="Fn(Count of People)", xlab="PPLCount", main="People Count (cdf)")
```



Here is a more sophisticated CDF for the PPLCount Variable

```
library(ggplot2)
ggplot(df3, aes(x = df3$PPLCount),) + stat_bin(aes(y = cumsum(..count..)), geom = "step", colour = "bla
```

```
## ymax not defined: adjusting position using y instead
```



Question 5 Part 2 Dataset 3:

Here is a summary of the data set

```
summary(df3)
```

```
##  X5MinInterval        Hour        PPLCount
##  Min.   : 0.00   Min.   :1.0   Min.   :0.00
##  1st Qu.: 2.75   1st Qu.:1.0   1st Qu.:3.00
##  Median : 5.50   Median :1.5   Median :4.00
##  Mean   : 5.50   Mean   :1.5   Mean   :4.50
##  3rd Qu.: 8.25   3rd Qu.:2.0   3rd Qu.:6.25
##  Max.   :11.00   Max.   :2.0   Max.   :9.00
```

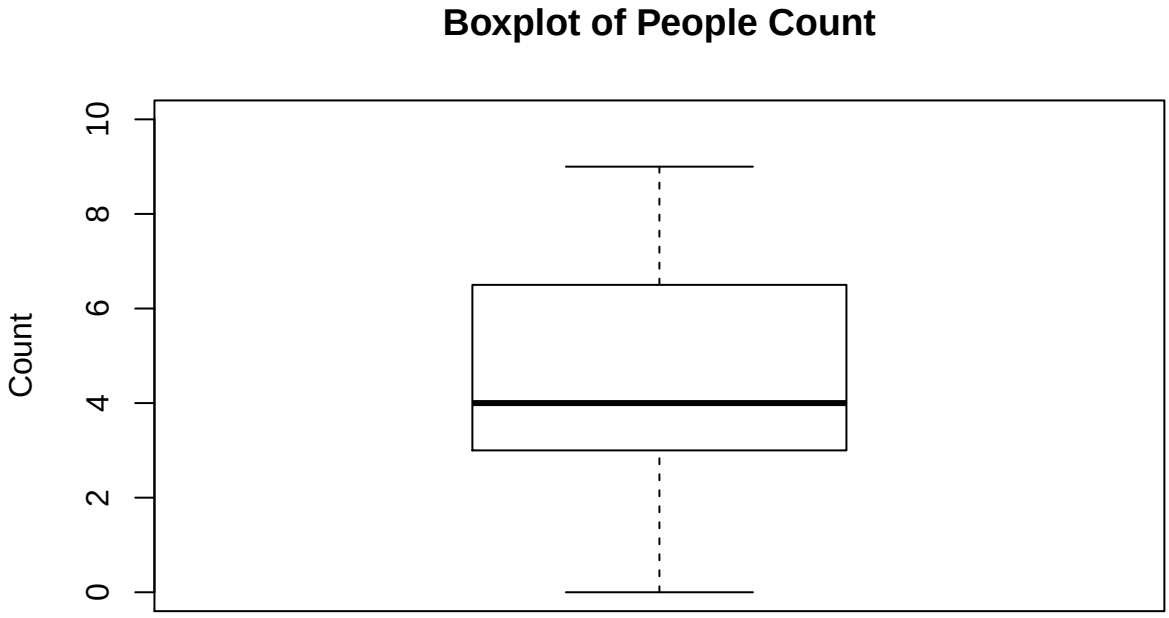
Here is a summary of the PPLCount variable

```
summary(df3$PPLCount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    3.00    4.00    4.50    6.25    9.00
```

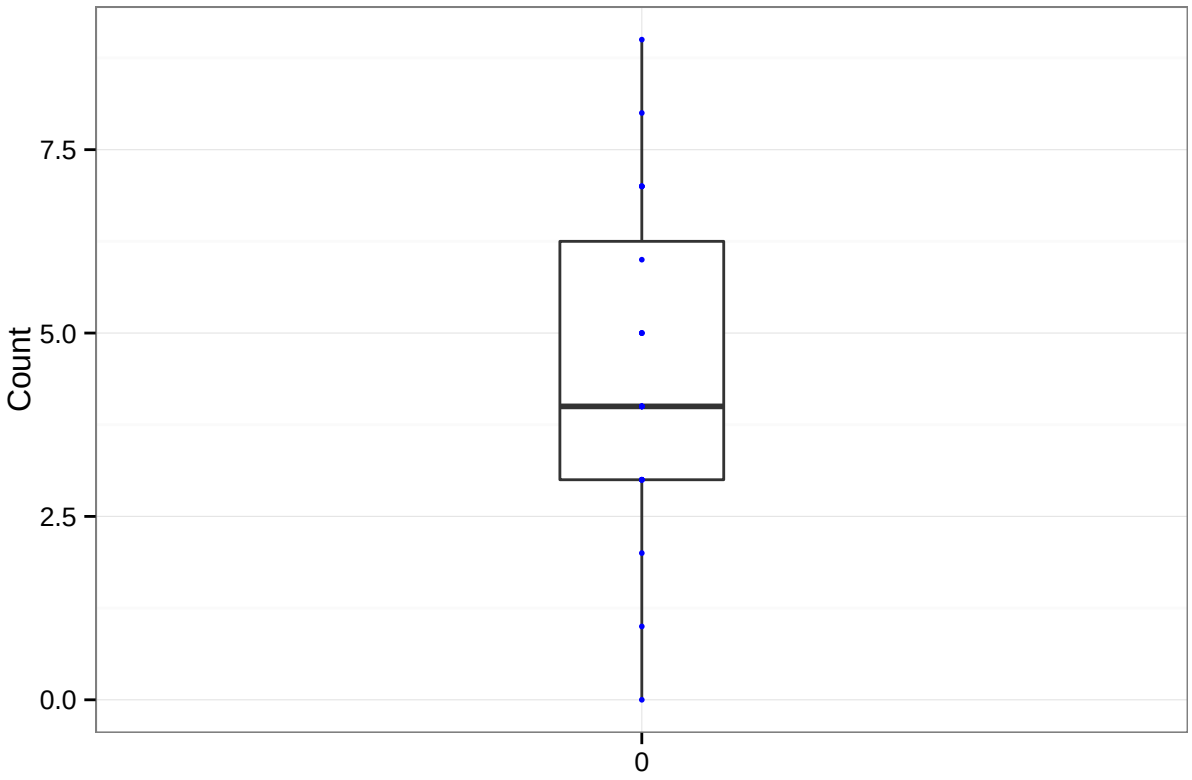
Here is a boxplot of the PPLCount variable

```
boxplot(df3$PPLCount, main="Boxplot of People Count", ylab="Count", ylim=c(0,10))
```



Here is a boxplot using ggplot2 with Data Points for the PPLCount variable

```
ggplot(df3, aes(x = factor(0), y = PPLCount)) + geom_boxplot(width = .2) + ylab("Count") + xlab("") + g
```



Question 5 Part 3 Dataset 3:

```
Here is the MLE for the PPLCount Variable
```

```
theta = c(0,1)
fn = function(theta) { sum( 0.5*(df3$PPLCount-theta[1])^2/theta[2] + 0.5* log(theta[2]) ) }
nlm(fn, c(0,2), hessian = TRUE)
```

```
## $minimum
## [1] 30.49
##
## $estimate
## [1] 4.500 4.667
##
## $gradient
## [1] 2.337e-07 7.232e-08
##
## $hessian
##            [,1]       [,2]
## [1,]  5.1428596 -0.0002455
## [2,] -0.0002455  0.5508011
##
## $code
## [1] 1
##
## $iterations
## [1] 17
```

I am not that sure on this part but from what I read online it looks as if the estimate of 4.49 is very close to my actual mean of:

```
mean(df3$PPLCount)
```

```
## [1] 4.5
```

Question 5 Part 4 Dataset 3:

I expected to see a normal distribution and that is what I saw after running the Histogram.
When collecting the data, I was kind of surprised by the number of 5 minute intervals where there were exactly 4 people. Learning from the 2nd dataset, I made sure that I collected data in 5 minute intervals. Doing this helped give me a better idea of the flow of people per time 5 minute interval. Even though this doesnt change the mean of the variable over the course of the 2 hours, it gives you a more granular look at the change per every 5 minutes. From this point you can perform calculations that tell you the rate of change from one 5 minute interval to the next.