HW3

Kristyan Dimitrov, Srividya Ganapathi, Shreyashi Ganguly, Greesham Simon, Joe Zhang

2/3/2020

Problem 5 : US Cell Phone Provider

Part (a)

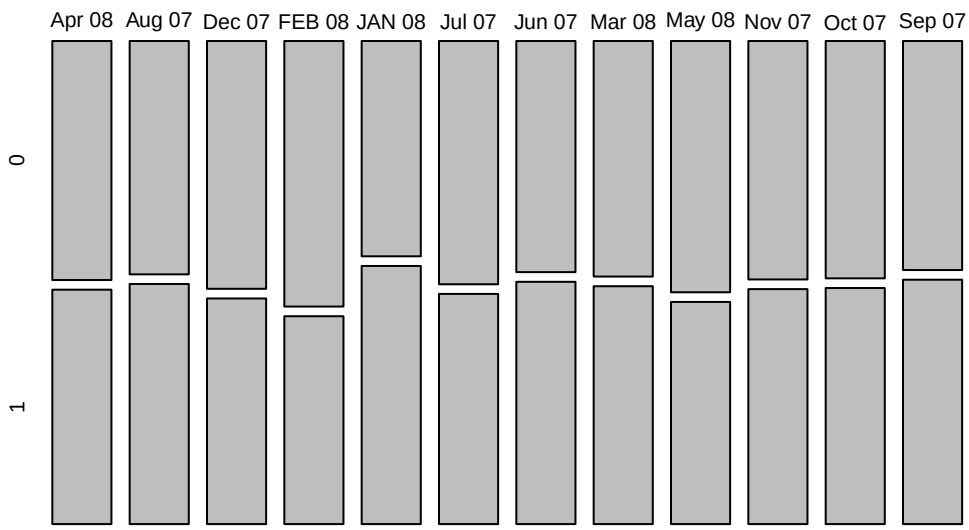
```
library(readr)
cell <- read_csv("/Users/shreyashiganguly/Documents/Northwestern_MSiA/Winter 2020/Data Mining/CLV & CRM/

crosstab <- table(cell$billmonth, cell$churn)
print("Cross tab of billmonth and churn")
```

```
## [1] "Cross tab of billmonth and churn"
```

```
crosstab
```

```
##
##            0   1
##   Apr 08 209 205
##   Aug 07 207 213
##   Dec 07 246 224
##   FEB 08 244 191
##   JAN 08 191 229
##   Jul 07 205 194
##   Jun 07 207 217
##   Mar 08 203 205
##   May 08 207 183
##   Nov 07 206 203
##   Oct 07 190 189
##   Sep 07 209 223
```

```
prop_table <- prop.table(crosstab,1)
plot(prop_table)
```



There is fairly constant proportion of attriters and non attriters in each bill month, except for may be Feb08 when the proportion of attriters falls to 44%. This seems like a stratified sampling design with equal proportion of churns and non churns picked from every month

Part (b)

```
total = sum(cell$t2, na.rm = TRUE)
churn = sum(cell$churn)
r = (1-churn/total)
sprintf("Number of people who cancelled = %d", churn)
```

```
## [1] "Number of people who cancelled = 2476"
```

```
sprintf("Number of opportunities of cancelling = %d", total)
```

```
## [1] "Number of opportunities of cancelling = 95624"
```

```
sprintf("Retention Rate = %f %%", (1-churn/total)*100)
```

```
## [1] "Retention Rate = 97.410692 %"
```

Part (c)

The sampling plan is such that we see almost 50% retention rate. However the actual retention rate as we saw is much higher at 97.4%. This would imply that the estimate of retention rate that we will get from this dataset will under estimate the actual retention rate.

Part (d)

```
cell$average_rev = rowMeans(cell[,7:21], na.rm = TRUE)
m = mean(cell$average_rev, na.rm = TRUE)
sprintf("Average monthly revenue = %f",m)
```

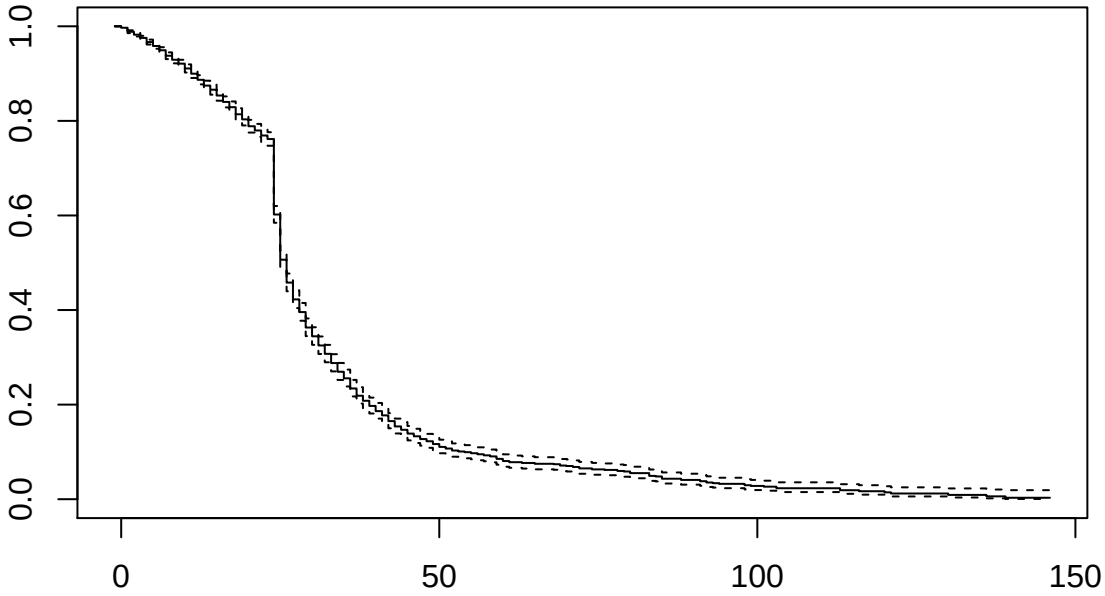
```
## [1] "Average monthly revenue = 56.422497"
```

```
d = 0.01
CLV = m*(1+d)/(1+d-r)
sprintf("Expected CLV = $%f",CLV)
```

```
## [1] "Expected CLV = $1587.679850"
```

Part (e)

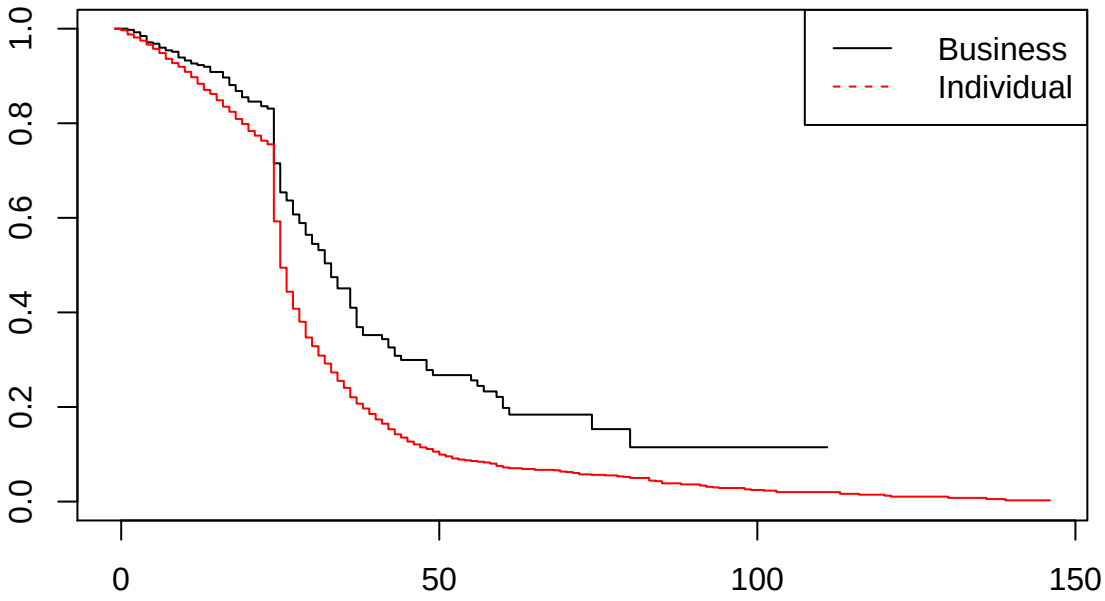
```
library(survival)
fit = survfit(Surv(t2, churn) ~ 1, data=cell)
plot(fit)
```



There is a sharp drop in survival rates around 23rd and 24th months. Seems like most customers must have a two year contract, hence the spike in churns around the two year period.

Part (f)

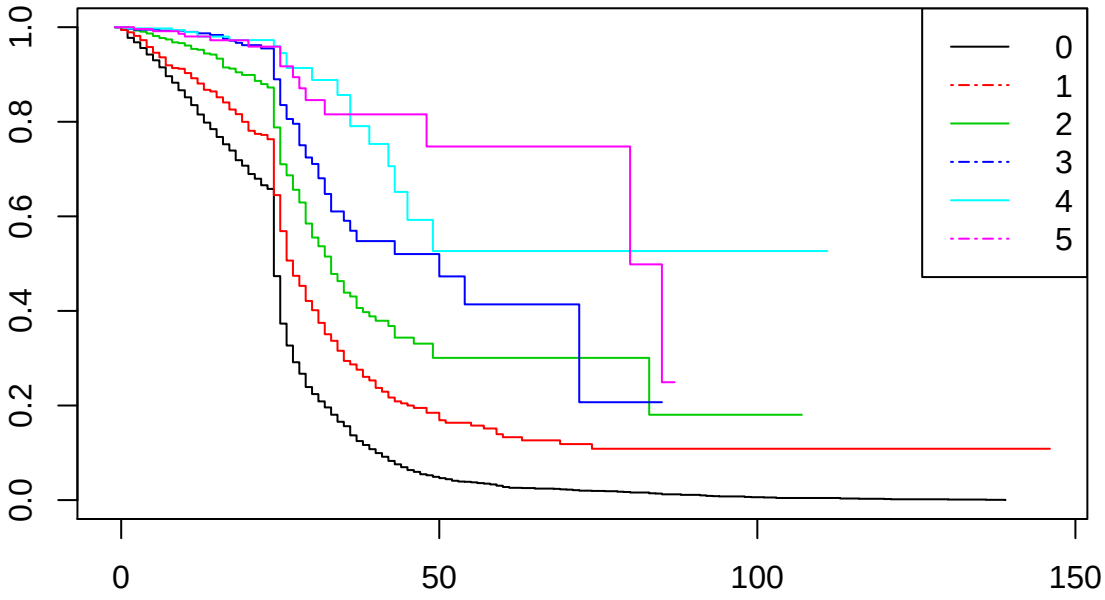
```
fit_str = survfit(Surv(t2, churn) ~ ACCOUNT_TYPE, data=cell)
plot(fit_str, col=1:2)
legend("topright",paste(" ",c("Business","Individual")), col=1:2, lty=c(1,2))
```



The 'Individual' is more likely to churn, specially post the two year period.

Part (g)

```
cell$line_cap = ifelse(cell$LINE_COUNT < 5, cell$LINE_COUNT, 5)
fit_str = survfit(Surv(t2, churn) ~ line_cap, data=cell)
plot(fit_str, col=1:6)
legend("topright",paste(" ",c(0,1,2,3,4,5)), col=1:6, lty=c(1,6))
```



Customers with more number of lines tend to be stickier, specially post the two year mark. The churn patterns of customers with 3, 4 and >=5 lines seem almost equal for the first 23 months.