HW3

Naira Maria Barseghyan

2023-11-30

Parametric Models

For this part of the assignment it was required to plot the survival curves of all distributions and make decision. From the plot we can see that best survival curve is the lognormal curve.

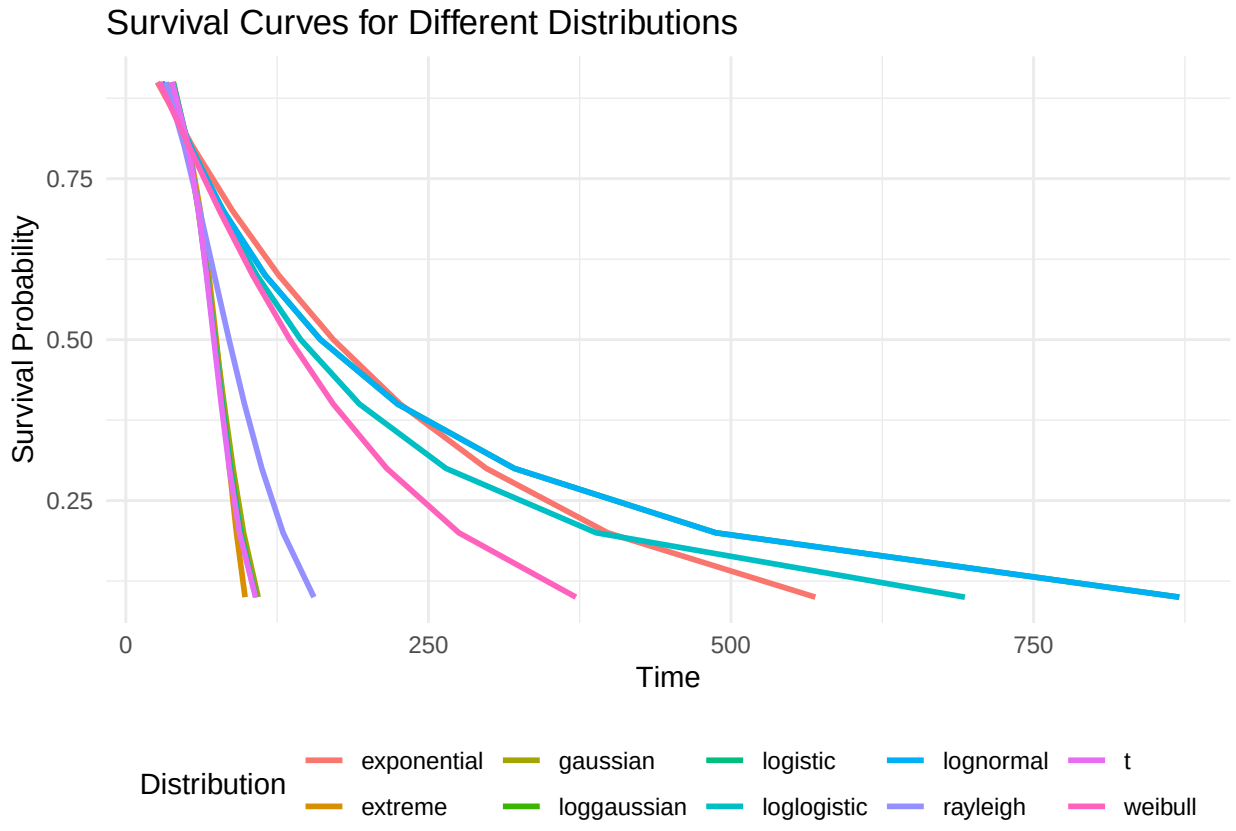


Figure 1

To make better choice of the model we can use other statistical measures such as AIC and BIC. Statistically best performing models are the models with lowest AIC and BIC values. As we can see from the results minimum AIC(2951.151) and BIC(3039.491) is generated by the model with lognormal distribution. Our final choice is the model with lognormal distribution.

```
## [1] 3039.491
```

```
## [1] 2951.151
```

```
##    Loglikelihood      AIC      BIC Distribution
## 1      -1747.194 3181.130 3269.470      extreme
## 2      -1572.565 3181.130 3269.470      extreme
## 3      -1734.223 3149.168 3237.507     logistic
## 4      -1556.584 3149.168 3237.507     logistic
## 5      -1714.485 3133.226 3221.565     gaussian
## 6      -1548.613 3133.226 3221.565     gaussian
## 7      -1606.431 2962.382 3050.721      weibull
## 8      -1463.191 2962.382 3050.721      weibull
## 9      -1606.980 2971.078 3054.510  exponential
## 10     -1468.539 2971.078 3054.510  exponential
## 11     -1739.723 3091.719 3175.151     rayleigh
## 12     -1528.859 3091.719 3175.151     rayleigh
## 13     -1602.518 2951.151 3039.491  loggaussian
## 14     -1457.576 2951.151 3039.491  loggaussian
## 15     -1602.518 2951.151 3039.491    lognormal
## 16     -1457.576 2951.151 3039.491    lognormal
## 17     -1605.208 2953.691 3042.030  loglogistic
## 18     -1458.845 2953.691 3042.030  loglogistic
## 19     -1748.062 3165.973 3254.312            t
## 20     -1564.986 3165.973 3254.312            t
```

Now lets find out which feutures are useful for the model. For the first model lets include all possible feutures and examine their significance. For significance level $\alpha = 0.1$ was choosen.

```
##
## Call:
## survreg(formula = surv_obj ~ age + marital + address + income +
##     ed + retire + gender + voice + internet + forward + custcat,
##     data = telco, dist = "lognormal")
##                                  Value Std. Error     z       p
## (Intercept)                   2.338870   0.281279  8.32 < 2e-16
## age                           0.032795   0.007247  4.53 6.0e-06
## maritalUnmarried             -0.459424   0.114720 -4.00 6.2e-05
## address                       0.042153   0.008882  4.75 2.1e-06
## income                        0.001387   0.000918  1.51   0.131
## edDid not complete high school  0.379168   0.200877  1.89   0.059
## edHigh school degree          0.315976   0.162495  1.94   0.052
## edPost-undergraduate degree  -0.019815   0.222366 -0.09   0.929
## edSome college                0.285140   0.164846  1.73   0.084
## retireYes                     0.031781   0.444440  0.07   0.943
## genderMale                    0.051108   0.114237  0.45   0.655
## voiceYes                     -0.424370   0.168551 -2.52   0.012
## internetYes                  -0.758597   0.142814 -5.31 1.1e-07
## forwardYes                   -0.196353   0.179535 -1.09   0.274
## custcatE-service              1.059925   0.170244  6.23 4.8e-10
## custcatPlus service           0.923373   0.214843  4.30 1.7e-05
## custcatTotal service          1.182016   0.249736  4.73 2.2e-06
## Log(scale)                    0.275904   0.045997  6.00 2.0e-09
##
## Scale= 1.32
##
## Log Normal distribution
## Loglik(model)= -1457.6   Loglik(intercept only)= -1602.5
##  Chisq= 289.88 on 16 degrees of freedom, p= 3.2e-52
## Number of Newton-Raphson Iterations: 5
## n= 1000
```

```
##                   (Intercept)                           age
##                          TRUE                          TRUE
##              maritalUnmarried                       address
##                          TRUE                          TRUE
##                        income edDid not complete high school
##                         FALSE                          TRUE
##          edHigh school degree    edPost-undergraduate degree
##                          TRUE                         FALSE
##                edSome college                     retireYes
##                          TRUE                         FALSE
##                    genderMale                      voiceYes
##                         FALSE                          TRUE
##                   internetYes                     forwardYes
##                          TRUE                         FALSE
##              custcatE-service           custcatPlus service
##                          TRUE                          TRUE
##          custcatTotal service                     Log(scale)
##                          TRUE                          TRUE
```

As we can see from the results P-values of some features are bigger than 0.1. Those features are forward, gender, income and retirement. To make the best model with good decisions without using non-useful features I eliminated mentioned features from the model. The regression summary of the final model is following.

```
##
## Call:
## survreg(formula = surv_obj ~ age + marital + address + ed + voice +
##     internet + custcat, data = telco, dist = "lognormal")
##                                  Value Std. Error     z       p
## (Intercept)                    2.30040    0.26658  8.63 < 2e-16
## age                            0.03672    0.00642  5.72 1.1e-08
## maritalUnmarried              -0.45111    0.11455 -3.94 8.2e-05
## address                        0.04228    0.00884  4.78 1.7e-06
## edDid not complete high school 0.32318    0.19886  1.63    0.10
## edHigh school degree           0.28346    0.16202  1.75    0.08
## edPost-undergraduate degree   -0.00704    0.22287 -0.03    0.97
## edSome college                 0.26066    0.16435  1.59    0.11
## voiceYes                      -0.43112    0.16788 -2.57    0.01
## internetYes                   -0.76976    0.14268 -5.40 6.8e-08
## custcatE-service               1.06378    0.17072  6.23 4.6e-10
## custcatPlus service            0.80252    0.16934  4.74 2.1e-06
## custcatTotal service           1.05892    0.21074  5.02 5.0e-07
## Log(scale)                     0.28004    0.04601  6.09 1.1e-09
##
## Scale= 1.32
##
## Log Normal distribution
## Loglik(model)= -1459.7   Loglik(intercept only)= -1602.5
##  Chisq= 285.71 on 12 degrees of freedom, p= 4.7e-54
## Number of Newton-Raphson Iterations: 5
## n= 1000
```

```
##                   (Intercept)                           age
##                     9.9781819                     1.0374031
##              maritalUnmarried                       address
##                     0.6369217                     1.0431842
## edDid not complete high school          edHigh school degree
##                     1.3815083                     1.3277135
##    edPost-undergraduate degree                edSome college
##                     0.9929849                     1.2977840
##                      voiceYes                    internetYes
##                     0.6497821                     0.4631241
##               custcatE-service           custcatPlus service
##                     2.8972934                     2.2311654
##           custcatTotal service
##                     2.8832641
```

For the interpretation of the coefficients we should look at the exponents of the coefficients which shwo the hazard ratio for each predictor. Coefficient of age is positive and HR is 1.0374031 which indicates that for each additional year of life of customer there is a 3% increase of hazard. HR of maritalUnmarried is 0.6369217 which indicates that Unmarried people have approximately 36 % lower hazard compared to Married. Education levle Hazard is compared to the College Degree, taregt group. HR of did not complete

high school is 1.3815083 which means that mentioned group have 38 % higher hazard compare to target group. HR of did high school is 1.3277135 which means that mentioned group have 32 % higher hazard compare to target group. HR of did post-Undergrad degree is 0.9929849 which means that mentioned group have approximately 1 % lower hazard compare to target group. HR of did some college is 1.2977840 which means that mentioned group have 29 % higher hazard compare to target group. HR of Voice yes is 0.6497821 which menas that mentioned group has approximately 35% lower hazard compared to Voice No group. HR of Internet yes is 0.4631241 which menas that mentioned group has approximately 55% lower hazard compared to internet No group. Customer category is comared to the Basic service, target group. HR of E-service is 2.8972934 which means that mentioned group have 189 % higher hazard compare to target group. HR of Plus Service is 2.2311654 which means that mentioned group have 123 % higher hazard compare to target group. HR of Total Service is 2.8832641 which means that mentioned group have 188 % higher hazard compare to target group.

CLV

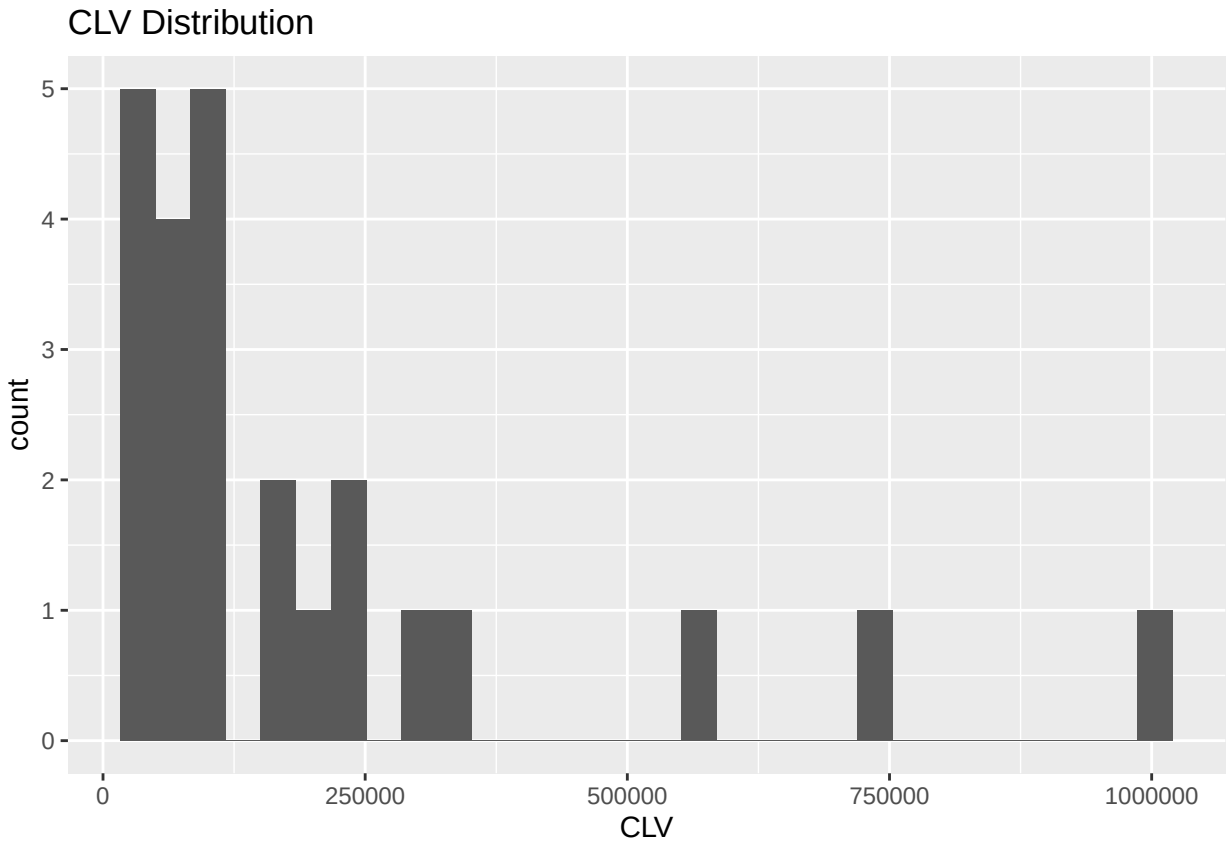
Based on the best model I made predictions and calculated CLV. For Calculating CLV I used formula

$$CLV = MM \sum_{i=1}^{t} \frac{p_i}{(1 + r/12)^{i-1}}$$

Assumption for monthly margin is 1300 AMD and asssumption for iscount rate(r) is 10 percent (retreived from the slides).

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6531   55138  117200  246071  266528 3843252
```

```
##   X.predictions.       CLV
## 1       73.44999  95484.99
## 2       83.83816 108989.61
## 3      572.62729 744415.47
## 4       47.08063  61204.82
## 5      135.39778 176017.12
## 6      161.75739 210284.60
```



Now lets look at CLVs and compare them by various features. For simplicity of interpretation I took only first 24 months. From the Figure two we can see difference of CLV among Males and Females. We can see that Males are less likely to make big purchases in early stages of their life as a customer compared to female, but later on males are making more consistent and highrer value purchases compared to the females. Female CLV have spikes while Male CLV doesn't have significant spikes. For both Males and females we can see that they make one single big purchase at start and consistent small purchases later.

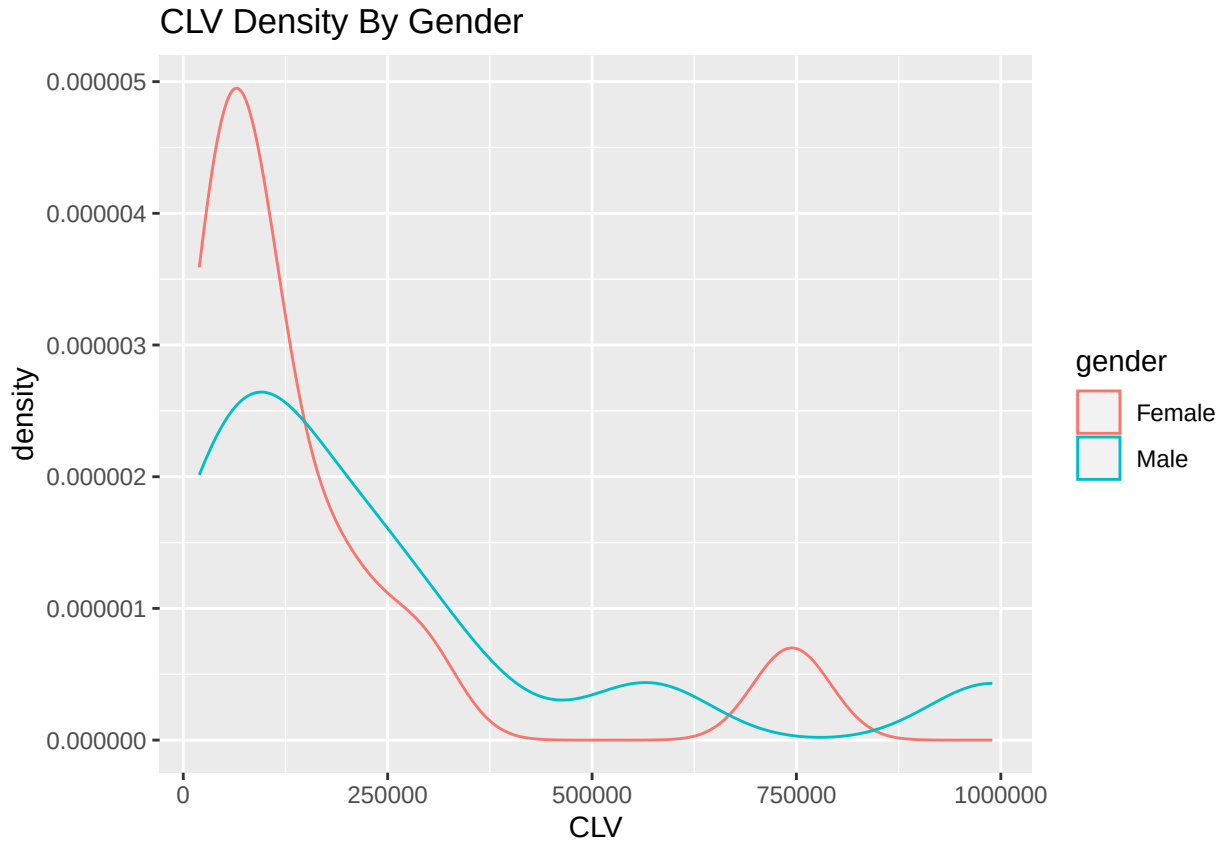


Figure: 2

On the figure three I am comparing CLV's of Marries and Unmarried people. We can see that Single people tend to make Big purchases at the start of their journey as a customer, but later on that disengage and do not make consistent purchases of high value. On the other side married people after initial big purchase, are making consistent little purchases later. The Spike on the end of the graph for unmarried people can be explained by them, not using serivices for long time and later on reengaging with again.

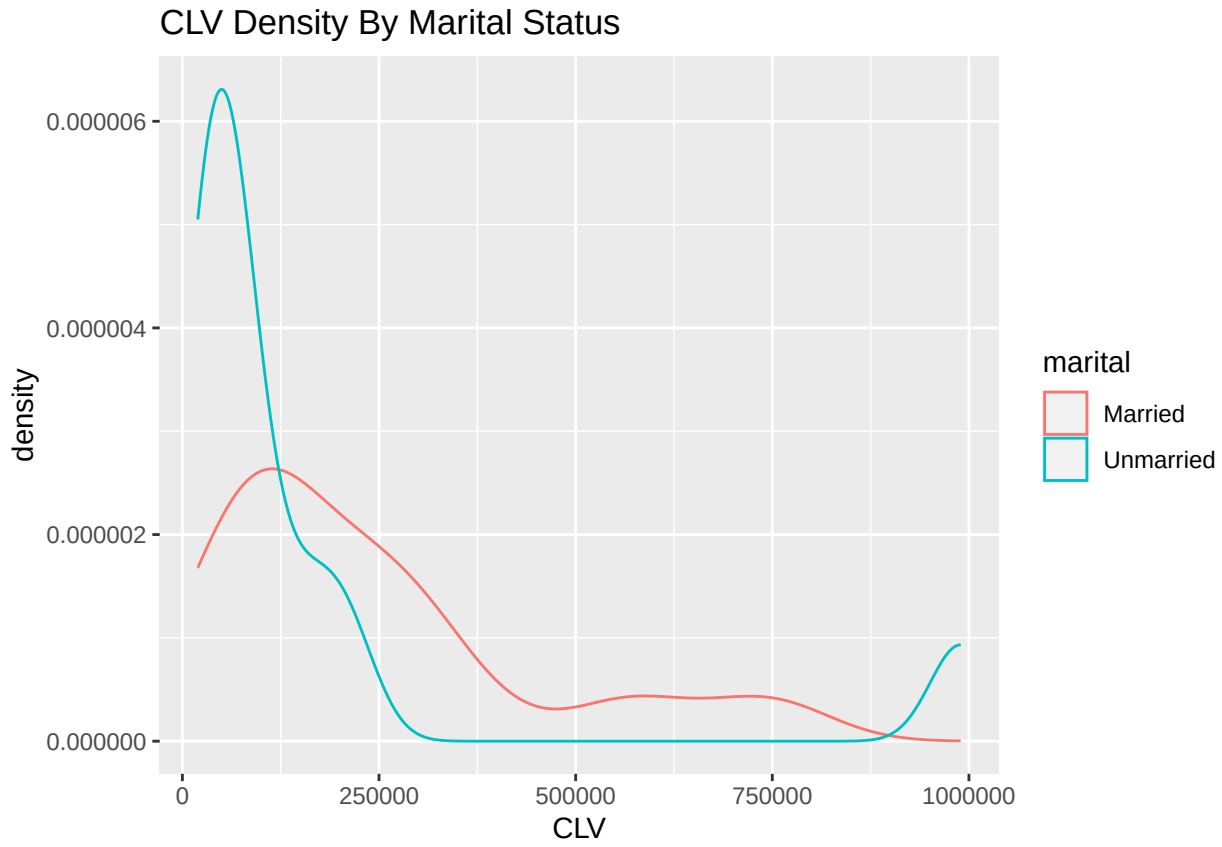


Figure: 3

For the third criterion of comparison I took education level of the customers. We can see the results on the fourth figure. As we can see most consistent customers are the ones that did not complete high school, since they are making persistent purchases over time. The customers with Post-Undergraduate digrees are the most likely to make high value purchases in the first stages and later on do not make any big purchases. This can be explained by high incomes of this type of customers, that choose the best products right on the start. The curve of people who did not complete high school indicates that they make non consistent purchase overtime, this can indicate that they are most likely to change plans and services, experimenting with various products. Customers with high school degrees have similar behaviour to post undergraduate degrees, with the difference tht they make lower price purchases at the start, but overall they are consistent.

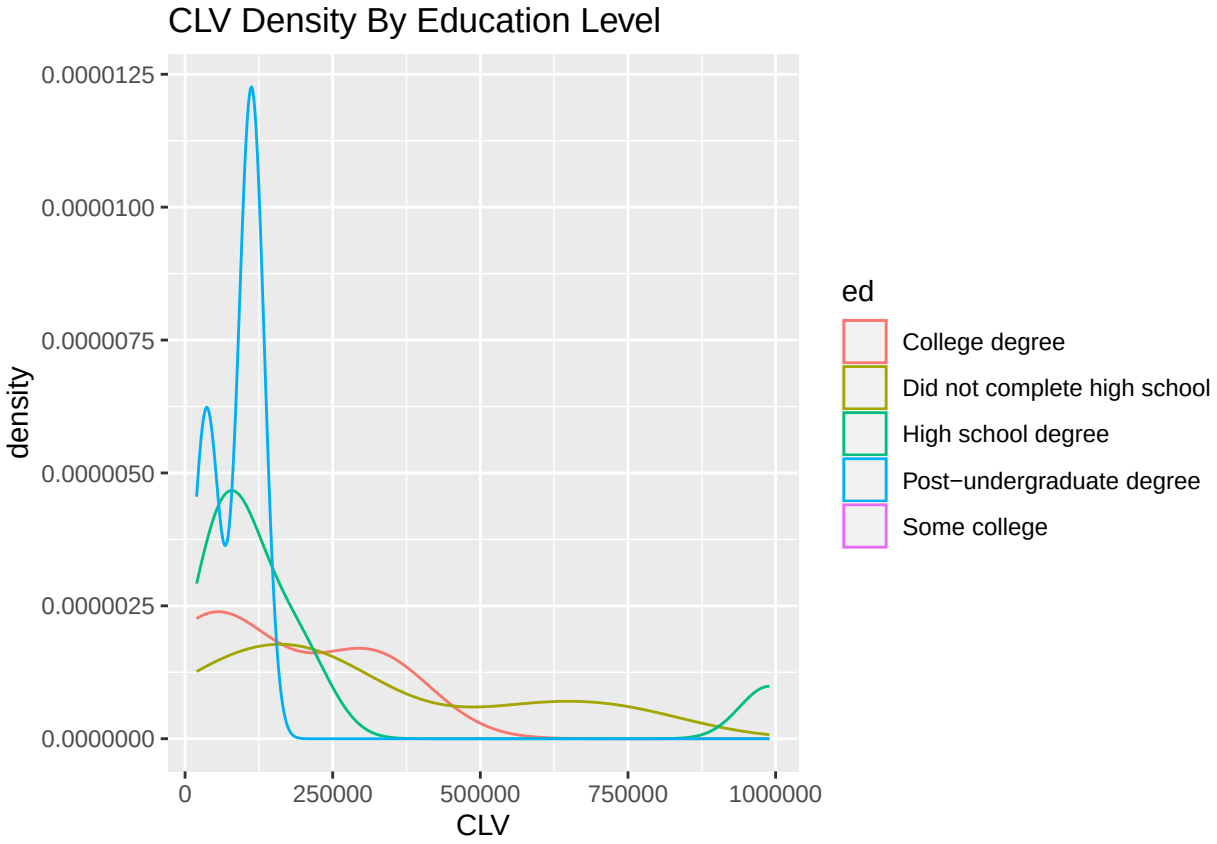


Figure: 4 Based on all of the findings I think that most valuable clients are Married people for the long run. They are most likely to have consistent purchases over time, and consistency of purchases is a good indicator for the business. Next Valuable group is male customers. They are also showing consistent purchase pattern. From the point of view of education, it seems like customer who didn't complete high school are more likely to have a lot of purchases. Customers with Post-Undergraduate degrees will also be valuable for the business, since they are making high value purchases, which will surely benefit the business. Overall I think the most valuable clients considering all of the facts combined(consistency, high value purchases) are Married Males.

Retention

For calculating retention rate of the customer on the first year first I calculated churn_rate and multiplied churn rate with overall numbers of customers(assuming that data includes all customers). To get the number of at risk customers I multiplied number of customers by churn rate. Then we get the retention budget by multiplying number of at risk customer with our average CLV. As a result I got that retention budget will be 3.937.142 drams for one year.

```
## [1] 3937142
```

Suggestions for retention. To decrease the retention rate it is important to segment at risk customers. After the segmentation we should find out whether those customers bring high value to the company. If the customers are not bringing high value to the company it is not reasonable to spend budget on retaining those customers. For the at risk customers who bring high value to the company, we need to construct specific offers in order to retain them. Specific orders can include specialized plans for narrow group of customers, for example Unlimited internet for customers with high internet usage. For customer retention we can offer specific offers and discounts to some groups. One more good strategy for customer consistent retention being in contact with the customers, throughout their life in the company. For example, making surveys od satisfication from time to time, or organizing events for customers to increase customer loyalty.