Marketing Analytics HW3

Nare Stepanyan

2024-05-01

```
##   ID region tenure age   marital address income                          ed
## 1  1 Zone 2     13  44   Married       9     64               College degree
## 2  2 Zone 3     11  33   Married       7    136    Post-undergraduate degree
## 3  3 Zone 3     68  52   Married      24    116 Did not complete high school
## 4  4 Zone 2     33  33 Unmarried      12     33           High school degree
## 5  5 Zone 2     23  30   Married       9     30 Did not complete high school
## 6  6 Zone 2     41  39 Unmarried      17     78           High school degree
##   retire gender voice internet forward       custcat churn
## 1     No   Male    No       No     Yes Basic service   Yes
## 2     No   Male   Yes       No     Yes Total service   Yes
## 3     No Female    No       No      No  Plus service    No
## 4     No Female    No       No      No Basic service   Yes
## 5     No   Male    No       No     Yes  Plus service    No
## 6     No Female    No       No      No  Plus service    No
```

Parametric Models

Now let's plot the survival curves of all distributions and make a decision. From the plot we can see that the best survival curve is the lognormal curve.

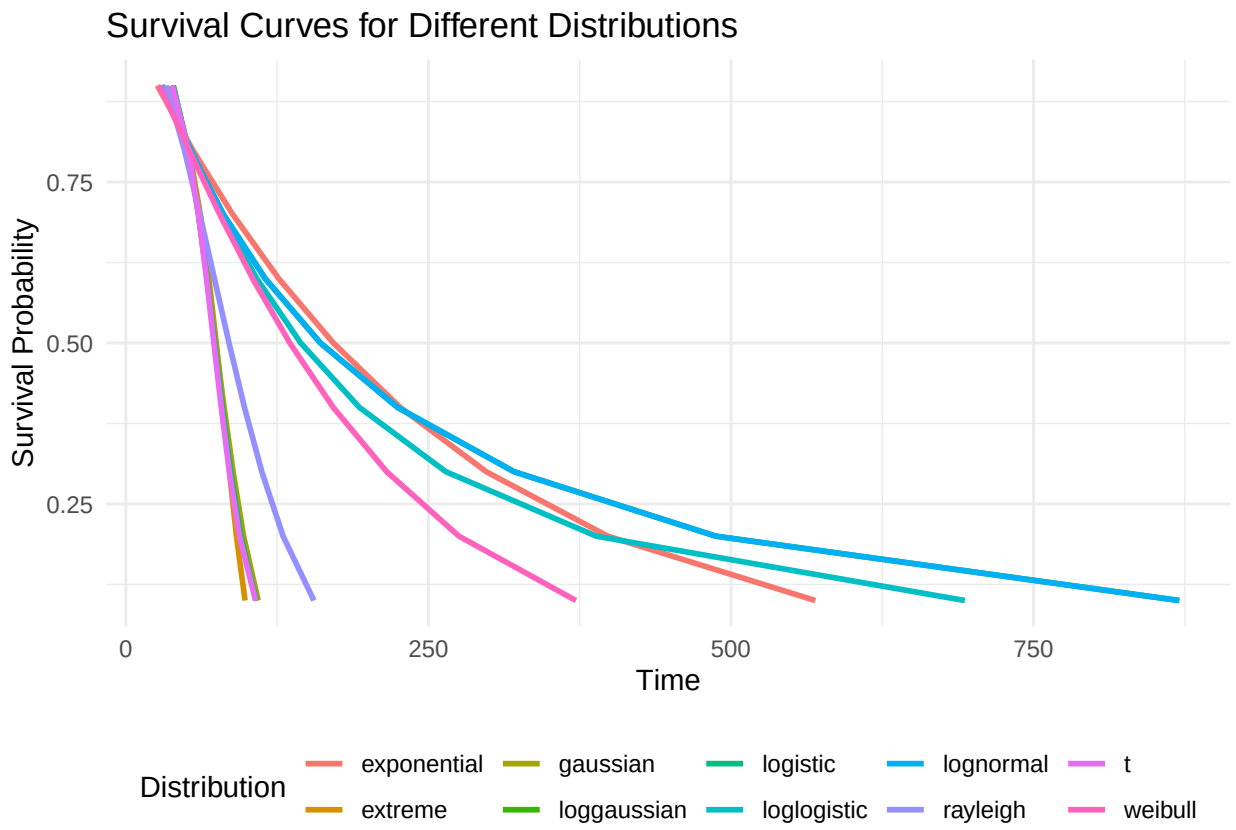


Figure 1

To select the best model, we can consider other statistical measures like AIC and BIC. The top-performing models have the lowest AIC and BIC values. From the results, we find that the model with a lognormal distribution has the minimum AIC (2951.151) and BIC (3039.491). Therefore, we choose the model with a lognormal distribution as our final option.

```
## [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
```

Let's see which features are useful for the model. For the first model, we'll incorporate all available features and assess their significance. We've selected a significance level of $\alpha = 0.1$ for this analysis.

```
##
## 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
```

From the results, we can see that the p-values of certain features exceed 0.1. These features include forward, gender, income, and retirement. To construct the best model and ensure sound decision-making without including non-informative features, I removed these mentioned features from the model. The regression summary of the final model is as follows.

```
##
## 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 show 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 marital Unmarried is 0.6369217 which indicates that single people have approximately 36 % lower hazard compared to married.

Education level Hazard is compared to the College Degree, the target group.
HR of did not complete high school is 1.3815083 which means that the mentioned group has 38 % higher hazard compare to the target group.
HR of did high school is 1.3277135 which means that the mentioned group has 32 % higher hazard compare to target group.
HR of did post-Undergrad degree is 0.9929849 which means that the mentioned group has approximately 1 % lower hazard compare to the target group.
HR of did some college is 1.2977840 which means that the mentioned group has 29 % higher hazard compared to the target group.
HR of Voice yes is 0.6497821 which means that the mentioned group has approximately 35% lower hazard compared to the Voice No group.
HR of Internet yes is 0.4631241 which means that the mentioned group has approximately 55% lower hazard compared to the internet No group.
Customer category is compared to the Basic service, the target group.
HR of E-service is 2.8972934 which means that the mentioned group has 189 % higher hazard compared to the target group.
HR of Plus Service is 2.2311654 which means that the mentioned group has 123 % higher hazard compared to the target group.
HR of Total Service is 2.8832641 which means that the mentioned group has 188 % higher hazard compared to the target group.

CLV

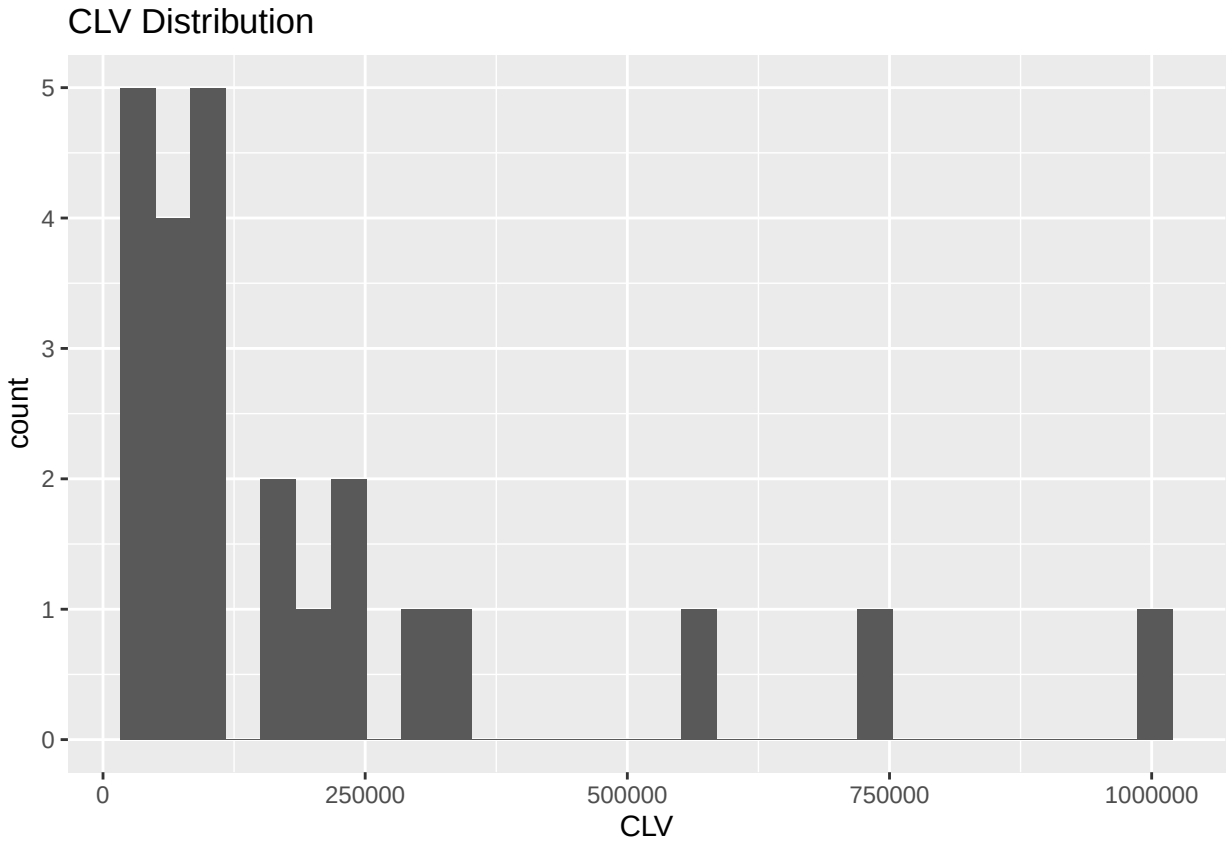
Based on the best model I made predictions and calculated CLV. For Calculating CLV I used the formula

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

Assumption for monthly margin is 1300 AMD and assumption for iscount rate(r) is 10 % (retrieved 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, let's examine CLV-s and compare them based on different features. To simplify our analysis, I've focused on the first 24 months.

From Figure 2, we observe variations in CLV-s between males and females. It's evident that males tend to make fewer substantial purchases during the initial stages of their customer journey compared to females. However, as time progresses, males demonstrate a pattern of consistent and higher-value purchases in contrast to females. While female CLV-s exhibit spikes, male CLV-s do not display significant fluctuations. Additionally, both males and females typically make one significant purchase initially followed by consistent smaller purchases thereafter.

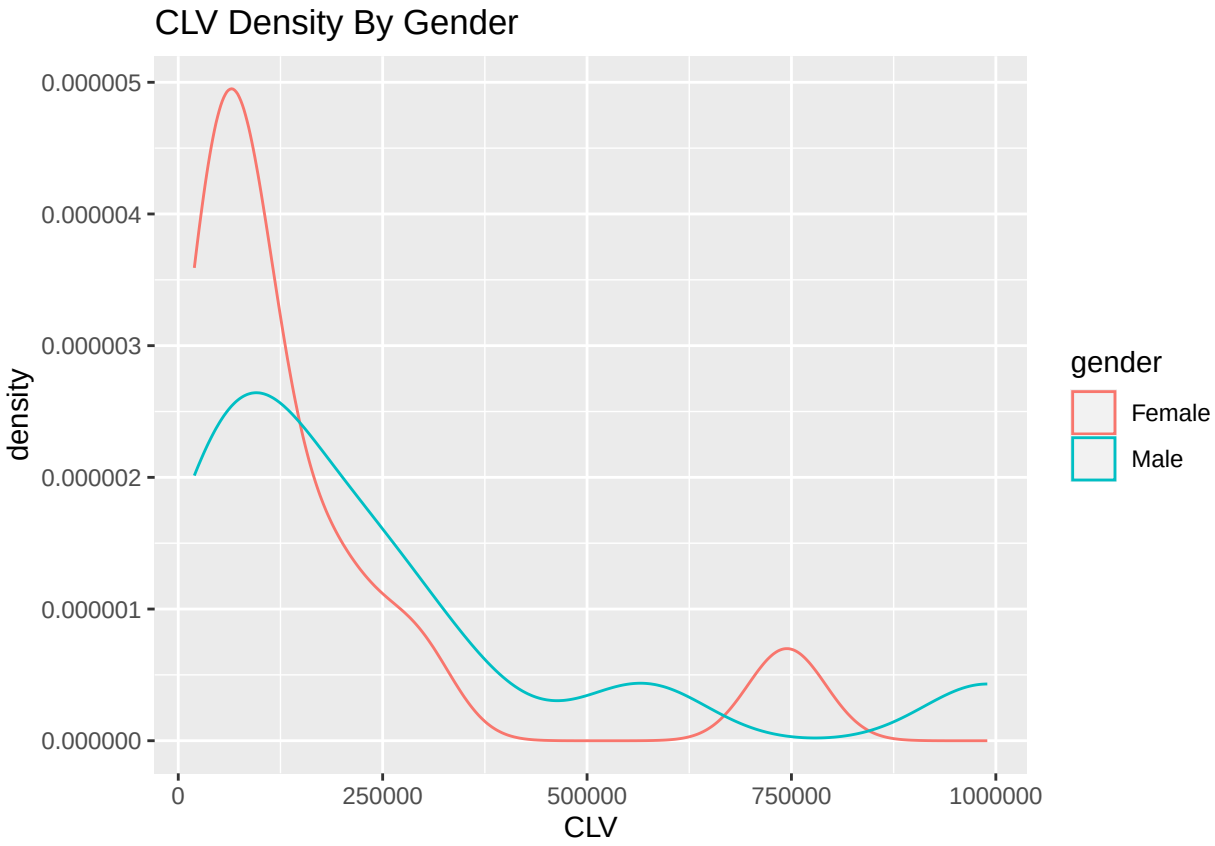


Figure: 2

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.

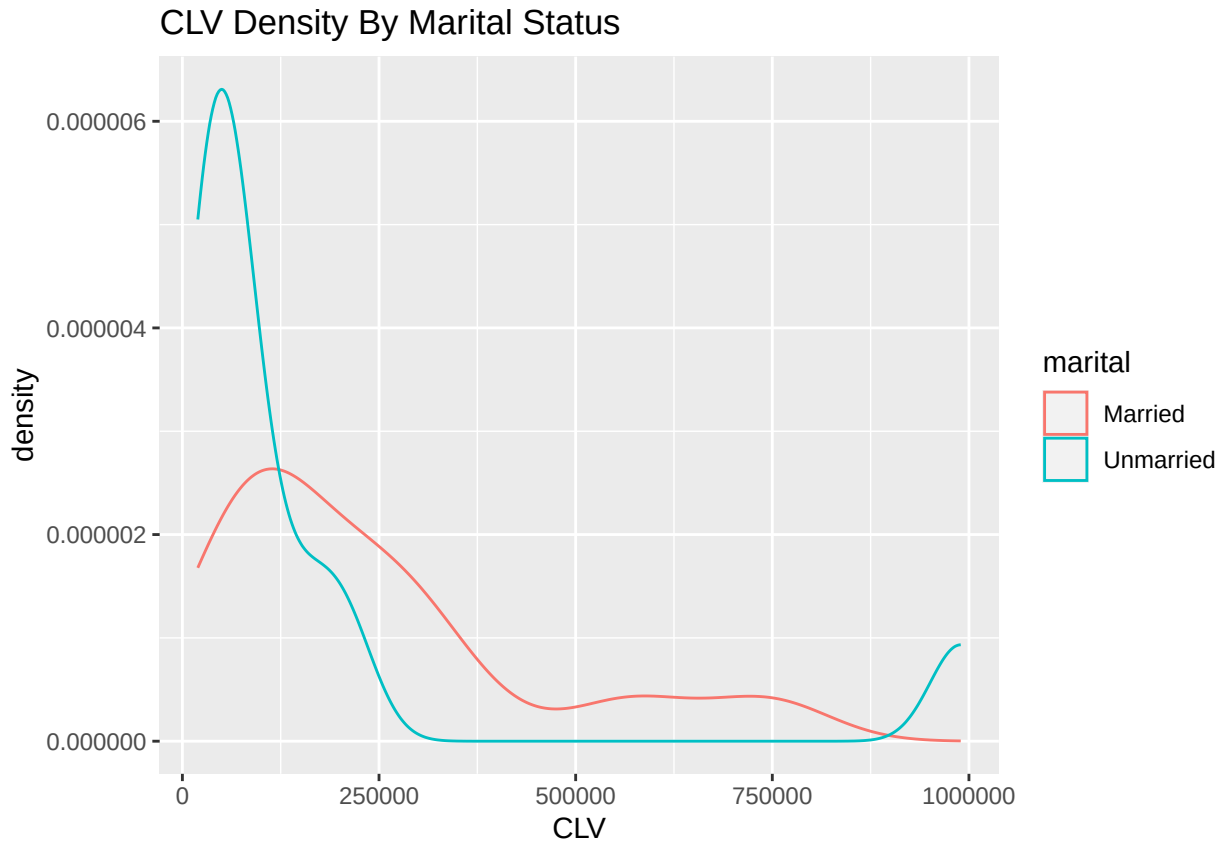


Figure: 3

For the third aspect of comparison, let's examine the education levels of customers, as depicted in the fourth figure. It's evident that customers who did not complete high school exhibit the most consistent purchasing behavior, consistently making purchases over time. On the other hand, customers with post-undergraduate degrees are more likely to make high-value purchases initially but do not continue this trend over time. This behavior may be attributed to their higher incomes, allowing them to opt for premium products from the outset. The curve for individuals who did not complete high school suggests inconsistent purchasing patterns, indicating a propensity to experiment with various products and services. Customers with high school degrees demonstrate a similar pattern to those with post-undergraduate degrees, with the exception of lower initial purchase prices, yet maintaining consistency overall.

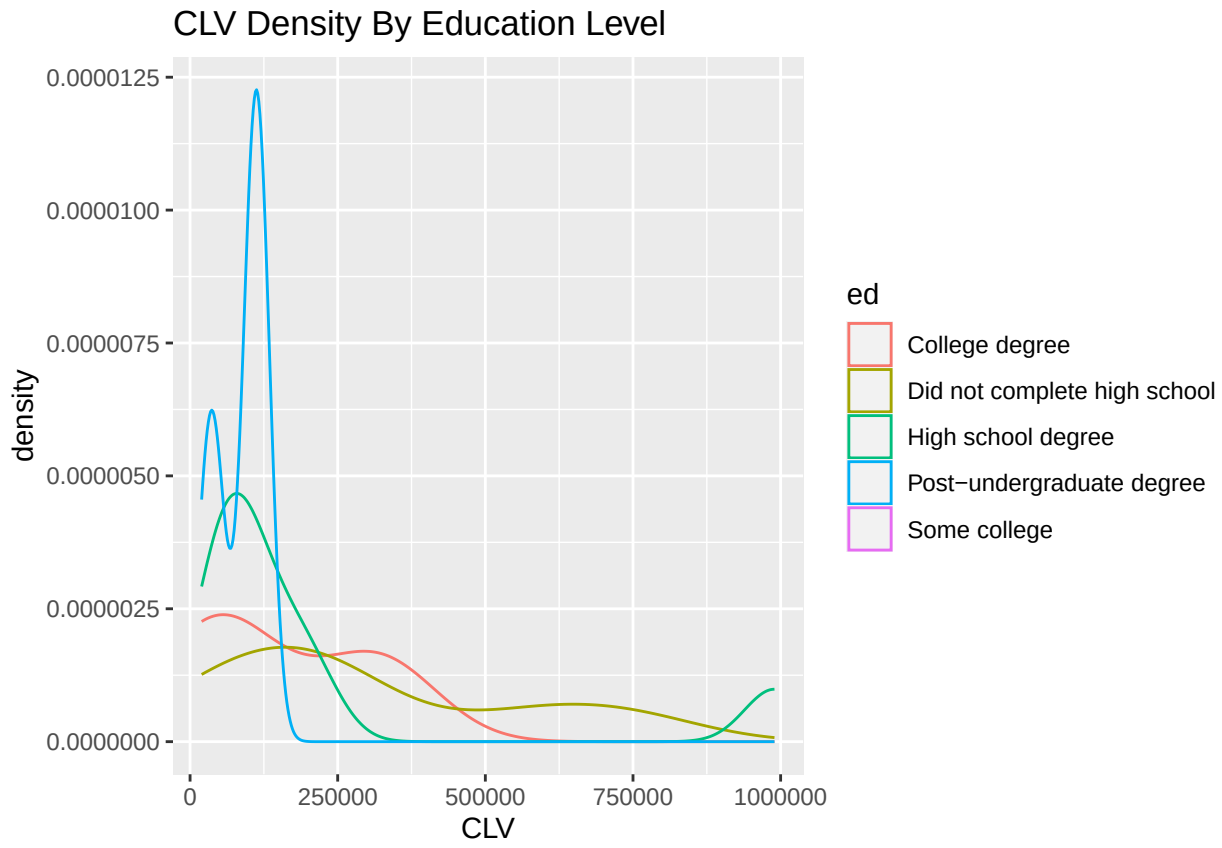


Figure: 4

Based on the findings, it appears that married individuals are the most valuable clients for long-term business success. They demonstrate consistent purchasing behavior over time, which is a positive indicator for business stability. Following closely are male customers, who also exhibit a consistent purchasing pattern. In terms of education, customers who did not complete high school tend to have a higher frequency of purchases. Additionally, customers with post-undergraduate degrees make high-value purchases, making them valuable for the business. Taking all factors into account, married males emerge as the most valuable clients due to their combined traits of consistency and high-value purchases.

Retention

To compute the customer retention rate for the first year, I initially determined the churn rate and then multiplied it by the total number of customers (assuming the dataset covers all customers). This provided the number of customers at risk of churn. Next, I calculated the retention budget by multiplying the number of at-risk customers by our average CLV. Consequently, I found that the retention budget for one year would be 3,937,142 drams.

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

Recommendations for improving customer retention: To reduce the retention rate, it's crucial to segment at-risk customers. After segmentation, it's essential to assess whether these customers contribute significantly to the company's value. If they don't, it may not be cost-effective to allocate budget for retaining them. However, for at-risk customers who bring substantial value to the company, customized retention strategies should be devised. These strategies may include offering specialized plans catering to specific customer needs, such as providing unlimited internet for customers with high internet usage. Additionally, personalized offers and discounts can be extended to certain customer groups to enhance retention. Another effective strategy for maintaining consistent customer retention involves maintaining regular communication with customers throughout their tenure with the company. This can be achieved by periodically conducting satisfaction surveys or organizing events aimed at nurturing customer loyalty.