# Final Final MA3

Anna Charchyan

2024-04-30

## 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.

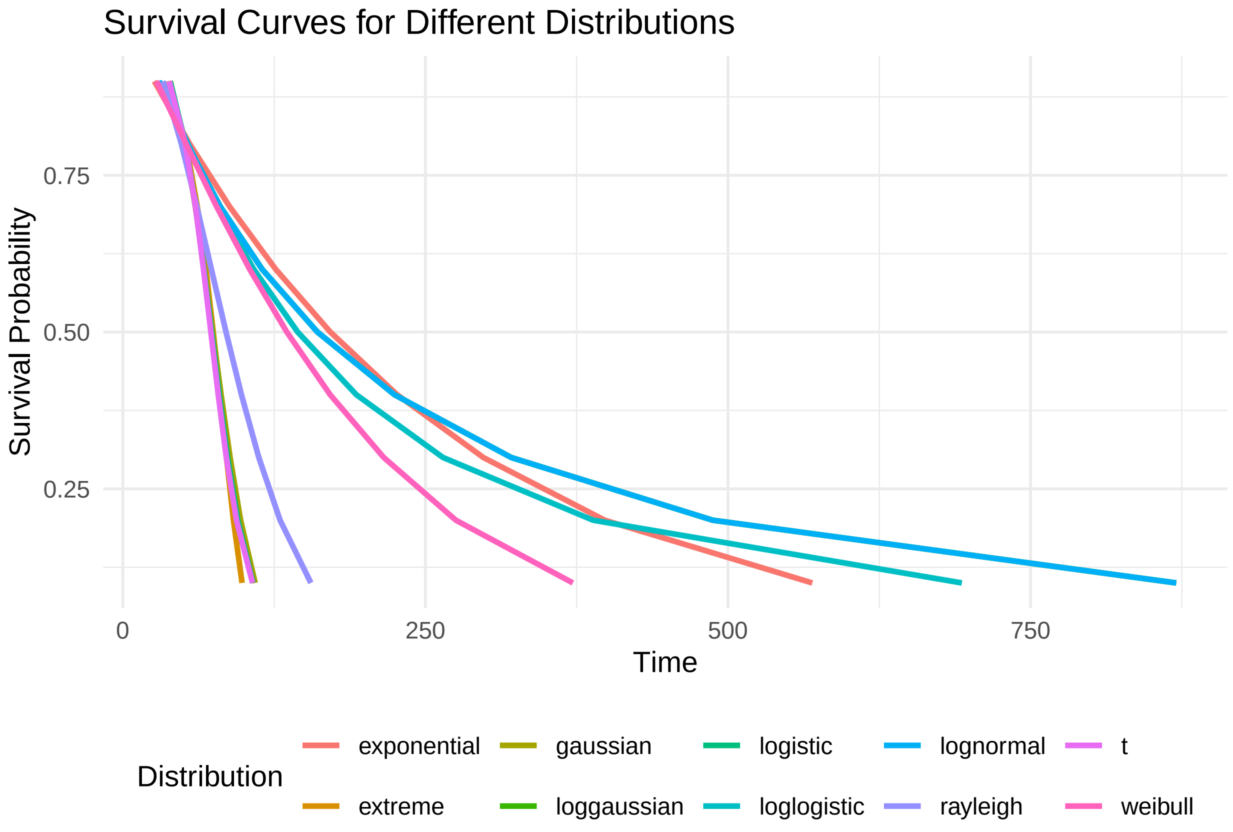


**Figure 1 : Survival Curves for Different Distributions**

Figure one illustrates plotting of survival curves of all distributions. Upon examining the plot, it is evident that the lognormal curve exhibits the most desirable survival characteristics compared to the other distributions.

```
## [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
```

To make a more accurate selection of the model, we utilize other statistical measures such as AIC and BIC. The statistically superior models are those with the lowest AIC and BIC values. From the results, we observe that the model with lognormal distribution and loggaussian distribution yields the minimum AIC (2951.151) and BIC (3039.491). Therefore, our final selection is the model with the lognormal distribution.

For the first model we will 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
```

From the results, it's evident that some features have p-values greater than 0.1. These features include forward, gender, income, and retirement. To optimize the model and ensure effective decision-making without incorporating non-useful features.

```
##
## 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
```

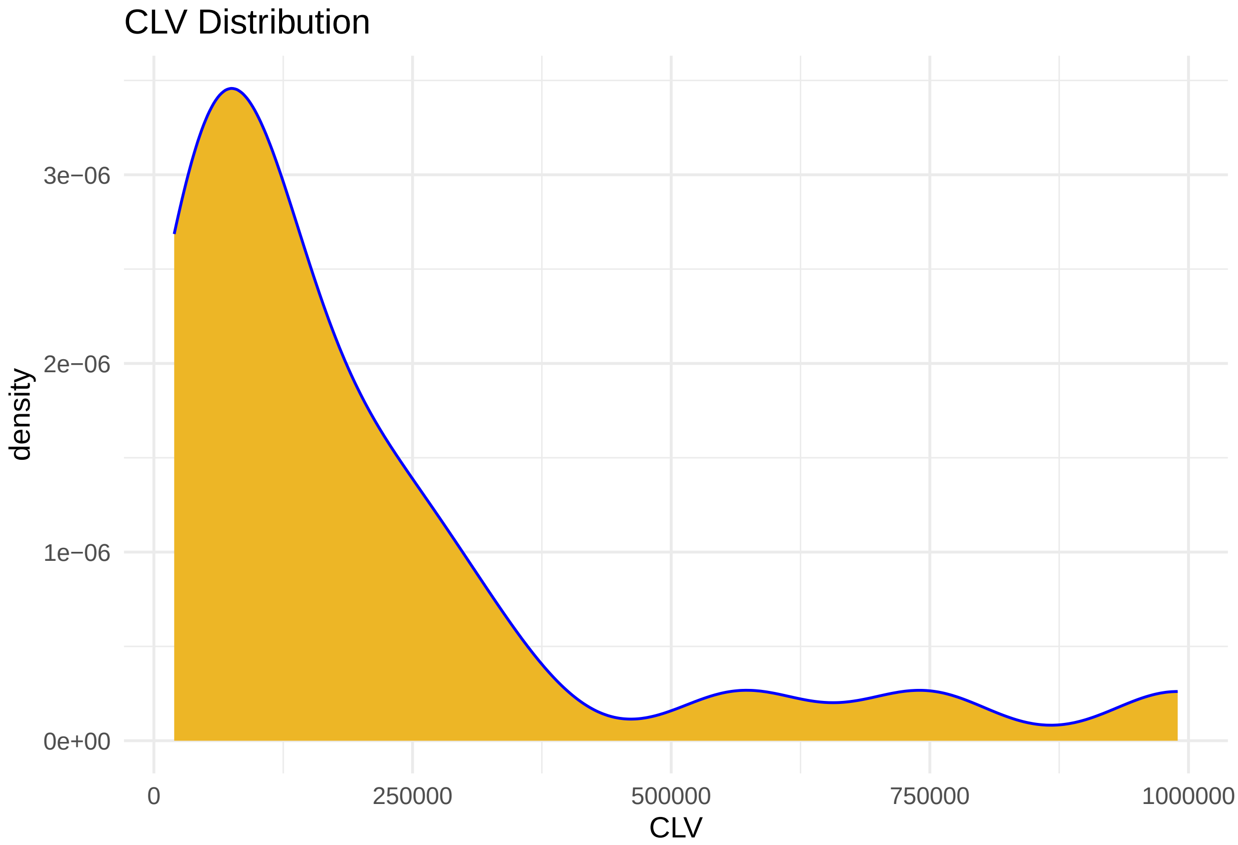
```
##                  (Intercept)                           age
##                    9.9781819                     1.0374031
##             maritalUnmarried                       address
##                    0.6369217                     1.0431842
## edDid not complete high school          edHigh 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
```

To interpret the coefficients, we examine their exponents, which represent the hazard ratios 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.

# CLV

```
##    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
```


CLV Distribution
density
3e−06
2e−06
1e−06
0e+00
0
250000
500000
750000
1000000
CLV

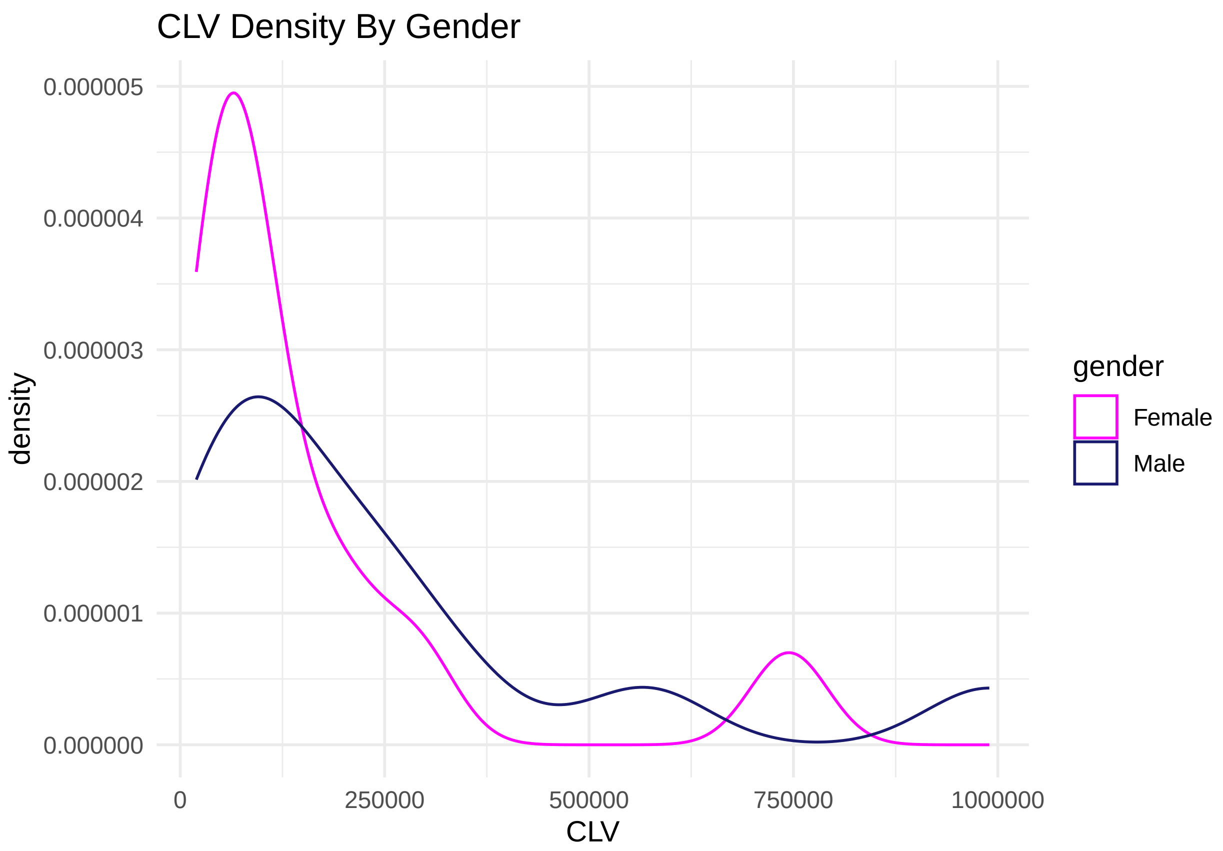


**Figure 2: CLV Density By Gender**

Figure two illustrates the disparity in customer lifetime values (CLVs) between males and females. Initially, males are less inclined to make substantial purchases compared to females. However, over time, males exhibit more consistent and higher-value purchasing behavior compared to females. While female CLVs show spikes, male CLVs do not exhibit significant spikes. Both genders tend to make a single large purchase at the beginning, followed by consistent smaller purchases later.

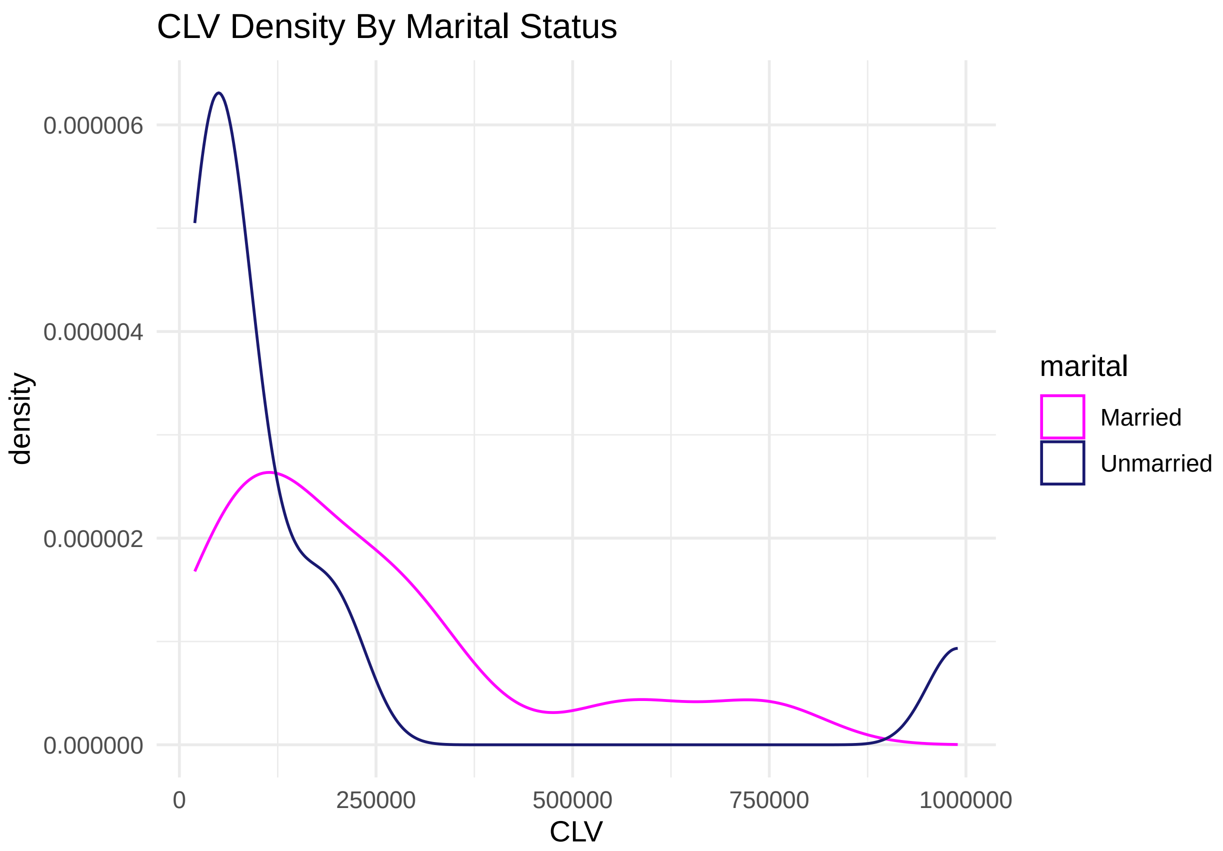


**Figure 3: CLV Density By Marital Statu**

Figure three compares the customer lifetime values (CLVs) of married and unmarried individuals. We observe that unmarried customers tend to make significant purchases early on, but their purchasing behavior becomes inconsistent over time, with fewer high-value purchases. In contrast, married customers make a large initial purchase followed by consistent smaller purchases. The spike in CLV for unmarried individuals at the end of the graph may indicate periods of disengagement followed by re-engagement with the services.

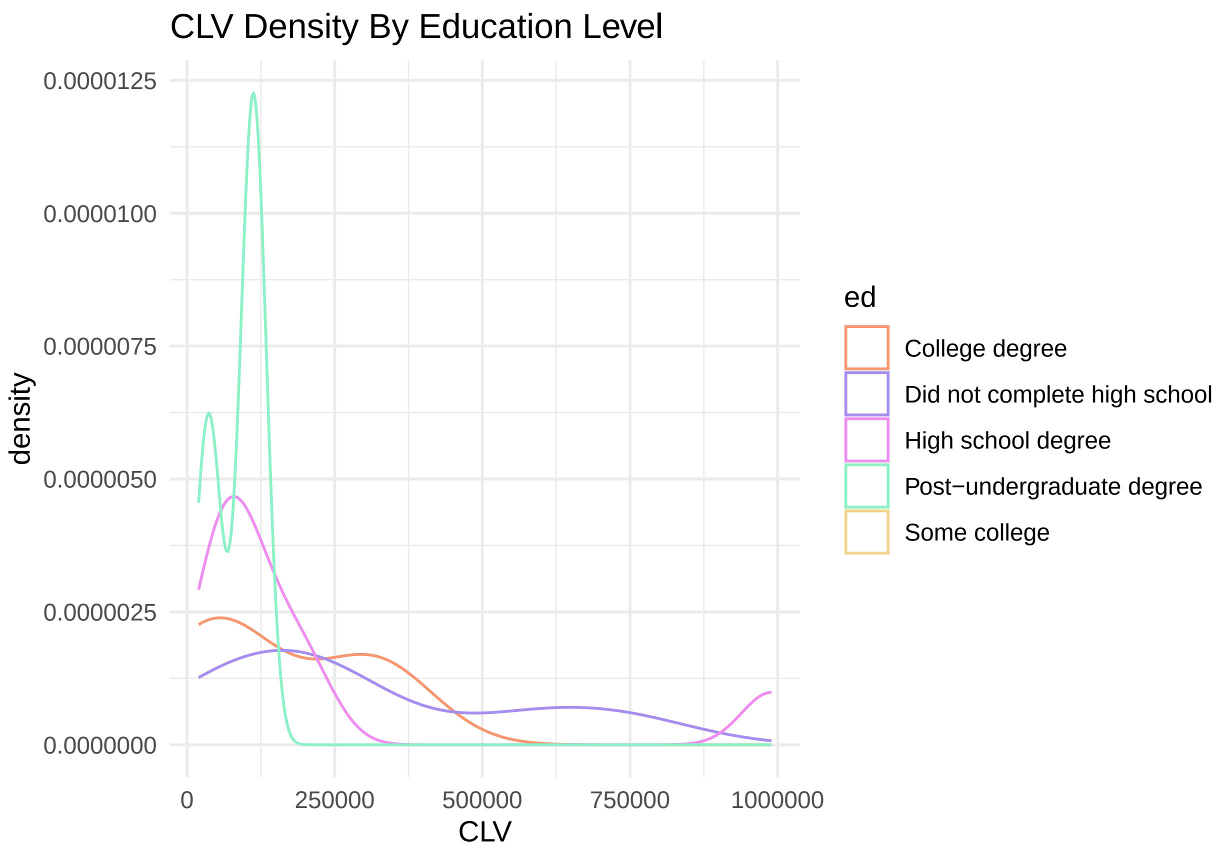


**Figure 4: CLV Density By Education Level**

In the third comparison, focusing on education level, we observed interesting patterns. Customers without high school diplomas show steady purchasing behavior over time, suggesting they stick with their plans. Those with post-undergraduate degrees tend to make significant purchases early on but decrease over time, likely due to their initial high spending. Those with high school diplomas behave similarly to post-undergraduate customers but start with smaller purchases. Overall, these insights help us understand how education level influences customer purchasing habits.

Based on our analysis, the most valuable clients appear to be married individuals, particularly males. They demonstrate consistent purchasing behavior over time, which is beneficial for the business. Additionally, customers who didn't complete high school show a tendency for frequent purchases. Those with post-undergraduate degrees also contribute significantly with high-value purchases. Taking all factors into account, married males emerge as the most valuable clients, combining both consistency and high-value purchases.

## Retention

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

## Suggestions for retention.

To reduce the retention rate effectively, it's crucial to identify at-risk customers through segmentation. Once segmented, it's important to assess whether these customers contribute significantly to the company's revenue. If not, it may not be cost-effective to allocate retention budget towards them. For at-risk customers who are valuable to the company, personalized retention strategies should be implemented. These strategies could involve tailored offers such as specialized plans based on individual needs, like offering unlimited internet to customers with high usage.

Furthermore, maintaining regular communication with customers throughout their tenure is essential for consistent retention. This could involve periodic satisfaction surveys or organizing customer events to enhance loyalty and foster long-term relationships.

In a similar vein, fostering a sense of community among customers can significantly contribute to retention efforts. Building a community around your brand can create a sense of belonging and loyalty among customers, which can in turn reduce churn rates. This can be achieved through various means such as online forums, social media groups, or exclusive events where customers can interact with each other and with the company.