Summary

Data Cleaning

Column name	Filter value	Total records	% of total data set
AgeFirstKill	99999	9	1.45
AgeFirstKill	<1900	35	5.66
Motive	NA	6	0.97

After cleaning and adding new variable career duration, we now have a total of 569 records and 10 columns.

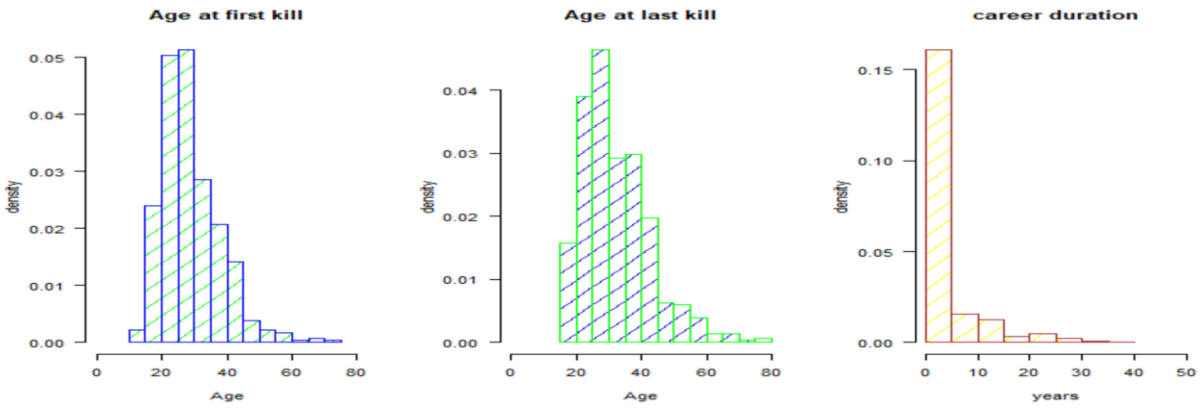
Data Exploration

Table 3: Numerical summary

parameters	variables		
	AgeFirstKill	AgeLastKill	CareerDuration
Standard Dev	8.91	10.47	6.28
Mean	29.41	32.64	3.22
length	569.00	569.00	569.00
Median	27.00	30.00	0.00
Minimum	13.00	15.00	0.00
Maximum	75.00	77.00	39.00
skewness	1.27	1.10	2.61

From numerical summary we can see that 50% of the observations for age at first kill is <= 27 years old Whereas age at last kill is <=30 years old. The maximum age that committed first kill is 75 and for last kill is 77. Another interesting trend we can see that, more than 50% of observations for career duration variable have zero median. This is due to the age at first and last kill is same. We can also observe that the youngest age that committed first kill is 13 and last kill is 15.

Histogram Plot: Graphical Summary

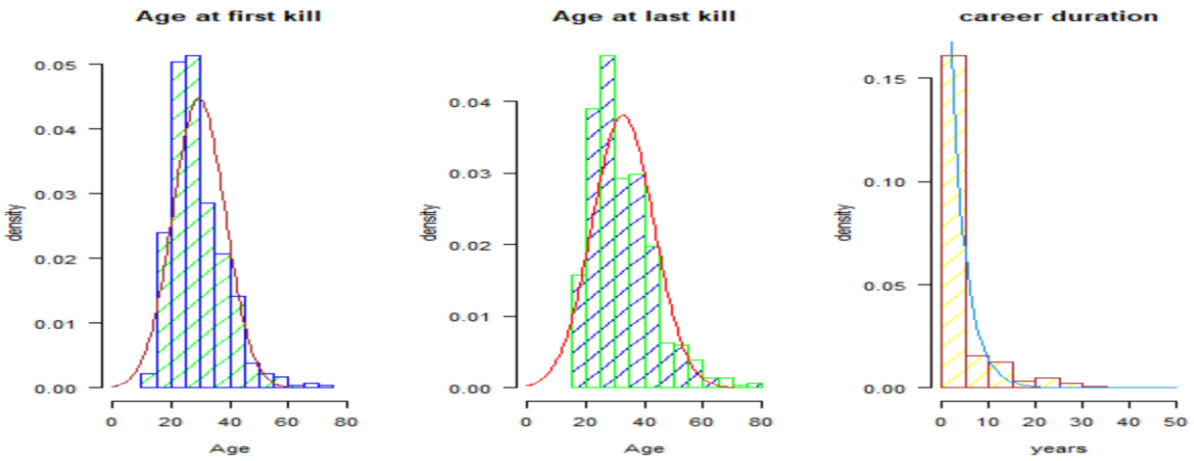


Relationship between variables.

Variables	AgeFirstKill	AgeLastKill	CareerDuration
AgeFirstKill	1.0	0.8	-0.1
AgeLastKill	0.8	1.0	0.5
CareerDuration	-0.1	0.5	1.0

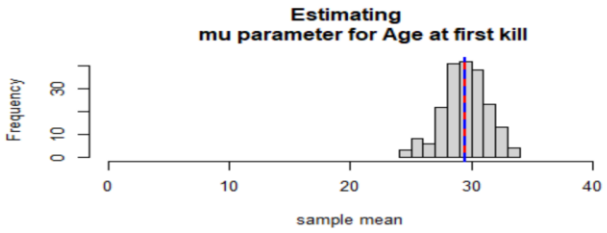
The above values is drawn by using cor() function. There is strong correlation between age first and last kill which means number of killers by age is approximately same at first and last kill. Weak correlation can be seen between career duration and Age at first kill. It is also observed that majority of the killings is driven by robbery or financial gain and majority of killers tend to fall in white racial group, followed by black race.

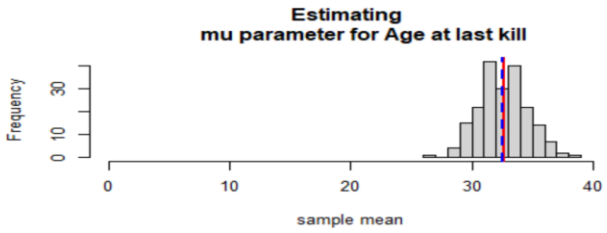
Modelling

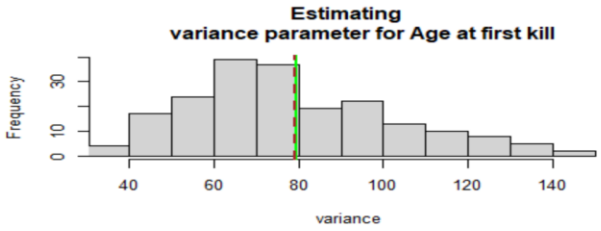


from above graph it is obvious that most of the histogram bins in age at first kill and last kill fits in the bell-shaped curve. Based on empirical rule (68-95-99), almost all observed data is within three standard deviation as per analysis we done. And there are not many outliners in data as well. So, for variables age at first and last kill we propose normal distribution. For career duration we have seen that in table 3, the skewness is high to the right side of the distribution with rate parameter lambda=1/mean. Which is 0.3 for this variable. Thus, we recommend exponential distribution for career duration.

Estimation







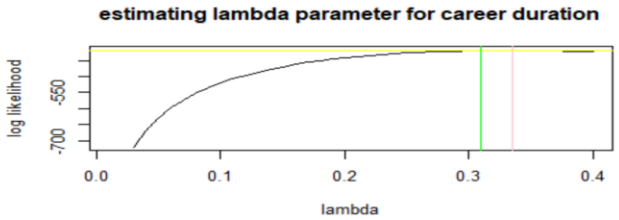


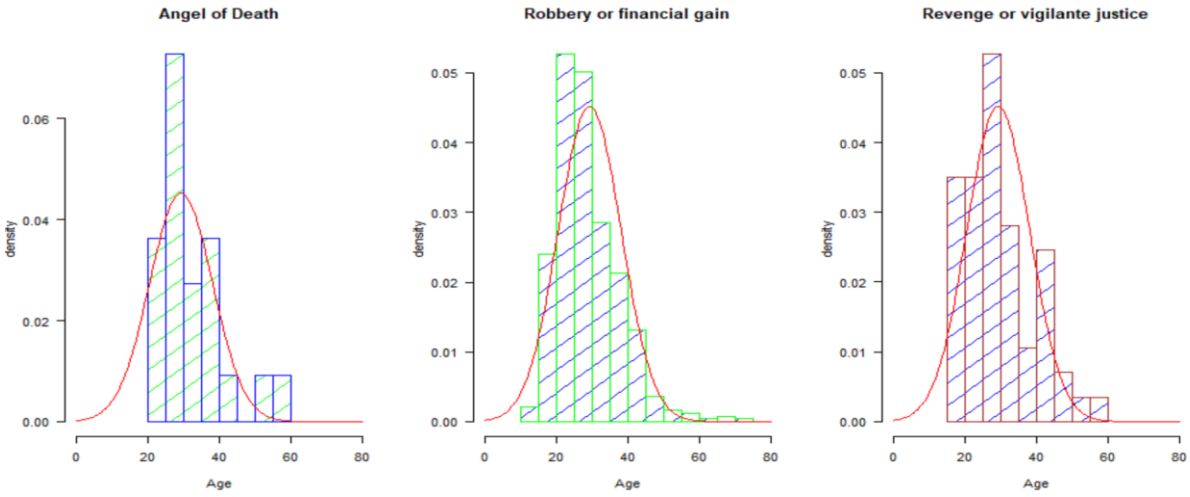
Figure 1: Mean of age at first kill mu =29.41, average estimator =29.55
Figure 2: Mean of age at last kill mu =32.64, average estimator =32.58.
Figure 3: variance of age at first kill sigma^2 = 79.26, average estimator = 79.03
Figure 4: true lambda of career duration = 0.30, MLE of lambda=0.33

Generated 200 samples with size n=25 by using mysample data set mean and standard deviation. We plotted 200 samples mean using histogram, we found that the average estimator (color blue) appears to be unbiased. Both true mean (color red) and average estimator is overlapping each other, suggesting it has very little difference. So MLE of mu for age at first kill and last kill is 29.55 and 32.58, respectively. The MLE for variance sigma^2 in figure 3 is 79.03. As you can see the sample variance appears to be unbiased.

From figure 4 we clearly see, the MLE of lambda (color pink) is directly below the peak of the curve where log likelihood is maximized at 1/mean. The horizontal line (color yellow) appears to confirm this. The green vertical line is true lambda.

Testing Hypothesis

motive	Summary						
	length	mean	median	Std dev	min	max	skewness
Angel of death	22	32.59	30	8.83	21	58	1.32
robbery or financial gain	490	29.20	27	8.82	13	75	1.34
Revenge or vigilante justice	57	30.09	28	9.64	15	59	0.79



Based on density curve all motives does not seem to be fit in standard normal distribution. However, they are approximately can fit in normal distribution.

We will choose Z-test for all three motives because sample size is greater than 20 and we know population variance and mean.

Using below Z test statistic, p-value and confidence interval formulas, table 6 is generated. We want to test null hypothesis H0: μ=27 against H1: μ≠27

Z=xbar- μ/sqrt(sigma^2/n)
p-value=P(|Z|>=c) = 2 Φ (<=-c)
CI= xbar – 1.96(sqrt(sigma^2/n), xbar +1.96(sqrt(sigma^2/n)

Table 6: Hypothesis testing summary

Test	Motives	Summary				
		size	Mean difference	Z-score	CI	p-value
Z-test	Angel of death	22	5.59	3.05	28.44,29.96)	0.002288414
Z-test	robbery or financial gain	490	2.20	5.67	(28.44,29.96)	0.000000014
Z-test	Revenge or vigilante justice	57	3.09	2.71	(27.86,32.32)	0.006728321

For all motives Since |Z| >=1.96, we can reject null hypothesis at the 5% significance level. p-value is also less than alpha (0.05), so we reject null hypothesis.
From 95% Confidence interval for all motives, we can see that the parameter mu is outside the interval so we can reject null hypothesis H0: μ=27.
For motive “Revenge or vigilante justice” mu is close to the interval but missing by margin.