Introduction-

The coursework entails analysing the dataset based on a sample of killers from the Radford/FGCU Serial Killer database. This report is a progress summary of the coursework so far.

The first step involves downloading the killersandmotives.Rdata from Minerva and loading into R. Each student has been assigned a code number as x in order to obtain their data subset from killersandmotives.Rdata. By using the number provided to me which is 14, I created my subset as mysample. The goal from here onwards is to analyse mysample.

Data Cleaning-

Data cleaning involved removing rows with missing values for the variables – AgeFirstKill (missing values are denoted as 99999 for this column), Motive(missing values are denoted as NA for this column). Data cleaning also involved removing any record of killers who first killed before the year 1900.

The following code were run and statistics were collected to meet the objectives of this stage –

- Total number of records in mysample before data cleaning steps were performed is 566 obtained by – *nrow(mysample)*

- Total number of records in mysample for missing values of AgeFirstKill is 9, obtained by – *nrow(mysample[mysample$AgeFirstKill == 99999,])*

- Total number of records in mysample for missing values of Motive is 6, obtained by – nrow(mysample[is.na(mysample$Motive),])

- Total number of records for killers who first killed before the 1900 is 8, obtained by – *nrow(mysample[(mysample$YearBorn + mysample$AgeFirstKill) < 1900,])*

A total of 23 records meeting the above criteria should be removed from the dataset mysample as a part of the data cleaning activity.

Sanity check performed using the below code to check the counts pertaining to the above criteria which should add up to 23.

nrow(mysample[mysample$AgeFirstKill == 99999 | is.na(mysample$Motive) | (mysample$YearBorn + mysample$AgeFirstKill) < 1900,])

Output – 23

The below code was run to create the dataset mysample_cleaned which contains the cleaned data.

```
mysample_cleaned <- mysample[which(mysample$AgeFirstKill != 99999),  ]
mysample_cleaned <- mysample_cleaned[!is.na(mysample_cleaned$Motive),  ]
mysample_cleaned <- mysample_cleaned[(mysample_cleaned$YearBorn + mysample_cleaned$AgeFirstKill) >=
1900, ]
```

Sanity check was performed again running the below code to make sure that the dataset mysample_cleaned does not contain any rows with missing data as described above.

nrow(mysample_cleaned[mysample_cleaned$AgeFirstKill == 99999 | is.na(mysample_cleaned$Motive) | (mysample_cleaned$YearBorn + mysample_cleaned$AgeFirstKill) < 1900,])

Output – 0

This way we ensured that the dataset mysample_cleaned was a cleaned dataset to be used for analysis and contained 543 records.

Data Exploration-

In this stage, numerical and graphical summaries where derived for the distribution of the three variables: AgeFirstKill, AgeLastKill and CareerDuration. The variable for career duration was not readily available in the provided the dataset and was added using the below code.

mysample_cleaned[mysample_cleaned$CareerDuration != (mysample_cleaned$AgeLastKill - mysample_cleaned$AgeFirstKill),]

Numerical Summaries -

Moment based summaries were derived for the mentioned variables using the below commands the stats for the same were captured –

1. **Mean-**

mean(mysample_cleaned$AgeFirstKill)
mean(mysample_cleaned$AgeLastKill)
mean(mysample_cleaned$CareerDuration)

AgeFirstKill	AgeLastKill	CareerDuration
29.70902	32.78821	32.78821

2. **Standard Deviation-**

sd(mysample_cleaned$AgeFirstKill)
sd(mysample_cleaned$AgeLastKill)
sd(mysample_cleaned$CareerDuration)

AgeFirstKill	AgeLastKill	CareerDuration
9.036828	10.76524	6.15087

Quantile based summaries were derived for the mentioned variables using the below commands the stats for the same were captured –

1. **Quantiles**

quantile(mysample_cleaned$AgeFirstKill, type = 1)
quantile(mysample_cleaned$AgeLastKill, type = 1)
quantile(mysample_cleaned$CareerDuration, type = 1)

	Min	Lower Quartile	Median	Upper Quartile	Max
AgeFirstKill	13	23	28	35	75
AgeLastKill	15	25	30	39	77
CareerDuration	0	0	0	3	39

2. **Inter Quartile Range (IQR)**

IQR(mysample_cleaned$AgeFirstKill, type = 1)
IQR(mysample_cleaned$AgeLastKill, type = 1)
IQR(mysample_cleaned$CareerDuration, type = 1)

AgeFirstKill	AgeLastKill	CareerDuration
12	14	13

Graphical Summaries –

Graphical summaries in form of boxplots and histograms have been plotted for the variables AgeFirstKill, AgeLastKill & CareerDuration as shown below.

1. **Boxplots**

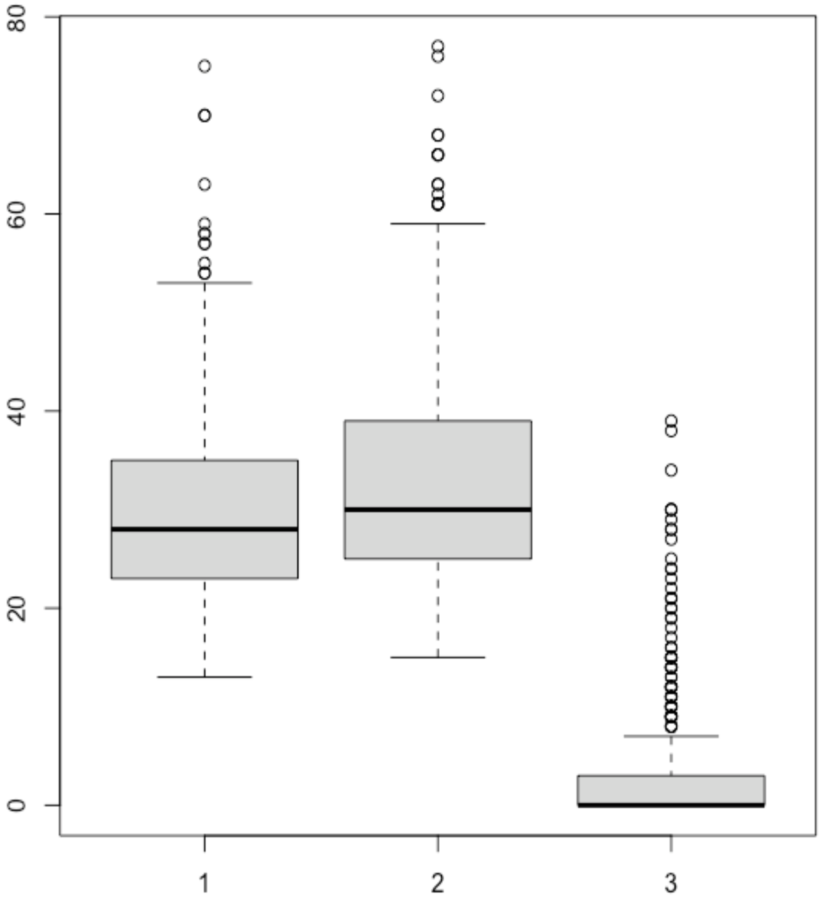

Fig 1.1

The boxplots above depict the distributions of the variables 1) AgeFirstKill 2) AgeLastKill 3) CareerDuration. It can be seen that the median for Age of First Kill and Age of Last Kill lie around the same range with the distribution being skewed towards the right whereas the median for career duration is at 0 with a long tail towards the right showing career duration for most killers in the dataset don't even last a year.

2. Histograms

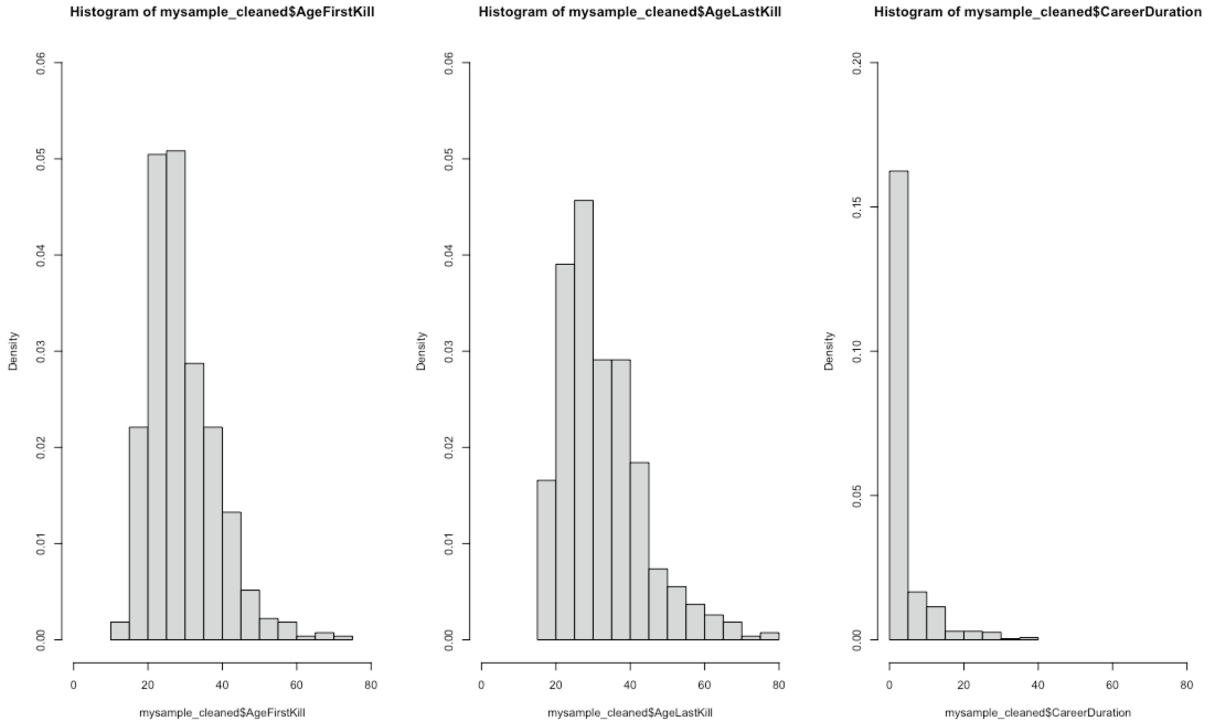

Fig 1.2

The histograms for Age of First Kill and Age of Last Kill clearly show that the data is right skewed. For career duration as well the data show a heavy level of skewness towards the right. The histograms for Age of First Kill and Age of Last Kill points towards a normal distribution for their distributions whereas the histogram for CareerDuration points towards an exponential distribution.

Strong +ve correlation of 0.8209063 observed between Age of First Kill & Age of Last Kill indicating killers starting at young age stop at a young age and killers starting at old age stop at an older age. Very weak -ve correlation -0.03244588 observed between Age of First Kill and Career duration. This can be observed in the below scatter plot as well.

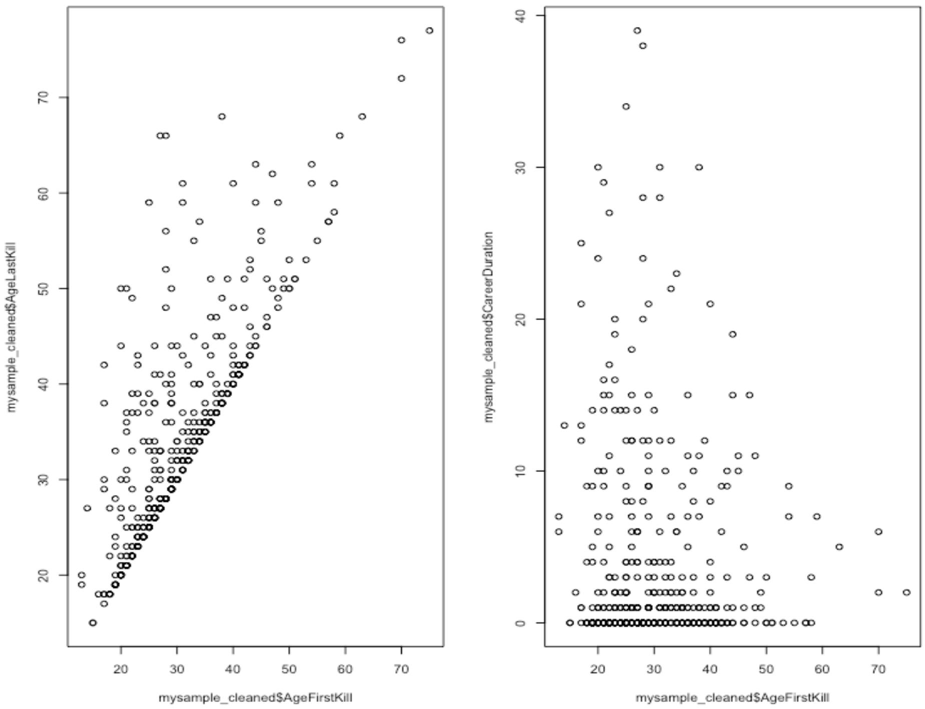

Fig 1.3