1940223\_CIA3.R

pranav

2020-03-06

data = read.csv("NTCA - TIGERNET.csv")  
#ABOUT THE DATASET  
#Records: Mortalities and seizures of tigers.  
#Country: India  
#Year: 2018  
  
#Libraries  
library(ggplot2)  
library(ggcorrplot)  
library(dplyr)

##   
## Attaching package: 'dplyr'

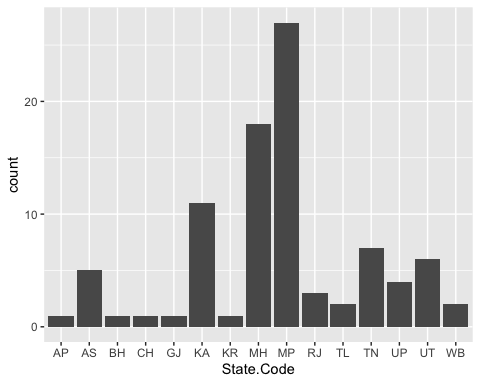
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

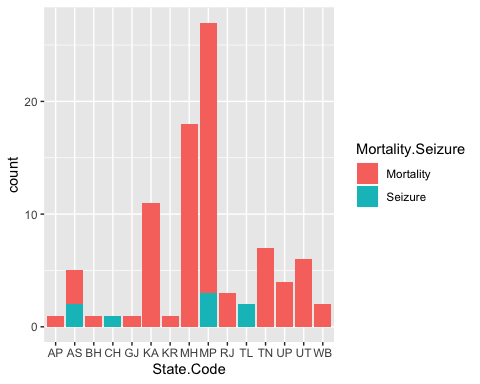
attach(data)  
  
head(select(data, 4:8), 6)

## Mortality.Seizure State.Code State Sex Age  
## 1 Mortality MH Maharashtra <NA> NA  
## 2 Mortality MH Maharashtra <NA> 2.3  
## 3 Mortality MP Mahya Pradesh <NA> 7.8  
## 4 Mortality MH Maharashtra <NA> 4.0  
## 5 Mortality KR Kerala Male 10.4  
## 6 Mortality MP Madhya Pradesh Male 2.0

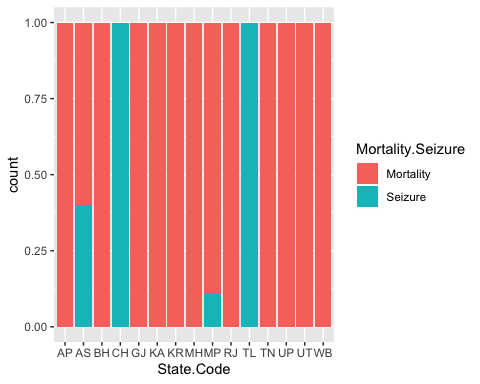
#1.1  
#Bar chart of observations according to state  
bp1 = ggplot(data, aes(x = State.Code))   
bp1 + geom\_bar()



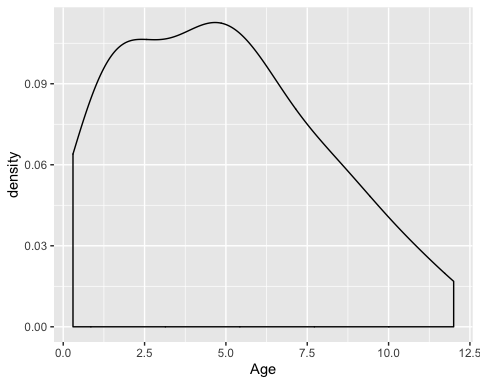
#INTERPRETATIONS  
#a. Maximum observations are from Madhya Pradesh, followed by Maharshtra and Karnataka  
#b. The range of frequencies is large   
# i.e. difference between least no. of observations and most no. of observations is large  
  
#1.2  
#Segmented bar chart to see mortalities and seizure per state  
bp2 = ggplot(data, aes(x = State.Code, fill = Mortality.Seizure))  
#1. In absolute values  
bp2 + geom\_bar()



#INTERPRETATION  
#a. Seizures recorded are very few compared to mortalities  
#b. Seizures are recorded only in 4 states  
#2. In proportion to the total observations from the state  
bp2 + geom\_bar(position = "fill")

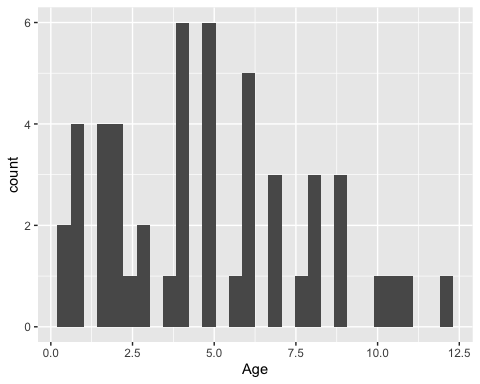


#INTERPRETATION  
#a. Chandigarh and Telangana recorded seizures in 100% of the observations  
#b. Most states recorded mortalities in 100% of the observations  
  
#2.1  
#Density plot and histogram according to tiger ages  
dp1 = ggplot(data, aes(x = Age))  
dp1 + geom\_density(na.rm = TRUE)

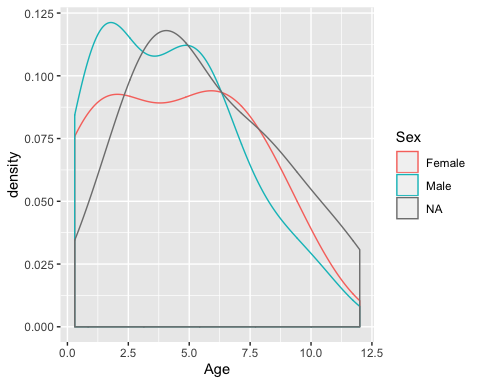


dp1 + geom\_histogram(na.rm = TRUE)

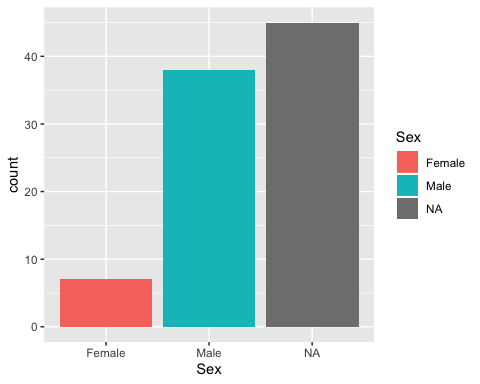
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#INTERPRETATION  
#a. Ages around 5 years is most highly recorded  
#b. The observations get almost linearly fewer as the age rises beyond 4-5 years  
# i.e. after around 4-5 years, the number of observed tigers in inversely propotional to age  
#c. Observations at the lower extreme of the ages are more numerous than at the higher extreme  
# i.e the density curve is positively skewed  
# This may indicate some combination of the following factors  
# -Infant mortality has risen  
# -Birth rate has fallen  
# (However, due to the significant number of unknown ages, the above conclusions are far from definitive)  
  
#2.2  
#Density plot of age considering the two sexes  
dp2 = ggplot(data, aes(x = Age, colour = Sex))  
dp2 + geom\_density(na.rm = TRUE)

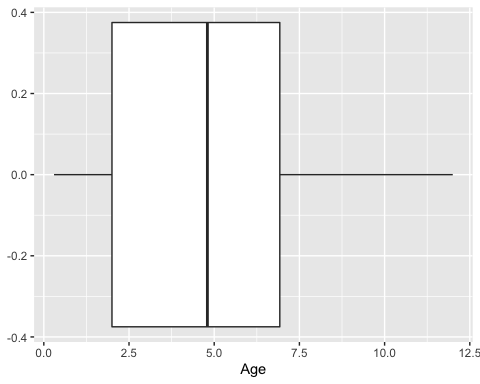


#Supporting graph  
bp3 = ggplot(data, aes(x = Sex, fill = Sex))  
bp3 + geom\_bar()



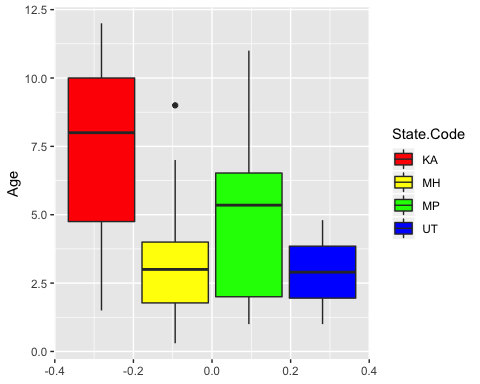
#INTERPRETATION  
#The supporting graph shows that the different density plots do not represent absolute values, only proportions  
#a. There are many more males with lower ages than with higher ages  
# i.e. between 0 and 7  
#b. The maximum number of males are with ages around 1.5  
#c. The female records follow a similar pattern to male records  
#d. The ages are more flatly spread than they are for males  
# Hence, a larger proportion of the female population has higher ages  
# i.e. between 7 and 12.5  
#e. The maximum number of females are with ages around 7  
  
#3.1  
#Boxplot for ages  
bxp1 = ggplot(data, aes(y = Age))  
bxp1 + geom\_boxplot() + coord\_flip()

## Warning: Removed 40 rows containing non-finite values (stat\_boxplot).



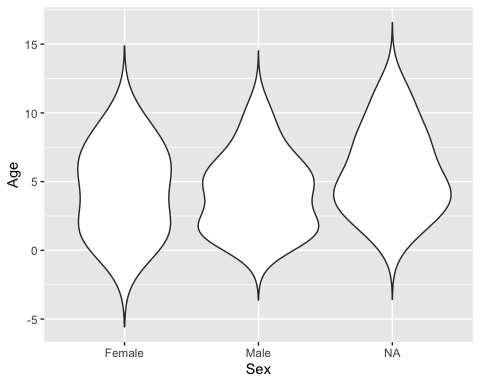
#INTERPRETATION  
#We can see that   
#a. The mean is between 4 to 5  
#b. There are no outliers   
# i.e. every value is within an interquartile range from the previous quartile  
#c. The 1st quartile is between 2 and 2.5  
#d. The 3rd quartile is between 7 and 7.5  
#e. The minimum is about 0.25  
#f. The maximum is about 12  
  
#3.2  
#Boxplot for ages with regard to 4 states  
tmp = data  
s1 = filter(data, State.Code == "MH")  
s2 = filter(data, State.Code == "KA")  
s3 = filter(data, State.Code == "MP")  
s4 = filter(data, State.Code == "UT")  
tmp = union(union(union(s1, s2), s3), s4)  
bxp2 = ggplot(tmp, aes(y = Age, fill = State.Code))  
bxp2 + geom\_boxplot() + scale\_fill\_manual(values = c("red", "yellow", "green", "blue"))

## Warning: Removed 25 rows containing non-finite values (stat\_boxplot).



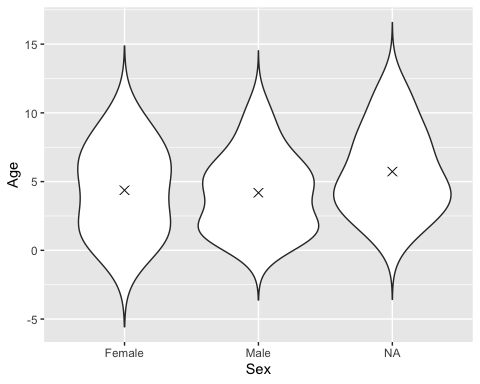
#INTERPRETATION  
#Among these four states  
#a. Tigers observed Karnataka have the highest median, minimum and maximum ages  
#b. Karnataka and Madhya Pradesh would have negatively skewed distributions  
# This means that  
# 1) Tigers below median age are more scattered accross the age spectrum  
# 2) Tigers above median age are more concentrated in a smaller range of ages  
  
#4.1  
#Violin plot for age per sex   
vp1 = ggplot(data, aes(x = Sex, y = Age))  
vp1 + geom\_violin(trim = FALSE)

## Warning: Removed 40 rows containing non-finite values (stat\_ydensity).



vp1 + geom\_violin(trim = FALSE) + stat\_summary(fun.y = mean, geom = "point", na.rm = TRUE, shape = "cross", size = 3)

## Warning: Removed 40 rows containing non-finite values (stat\_ydensity).



#INTERPRETATION  
#The interpretations are similar to the density plot.  
#a. As seen with the density plot, females are distributed similar to males accross ages  
# (They are also more evenly spread accross ages)  
#b. The means of males, females and unrecorded genders are very similar  
#c. The population is more concentrated around the young to mid-range ages  
  
#4.2  
#Violin and box plot for age with regard to mortalities and sex  
vp2 = ggplot(data, aes(x = Mortality.Seizure, y = Age))  
vp2 + geom\_violin(trim = TRUE) + geom\_boxplot( width = 0.1, aes(y = Age, fill = Sex)) + scale\_fill\_brewer(palette = "Dark2")

## Warning: Removed 40 rows containing non-finite values (stat\_ydensity).  
  
## Warning: Removed 40 rows containing non-finite values (stat\_boxplot).



#INTERPRETATION  
#a. Recorded mortalities are higher at lower to mid-range ages  
#b. Recorded seizures are only around mid-range ages  
#c. No female seizures recorded  
#d. From the box plot, the recorded ages of mortalities of females is lower on average than males  
  
#5.1  
#Classifying tigers as mature and cubs  
data = mutate(data, Age.Group = factor(Age > 2, labels = c("Cub", "Mature")))  
  
#6.1  
#Finding age-wise data for each sex and mortality / seizure class  
#Mean and standard deviation of ages  
group\_by(data, Sex, Mortality.Seizure) %>% summarise(mean(Age, na.rm = TRUE), sd(Age, na.rm = TRUE))

## Warning: Factor `Sex` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

## Warning: Factor `Sex` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

## # A tibble: 5 x 4  
## # Groups: Sex [3]  
## Sex Mortality.Seizure `mean(Age, na.rm = TRUE)` `sd(Age, na.rm = TRUE)`  
## <fct> <fct> <dbl> <dbl>  
## 1 Female Mortality 4.37 3.20  
## 2 Male Mortality 3.86 2.71  
## 3 Male Seizure 7.5 2.12  
## 4 <NA> Mortality 5.72 3.33  
## 5 <NA> Seizure 5.75 2.35

#Minimum and maximum of ages  
group\_by(data, Sex, Mortality.Seizure) %>% summarise(min(Age, na.rm = TRUE), max(Age, na.rm = TRUE))

## Warning: Factor `Sex` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`  
  
## Warning: Factor `Sex` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

## # A tibble: 5 x 4  
## # Groups: Sex [3]  
## Sex Mortality.Seizure `min(Age, na.rm = TRUE)` `max(Age, na.rm = TRUE)`  
## <fct> <fct> <dbl> <dbl>  
## 1 Female Mortality 0.3 9   
## 2 Male Mortality 0.5 10.4  
## 3 Male Seizure 6 9   
## 4 <NA> Mortality 1 12   
## 5 <NA> Seizure 3.5 9

#7.1  
#Correlation may not be meaningful or accurate as there are many unknown values, hence we may not have sufficient to indicate relations.  
#Also, I am calculating correlation for binary variables, and the available correlation methods may not be appropriate.  
  
#Correlation between age and mortality  
j = 1  
tmp.age = c(0)  
tmp.mortality = c(0)  
for(i in 1:90)  
{  
 if(!is.na(Age[i]) && !is.na(Mortality.Seizure))  
 {  
 tmp.age[j] = Age[i] > 10  
 tmp.mortality[j] = Mortality.Seizure[i] == "Mortality"  
 j = j + 1  
 }  
}  
cor(tmp.age, tmp.mortality)

## [1] 0.09329556

#INTERPRETATION  
#We see a very weak correlation between age being above 10 and mortality  
#However, this could be because of the many missing values  
  
#7.2  
#Confirming that the vectors are of the same size  
length(tmp.age)

## [1] 50

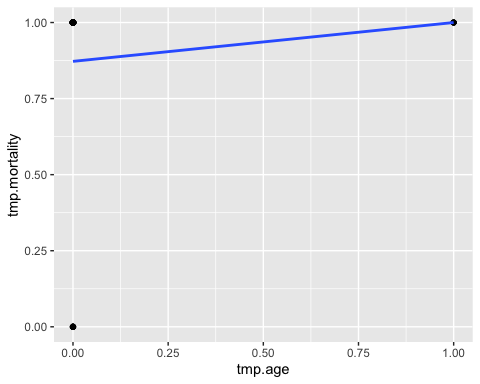
length(tmp.mortality)

## [1] 50

#Making a dataframe from the vectors  
df = data.frame(tmp.age, tmp.mortality)  
#Finding the regression  
lm(tmp.mortality~tmp.age, data = df)

##   
## Call:  
## lm(formula = tmp.mortality ~ tmp.age, data = df)  
##   
## Coefficients:  
## (Intercept) tmp.age   
## 0.8723 0.1277

#According to the result, mortality = 0.8723 + 0.1277\*age  
ggplot(df, aes(x = tmp.age, y = tmp.mortality)) + geom\_point() + geom\_smooth(method = "lm", se = FALSE)



#INTERPRETATION  
#The above result is not meaningful, and is done mainly to show off my skills  
#It is not meaningul, as it is not between two continuous variables