MT project

Shangchen Han

10/19/2019

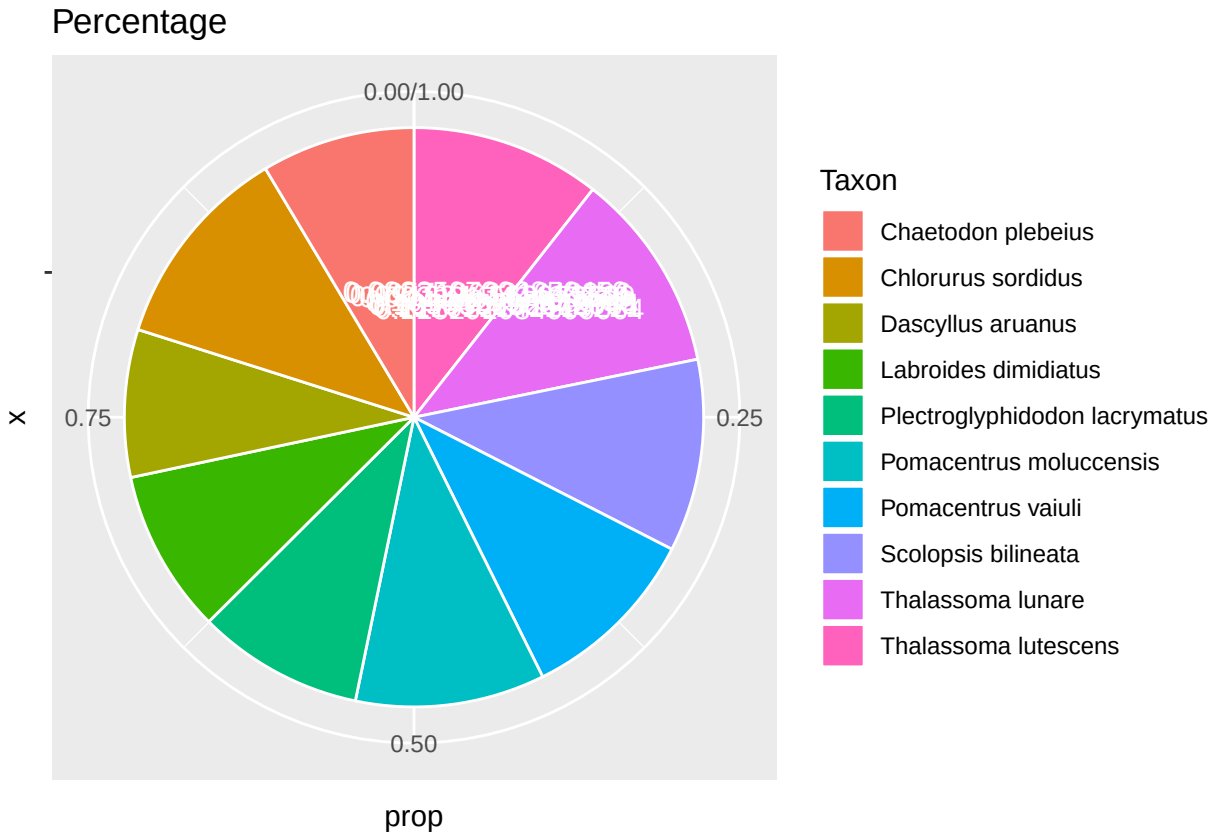
```
df <- read.csv(file = "Bar-chart.csv",header = TRUE,sep = ",")
MyData <- read.csv(file = "data1.csv" ,header = TRUE, sep = ",")
Data1 <- unique(MyData$Ecoregion)
num_Taxon <- MyData %>% count(Taxon)
num_Date <- MyData %>% count(SurveyDate)
Total_Ningaloo <- MyData %>% filter(Ecoregion == "Ningaloo")
Total_Broome <- MyData %>% filter(Ecoregion == "Exmouth to Broome")
Ning_Taxon <- Total_Ningaloo %>% count(Taxon)
Broo_Taxon <- Total_Broome %>% count(Taxon)
```

Extract data from Global reef fish dataset. After explored data features, I found that the whole data has two ecoregions.So, I wanted to find the differences of taxon between these two ecoregions.Then, I divided the whole data into two parts.

```
Ning_Taxon <- Ning_Taxon[with(Ning_Taxon,order(-n)),]
Ning_Taxon_10 <- Ning_Taxon[1:10,]
New_Ning_Taxon_10 <- Ning_Taxon_10 %>% mutate(prop = n/sum(n))
New_Ning_Taxon_10
```

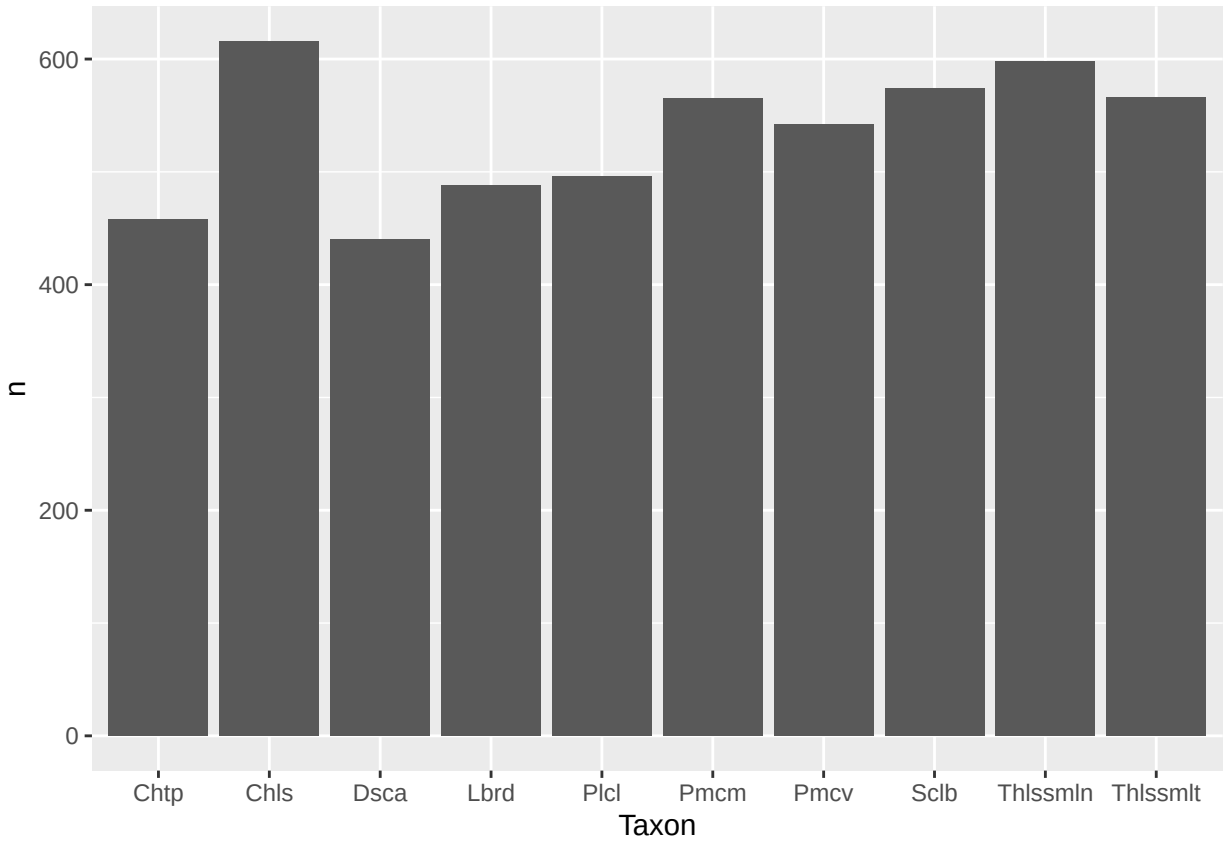
```
## # A tibble: 10 x 3
##    Taxon                          n   prop
##    <fct>                      <int>  <dbl>
##  1 Chlorurus sordidus           616 0.115
##  2 Thalassoma lunare            598 0.112
##  3 Scolopsis bilineata          574 0.107
##  4 Thalassoma lutescens         566 0.106
##  5 Pomacentrus moluccensis      565 0.106
##  6 Pomacentrus vaiuli           542 0.101
##  7 Plectroglyphidodon lacrymatus  496 0.0928
##  8 Labroides dimidiatus         488 0.0913
##  9 Chaetodon plebeius           458 0.0857
## 10 Dascyllus aruanus            440 0.0824
```

```
ggplot(New_Ning_Taxon_10,aes(x = "", y = prop, fill = Taxon))+
  geom_bar(width = 1, stat = "identity", color = "white")+
  coord_polar("y", start = 0)+
  geom_text(aes(y = prop,label = prop), color = "white")+
  labs(title = "Percentage")
```



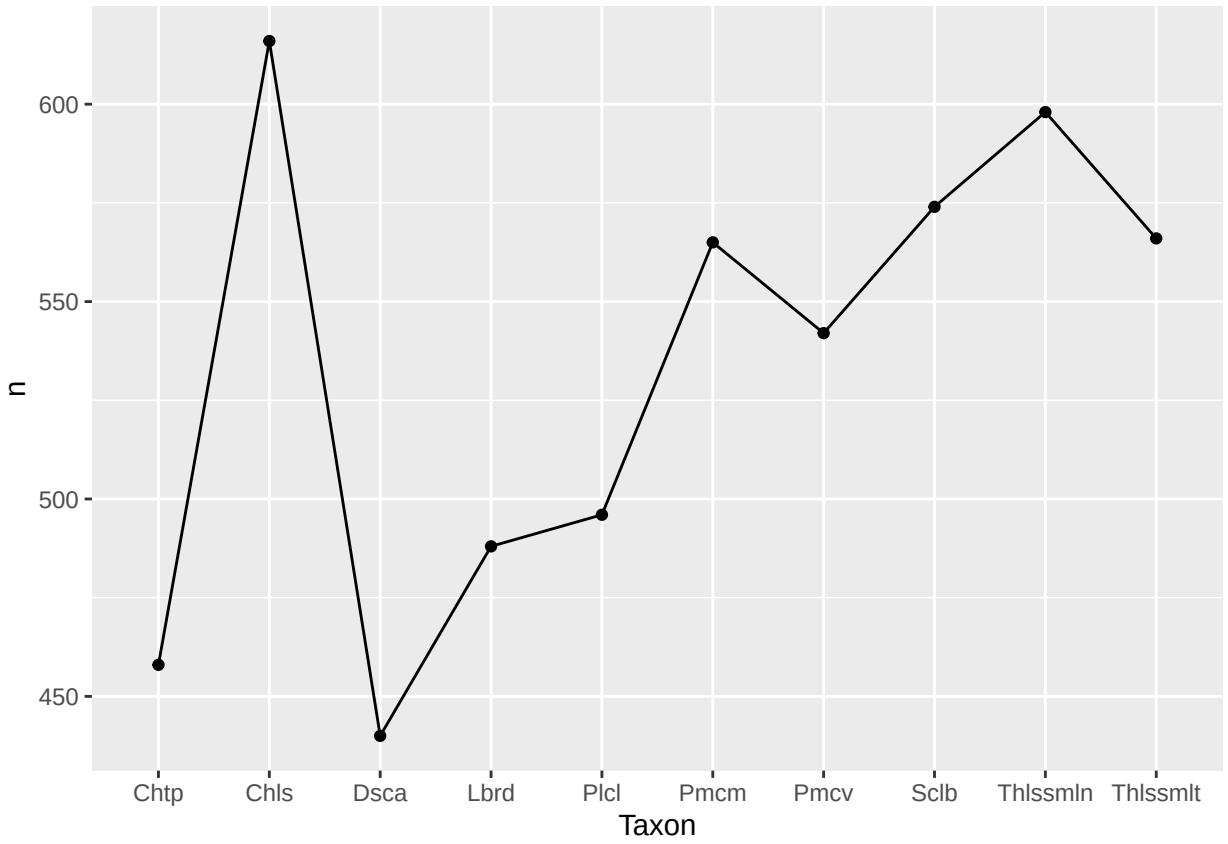
Pick the top 10 taxon in each regions, because there are lots of taxon in each, and then use graphs to analyze. But, unfortunately, the pie chart of ggplot has problems (the data overlapped).

```
ggplot(New_Ning_Taxon_10,aes(x=Taxon,y=n))+
  geom_bar(stat = "identity")+
  scale_x_discrete(labels = abbreviate)
```



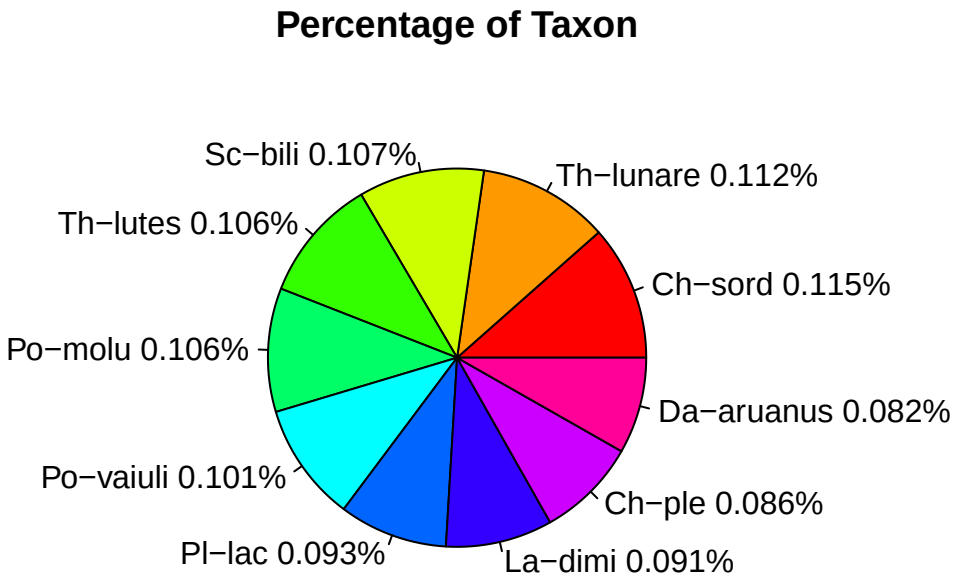
Using bar-chart to find out the differences of numbers between 10 taxon.

```
ggplot(data = New_Ning_Taxon_10, aes(x=Taxon, y= n, group = 1))+
  geom_line()+
  geom_point()+
  scale_x_discrete(labels = abbreviate)
```



Line chart is likely to show the differences clearly.

```
slices <- New_Ning_Taxon_10$prop
lbls <- c("Ch-sord","Th-lunare","Sc-bili","Th-lutes","Po-molu","Po-vaiuli","Pl-lac","La-dimi","Ch-ple",
pct <-round(New_Ning_Taxon_10$prop,3)
lbls <- paste(lbls,pct)
lbls <- paste(lbls,"%",sep = "")
pie(slices,labels = lbls, col = rainbow(length(lbls)),main = "Percentage of Taxon")
```

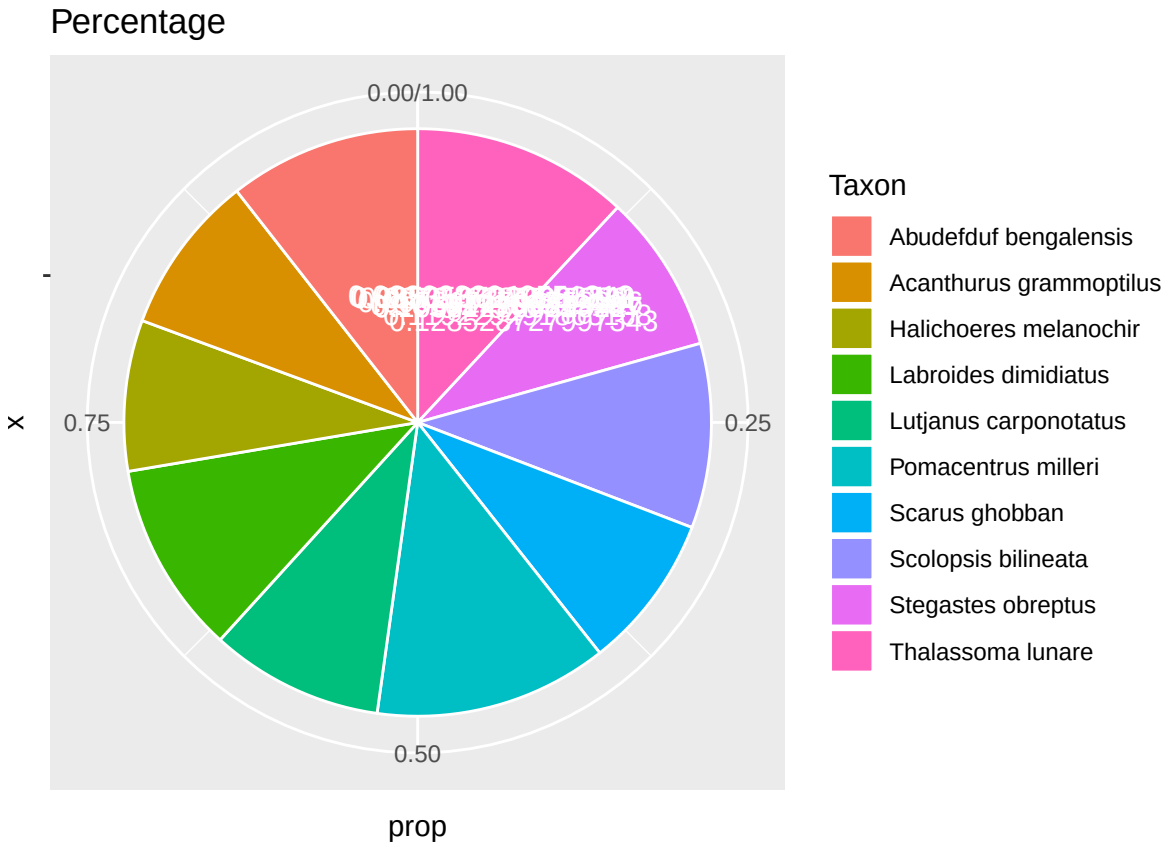


If the subset coveres only 10 taxon, and find out the proportion of these 10 taxon.

```
Broo_Taxon <- Broo_Taxon[with(Broo_Taxon,order(-n)),]
Broo_Taxon_10 <- Broo_Taxon[1:10,]
New_Broo_Taxon_10 <- Broo_Taxon_10 %>% mutate(prop = n/sum(n))
New_Broo_Taxon_10
```

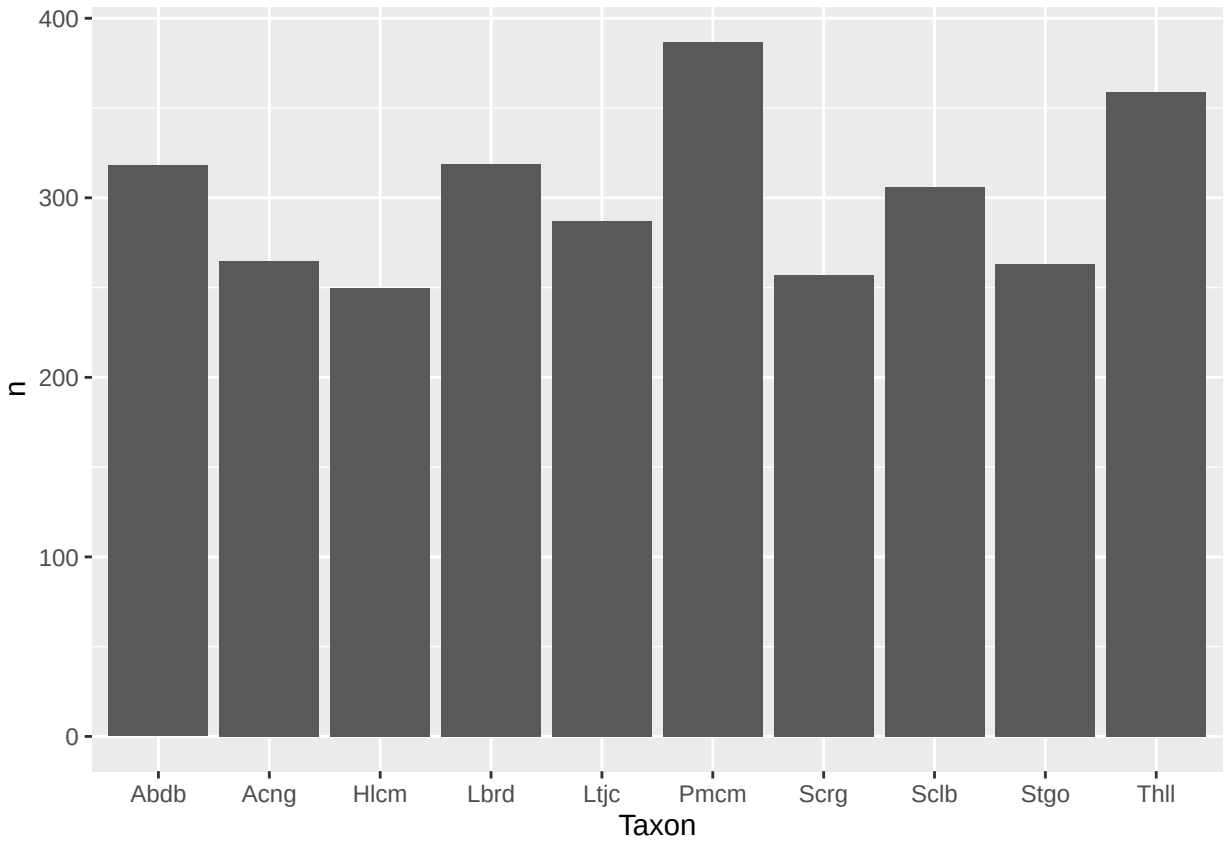
```
## # A tibble: 10 x 3
##    Taxon                      n   prop
##    <fct>                  <int>  <dbl>
##  1 Pomacentrus milleri      387 0.129
##  2 Thalassoma lunare        359 0.119
##  3 Labroides dimidiatus     319 0.106
##  4 Abudefduf bengalensis    318 0.106
##  5 Scolopsis bilineata      306 0.102
##  6 Lutjanus carponotatus    287 0.0953
##  7 Acanthurus grammoptilus  265 0.0880
##  8 Stegastes obreptus       263 0.0873
##  9 Scarus ghobban           257 0.0854
## 10 Halichoeres melanochir   250 0.0830
```

```
ggplot(New_Broo_Taxon_10,aes(x = "", y = prop, fill = Taxon))+
  geom_bar(width = 1, stat = "identity", color = "white")+
  coord_polar("y", start = 0)+
  geom_text(aes(y = prop,label = prop), color = "white")+
  labs(title = "Percentage")
```



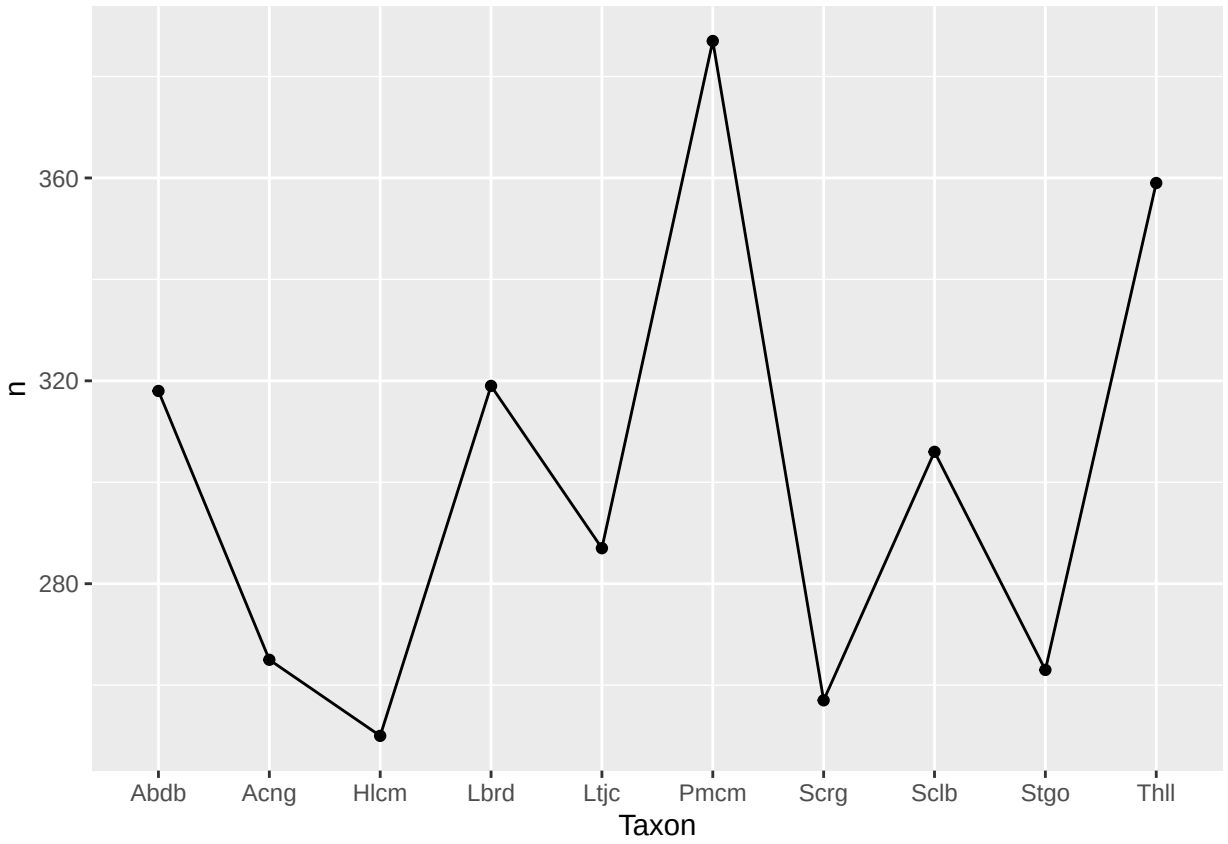
The same as above mentioned, the ggplot of pie-chart has problem of overlapped.

```
ggplot(New_Broo_Taxon_10,aes(x=Taxon,y=n))+
  geom_bar(stat = "identity")+
  scale_x_discrete(labels = abbreviate)
```

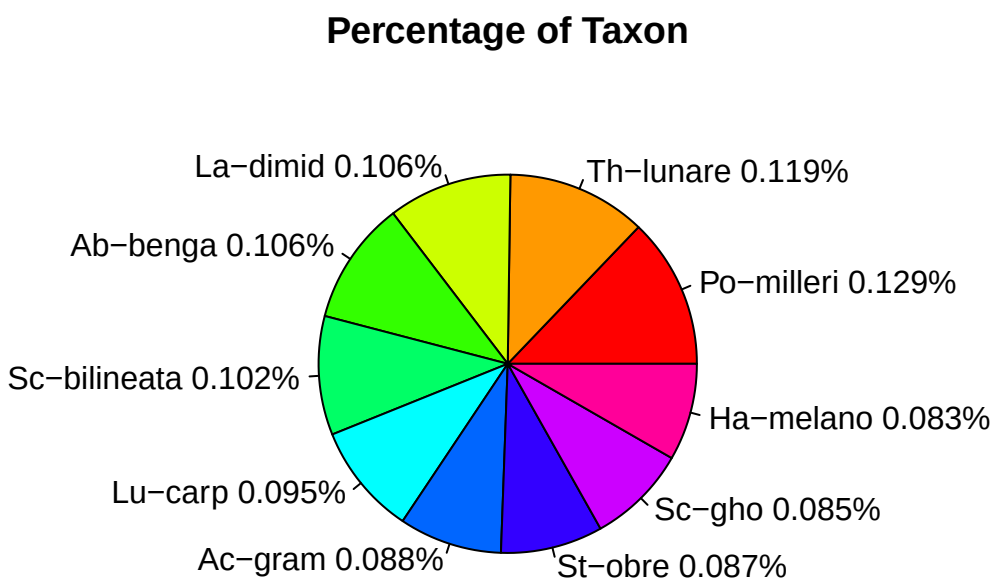


Using bar chart to find out the features of top 10 taxon.

```
ggplot(data = New_Broo_Taxon_10, aes(x=Taxon, y= n, group = 1))+
  geom_line()+
  geom_point()+
  scale_x_discrete(labels = abbreviate)
```




```
slices <- New_Broo_Taxon_10$prop
lbls <- c("Po-milleri","Th-lunare","La-dimid","Ab-benga","Sc-bilineata","Lu-carp","Ac-gram","St-obre","
pct <-round(New_Broo_Taxon_10$prop,3)
lbls <- paste(lbls,pct)
lbls <- paste(lbls,"%",sep = "")
pie(slices,labels = lbls, col = rainbow(length(lbls)),main = "Percentage of Taxon")
```



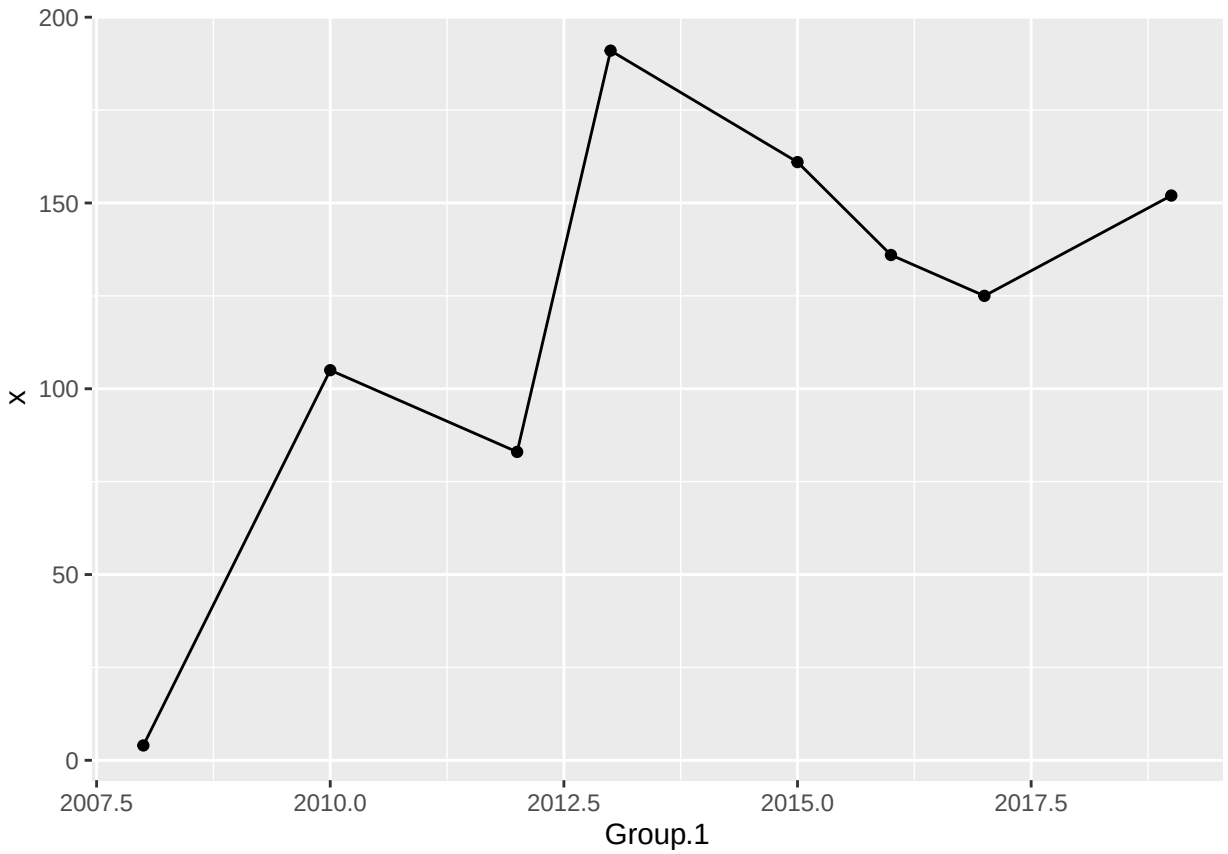
Using pie-chart to find out the proportion of these 10 taxon in Exmouth to Broome region.

```
data_Th_lunare <- MyData %>% filter(Taxon == "Thalassoma lunare")
num_date_Th_lunare <- data_Th_lunare %>% count(SurveyDate)
New_num_date_Th_lunare <- num_date_Th_lunare %>% mutate(year(num_date_Th_lunare$SurveyDate))
names(New_num_date_Th_lunare)[names(New_num_date_Th_lunare) == "year(num_date_Th_lunare$SurveyDate)"] <-
year_Th <- unique(New_num_date_Th_lunare$year)
sum_Th <- aggregate(New_num_date_Th_lunare$n, by=list(New_num_date_Th_lunare$year), FUN = sum)
sum_Th
```

```
##   Group.1   x
## 1    2008   4
## 2    2010 105
## 3    2012  83
## 4    2013 191
## 5    2015 161
## 6    2016 136
## 7    2017 125
## 8    2019 152
```

After compared graphs in Ningaloo and Exmouth to Broome, I found that a taxon named "Thalassoma lunare appeared in both region, so I extracted the data of"Thalassoma lunare", and wanted to find the trend of this species.

```
ggplot(data = sum_Th,aes(x=Group.1,y=x,group=1))+
  geom_line()+
  geom_point()
```



The trend of "Thalassoma lunare" in 11 years.

```r
Data <- read.csv(file = "Data_Th.csv",header = TRUE, sep = ",")
summary(Data)
```

```
##              Ecoregion                   Site       SurveyDate
##  Exmouth to Broome:359   Bundegi SZ North : 32   Min.   :2008
##  Ningaloo         :598   Bundegi SZ South : 31   1st Qu.:2013
##                          Dugong           : 31   Median :2015
##                          Bruboodjoo       : 30   Mean   :2015
##                          Coral Bay Central: 28   3rd Qu.:2017
##                          Monck Head       : 28   Max.   :2019
##                          (Other)          :777
##      Depth                       Taxon
##  Min.   : 1.000   Thalassoma lunare:957
##  1st Qu.: 2.300
##  Median : 3.300
##  Mean   : 4.126
##  3rd Qu.: 5.100
##  Max.   :18.000
##
```

```r
a <- Data$Ecoregion
b <- Data$Site
c <- Data$SurveyDate
d <- Data$Depth
glm.fit <- glm(Taxon~a+b+c+d,family = binomial(link = "logit"),data = Data,control = list(maxit=100))
glm.fit
```

```
##
## Call:  glm(formula = Taxon ~ a + b + c + d, family = binomial(link = "logit"),
##     data = Data, control = list(maxit = 100))
##
## Coefficients:
##                      (Intercept)                          aNingaloo
##                        -2.957e+01                         -3.688e-14
##                   bAirlie island         bAirlie Island inner reef
##                         4.043e-14                          3.337e-14
##        bBarrow Island deep south               bBarrow Island East
##                        -2.112e-15                          1.196e-13
##     bBarrow Island Port Central                 bBarrow NE bommie
##                         2.585e-14                          1.196e-13
##                    bBedout north            bBedout porites bommie
##                        -1.434e-14                          3.549e-14
##                bBedout reef edge               bBedout Reef sunset
##                         3.549e-14                         -6.464e-15
##                 bBedout reef top                       bBills Bay 1
##                        -3.878e-14                          8.408e-14
##                      bBruboodjoo                          bBundegi
##                         4.738e-14                          2.062e-14
##                     bBundegi RZ                 bBundegi SZ North
##                         1.377e-14                          2.408e-14
##                bBundegi SZ South                  bCardabia Patch
##                         5.577e-14                          1.756e-13
##     bCoral Bay (Sanctuary zone)                bCoral Bay Central
##                         4.990e-14                          1.130e-13
```

```
##                bCoral Bay DEC1             bCoral Bay Offshore
##                       1.228e-13                       3.020e-13
##             bCorneliesse Shoal                bDead Tree Beach
##                       3.695e-13                       1.180e-15
##              bDelambre Inner 1               bDelambre Inner 3
##                       9.049e-14                       7.275e-14
##             bDelambre Outside 3                    bDelay Point
##                       3.549e-14                       3.445e-14
##   bDons Point West Lewis Island                         bDugong
##                      -2.701e-14                       5.035e-14
##              bEnderby I West Pt                     bEnderby NE
##                       7.370e-14                      -9.197e-15
##                bEnderby SW Bay              bEnderby West Rock
##                       2.209e-14                       3.468e-14
##               bGeneva Bay North             bGeographe Shoals N
##                       3.737e-14                       2.231e-13
##            bGeographe Shoals NW                     bGoodwyn N
##                       2.844e-13                       3.549e-14
##                     bGoodwyn NW              bGoodwyn South Bay
##                       6.806e-14                       2.585e-14
##                    bGoodwyn Sth             bHamersley Inside 1
##                       7.310e-14                       1.692e-14
##                 bJonquil Island                  bKate s Corner
##                       9.167e-15                      -1.411e-15
##                     bKendrew 2               bLegendre Inside 3
##                       2.585e-14                       1.034e-13
##             bLegendre Outside 1            bLittle Mangrove Bay
##                       1.034e-13                       5.200e-14
##                       bMalus Is                       bMaud RZ1
##                       2.370e-14                       2.037e-12
##                       bMaud RZ2               bMaud SZ external
##                       1.343e-13                       6.382e-14
##                    bMaud SZ Nth                     bMonck Bowl
##                       4.104e-14                       1.327e-13
##                     bMonck Head               bMonck Head Inner
##                       4.143e-14                       6.878e-14
##                bMonck Head North               bMonck Head Outer
##                       4.946e-14                       8.015e-14
##                bMonck Head Sth                     bMonck Wall
##                       5.688e-14                       1.526e-13
##     bMontebellos Central Lagoon          bMontebellos East Rock
##                      -6.011e-27                       3.549e-14
##                bMontebellos SE              bMontebellos South
##                       1.401e-13                       2.585e-14
##         bNorth Cormorant Island        bNorth Muiron Island East
##                      -6.475e-27                       7.174e-14
##         bNorth Muiron Island NE              bNorth West Island
##                       7.799e-14                      -6.236e-27
##           bNorth West Island NW                    bNth Bundegi
##                       3.549e-14                      -3.688e-14
##                     bNth Muiron                bNW Island Lagoon
##                       1.322e-13                       2.750e-14
##                bOld 9 Buoy Reef           bOutside Yalobia North
##                       2.271e-13                       2.721e-13
```

```
##          bOutside Yalobia South                bOyster rocks buoy
##                        3.141e-13                         2.131e-13
##             bOyster rocks south                          bPelican
##                        2.192e-13                         5.798e-14
##                     bPelican Nth                bPoint Maud Inner
##                        6.267e-14                         7.697e-14
##                bPoint Maud Outer                    bRick s Folly
##                        6.012e-14                         2.585e-14
##                       bRubble Is                bSailfish North 1
##                        5.361e-14                         6.404e-14
##                bSailfish South 2  bSouth Muiron Island NE Channel
##                        2.585e-14                         6.727e-14
##                bSouth West Reef              bStephenson Channel
##                        7.520e-15                         1.293e-14
##                     bSth Bundegi                       bSth Muiron
##                       -1.338e-14                         1.856e-13
##                bSW West Lewis Is            bTrimouille Island NE
##                       -6.805e-15                         2.585e-14
##           bTrimouille Island SE  bTurquoise Bay (Sanctuary zone)
##                        2.585e-14                         6.279e-15
##                  bVaranus Island            bVaranus Island north
##                        1.401e-13                         1.196e-13
##           bVaranus North bommie              bVaranus Port limit
##                        8.567e-14                         1.196e-13
##                  bW Lewis Is NW                       bWest Reef
##                       -8.443e-15                                NA
##                 bYalobia Bommie                  bYalobia Passage
##                        1.800e-13                         1.118e-13
##                  bYalobia South                                  c
##                        1.488e-13                        -9.788e-15
##                                d
##                       -2.585e-14
##
## Degrees of Freedom: 956 Total (i.e. Null);  851 Residual
## Null Deviance:       0
## Residual Deviance: 2.767e-10     AIC: 212
```

```
glm.fit2 <- glm(Taxon~a+c+d,family = binomial(link = "logit"),data = Data,control = list(maxit=100))
glm.fit2
```

```
##
## Call:  glm(formula = Taxon ~ a + c + d, family = binomial(link = "logit"),
##     data = Data, control = list(maxit = 100))
##
## Coefficients:
## (Intercept)    aNingaloo            c            d
##  -2.957e+01    4.208e-14   -2.217e-15   -6.299e-15
##
## Degrees of Freedom: 956 Total (i.e. Null);  953 Residual
## Null Deviance:       0
## Residual Deviance: 2.767e-10     AIC: 8
```

After clean data, I picked 4 variables to fix the logistic model. But from the output, the coefficients of Ecoregion,Site, SurveyDate and Depth are pretty tiny. So, these variables do not have siginificant influence on Thalassoma lunare.But only for these two models, the second one may be better, because of smaller AIC value.