```
# How-to-Report-the-Distribution-of-Attributes-per-Cluster

Let’s say that you have applied your Clustering algorithm and you would like to report the distribution of the categorical variables per cluster in a “tidy” report. Below you can see a suggestion of how you can do it in R.
Generate the Data

Let’s assume that we came up with 3 clusters such as “C1, C2 and C3” and that we have 3 attribute such as:

    Gender: “M”, “F”
    Type: “A”, “B”, “C”, “D”
    Category: “High”, “Medium”, “Low”

library(tidyverse)

set.seed(5)

df1<-tibble(ID=seq_len(500))%>%
     mutate(Cluster = "C1",
            Gender=sample(c("M", "F"), n(), replace=TRUE, prob=c(0.6, 0.4)),
            Type=sample(c("A", "B", "C", "D"), n(), replace=TRUE, prob=c(0.20, 0.3, 0.4, 0.1)),
            Category=sample(c("High", "Medium", "Low"), n(), replace=TRUE, prob=c(0.1, 0.6, 0.3))

df2<-tibble(ID=seq_len(300))%>%
  mutate(Cluster = "C2",
         Gender=sample(c("M", "F"), n(), replace=TRUE, prob=c(0.4, 0.6)),
         Type=sample(c("A", "B", "C", "D"), n(), replace=TRUE, prob=c(0.40, 0.1, 0.2, 0.3)),
         Category=sample(c("High", "Medium", "Low"), n(), replace=TRUE, prob=c(0.7, 0.2, 0.1)))

df3<-tibble(ID=seq_len(200))%>%
  mutate(Cluster = "C3",
         Gender=sample(c("M", "F"), n(), replace=TRUE, prob=c(0.2, 0.8)),
         Type=sample(c("A", "B", "C", "D"), n(), replace=TRUE, prob=c(0.5, 0.3, 0.1, 0.1)),
         Category=sample(c("High", "Medium", "Low"), n(), replace=TRUE, prob=c(0.1, 0.2, 0.7)))

df<-rbind.data.frame(df1, df2, df3)

df


# A tibble: 1,000 x 5
      ID Cluster Gender Type  Category

 1     1 C1      M      C     Medium
 2     2 C1      F      C     Medium
 3     3 C1      F      C     Medium
 4     4 C1      M      B     Low
 5     5 C1      M      B     Low
 6     6 C1      F      C     Medium
 7     7 C1      M      C     Medium
 8     8 C1      F      B     High
 9     9 C1      F      C     Medium
10    10 C1      M      A     Medium
# ... with 990 more rows
```

```
53
54  Report the Distribution of Attributes
55
56
57
58  attributes <- names(df[3:dim(df)[2]])
59
60
61  output<-NULL
62
63  for (a in attributes) {
64
65    tmp<-df%>%group_by_(a, "Cluster")%>% summarise(n = n())%>%
66      group_by(Cluster)%>%mutate(Prop=n/(sum(n)))%>%
67      ungroup()%>%select(-n)%>%
68      spread(Cluster, Prop)%>%mutate(Attribute = a)%>%select(Attribute, everything())
69    colnames(tmp)[1:2]<-c("attribute", "values")
70
71    output<-rbind(output, tmp)
72
73  }
74
75  output
76
77
78  # A tibble: 9 x 5
79    attribute values    C1    C2    C3
80
81  1 Gender    F      0.398 0.593 0.78
82  2 Gender    M      0.602 0.407 0.22
83  3 Type      A      0.188 0.413 0.425
84  4 Type      B      0.318 0.1   0.365
85  5 Type      C      0.39  0.193 0.105
86  6 Type      D      0.104 0.293 0.105
87  7 Category  High   0.114 0.683 0.065
88  8 Category  Low    0.312 0.103 0.75
```