607_HW1

Nathan Cooper

August 28, 2017

In this assignment we are tasked with downloading a famous dataset about mushrooms from : https://archive.ics.uci.edu/ml/datasets/Mushroom, putting the data into a data frame such as

```
mushrooms <- read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/agaricus-lepi
View(mushrooms)
mushrooms_df <- data.frame(mushrooms)
View(mushrooms_df)
```

We can see that this is entirely catagorical data. A key is needed to make sense of it. From the website: https://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/agaricus-lepiota.names

7. Attribute Information: (classes: edible=e, poisonous=p)
    1. cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
    2. cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
    3. cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
    4. bruises?: bruises=t,no=f
    5. odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
    6. gill-attachment: attached=a,descending=d,free=f,notched=n
    7. gill-spacing: close=c,crowded=w,distant=d
    8. gill-size: broad=b,narrow=n
    9. gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y
    10. stalk-shape: enlarging=e,tapering=t
    11. stalk-root: bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?
    12. stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
    13. stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
    14. stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
    15. stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
    16. veil-type: partial=p,universal=u
    17. veil-color: brown=n,orange=o,white=w,yellow=y
    18. ring-number: none=n,one=o,two=t
    19. ring-type: cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
    20. spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
    21. population: abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
    22. habitat: grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d We can use this information to rename the columns of the data frame:

```
mushrooms_df <- setNames(mushrooms_df, c("Edibility", "Cap Shape", "Cap Surface", "Cap Color", "Bruises
```

My subset will be based on Edibility, as required, and also odor, population, and habitat.

```
subshroom_df <- subset(mushrooms_df, select= c("Edibility", "Odor", "population", "habitat"))
```

To reassign values in the dataframe the library plyr is usefull

```
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.4.1
```

I will use the reassign function first for Edibility

```
subshroom_df$Edibility <- revalue(subshroom_df$Edibility, c("e" = "edible"))
subshroom_df$Edibility <- revalue(subshroom_df$Edibility, c("p" = "poisonous"))
```

Next for Odor

```
subshroom_df$Odor <- revalue(subshroom_df$Odor, c("a" = "almond"))
subshroom_df$Odor <- revalue(subshroom_df$Odor, c("l" = "anise"))
subshroom_df$Odor <- revalue(subshroom_df$Odor, c("c" = "creosote"))
subshroom_df$Odor <- revalue(subshroom_df$Odor, c("y" = "fishy"))
subshroom_df$Odor <- revalue(subshroom_df$Odor, c("f" = "foul"))
subshroom_df$Odor <- revalue(subshroom_df$Odor, c("m" = "musty"))
subshroom_df$Odor <- revalue(subshroom_df$Odor, c("n" = "none"))
subshroom_df$Odor <- revalue(subshroom_df$Odor, c("p" = "pungent"))
subshroom_df$Odor <- revalue(subshroom_df$Odor, c("s" = "spicy"))
```

Now for Population

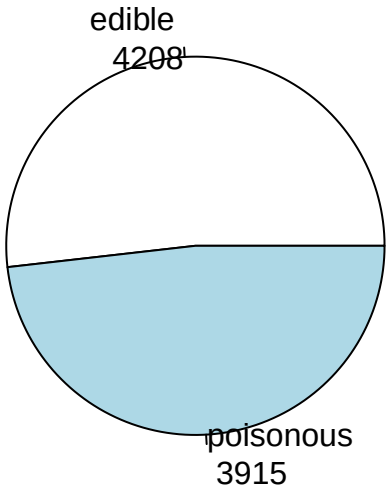
```
subshroom_df$population <- revalue(subshroom_df$population, c("a" = "abundant"))
subshroom_df$population <- revalue(subshroom_df$population, c("c" = "clustered"))
subshroom_df$population <- revalue(subshroom_df$population, c("n" = "numerous"))
subshroom_df$population <- revalue(subshroom_df$population, c("s" = "scattered"))
subshroom_df$population <- revalue(subshroom_df$population, c("v" = "several"))
subshroom_df$population <- revalue(subshroom_df$population, c("y" = "solitary"))
```

Finally for habitat

```
subshroom_df$habitat <- revalue(subshroom_df$habitat, c("g" = "grasses"))
subshroom_df$habitat <- revalue(subshroom_df$habitat, c("l" = "leaves"))
subshroom_df$habitat <- revalue(subshroom_df$habitat, c("m" = "meadows"))
subshroom_df$habitat <- revalue(subshroom_df$habitat, c("p" = "pathes"))
subshroom_df$habitat <- revalue(subshroom_df$habitat, c("u" = "urban"))
subshroom_df$habitat <- revalue(subshroom_df$habitat, c("w" = "waste"))
subshroom_df$habitat <- revalue(subshroom_df$habitat, c("d" = "woods"))
```

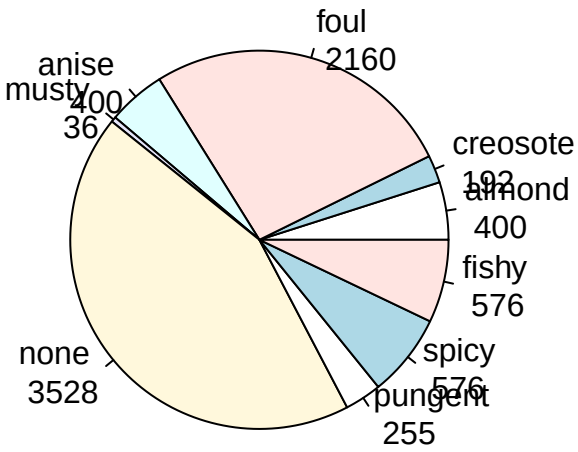
Let's take a look at the data to see if there are any interesting patterns. Since these data are catagorical, pie charts might be handy in looking at how the percentages edible vs poisonous is distributed. I used http://www.statmethods.net/graphs/pie.html as a guide.

```
Edibility_pie <- table(subshroom_df$Edibility)
lbls <- paste(names(Edibility_pie), "\n", Edibility_pie, sep = " ")
pie(Edibility_pie, labels = lbls, main = "Pie Chart of Mushroom Edibility\n (with sample sizes)")
```



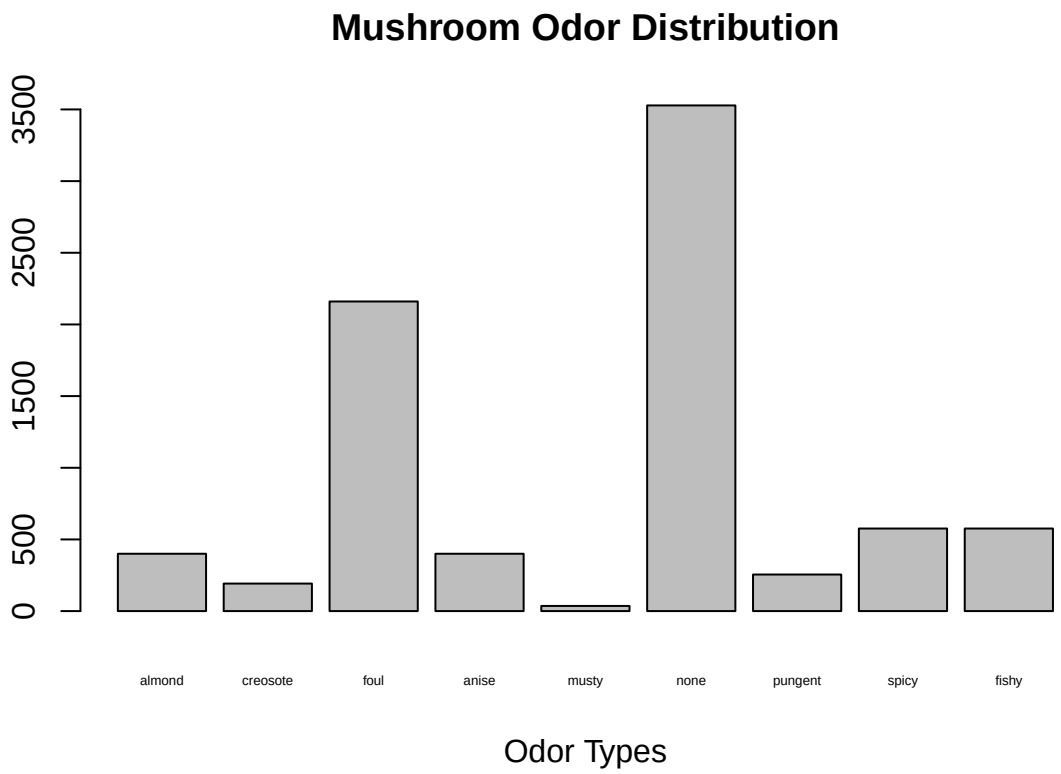
Next we will look at Odor

```
Odor_pie <- table(subshroom_df$Odor)
lbls <- paste(names(Odor_pie) ,"\n", Odor_pie, sep = " ")
pie(Odor_pie, labels = lbls, main = "Mushroom Odor Pie Chart\n (with sample sizes")
```



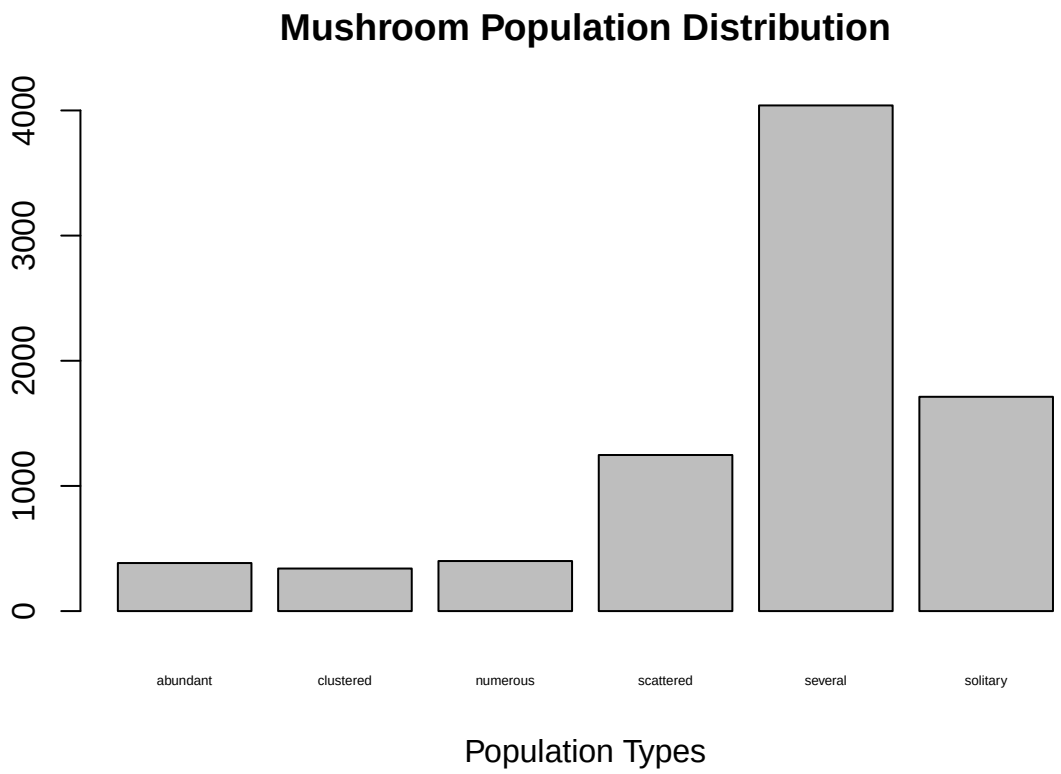
That's a little crowded, though I am surprised to find "none" as the largest catagory. Let's try a bar chart.

```
odor_barplot <- table(subshroom_df$Odor)
barplot(odor_barplot, main = "Mushroom Odor Distribution" , xlab = "Odor Types", cex.names = 0.45)
```



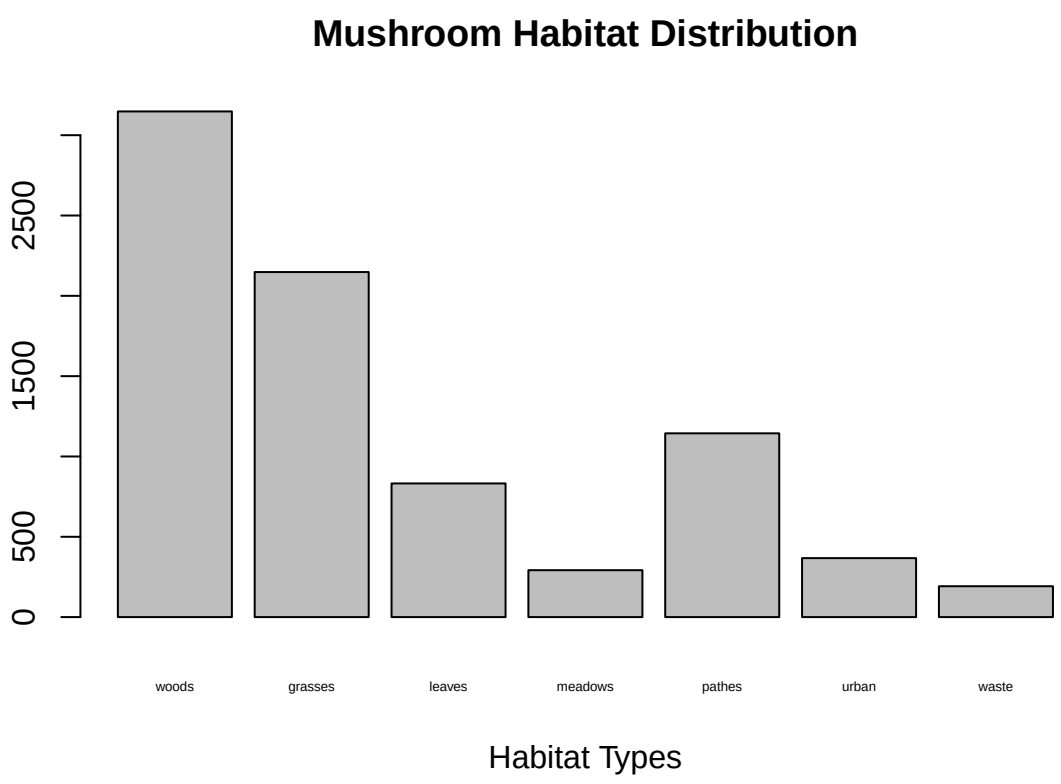
I had to make the Odor type print small so all types would be visible. We will look at population next.

```
pop_barplot <- table(subshroom_df$population)
barplot(pop_barplot, main = "Mushroom Population Distribution" , xlab = "Population Types", cex.names =
```



Finally we will look at habitat.

```
hab_barplot <- table(subshroom_df$habitat)
barplot(hab_barplot, main = "Mushroom Habitat Distribution" , xlab = "Habitat Types", cex.names = 0.45)
```


Mushroom Habitat Distribution
0
500
1500
2500
woods
grasses
leaves
meadows
pathes
urban
waste
Habitat Types