1. **Before the homework, I read the file:**

Read a csv file:

hydropower <- read.csv("hydropower.csv")

1. **In your own words, describe the data and how you obtained the data.**

Renewable energy becomes trendy nowadays and hydropower is an option that provides people with clean energy. In the USA, they are many dams that provide hydropower According to the 2010 Federal Memorandum of Understanding for Hydropower and the Energy Policy Act of 2005, The United States government led the investigation in all the dams in the USA. I obtain the raw data from the website: <https://corgis-edu.github.io/corgis/csv/hydropower/>

1. **How did you address cleaning the data? What did you do for missing data (if any)? What rows and columns did you delete? Provide the R code, if any. If you did the cleaning in Excel, describe your step-by-step approach.**

First of all, I use the summary function code to check my statistics results by columns.

summary(hydropower)

**A screenshot of a computer

Description automatically generated with medium confidence**

**Figure 2.1**

Then, based on different columns, I glimpse the data to check if the columns I want are abnormal. Hence, In Figure 2.1, when I check Identity.Project.Year, it has something wrong because years in normal understanding should not be zero. So, I decide to check it further.

hydropower$Identity.Project.Year

**Text

Description automatically generated with low confidence**

**Figure 2.2**

dim(hydropower)

****

**Figure 2.3**

Bingo! some rows in the Identity.Project.Year are obviously type error (Figure 2.2), so I decide to delete all the rows.

hydropower$Identity.Project.Year[hydropower$Identity.Project.Year == 0] <-- NA

hydropower$Identity.Project.Year[hydropower$Identity.Project.Year == 19] <-- NA

hydropower$Identity.Project.Year[hydropower$Identity.Project.Year == 1076] <-- NA

I change the Identity.Project.Year row which is not ‘19XX’ to be NA at first.

Then, I double-check if my correction is wrong (Figure 2.4).

hydropower$Identity.Project.Year

Table

Description automatically generated

**Figure 2.4**

Now, I delete the NA rows.

which(!is.na(hydropower$Identity.Project.Year))

hydropower <- hydropower[which(!is.na(hydropower$Identity.Project.Year)),]

hydropower$Identity.Project.Year

Table

Description automatically generated

**Figure 2.5**

dim(hydropower)



**Figure 2.6**

Finally, the result is as below: all the wrong data in Identity.Project.Year is cleaned (Figure 2.5), and twelve rows are deleted (Figure 2.3 and Figure 2.6).

1. **Use summary() to generate summary statistics of your three numeric variables (e.g. var1, var2, var3) by one of your categorical variables (e.g. summary statistics for "region 1" versus "region 2" versus "region 3"). Use boxplot() to visually show the summaries.**
   1. **Provide the R code**
   2. **Followed by graphs**
   3. **Followed by a brief description of what you see (in complete sentences).**

I define one categorical variable from the year into three different periods: ‘Before WW1’, ‘WW1 to WW2’, ‘After WW2’ to check if the norms to build a dam have changed.

hydropower$yearcheck[hydropower$Identity.Project.Year < 1918] <- 'Before WW1'

hydropower$yearcheck[hydropower$Identity.Project.Year >= 1918 & hydropower$Identity.Project.Year < 1945] <- 'WW1 to WW2'

hydropower$yearcheck[hydropower$Identity.Project.Year >= 1945] <- 'After WW2'

hydropower$yearcheck

Then, I sort the categorical variable from year old to new. That is, it will follow the order: ‘Before WW1’ 🡪 ‘WW1 to WW2’ 🡪 ‘After WW2’

hydropower$yearcheck <- factor(hydropower$yearcheck, labels = c('Before WW1','WW1 to WW2', 'After WW2' ))

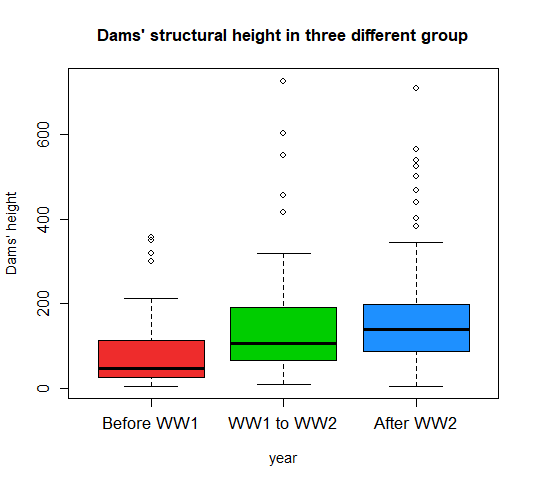
I draw a box plot to show the three different periods of the dam’s structural height (Figure 3.1):

boxplot(Dimensions.Structural.Height~yearcheck, data = hydropower,

col = c('firebrick2','green3', 'dodgerblue'),

main = "Dams' structural height in three different group", ylab = "Dams' height",

xlab = 'year', cex.axis = 1.2)



**Figure 3.1**

In the plot (Figure 3.1), the group ‘Before WW1’ has the lowest height of the three groups, and the group ‘After WW2’ has the highest average height of dams. I can assume that when the technology is more advanced, engineers can build a dam with a higher height.

1. **Generate a histogram with an overlaid density plot of your three numeric variables by one of your categorical variables.**
   1. **Provide the R code**
   2. **Followed by graphs**
   3. **Followed by a brief description of what you see (in complete sentences). You can compare the histograms/density plots of the categories.**

I chose structural height as a numeric variable and grouped it by different periods of the year:

The density histogram and the density lines are as below (Figure 4.1):

hist(hydropower$Dimensions.Structural.Height[hydropower$yearcheck == 'Before WW1'],

breaks = seq(0, 800, 50), main = "Dams hight distribution in three different group of year",

xlab = 'Height', col = rgb(1,0,0,0.25), ylim = c(0,0.012), freq = F)

hist(hydropower$Dimensions.Structural.Height[hydropower$yearcheck == 'WW1 to WW2'],

breaks = seq(0, 800, 50), col = rgb(0,1,0,0.25), add = T, freq = F)

hist(hydropower$Dimensions.Structural.Height[hydropower$yearcheck == 'After WW2'],

breaks = seq(0, 800, 50), col = rgb(0,0,1,0.25), add = T, freq = F)

dens1 <- density(hydropower$Dimensions.Structural.Height[hydropower$yearcheck == 'Before WW1'],

bw = 30,)

dens2 <- density(hydropower$Dimensions.Structural.Height[hydropower$yearcheck == 'WW1 to WW2'],

bw = 30)

dens3 <- density(hydropower$Dimensions.Structural.Height[hydropower$yearcheck == 'After WW2'],

bw = 30)

lines(dens1, col = 'deeppink1', lty = 1, lwd = 2)

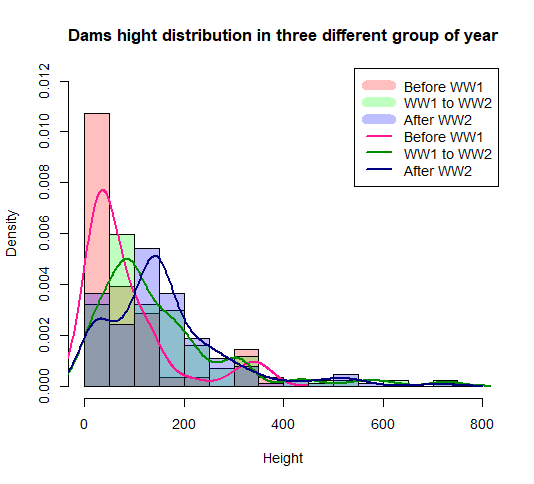
lines(dens2, col = 'green4', lty = 1, lwd = 2)

lines(dens3, col = 'navy', lty = 1, lwd = 2)

legend('topright', c(rep(c('Before WW1','WW1 to WW2', 'After WW2'),2)),

col = c(rgb(1,0,0,0.25), rgb(0,1,0,0.25), rgb(0,0,1,0.25),'deeppink1',

'green4', 'navy'), lwd =c(rep(10,3), rep(2,3)))



**Figure 4.1**

In the density histogram and density line (Figure 4.1), I can double-confirm that most dams before WW1 are built between 0 and 50, and when we have more technic skills, the dam’s height becomes higher.

The histogram frequency plot is as below:

hist(hydropower$Dimensions.Structural.Height[hydropower$yearcheck == 'Before WW1'],

breaks = seq(0, 800, 50), main = "Dams hight distribution in three different group of year",

xlab = 'Height', col = rgb(1,0,0,0.25), freq = T, ylim = c(0,50))

hist(hydropower$Dimensions.Structural.Height[hydropower$yearcheck == 'WW1 to WW2'],

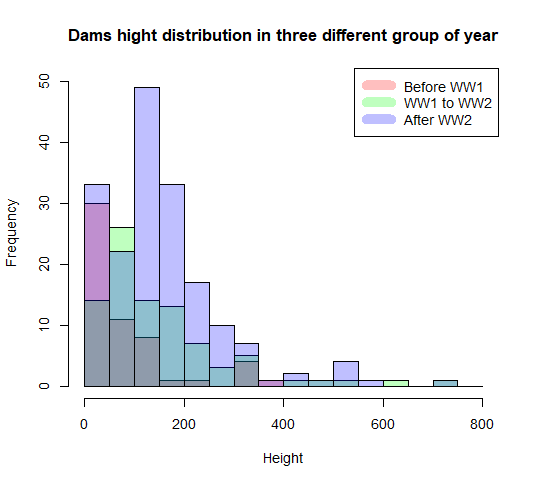
breaks = seq(0, 800, 50), col = rgb(0,1,0,0.25), add = T, freq = T)

hist(hydropower$Dimensions.Structural.Height[hydropower$yearcheck == 'After WW2'],

breaks = seq(0, 800, 50), col = rgb(0,0,1,0.25), add = T, freq = T)

legend('topright', c(rep(c('Before WW1','WW1 to WW2', 'After WW2'))),

col = c(rgb(1,0,0,0.25), rgb(0,1,0,0.25), rgb(0,0,1,0.25)), lwd = 10 )



**Figure 4.2**

Compared to density lines (Figure 4.2), the histogram by frequency is apparently more difficult to figure out when all the categorical variables are overlapped. Luckily, when I project them step by step (Figure 4.3), we still can get the same conclusion from the density histogram plot.

Chart, histogram

Description automatically generated

**Figure 4.3**

1. **Create a bar plot showing the count of one of your categorical variables.**
   1. **Provide the R code**
   2. **Followed by graphs**
   3. **Followed by a brief description of any patterns that you might see.**

hydropower$Location.State <- as.factor(hydropower$Location.State)

library("viridis")

state <- table(hydropower$Location.State)

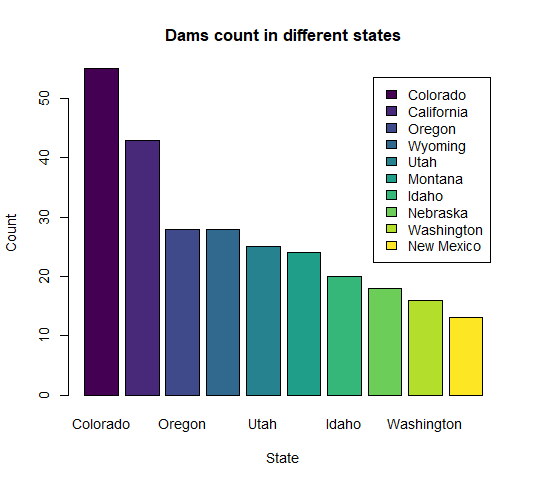
state

order(state, decreasing = T)

stateorder<- head(state[order(state, decreasing = T)],10)

barplot(stateorder, col = viridis(10), ylab = 'Count', xlab = 'State', legend.text = T,

main = 'Dams count in different states')

****

**Figure 5.1**

In the bar plot (Figure 5.1), we can find that Colorado has the most dam counts in the United States and the second-highest in California.

By the way, Because I am inert, so I use the color pattern to help me fill in the color by the different bars. (Use library("viridis") )

1. **Create a bar plot showing the by-group (category) average of one of your numeric variables:**
   1. **Provide the R code**
   2. **Followed by graphs**
   3. **Followed by a brief description of any patterns that you might see.**

library(RColorBrewer)

stateelev <- aggregate(Dimensions.Structural.Height~Location.State, data = hydropower,

FUN = mean)

stateelev

stateelevorder <- head(stateelev[order(stateelev$Dimensions.Structural.Height, decreasing = T),],10)

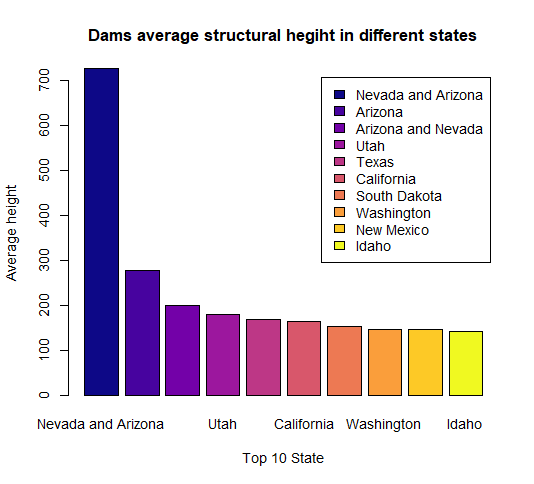
stateelevorder

barplot(stateelevorder$Dimensions.Structural.Height, names.arg = stateelevorder$Location.State,

col = plasma(10), ylab = 'Average height', xlab = 'Top 10 State', legend.text = stateelevorder$Location.State,

main = "Dams average structural hegiht in different states")

I chose dams height in all states to do math, and then draw a bar plot:



**Figure 6.1**

Interestingly, Nevada and Arizona have abnormal heights in dams in different states (Figure 6.1). So, I decided to check the raw data of Nevada and Arizona.

hydropower[hydropower$Location.State == 'Nevada and Arizona',]

It shows that Nevada and Arizona only have one dam (Figure 6.2), that’s why it’s an abnormal high. That makes me realize that sometimes the bar plot will be distorted when I use aggregation.

Text

Description automatically generated

**Figure 6.2**

So, I decide to change my strategy. I group the dataset the counts in different states at first, then sort the ten states by the average height of dams.

First of all, I combine two values to calculate count(state) and to calculate average height(statelev) into a new table. Then I do two orders: the first order is to extract the top 10 states which have the most dams. The second order is to order the average height by state. Finally, I draw a bar plot again:

stateorder <- data.frame(stateorder)

top10state <- cbind(state, stateelev)

top10state <- top10state[,-3]

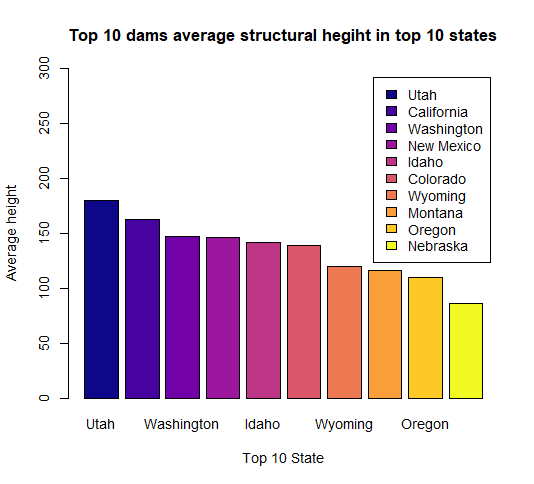
top10stateorder <- head(top10state[order(top10state$Freq, decreasing = T),],10)

top10heightorfer <- top10stateorder[order(top10stateorder$Dimensions.Structural.Height, decreasing = T),]

barplot(top10heightorfer$Dimensions.Structural.Height, names.arg = top10heightorfer$Location.State,

col = plasma(10), ylab = 'Average height', xlab = 'Top 10 State', legend.text = top10heightorfer$Location.State,

main = "Top 10 dams average structural hegiht in top 10 states", ylim = c(0,300))



**Figure 6.3**

The top 10 states show that in Utah has the highest height in dams (Figure 6.3), and the last one is in Nebraska.

1. **Create an interesting pie chart.**
   1. **Provide the R code**
   2. **Followed by graphs**
   3. **Followed by a brief description of any patterns that you might see.**

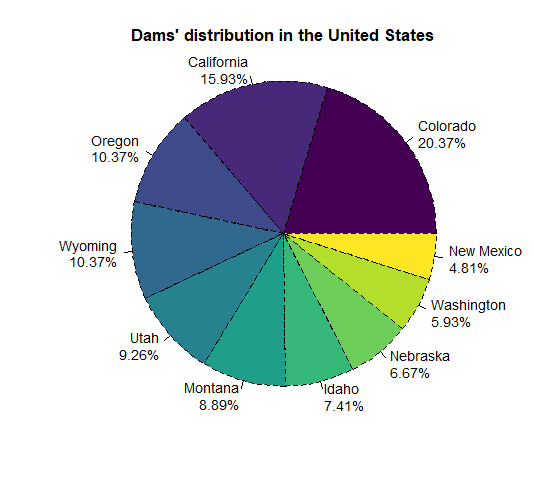
stateorder <- data.frame(stateorder)

propelev <- round(prop.table(stateorder$Freq)\*100,2)

heightprop <- paste(stateorder$Var1, "\n", propelev, "%", sep ="")

pie(stateorder$Freq, labels = heightprop, col = viridis(10), lty =2, radius = 1,

main = "Dams' distribution in the United States")



**Figure 7.1**

In the pie chart (Figure 7.1), I check the percent of dams in top 10 states. The pie chart also tells me that approximately 20% of dams is in Colorado.

~The End~