Week7RScript.R

Tyler House

2023-02-25

##Module # 7 Visual Analytics Assignment  
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

library(tidyverse)

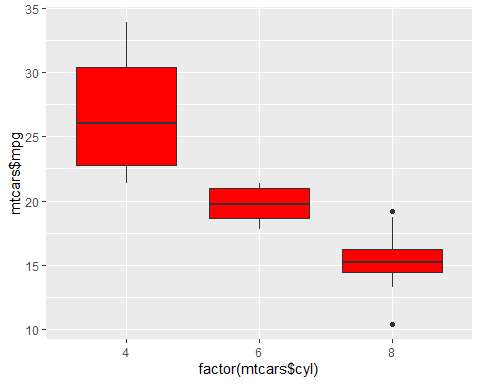
## ── Attaching packages  
## ───────────────────────────────────────  
## tidyverse 1.3.2 ──

## ✔ ggplot2 3.4.0 ✔ purrr 1.0.1  
## ✔ tibble 3.1.8 ✔ stringr 1.5.0  
## ✔ tidyr 1.3.0 ✔ forcats 0.5.2  
## ✔ readr 2.1.3   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

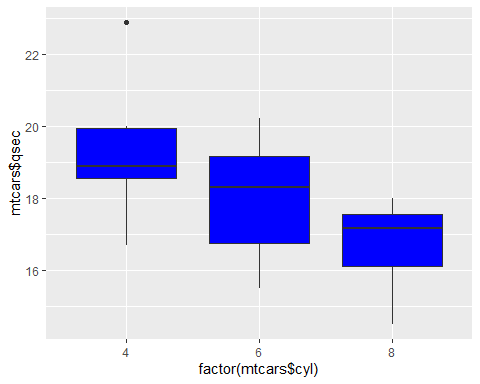
library(ggplot2)  
## Step 1: Load the dataset. I will be using the built in 'mtcars' dataset  
  
data(mtcars)  
mtcars

## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160.0 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160.0 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108.0 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258.0 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 8 360.0 175 3.15 3.440 17.02 0 0 3 2  
## Valiant 18.1 6 225.0 105 2.76 3.460 20.22 1 0 3 1  
## Duster 360 14.3 8 360.0 245 3.21 3.570 15.84 0 0 3 4  
## Merc 240D 24.4 4 146.7 62 3.69 3.190 20.00 1 0 4 2  
## Merc 230 22.8 4 140.8 95 3.92 3.150 22.90 1 0 4 2  
## Merc 280 19.2 6 167.6 123 3.92 3.440 18.30 1 0 4 4  
## Merc 280C 17.8 6 167.6 123 3.92 3.440 18.90 1 0 4 4  
## Merc 450SE 16.4 8 275.8 180 3.07 4.070 17.40 0 0 3 3  
## Merc 450SL 17.3 8 275.8 180 3.07 3.730 17.60 0 0 3 3  
## Merc 450SLC 15.2 8 275.8 180 3.07 3.780 18.00 0 0 3 3  
## Cadillac Fleetwood 10.4 8 472.0 205 2.93 5.250 17.98 0 0 3 4  
## Lincoln Continental 10.4 8 460.0 215 3.00 5.424 17.82 0 0 3 4  
## Chrysler Imperial 14.7 8 440.0 230 3.23 5.345 17.42 0 0 3 4  
## Fiat 128 32.4 4 78.7 66 4.08 2.200 19.47 1 1 4 1  
## Honda Civic 30.4 4 75.7 52 4.93 1.615 18.52 1 1 4 2  
## Toyota Corolla 33.9 4 71.1 65 4.22 1.835 19.90 1 1 4 1  
## Toyota Corona 21.5 4 120.1 97 3.70 2.465 20.01 1 0 3 1  
## Dodge Challenger 15.5 8 318.0 150 2.76 3.520 16.87 0 0 3 2  
## AMC Javelin 15.2 8 304.0 150 3.15 3.435 17.30 0 0 3 2  
## Camaro Z28 13.3 8 350.0 245 3.73 3.840 15.41 0 0 3 4  
## Pontiac Firebird 19.2 8 400.0 175 3.08 3.845 17.05 0 0 3 2  
## Fiat X1-9 27.3 4 79.0 66 4.08 1.935 18.90 1 1 4 1  
## Porsche 914-2 26.0 4 120.3 91 4.43 2.140 16.70 0 1 5 2  
## Lotus Europa 30.4 4 95.1 113 3.77 1.513 16.90 1 1 5 2  
## Ford Pantera L 15.8 8 351.0 264 4.22 3.170 14.50 0 1 5 4  
## Ferrari Dino 19.7 6 145.0 175 3.62 2.770 15.50 0 1 5 6  
## Maserati Bora 15.0 8 301.0 335 3.54 3.570 14.60 0 1 5 8  
## Volvo 142E 21.4 4 121.0 109 4.11 2.780 18.60 1 1 4 2

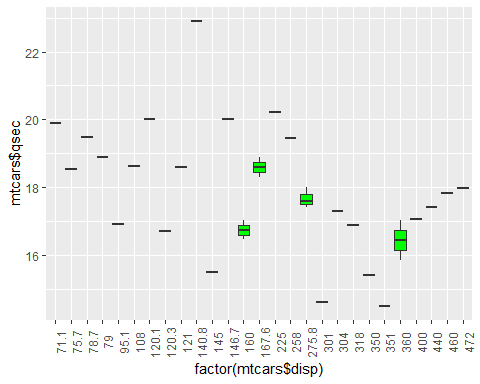
## Step 2: Let's generate some visuals. If we wanted to visualize the mpg compared to the engines layout, we can use ggplot  
## to better visualize. Remember to use factor() so that ggplot knows to seperate each visual to each value and the range of outputs  
  
ggplot(mtcars, aes(x = factor(mtcars$cyl), y = mtcars$mpg)) +  
 geom\_boxplot(fill = "red")



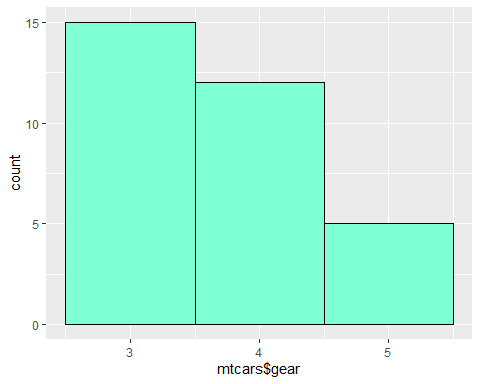
##Let's try something new, how about visualizing whether or not that the displacement and quarter mile times (qsec) are related  
## We might see instances of the old "power to weight ratio" concept on some of these cars. Some of the models, like Lincoln, Mercedes  
## and Hornet tend to be heavier, but with a larger engine. Some of the other brands like Datsun(Nissan) are lighter with boosted engines that are smaller  
  
ggplot(mtcars, aes(x = factor(mtcars$cyl), y = mtcars$qsec)) +  
 geom\_boxplot(fill = "blue")



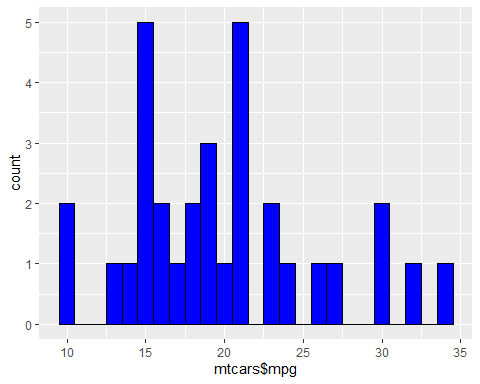
##Note: I had to use theme() to rotate the axial increments on the x axis because it looked pretty weird before but not anymore.  
  
ggplot(mtcars, aes(x = factor(mtcars$disp), y = mtcars$qsec)) +  
 geom\_boxplot(fill = "green") +  
 theme(axis.text.x = element\_text(angle = 90))



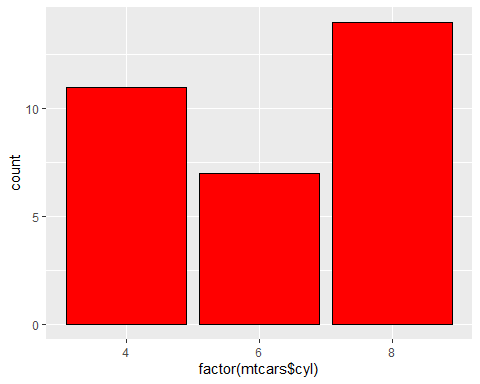
##Let's try something new, how about some histograms and piecharts?  
# This is a histogram on how many cars of speed gearbox there are  
#binwidth is going to be set at 1, and "color" will be black to define a better visual aesthetic  
ggplot(mtcars, aes(x = mtcars$gear)) +  
 geom\_histogram(fill = "aquamarine", color = "black", binwidth = 1)



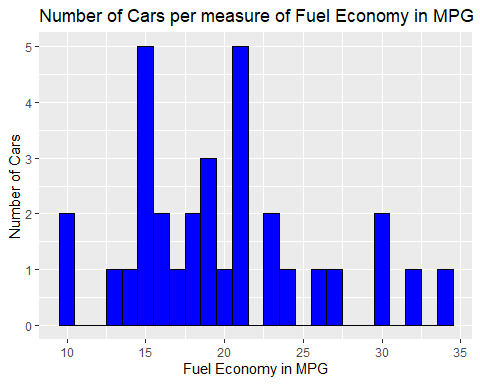
##Setting the binwidth is essential to accurately displying data  
#Note how my histograms values are centered on the x-axis increment, so that there are equal sized and ammounts of spaces that constitute for zero values  
  
ggplot(mtcars, aes(x = mtcars$mpg)) +  
 geom\_histogram(fill = "blue", color = "black", binwidth = 1)



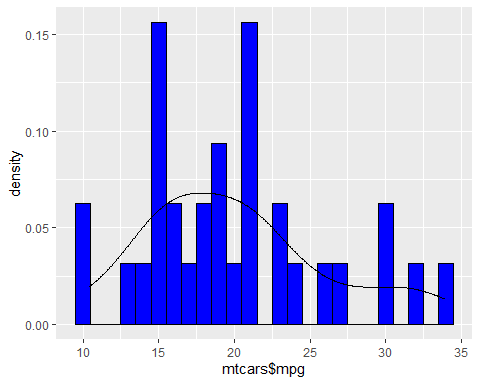
#You can also visualize using slightly different function calls to achieve the same effect. Earlier I called for a histogram about cylinder count, now I will call a barchart  
  
ggplot(mtcars, aes(x = factor(mtcars$cyl))) +  
 geom\_bar(fill = "red", color = "black")



#You can even add lines and points to pre-existing visuals to add more information in one place.  
  
##These are some basic visuals with ggplot, now lets introduce some Distribution Analysis concepts to these visuals  
#Edit and clean the visual and its text.  
#Lets recall histogram of fuel and edit using labs() with is in ggplot2  
  
ggplot(mtcars, aes(x = mtcars$mpg)) +  
 geom\_histogram(fill = "blue", color = "black", binwidth = 1) +  
 labs(  
 title = "Number of Cars per measure of Fuel Economy in MPG",  
 x = "Fuel Economy in MPG",  
 y = "Number of Cars",  
 )

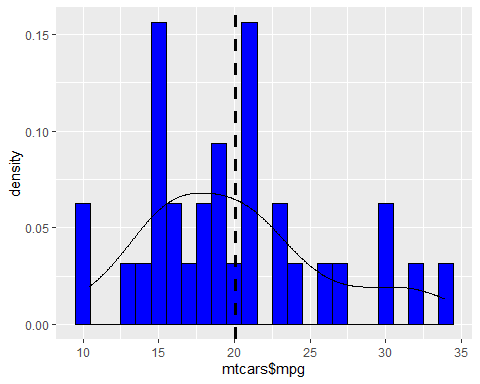


#Now that the chart is cleaned, lets add a distribution line or whats regarded as a kernel density curve in ggplot, to the visual  
##ggplot(mtcars, aes(x = mtcars$mpg)) +  
## geom\_histogram(fill = "blue", color = "black", binwidth = 1) +  
## labs(  
## title = "Number of Cars per measure of Fuel Economy in MPG",  
## x = "Fuel Economy in MPG",  
## y = "Number of Cars")   
#Above is the histogram we will use, it is commented out so we can build the distribustion curve.  
  
ggplot(mtcars, aes(x = mtcars$mpg)) +  
 geom\_histogram(aes(y = after\_stat(density)),  
 binwidth = 1,  
 color = "black",  
 fill = "blue") +  
 geom\_density(alpha = 0.1)

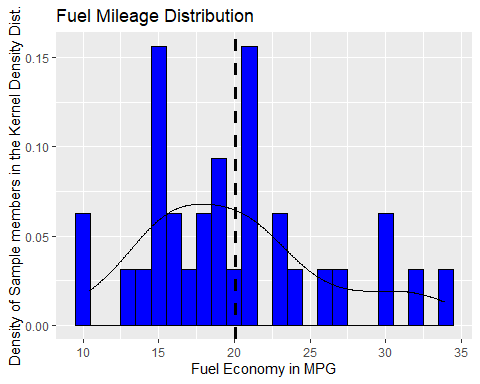


#For fun, we can add a dashed mean line to better visualize this density distribution  
ggplot(mtcars, aes(x = mtcars$mpg)) +  
 geom\_histogram(aes(y = after\_stat(density)),  
 binwidth = 1,  
 color = "black",  
 fill = "blue") +  
 geom\_density(alpha = 0.1) +  
 geom\_vline(aes(xintercept=mean(mtcars$mpg, na.rm = TRUE)),  
 color = "black",  
 linetype = "dashed",  
 size = 1.1)

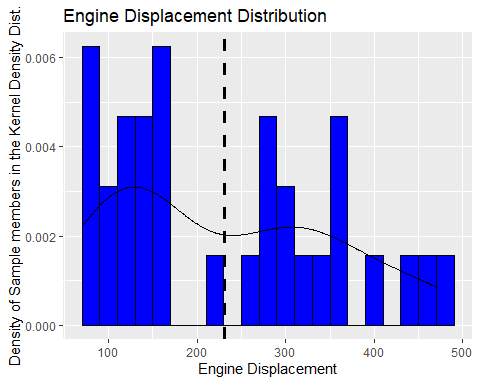
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.



## And dont forget some labels!!  
ggplot(mtcars, aes(x = mtcars$mpg)) +  
 geom\_histogram(aes(y = after\_stat(density)),  
 binwidth = 1,  
 color = "black",  
 fill = "blue") +  
 labs(  
 title = "Fuel Mileage Distribution",  
 x = "Fuel Economy in MPG",  
 y = "Density of Sample members in the Kernel Density Dist.",  
 ) +  
 geom\_density(alpha = 0.1) +  
 geom\_vline(aes(xintercept=mean(mtcars$mpg, na.rm = TRUE)),  
 color = "black",  
 linetype = "dashed",  
 size = 1.1)



##As we can see from the completed visual, the fuel economy in MPG is on the x-axis and the relative density is on the y-axis  
#It is important to remember that when conducting visualizations such as this which is a kind of overlay. that the y axis will change  
# to the density rather than the count of the ammount of cars per MPG. The mean is around 20 MPG with more members of the sample  
#being located below the mean, causing a positive skew and distribution that is more dense, and heavily populated on the low end  
  
#How could this be? Lets do the same for the engine type and find out   
ggplot(mtcars, aes(x = mtcars$disp)) +  
 geom\_histogram(aes(y = after\_stat(density)),  
 binwidth = 20,  
 color = "black",  
 fill = "blue") +  
 labs(  
 title = "Engine Displacement Distribution",  
 x = "Engine Displacement",  
 y = "Density of Sample members in the Kernel Density Dist.",  
 ) +  
 geom\_density(alpha = 0.1) +  
 geom\_vline(aes(xintercept=mean(mtcars$disp, na.rm = TRUE)),  
 color = "black",  
 linetype = "dashed",  
 size = 1.1)



##It can be shown in this distribution that, possible reasons for there being more cars of a higher fuel consumption would be because   
#Of the high average displacement of the engine. The mean is about 230 CI which is quite large and would have an average mpg of around 25-27 in real life  
#But it is because of having more high fuel consuming vehichles in the data, that the distribution is skewed to have a density centered lower.  
#In the Displacement Distribution, you will notice that a large portion of vehichles over the mean are well over 350 CI or 5.7 Litres  
#Engines like that tend to produce about 8-13 MPG in the city and 20 max on the highway on a cold day with cruise control.  
#Some inferential thought is required here to be able to make sense of why having data distributed in certain places may create differences in the end result, and these inferences  
#Might lie outside of the realm of R or Statistics.