##Note: Using the mpg data set (base in R) we will explore ggplot2 and what the added plotting features can do. Please remember to comment all your code and describe your findings in detail. This assignment is intended for you to work alone on, please do not use out side resources like code we do not cover in class or other students. Everything in this assignment is covered in week 3 and prior weeks. If you have any questions please reach out via email.

##Please Load packages and data:

library(ggplot2) # loads the ggplot2 package  
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

# loads the dplyr package   
data <- mpg   
# loads the data set mpg and assigns to data variable  
data

## # A tibble: 234 × 11  
## manufacturer model displ year cyl trans drv cty hwy fl class  
## <chr> <chr> <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>  
## 1 audi a4 1.8 1999 4 auto… f 18 29 p comp…  
## 2 audi a4 1.8 1999 4 manu… f 21 29 p comp…  
## 3 audi a4 2 2008 4 manu… f 20 31 p comp…  
## 4 audi a4 2 2008 4 auto… f 21 30 p comp…  
## 5 audi a4 2.8 1999 6 auto… f 16 26 p comp…  
## 6 audi a4 2.8 1999 6 manu… f 18 26 p comp…  
## 7 audi a4 3.1 2008 6 auto… f 18 27 p comp…  
## 8 audi a4 quattro 1.8 1999 4 manu… 4 18 26 p comp…  
## 9 audi a4 quattro 1.8 1999 4 auto… 4 16 25 p comp…  
## 10 audi a4 quattro 2 2008 4 manu… 4 20 28 p comp…  
## # ℹ 224 more rows

# prints the data set

1. Explore the data set, what are the variables?

# Load required library  
library(ggplot2)  
data <- mpg   
# Print the column names (variables) in the dataset  
print(names(data)) # Displays the names of all columns (variables)

## [1] "manufacturer" "model" "displ" "year" "cyl"   
## [6] "trans" "drv" "cty" "hwy" "fl"   
## [11] "class"

# Display the structure of the dataset (data types, sample values)  
str(data)

## tibble [234 × 11] (S3: tbl\_df/tbl/data.frame)  
## $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...  
## $ model : chr [1:234] "a4" "a4" "a4" "a4" ...  
## $ displ : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...  
## $ year : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...  
## $ cyl : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...  
## $ trans : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...  
## $ drv : chr [1:234] "f" "f" "f" "f" ...  
## $ cty : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...  
## $ hwy : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...  
## $ fl : chr [1:234] "p" "p" "p" "p" ...  
## $ class : chr [1:234] "compact" "compact" "compact" "compact" ...

# Display the first few rows of the dataset  
head(data)

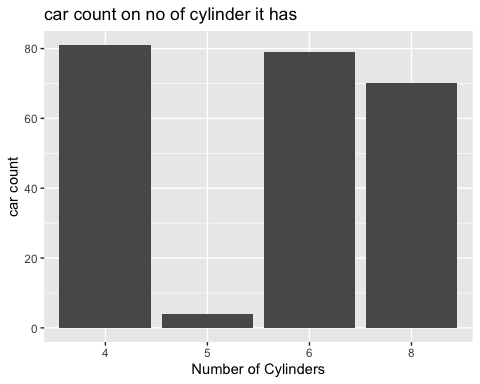
## # A tibble: 6 × 11  
## manufacturer model displ year cyl trans drv cty hwy fl class   
## <chr> <chr> <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>   
## 1 audi a4 1.8 1999 4 auto(l5) f 18 29 p compa…  
## 2 audi a4 1.8 1999 4 manual(m5) f 21 29 p compa…  
## 3 audi a4 2 2008 4 manual(m6) f 20 31 p compa…  
## 4 audi a4 2 2008 4 auto(av) f 21 30 p compa…  
## 5 audi a4 2.8 1999 6 auto(l5) f 16 26 p compa…  
## 6 audi a4 2.8 1999 6 manual(m5) f 18 26 p compa…

# Summary statistics of the dataset  
summary(data)

## manufacturer model displ year   
## Length:234 Length:234 Min. :1.600 Min. :1999   
## Class :character Class :character 1st Qu.:2.400 1st Qu.:1999   
## Mode :character Mode :character Median :3.300 Median :2004   
## Mean :3.472 Mean :2004   
## 3rd Qu.:4.600 3rd Qu.:2008   
## Max. :7.000 Max. :2008   
## cyl trans drv cty   
## Min. :4.000 Length:234 Length:234 Min. : 9.00   
## 1st Qu.:4.000 Class :character Class :character 1st Qu.:14.00   
## Median :6.000 Mode :character Mode :character Median :17.00   
## Mean :5.889 Mean :16.86   
## 3rd Qu.:8.000 3rd Qu.:19.00   
## Max. :8.000 Max. :35.00   
## hwy fl class   
## Min. :12.00 Length:234 Length:234   
## 1st Qu.:18.00 Class :character Class :character   
## Median :24.00 Mode :character Mode :character   
## Mean :23.44   
## 3rd Qu.:27.00   
## Max. :44.00

1. Create a bar chart of ‘cyl’ count for the data set.

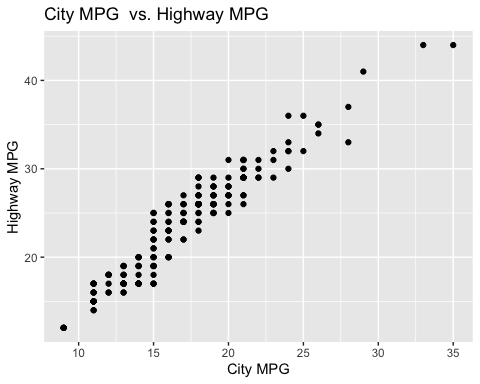
ggplot(data, aes(x = factor(cyl))) +   
 # dataset as data and aes mapping to "cyl" as x axis  
 geom\_bar() + # plots the bar graph  
 labs(x = "Number of Cylinders", y = "car count", title = "car count on no of cylinder it has")



# main title is added and the description for X and Y axis

1. Plot two variables (cty vs hwy) fuel economy as points, do we see a pattern?

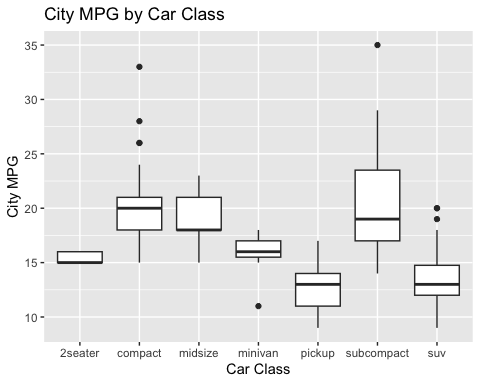
ggplot(data, aes(x = cty, y = hwy)) +   
 # aes() mapping "cty" for x-axis and "hwy" for y- axis   
 geom\_point() + # creates scatter plot   
 labs(x = "City MPG", # x-axis label  
 y = "Highway MPG", # y-axis label  
 title = "City MPG vs. Highway MPG") # title for chart



# As city fuel economy increases, highway fuel economy also increases.  
#the cars that provides good mileage in the city also gives good mileage on the highway.  
#Some variation may exist due to vehicle types and driving conditions.

1. Build a Box and wisker plot of the car ‘class’ and cty fuel economy

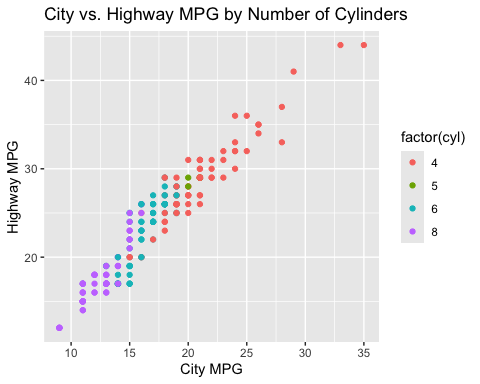
ggplot(data, aes(x = class, y = cty)) +  
 # mapping "class" on x and "cty" on y axes  
 geom\_boxplot() +   
 # boxplot is plotted  
 labs(x = "Car Class", # x-axis label  
 y = "City MPG", # y-axis label  
 title = "City MPG by Car Class") # title for graph



#Compact and subcompact cars tend to have higher city MPG.  
# SUVs and pickups have lower city MPG, showing less fuel efficiency

1. Lets plot cty vs hwy again but now include the class “cyl” and color code it by number of cyl

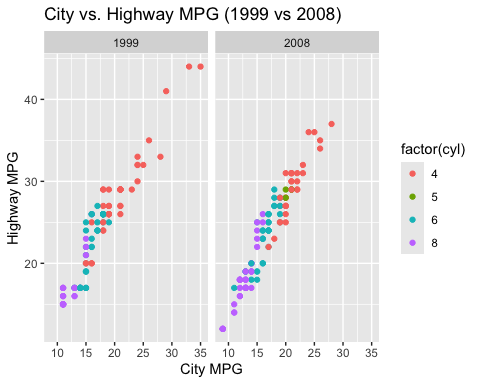
ggplot(data, aes(x = cty, y = hwy, color = factor(cyl))) +   
 # plotting cty for x axis, hwy for y axis and color as cyl)  
 geom\_point() + # adds points  
 labs(x = "City MPG", # x-axis label  
 y = "Highway MPG", # y-axis label  
 title = "City vs. Highway MPG by Number of Cylinders") # title of the chart



# the color variation by the number of cylinders explains that cars with fewer normally have higher efficiency in both city and highway driving situations.

1. Let’s build on chart 5, using ‘facet\_wrap()’ lets compare 1999 to 2008.

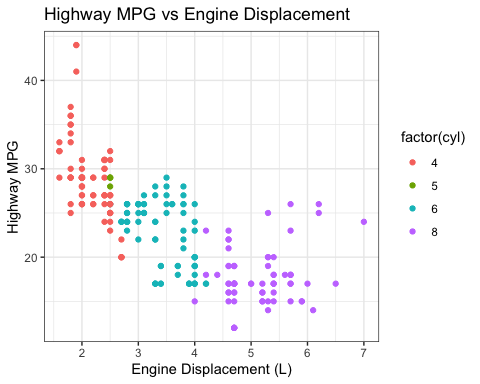
ggplot(data, aes(x = cty, y = hwy)) +   
 geom\_point(aes(color = factor(cyl))) +  
 # specifying color to factor cyl  
 facet\_wrap(~year) +   
 # facet\_wrap() compares year  
 labs(x = "City MPG", # x-axis label  
 y = "Highway MPG", # y-axis label   
 title = "City vs. Highway MPG (1999 vs 2008)") # title for the graph



# It compares the effectiveness of cars on city and highways. It is better in 2008 compared to 1999 maybe because of advancement of technology.

1. Come up with a chat by ploting two or more variables and tell me what it shows us. This can be a point plot, box, histogram, or another plot we covered.

ggplot(data, aes(x = displ, y = hwy)) +   
 geom\_point(aes(color = factor(cyl))) +   
 # specifying color to factor cyl  
 labs(x = "Engine Displacement (L)", # x-axis label  
 y = "Highway MPG", # y-axis label   
 title = "Highway MPG vs Engine Displacement") + # title for the graph  
  
 theme\_bw()



#we can observe that larger engines tend to have lower highway MPG, while smaller engines generally perform better on the highway in terms of fuel efficiency. The points are clustered to show this inverse relationship.

##Part 2: Here we will use the data set ‘gapminder’ data set. This data set has basic country information.

library(gapminder)  
head(gapminder)

## # A tibble: 6 × 6  
## country continent year lifeExp pop gdpPercap  
## <fct> <fct> <int> <dbl> <int> <dbl>  
## 1 Afghanistan Asia 1952 28.8 8425333 779.  
## 2 Afghanistan Asia 1957 30.3 9240934 821.  
## 3 Afghanistan Asia 1962 32.0 10267083 853.  
## 4 Afghanistan Asia 1967 34.0 11537966 836.  
## 5 Afghanistan Asia 1972 36.1 13079460 740.  
## 6 Afghanistan Asia 1977 38.4 14880372 786.

1. Explore the data and look at the variables listed, what are they?

# Print the column names (variables) in the dataset  
print(names(gapminder)) # Displays the names of all columns (variables)

## [1] "country" "continent" "year" "lifeExp" "pop" "gdpPercap"

# Display the first few rows of the dataset  
head(gapminder)

## # A tibble: 6 × 6  
## country continent year lifeExp pop gdpPercap  
## <fct> <fct> <int> <dbl> <int> <dbl>  
## 1 Afghanistan Asia 1952 28.8 8425333 779.  
## 2 Afghanistan Asia 1957 30.3 9240934 821.  
## 3 Afghanistan Asia 1962 32.0 10267083 853.  
## 4 Afghanistan Asia 1967 34.0 11537966 836.  
## 5 Afghanistan Asia 1972 36.1 13079460 740.  
## 6 Afghanistan Asia 1977 38.4 14880372 786.

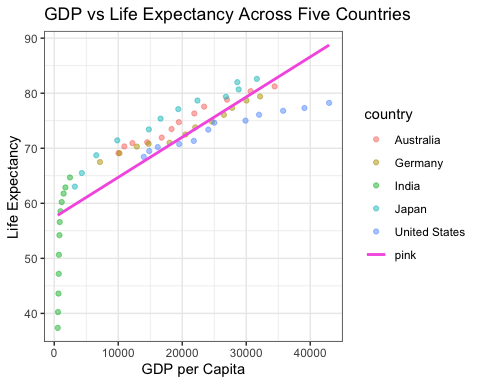
# Summary statistics of the dataset  
summary(gapminder)

## country continent year lifeExp   
## Afghanistan: 12 Africa :624 Min. :1952 Min. :23.60   
## Albania : 12 Americas:300 1st Qu.:1966 1st Qu.:48.20   
## Algeria : 12 Asia :396 Median :1980 Median :60.71   
## Angola : 12 Europe :360 Mean :1980 Mean :59.47   
## Argentina : 12 Oceania : 24 3rd Qu.:1993 3rd Qu.:70.85   
## Australia : 12 Max. :2007 Max. :82.60   
## (Other) :1632   
## pop gdpPercap   
## Min. :6.001e+04 Min. : 241.2   
## 1st Qu.:2.794e+06 1st Qu.: 1202.1   
## Median :7.024e+06 Median : 3531.8   
## Mean :2.960e+07 Mean : 7215.3   
## 3rd Qu.:1.959e+07 3rd Qu.: 9325.5   
## Max. :1.319e+09 Max. :113523.1   
##

1. Create a scatter plot of gdp v life expectancy for five (you pick) countries over a period of years. Plot these contries on the same chart. Show each country in a different color. Add a trend line (hint lm) for the set as a whole. Explain what you see.

countries <- gapminder %>% # data set to filter and pipe  
 filter(country %in% c("Japan", "India", "United States", "Australia", "Germany"))  
 #Filter function filters countries from c vector provided   
#Creating the scatter plot with filtered countries  
ggplot(countries, aes(x = gdpPercap, y = lifeExp, color = country)) + # aes() with x and y axes  
 geom\_point(alpha = 0.5) + # to see overlap used alpha (0.5) for semi-transparent points  
 geom\_smooth(method = "lm", se = FALSE, aes(group = 1, colour = "pink")) +   
 # using linear model line with color yellow and group = 1 for trend  
 labs(x = "GDP per Capita", y = "Life Expectancy", title = "GDP vs Life Expectancy Across Five Countries") +  
 # labels for x and y axes with title  
 theme\_bw()

## `geom\_smooth()` using formula = 'y ~ x'



#for good looking  
#The trend line indicates a positive correlation between GDP and life expectancy, meaning that, in general, countries with higher GDP tend to have higher life expectancy. However, the slope of the trend line can vary, reflecting different economic growth patterns in these countries.  
# There is a positive relationship involving GDP per capita and life expectation which indicates the country with higher gdp has extended life probability. As well, there is a substantial sum of flexibility in the plot which says other factors can be affecting the correlation like weather, lifestyle, diet, etc.

1. (Not graded) What was one thing you found interesting this week? What would you like changed?