> demo\_table3 <- read.csv('demo2.csv',check.names = F,stringsAsFactors = F)

Error in file(file, "rt") : cannot open the connection

In addition: Warning message:

In file(file, "rt") :

cannot open file 'demo2.csv': No such file or directory

> setwd("~/Desktop/Data Analysis Class/Module 15/R\_Analysis")

> demo\_table3 <- read.csv('demo2.csv',check.names = F,stringsAsFactors = F)

> View(demo\_table3)

> long\_table <- gather(demo\_table3,key="Metric",value="Score",buying\_price:popularity)

> View(long\_table)

> long\_table <- demo\_table3 %>% gather(key="Metric",value="Score",buying\_price:popularity)

> View(long\_table)

> ?spread()

> wide\_table <- long\_table %>% spread(key="Metric",value="Score")

> View(wide\_table)

> all.equal(demo\_table3, wide\_table)

[1] "Names: 7 string mismatches"

[2] "Component 2: Mean relative difference: 0.5808824"

[3] "Component 3: Mean relative difference: 0.8276762"

[4] "Component 4: Mean relative difference: 0.5555556"

[5] "Component 5: Mean relative difference: 0.4978166"

[6] "Component 6: Mean relative difference: 0.4863636"

[7] "Component 7: Mean relative difference: 1.818182"

[8] "Component 8: Mean relative difference: 0.5152355"

> setwd("~/Desktop/Data Analysis Class/Module 15/R\_Analysis/01\_Demo")

> ?ggplot()

> head(mpg)

# A tibble: 6 x 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 manua… f 21 29 p comp…

3 audi a4 2 2008 4 manua… 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 manua… f 18 26 p comp…

> plt <- ggplot(mpg,aes(x=class)) #import dataset into ggplot2

> View(plt)

> plt + geom\_bar() #plot a bar plot

> ?geom\_bar()

> mpg\_summary <- mpg %>% group\_by(manufacturer) %>% summarize(Vehicle\_Count=n()) #create summary table

`summarise()` ungrouping output (override with `.groups` argument)

> View(mpg\_summary)

> plt <- ggplot(mpg\_summary,aes(x=manufacturer,y=Vehicle\_Count)) #import dataset into ggplot2

> plt + geom\_col() #plot a bar plot

> plt + geom\_col() + xlab("Manufacturing Company") + ylab("Number of Vehicles in Dataset") #plot bar plot with labels

> plt + geom\_col() + xlab("Manufacturing Company") + ylab("Number of Vehicles in Dataset") + #plot a boxplot with labels

+ theme(axis.text.x=element\_text(angle=45,hjust=1)) #rotate the x-axis label 45 degrees

> mpg\_summary <- subset(mpg,manufacturer=="toyota") %>% group\_by(cyl) %>% summarize(Mean\_Hwy=mean(hwy)) #create summary table

`summarise()` ungrouping output (override with `.groups` argument)

> plt <- ggplot(mpg\_summary,aes(x=cyl,y=Mean\_Hwy)) #import dataset into ggplot2

> plt + geom\_line()

> plt + geom\_line() + scale\_x\_discrete(limits=c(4,6,8)) + scale\_y\_continuous(breaks = c(15:30)) #add line plot with labels

Warning message:

Continuous limits supplied to discrete scale.

Did you mean `limits = factor(...)` or `scale\_\*\_continuous()`?

> plt <- ggplot(mpg,aes(x=displ,y=cty)) #import dataset into ggplot2

> plt + geom\_point() + xlab("Engine Size (L)") + ylab("City Fuel-Efficiency (MPG)") #add scatter plot with labels

> plt <- ggplot(mpg,aes(x=displ,y=cty,color=class)) #import dataset into ggplot2

> plt + geom\_point() + labs(x="Engine Size (L)", y="City Fuel-Efficiency (MPG)", color="Vehicle Class") #add scatter plot with labels

> plt <- ggplot(mpg,aes(x=displ,y=cty,color=class,shape=drv)) #import dataset into ggplot2

> plt + geom\_point() + labs(x="Engine Size (L)", y="City Fuel-Efficiency (MPG)", color="Vehicle Class",shape="Type of Drive") #add scatter plot with multiple aesthetics

> View(summarize\_demo)

> View(plt)

> View(plt)

> View(long\_table)

> View(summarize\_demo)

> View(mpg\_summary)

> View(wide\_table)

> View(summarize\_demo)

> View(mpg\_summary)

> View(plt)

> View(long\_table)

> View(demo\_table3)

> View(summarize\_demo)

> View(mpg\_summary)

> View(mpg\_summary)

> View(plt)

> View(plt)

> plt <- ggplot(mpg,aes(x=year,y=cty)) #import dataset into ggplot2

> plt + geom\_point() + xlab("year") + ylab("City Fuel-Efficiency (MPG)") #add scatter plot with labels

> plt <- ggplot(mpg\_summary,aes(x=Mean\_Hwy,y=cty)) #import dataset into ggplot2

> plt <- ggplot(mpg,aes(x=displ,y=cty,color=class,shape=model)) #import dataset into ggplot2

> plt + geom\_point() + labs(x="Engine Size (L)", y="City Fuel-Efficiency (MPG)", color="Vehicle Class",shape="Model") #add scatter plot with multiple aesthetics

Warning messages:

1: The shape palette can deal with a maximum of 6 discrete values because more than

6 becomes difficult to discriminate; you have 38. Consider specifying shapes

manually if you must have them.

2: Removed 199 rows containing missing values (geom\_point).

> plt <- ggplot(mpg,aes(x=displ,y=cty,color=class,shape=year)) #import dataset into ggplot2

> plt + geom\_point() + labs(x="Engine Size (L)", y="City Fuel-Efficiency (MPG)", color="Vehicle Class",shape="Year") #add scatter plot with multiple aesthetics

Error: A continuous variable can not be mapped to shape

Run `rlang::last\_error()` to see where the error occurred.

> plt <- ggplot(mpg,aes(x=displ,y=cty,color=class,shape=trans)) #import dataset into ggplot2

> plt + geom\_point() + labs(x="Engine Size (L)", y="City Fuel-Efficiency (MPG)", color="Vehicle Class",shape="Transmission") #add scatter plot with multiple aesthetics

Warning messages:

1: The shape palette can deal with a maximum of 6 discrete values because more than

6 becomes difficult to discriminate; you have 10. Consider specifying shapes

manually if you must have them.

2: Removed 96 rows containing missing values (geom\_point).

> plt <- ggplot(mpg,aes(y=hwy)) #import dataset into ggplot2

> plt + geom\_boxplot() #add boxplot

> plt <- ggplot(mpg,aes(x=manufacturer,y=hwy)) #import dataset into ggplot2

> plt + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,hjust=1)) #add boxplot and rotate x-axis labels 45 degrees

> plt <- ggplot(mpg,aes(x=manufacturer,y=hwy,fill="teal", color="gray")) #import dataset into ggplot2

> plt + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,hjust=1)) #add boxplot and rotate x-axis labels 45 degrees

> mpg\_summary <- mpg %>% group\_by(class,year) %>% summarize(Mean\_Hwy=mean(hwy)) #create summary table

`summarise()` regrouping output by 'class' (override with `.groups` argument)

> plt <- ggplot(mpg\_summary, aes(x=class,y=factor(year),fill=Mean\_Hwy))

> plt + geom\_tile() + labs(x="Vehicle Class",y="Vehicle

+ Year",fill="Mean Highway (MPG)") #create heatmap with labels

> mpg\_summary <- mpg %>% group\_by(model,year) %>% summarize(Mean\_Hwy=mean(hwy)) #create summary table

`summarise()` regrouping output by 'model' (override with `.groups` argument)

> plt <- ggplot(mpg\_summary, aes(x=model,y=factor(year),fill=Mean\_Hwy)) #import dataset into ggplot2

> plt + geom\_tile() + labs(x="Model",y="Vehicle Year",fill="Mean Highway (MPG)") + #add heatmap with labels > theme(axis.text.x = element\_text(angle=90,hjust=1,vjust=.5)) #rotate x-axis labels 90 degrees

+

+ mpg\_summary <- mpg %>% group\_by(class,year) %>% summarize(Mean\_Hwy=mean(hwy)) #create summary table

`summarise()` regrouping output by 'class' (override with `.groups` argument)

Error in plt + geom\_tile() + labs(x = "Model", y = "Vehicle Year", fill = "Mean Highway (MPG)") + :

could not find function "+<-"

> plt + geom\_tile() + labs(x="Model",y="Vehicle Year",fill="Mean Highway (MPG)") + #add heatmap with labels > theme(axis.text.x = element\_text(angle=90,hjust=1,vjust=.5)) #rotate x-axis labels 90 degrees

+ plt + geom\_tile() + labs(x="Model",y="Vehicle Year",fill="Mean Highway (MPG)") + #add heatmap with labels > theme(axis.text.x = element\_text(angle=90,hjust=1,vjust=.5)) #rotate x-axis labels 90 degrees

+

+ plt + geom\_tile() + labs(x="Model",y="Vehicle Year",fill="Mean Highway (MPG)") + #add heatmap with labels > theme(axis.text.x = element\_text(angle=90,hjust=1,vjust=.5)) #rotate x-axis labels 90 degrees

+ mpg\_summary <- mpg %>% group\_by(class,year) %>% summarize(Mean\_Hwy=mean(hwy)) #create summary table

`summarise()` regrouping output by 'class' (override with `.groups` argument)

Error: Can't add `plt` to a ggplot object.

Run `rlang::last\_error()` to see where the error occurred.

> plt + geom\_tile() + labs(x="Model",y="Vehicle Year",fill="Mean Highway (MPG)") + #add heatmap with labels

+ theme(axis.text.x = element\_text(angle=90,hjust=1,vjust=.5)) #rotate x-axis labels 90 degrees

>

> plt <- ggplot(mpg,aes(x=manufacturer,y=hwy)) #import dataset into ggplot2

> plt + geom\_boxplot() + #add boxplot\

+ ss

Error: object 'ss' not found

> plt + geom\_boxplot() + #add boxplot

+ theme(axis.text.x=element\_text(angle=45,hjust=1)) + #rotate x-axis labels 45 degrees

+ geom\_point() #overlay scatter plot on top

> mpg\_summary <- mpg %>% group\_by(class) %>% summarize(Mean\_Engine=mean(displ)) #create summary table

`summarise()` ungrouping output (override with `.groups` argument)

> plt <- ggplot(mpg\_summary,aes(x=class,y=Mean\_Engine)) #import dataset into ggplot2

> plt + geom\_point(size=4) + labs(x="Vehicle Class",y="Mean Engine Size") #add scatter plot

> mpg\_summary <- mpg %>% group\_by(class) %>% summarize(Mean\_Engine=mean(displ),SD\_Engine=sd(displ))

`summarise()` ungrouping output (override with `.groups` argument)

> plt <- ggplot(mpg\_summary,aes(x=class,y=Mean\_Engine)) #import dataset into ggplot2

> plt + geom\_point(size=4) + labs(x="Vehicle Class",y="Mean Engine Size") + #add scatter plot with labels

+ geom\_errorbar(aes(ymin=Mean\_Engine-SD\_Engine,ymax=Mean\_Engine+SD\_Engine)) #overlay with error bars

> mpg\_long <- mpg %>% gather(key="MPG\_Type",value="Rating",c(cty,hwy)) #convert to long format

> head(mpg\_long)

# A tibble: 6 x 11

manufacturer model displ year cyl trans drv fl class MPG\_Type

<chr> <chr> <dbl> <int> <int> <chr> <chr> <chr> <chr> <chr>

1 audi a4 1.8 1999 4 auto… f p comp… cty

2 audi a4 1.8 1999 4 manu… f p comp… cty

3 audi a4 2 2008 4 manu… f p comp… cty

4 audi a4 2 2008 4 auto… f p comp… cty

5 audi a4 2.8 1999 6 auto… f p comp… cty

6 audi a4 2.8 1999 6 manu… f p comp… cty

# … with 1 more variable: Rating <int>

> plt <- ggplot(mpg\_long,aes(x=manufacturer,y=Rating,color=MPG\_Type))

> plt + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,hjust=1)) #add boxplot with labels rotated 45 degrees

> ?facet\_wrap()

> plt <- ggplot(mpg\_long,aes(x=manufacturer,y=Rating,color=MPG\_Type)) #import dataset into ggplot2

> plt + geom\_boxplot() + facet\_wrap(vars(MPG\_Type)) + #create multiple boxplots, one for each MPG type

+ theme(axis.text.x=element\_text(angle=45,hjust=1),legend.position = "none") + xlab("Manufacturer") #rotate x-axis labels

> plt <- ggplot(mpg\_long,aes(x=manufacturer,y=Rating,color=class)) #import dataset into ggplot2

> plt + geom\_boxplot() + facet\_wrap(vars(class)) + #create multiple boxplots, one for each MPG type

+ theme(axis.text.x=element\_text(angle=45,hjust=1),legend.position = "none") + xlab("Manufacturer") #rotate x-axis labels

> ggplot(mtcars,aes(x=wt)) + geom\_density() #visualize distribution using density plot

> ?shapiro.test()

> shapiro.test(mtcars$wt)

Shapiro-Wilk normality test

data: mtcars$wt

W = 0.94326, p-value = 0.09265

> ?sample\_n()

> population\_table <- read.csv('used\_car\_data.csv',check.names = F,stringsAsFactors = F) #import used car dataset

> plt <- ggplot(population\_table,aes(x=log10(Miles\_Driven))) #import dataset into ggplot2

> plt + geom\_density() #visualize distribution using density plot

> sample\_table <- population\_table %>% sample\_n(50) #randomly sample 50 data points

> plt <- ggplot(sample\_table,aes(x=log10(Miles\_Driven))) #import dataset into ggplot2

> plt + geom\_density() #visualize distribution using density plot

> ?t.test()

> t.test(log10(sample\_table$Miles\_Driven),mu=mean(log10(population\_table$Miles\_Driven))) #compare sample versus population means

One Sample t-test

data: log10(sample\_table$Miles\_Driven)

t = -1.0012, df = 49, p-value = 0.3216

alternative hypothesis: true mean is not equal to 4.39449

95 percent confidence interval:

4.189671 4.463085

sample estimates:

mean of x

4.326378

> sample\_table <- population\_table %>% sample\_n(50) #generate 50 randomly sampled data points

> sample\_table2 <- population\_table %>% sample\_n(50) #generate another 50 randomly sampled data points

> View(sample\_table)

> View(sample\_table2)

> t.test(log10(sample\_table$Miles\_Driven),log10(sample\_table2$Miles\_Driven)) #compare means of two samples

Welch Two Sample t-test

data: log10(sample\_table$Miles\_Driven) and log10(sample\_table2$Miles\_Driven)

t = 0.021351, df = 96.82, p-value = 0.983

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.1778096 0.1816768

sample estimates:

mean of x mean of y

4.395503 4.393569

> ?t.test()

> mpg\_data <- read.csv('mpg\_modified.csv') #import dataset

> mpg\_1999 <- mpg\_data %>% filter(year==1999) #select only data points where the year is 1999

> View(mpg\_1999)

> View(mpg\_data)

> View(mpg\_1999)

> mpg\_2008 <- mpg\_data %>% filter(year==2008) #select only data points where the year is 2008

> t.test(mpg\_1999$hwy,mpg\_2008$hwy,paired = T) #compare the mean difference between two samples

Paired t-test

data: mpg\_1999$hwy and mpg\_2008$hwy

t = -1.1165, df = 37, p-value = 0.2714

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-2.1480860 0.6217702

sample estimates:

mean of the differences

-0.7631579

> ?aov()

> mtcars\_filt <- mtcars[,c("hp","cyl")] #filter columns from mtcars dataset

> View(mtcars\_filt)

> mtcars\_filt$cyl <- factor(mtcars\_filt$cyl) #convert numeric column to factor

> View(mtcars\_filt)

> aov(hp ~ cyl,data=mtcars\_filt) #compare means across multiple levels

Call:

aov(formula = hp ~ cyl, data = mtcars\_filt)

Terms:

cyl Residuals

Sum of Squares 104030.54 41696.33

Deg. of Freedom 2 29

Residual standard error: 37.91839

Estimated effects may be unbalanced

> summary(aov(hp ~ cyl,data=mtcars\_filt))

Df Sum Sq Mean Sq F value Pr(>F)

cyl 2 104031 52015 36.18 1.32e-08 \*\*\*

Residuals 29 41696 1438

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> ?cor()

> head(mtcars)

mpg cyl disp hp drat wt qsec vs am gear carb

Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4

Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4

Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1

Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1

Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2

Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

> plt <- ggplot(mtcars,aes(x=hp,y=qsec)) #import dataset into ggplot2

> View(plt)

> View(plt)

> plt + geom\_point() #create scatter plot

> cor(mtcars$hp,mtcars$qsec) #calculate correlation coefficient

[1] -0.7082234

> used\_cars <- read.csv('used\_car\_data.csv',stringsAsFactors = F) #read in dataset

> head(used\_cars)

Car\_Name Year Selling\_Price Present\_Price Miles\_Driven Fuel\_Type

1 ritz 2014 3350 5590 27000 Petrol

2 sx4 2013 4750 9540 43000 Diesel

3 ciaz 2017 7250 9850 6900 Petrol

4 wagon r 2011 2850 4150 5200 Petrol

5 swift 2014 4600 6870 42450 Diesel

6 vitara brezza 2018 9250 9830 2071 Diesel

Seller\_Type Transmission Owner

1 Dealer Manual 0

2 Dealer Manual 0

3 Dealer Manual 0

4 Dealer Manual 0

5 Dealer Manual 0

6 Dealer Manual 0

> plt <- ggplot(used\_cars,aes(x=Miles\_Driven,y=Selling\_Price)) #import dataset into ggplot2

> plt + geom\_point() #create a scatter plot

> cor(used\_cars$Miles\_Driven,used\_cars$Selling\_Price) #calculate correlation coefficient

[1] 0.02918709

> used\_matrix <- as.matrix(used\_cars[,c("Selling\_Price","Present\_Price","Miles\_Driven")]) #convert data frame into numeric matrix

> View(used\_matrix)

> cor(used\_matrix)

Selling\_Price Present\_Price Miles\_Driven

Selling\_Price 1.00000000 0.8789825 0.02918709

Present\_Price 0.87898255 1.0000000 0.20364703

Miles\_Driven 0.02918709 0.2036470 1.00000000

> ?lm()

> lm(qsec ~ hp,mtcars) #create linear model

Call:

lm(formula = qsec ~ hp, data = mtcars)

Coefficients:

(Intercept) hp

20.55635 -0.01846

> summary(lm(qsec~hp,mtcars)) #summarize linear model

Call:

lm(formula = qsec ~ hp, data = mtcars)

Residuals:

Min 1Q Median 3Q Max

-2.1766 -0.6975 0.0348 0.6520 4.0972

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 20.556354 0.542424 37.897 < 2e-16 \*\*\*

hp -0.018458 0.003359 -5.495 5.77e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.282 on 30 degrees of freedom

Multiple R-squared: 0.5016, Adjusted R-squared: 0.485

F-statistic: 30.19 on 1 and 30 DF, p-value: 5.766e-06

> model <- lm(qsec ~ hp,mtcars) #create linear model

> View(model)

> yvals <- model$coefficients['hp']\*mtcars$hp +

+ model$coefficients['(Intercept)'] #determine y-axis values from linear model

> plt <- ggplot(mtcars,aes(x=hp,y=qsec)) #import dataset into ggplot2

> plt + geom\_point() + geom\_line(aes(y=yvals), color = "red") #plot scatter and linear model

> lm(qsec ~ mpg + disp + drat + wt + hp,data=mtcars) #generate multiple linear regression model

Call:

lm(formula = qsec ~ mpg + disp + drat + wt + hp, data = mtcars)

Coefficients:

(Intercept) mpg disp drat wt

16.541651 0.108579 -0.008076 -0.578953 1.792793

hp

-0.018383

> summary(lm(qsec ~ mpg + disp + drat + wt + hp,data=mtcars)) #generate summary statistics

Call:

lm(formula = qsec ~ mpg + disp + drat + wt + hp, data = mtcars)

Residuals:

Min 1Q Median 3Q Max

-1.6628 -0.6138 0.0706 0.4087 3.3885

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 16.541651 3.413109 4.847 5.04e-05 \*\*\*

mpg 0.108579 0.077911 1.394 0.17523

disp -0.008076 0.004384 -1.842 0.07689 .

drat -0.578953 0.551771 -1.049 0.30371

wt 1.792793 0.513897 3.489 0.00175 \*\*

hp -0.018383 0.005421 -3.391 0.00223 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.053 on 26 degrees of freedom

Multiple R-squared: 0.7085, Adjusted R-squared: 0.6524

F-statistic: 12.64 on 5 and 26 DF, p-value: 2.767e-06

> ?chisq.test()

> table(mpg$class,mpg$year) #generate contingency table

1999 2008

2seater 2 3

compact 25 22

midsize 20 21

minivan 6 5

pickup 16 17

subcompact 19 16

suv 29 33

> tbl <- table(mpg$class,mpg$year) #generate contingency table

> chisq.test(tbl) #compare categorical distributions

Pearson's Chi-squared test

data: tbl

X-squared = 1.0523, df = 6, p-value = 0.9836

Warning message:

In chisq.test(tbl) : Chi-squared approximation may be incorrect

> setwd("~/Desktop/Data Analysis Class/Module 15/R\_Analysis")

>