Assignment 8

Grisham Nathan

November 15, 2017

# 11

# a)

library(tidyverse)

## Loading tidyverse: ggplot2  
## Loading tidyverse: tibble  
## Loading tidyverse: tidyr  
## Loading tidyverse: readr  
## Loading tidyverse: purrr  
## Loading tidyverse: dplyr

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats  
## lag(): dplyr, stats

url="http://stat405.had.co.nz/data/billboard.csv"  
billboard=read\_csv(url)

## Parsed with column specification:  
## cols(  
## .default = col\_integer(),  
## artist.inverted = col\_character(),  
## track = col\_character(),  
## time = col\_time(format = ""),  
## genre = col\_character(),  
## date.entered = col\_date(format = ""),  
## date.peaked = col\_date(format = ""),  
## x66th.week = col\_character(),  
## x67th.week = col\_character(),  
## x68th.week = col\_character(),  
## x69th.week = col\_character(),  
## x70th.week = col\_character(),  
## x71st.week = col\_character(),  
## x72nd.week = col\_character(),  
## x73rd.week = col\_character(),  
## x74th.week = col\_character(),  
## x75th.week = col\_character(),  
## x76th.week = col\_character()  
## )

## See spec(...) for full column specifications.

billboard

## # A tibble: 317 x 83  
## year artist.inverted track  
## <int> <chr> <chr>  
## 1 2000 Destiny's Child Independent Women Part I  
## 2 2000 Santana Maria, Maria  
## 3 2000 Savage Garden I Knew I Loved You  
## 4 2000 Madonna Music  
## 5 2000 Aguilera, Christina Come On Over Baby (All I Want Is You)  
## 6 2000 Janet Doesn't Really Matter  
## 7 2000 Destiny's Child Say My Name  
## 8 2000 Iglesias, Enrique Be With You  
## 9 2000 Sisqo Incomplete  
## 10 2000 Lonestar Amazed  
## # ... with 307 more rows, and 80 more variables: time <time>, genre <chr>,  
## # date.entered <date>, date.peaked <date>, x1st.week <int>,  
## # x2nd.week <int>, x3rd.week <int>, x4th.week <int>, x5th.week <int>,  
## # x6th.week <int>, x7th.week <int>, x8th.week <int>, x9th.week <int>,  
## # x10th.week <int>, x11th.week <int>, x12th.week <int>,  
## # x13th.week <int>, x14th.week <int>, x15th.week <int>,  
## # x16th.week <int>, x17th.week <int>, x18th.week <int>,  
## # x19th.week <int>, x20th.week <int>, x21st.week <int>,  
## # x22nd.week <int>, x23rd.week <int>, x24th.week <int>,  
## # x25th.week <int>, x26th.week <int>, x27th.week <int>,  
## # x28th.week <int>, x29th.week <int>, x30th.week <int>,  
## # x31st.week <int>, x32nd.week <int>, x33rd.week <int>,  
## # x34th.week <int>, x35th.week <int>, x36th.week <int>,  
## # x37th.week <int>, x38th.week <int>, x39th.week <int>,  
## # x40th.week <int>, x41st.week <int>, x42nd.week <int>,  
## # x43rd.week <int>, x44th.week <int>, x45th.week <int>,  
## # x46th.week <int>, x47th.week <int>, x48th.week <int>,  
## # x49th.week <int>, x50th.week <int>, x51st.week <int>,  
## # x52nd.week <int>, x53rd.week <int>, x54th.week <int>,  
## # x55th.week <int>, x56th.week <int>, x57th.week <int>,  
## # x58th.week <int>, x59th.week <int>, x60th.week <int>,  
## # x61st.week <int>, x62nd.week <int>, x63rd.week <int>,  
## # x64th.week <int>, x65th.week <int>, x66th.week <chr>,  
## # x67th.week <chr>, x68th.week <chr>, x69th.week <chr>,  
## # x70th.week <chr>, x71st.week <chr>, x72nd.week <chr>,  
## # x73rd.week <chr>, x74th.week <chr>, x75th.week <chr>, x76th.week <chr>

# b)

billboard %>% gather(Week,Rank,x1st.week:x76th.week,na.rm=T)

## # A tibble: 5,307 x 9  
## year artist.inverted track  
## \* <int> <chr> <chr>  
## 1 2000 Destiny's Child Independent Women Part I  
## 2 2000 Santana Maria, Maria  
## 3 2000 Savage Garden I Knew I Loved You  
## 4 2000 Madonna Music  
## 5 2000 Aguilera, Christina Come On Over Baby (All I Want Is You)  
## 6 2000 Janet Doesn't Really Matter  
## 7 2000 Destiny's Child Say My Name  
## 8 2000 Iglesias, Enrique Be With You  
## 9 2000 Sisqo Incomplete  
## 10 2000 Lonestar Amazed  
## # ... with 5,297 more rows, and 6 more variables: time <time>,  
## # genre <chr>, date.entered <date>, date.peaked <date>, Week <chr>,  
## # Rank <chr>

I have fewer columns after converting the columns from before. I have alot more rows because each song must indicate each week and the corresponding rank of the week.

# c)

billboard %>% gather(Week,Rank,x1st.week:x76th.week,na.rm=T) %>% select(Week,Rank) %>% head

## # A tibble: 6 x 2  
## Week Rank  
## <chr> <chr>  
## 1 x1st.week 78  
## 2 x1st.week 15  
## 3 x1st.week 71  
## 4 x1st.week 41  
## 5 x1st.week 57  
## 6 x1st.week 59

# d)

billboard %>% gather(Week,Rank,x1st.week:x76th.week,na.rm=T) %>% mutate(Week\_numeric=parse\_number(Week), Rank\_numeric=as.numeric(Rank)) %>% select(Week\_numeric, Rank\_numeric)

## # A tibble: 5,307 x 2  
## Week\_numeric Rank\_numeric  
## <dbl> <dbl>  
## 1 1 78  
## 2 1 15  
## 3 1 71  
## 4 1 41  
## 5 1 57  
## 6 1 59  
## 7 1 83  
## 8 1 63  
## 9 1 77  
## 10 1 81  
## # ... with 5,297 more rows

# e)

billboard %>% gather(Week,Rank,x1st.week:x76th.week,na.rm=T) %>% mutate(Week\_numeric=parse\_number(Week), Rank\_numeric=as.numeric(Rank)) %>% mutate(current=parse\_date(date.entered)+(Week\_numeric-1)\*7) %>% select(date.entered, Week\_numeric, current)

## # A tibble: 5,307 x 3  
## date.entered Week\_numeric current  
## <date> <dbl> <date>  
## 1 2000-09-23 1 2000-09-23  
## 2 2000-02-12 1 2000-02-12  
## 3 1999-10-23 1 1999-10-23  
## 4 2000-08-12 1 2000-08-12  
## 5 2000-08-05 1 2000-08-05  
## 6 2000-06-17 1 2000-06-17  
## 7 1999-12-25 1 1999-12-25  
## 8 2000-04-01 1 2000-04-01  
## 9 2000-06-24 1 2000-06-24  
## 10 1999-06-05 1 1999-06-05  
## # ... with 5,297 more rows

# f)

billboard %>% gather(Week,Rank,x1st.week:x76th.week,na.rm=T) %>% mutate(Week\_numeric=parse\_number(Week), Rank\_numeric=as.numeric(Rank)) %>% mutate(current=parse\_date(date.entered)+(Week\_numeric-1)\*7) %>% filter(Rank\_numeric == 1) %>% select(artist.inverted, track,current) %>% arrange(current)

## # A tibble: 55 x 3  
## artist.inverted track current  
## <chr> <chr> <date>  
## 1 Aguilera, Christina What A Girl Wants 2000-01-15  
## 2 Aguilera, Christina What A Girl Wants 2000-01-22  
## 3 Savage Garden I Knew I Loved You 2000-01-29  
## 4 Savage Garden I Knew I Loved You 2000-02-05  
## 5 Savage Garden I Knew I Loved You 2000-02-12  
## 6 Carey, Mariah Thank God I Found You 2000-02-19  
## 7 Savage Garden I Knew I Loved You 2000-02-26  
## 8 Lonestar Amazed 2000-03-04  
## 9 Lonestar Amazed 2000-03-11  
## 10 Destiny's Child Say My Name 2000-03-18  
## # ... with 45 more rows

# g)

billboard %>% gather(Week,Rank,x1st.week:x76th.week,na.rm=T) %>% mutate(Week\_numeric=parse\_number(Week), Rank\_numeric=as.numeric(Rank)) %>% mutate(current=parse\_date(date.entered)+Week\_numeric\*7-1) %>% filter(Rank\_numeric == 1) %>% group\_by(artist.inverted,track) %>% count(track)%>% arrange(desc(n)) %>% head(1)

## # A tibble: 1 x 3  
## # Groups: artist.inverted, track [1]  
## artist.inverted track n  
## <chr> <chr> <int>  
## 1 Destiny's Child Independent Women Part I 11

## 12)

# a)

url = "http://www.utsc.utoronto.ca/~butler/c32/pinetrees.txt"  
pine=read\_delim(url, " ")

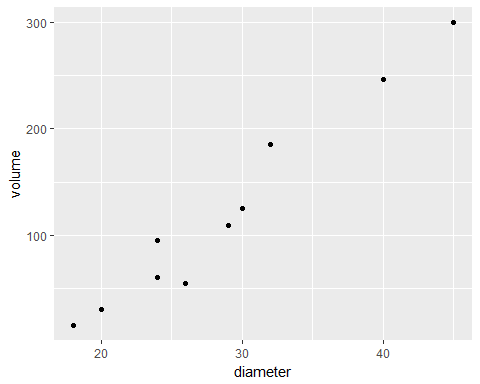
## Parsed with column specification:  
## cols(  
## diameter = col\_integer(),  
## volume = col\_integer()  
## )

pine

## # A tibble: 10 x 2  
## diameter volume  
## <int> <int>  
## 1 32 185  
## 2 29 109  
## 3 24 95  
## 4 45 300  
## 5 20 30  
## 6 30 125  
## 7 26 55  
## 8 40 246  
## 9 24 60  
## 10 18 15

# b)

ggplot(pine,aes(x=diameter, y=volume)) + geom\_point()



# c)

There is positive relationship between diameter and volume. As diameter of tree increases, the volume of tree also increases.

# d)

reg=lm(volume~diameter,pine)  
summary(reg)

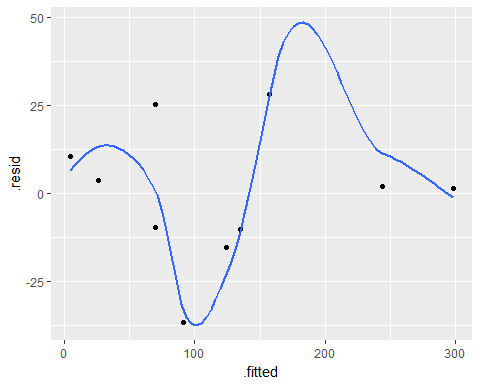
##   
## Call:  
## lm(formula = volume ~ diameter, data = pine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -36.497 -9.982 1.751 8.959 28.139   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -191.749 23.954 -8.005 4.35e-05 \*\*\*  
## diameter 10.894 0.801 13.600 8.22e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 20.38 on 8 degrees of freedom  
## Multiple R-squared: 0.9585, Adjusted R-squared: 0.9534   
## F-statistic: 185 on 1 and 8 DF, p-value: 8.217e-07

R-Squared is 0.9585, which means there is a very strong positive linear relationship between diameter and volume.

# e)

ggplot(reg,aes(x=.fitted,y=.resid))+geom\_point() +geom\_smooth(se=F)

## `geom\_smooth()` using method = 'loess'

 I do not trust the linearity of this regression. The residual plot does not show a straight line more or less at 0. As a result, we conclude that linear regression is not suitable for modeling the relationship between diameter and volume.

# f)

The volume of a tree of diameter of zero should be 0.No it is not what the regression predicts. According to my regression, when diameter is 0, volume is -191.749. Volume is suppose to be zero when diameter is 0.

# g)

cone\_volume = (pi*((d/2)^2)*h)/3

where h = height and d = diameter

# h)

h=2  
cone\_volume = (pi\*((pine$diameter/2)^2)\*h)/3  
reg\_cone=lm(cone\_volume~I((pine$diameter)^2))  
summary(reg\_cone)

## Warning in summary.lm(reg\_cone): essentially perfect fit: summary may be  
## unreliable

##   
## Call:  
## lm(formula = cone\_volume ~ I((pine$diameter)^2))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.437e-13 -7.417e-14 6.045e-15 3.685e-14 2.243e-13   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.438e-13 7.079e-14 -2.031e+00 0.0767 .   
## I((pine$diameter)^2) 5.236e-01 6.872e-17 7.619e+15 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.111e-13 on 8 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 5.805e+31 on 1 and 8 DF, p-value: < 2.2e-16