Module 3 Assignment 1

## Silas, Kimberly

#install.packages("caret")  
tidyverse.quiet = TRUE  
library(tidyverse)

## -- Attaching packages ------------------------------------------------------------------------------------------------------------------ tidyverse 1.2.1 --

## v ggplot2 3.1.0 v purrr 0.3.2   
## v tibble 2.1.1 v dplyr 0.8.0.1  
## v tidyr 0.8.3 v stringr 1.4.0   
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts --------------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

bike <- read\_csv("hour.csv")

## Parsed with column specification:  
## cols(  
## instant = col\_double(),  
## dteday = col\_date(format = ""),  
## season = col\_double(),  
## yr = col\_double(),  
## mnth = col\_double(),  
## hr = col\_double(),  
## holiday = col\_double(),  
## weekday = col\_double(),  
## workingday = col\_double(),  
## weathersit = col\_double(),  
## temp = col\_double(),  
## atemp = col\_double(),  
## hum = col\_double(),  
## windspeed = col\_double(),  
## casual = col\_double(),  
## registered = col\_double(),  
## count = col\_double()  
## )

bike = bike %>% mutate(season = as\_factor(as.character(season))) %>%  
mutate(season = fct\_recode(season,  
"Spring" = "1",  
"Summer" = "2",  
"Fall" = "3",  
"Winter" = "4"))%>%   
mutate(yr = as\_factor(as.character(yr)))%>%   
mutate(mnth = as\_factor(as.character(mnth))) %>%   
mutate(hr = as\_factor(as.character(hr)))%>%   
mutate(holiday = as\_factor(as.character(holiday))) %>%  
mutate(holiday = fct\_recode(holiday,  
"NotHoliday" = "0",  
"Holiday" = "1")) %>%  
mutate(workingday= as\_factor(as.character(workingday))) %>%  
mutate(workingday = fct\_recode(workingday,  
"NotWorkingDay" = "0",  
"WorkingDay" = "1")) %>%  
mutate(weathersit = as\_factor(as.character(weathersit))) %>%  
mutate(weathersit = fct\_recode(weathersit,  
"NoPrecip" = "1",  
"Misty" = "2",  
"LightPrecip" = "3",  
"HeavyPrecip" = "4"))%>%  
mutate(weekday = as\_factor(as.character(weekday))) %>%  
mutate(weekday = fct\_recode(weekday,  
"Sunday" = "0",  
"Monday" = "1",  
"Tuesday" = "2",  
"Wednesday" = "3",  
"Thursday" = "4",  
"Friday" = "5",  
"Saturday" = "6"  
))  
 bike

## # A tibble: 17,379 x 17  
## instant dteday season yr mnth hr holiday weekday workingday  
## <dbl> <date> <fct> <fct> <fct> <fct> <fct> <fct> <fct>   
## 1 1 2011-01-01 Spring 0 1 0 NotHol~ Saturd~ NotWorkin~  
## 2 2 2011-01-01 Spring 0 1 1 NotHol~ Saturd~ NotWorkin~  
## 3 3 2011-01-01 Spring 0 1 2 NotHol~ Saturd~ NotWorkin~  
## 4 4 2011-01-01 Spring 0 1 3 NotHol~ Saturd~ NotWorkin~  
## 5 5 2011-01-01 Spring 0 1 4 NotHol~ Saturd~ NotWorkin~  
## 6 6 2011-01-01 Spring 0 1 5 NotHol~ Saturd~ NotWorkin~  
## 7 7 2011-01-01 Spring 0 1 6 NotHol~ Saturd~ NotWorkin~  
## 8 8 2011-01-01 Spring 0 1 7 NotHol~ Saturd~ NotWorkin~  
## 9 9 2011-01-01 Spring 0 1 8 NotHol~ Saturd~ NotWorkin~  
## 10 10 2011-01-01 Spring 0 1 9 NotHol~ Saturd~ NotWorkin~  
## # ... with 17,369 more rows, and 8 more variables: weathersit <fct>,  
## # temp <dbl>, atemp <dbl>, hum <dbl>, windspeed <dbl>, casual <dbl>,  
## # registered <dbl>, count <dbl>

set.seed(1234)   
train.rows = createDataPartition(y = bike$count, p=0.7, list = FALSE)   
train = bike[train.rows,]   
test = bike[-train.rows,]

Task 2 train has 12167 rows while test has 5212

allmod = lm(count ~ season + mnth + holiday + weekday + temp+ weathersit , train)   
summary(allmod)

##   
## Call:  
## lm(formula = count ~ season + mnth + holiday + weekday + temp +   
## weathersit, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -335.30 -108.21 -27.23 70.95 725.46   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -43.109 7.150 -6.029 1.69e-09 \*\*\*  
## seasonSummer 35.788 9.042 3.958 7.60e-05 \*\*\*  
## seasonFall 3.969 10.678 0.372 0.710117   
## seasonWinter 45.199 9.117 4.958 7.23e-07 \*\*\*  
## mnth2 -29.617 7.258 -4.081 4.52e-05 \*\*\*  
## mnth3 -59.047 8.081 -7.307 2.91e-13 \*\*\*  
## mnth4 -108.435 11.977 -9.054 < 2e-16 \*\*\*  
## mnth5 -158.195 12.671 -12.485 < 2e-16 \*\*\*  
## mnth6 -195.032 12.844 -15.184 < 2e-16 \*\*\*  
## mnth7 -231.195 14.384 -16.073 < 2e-16 \*\*\*  
## mnth8 -194.367 14.158 -13.728 < 2e-16 \*\*\*  
## mnth9 -137.858 12.650 -10.898 < 2e-16 \*\*\*  
## mnth10 -93.855 11.947 -7.856 4.30e-15 \*\*\*  
## mnth11 -58.841 11.612 -5.067 4.10e-07 \*\*\*  
## mnth12 -42.757 9.211 -4.642 3.49e-06 \*\*\*  
## holidayHoliday -30.296 9.254 -3.274 0.001064 \*\*   
## weekdaySunday -15.990 5.368 -2.979 0.002900 \*\*   
## weekdayMonday -11.065 5.548 -1.994 0.046142 \*   
## weekdayTuesday -17.960 5.407 -3.322 0.000898 \*\*\*  
## weekdayWednesday -18.481 5.405 -3.420 0.000629 \*\*\*  
## weekdayThursday -9.038 5.386 -1.678 0.093372 .   
## weekdayFriday -4.605 5.380 -0.856 0.392031   
## temp 694.516 15.096 46.006 < 2e-16 \*\*\*  
## weathersitMisty -22.348 3.352 -6.667 2.73e-11 \*\*\*  
## weathersitLightPrecip -73.545 5.319 -13.826 < 2e-16 \*\*\*  
## weathersitHeavyPrecip -31.123 158.540 -0.196 0.844369   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 158.4 on 12141 degrees of freedom  
## Multiple R-squared: 0.2312, Adjusted R-squared: 0.2296   
## F-statistic: 146 on 25 and 12141 DF, p-value: < 2.2e-16

The adjusted r squared is low at .23, the pvalues for heavy precip, fall, thursday and friday were not significant the rest had significant pvalues

train\_preds = head(predict(allmod, newdata = train),6)  
  
train\_preds

## 1 2 3 4 5 6   
## 109.6845 109.6845 123.5749 123.5749 101.2268 109.6845

Task 4 The predictions are really high positive numbers compared to the linear regression model created in task 3, not sure this model is an appropriate fit.

test\_preds = head(predict(allmod, newdata = test),6)

SSE = sum((test$count - test\_preds)^2) #sum of squared residuals from model

## Warning in test$count - test\_preds: longer object length is not a multiple  
## of shorter object length

SST = sum((test$count - mean(test$count))^2) #sum of squared residuals from a "naive" model  
1 - SSE/SST #definition of R squared

## [1] -0.06190385

Task 6 the adjusted r squared value was .23 in our regression model and our test set returned a -0.061 so very far off from our task 3 adjusted r squared I think the model did not reduce the overfitting.  
Task 7 the process for k-fold cross validation is more complicated and splits the data into partitions of multiple train and test examples. K fold gives the opportunity to run the process over and over to compare how the performance changes. Model validation does not have the same in-depth capabilities.