SLRC Assignment

Stuart Broach

January 30, 2019

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.5.2

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

## v ggplot2 3.1.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.7  
## v tidyr 0.8.2 v stringr 1.3.1  
## v readr 1.1.1 v forcats 0.3.0

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

library(GGally)

## Warning: package 'GGally' was built under R version 3.5.2

##   
## Attaching package: 'GGally'

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

air = airquality  
glimpse(air)

## Observations: 153  
## Variables: 6  
## $ Ozone <int> 41, 36, 12, 18, NA, 28, 23, 19, 8, NA, 7, 16, 11, 14, ...  
## $ Solar.R <int> 190, 118, 149, 313, NA, NA, 299, 99, 19, 194, NA, 256,...  
## $ Wind <dbl> 7.4, 8.0, 12.6, 11.5, 14.3, 14.9, 8.6, 13.8, 20.1, 8.6...  
## $ Temp <int> 67, 72, 74, 62, 56, 66, 65, 59, 61, 69, 74, 69, 66, 68...  
## $ Month <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ...  
## $ Day <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,...

summary(air)

## Ozone Solar.R Wind Temp   
## Min. : 1.00 Min. : 7.0 Min. : 1.700 Min. :56.00   
## 1st Qu.: 18.00 1st Qu.:115.8 1st Qu.: 7.400 1st Qu.:72.00   
## Median : 31.50 Median :205.0 Median : 9.700 Median :79.00   
## Mean : 42.13 Mean :185.9 Mean : 9.958 Mean :77.88   
## 3rd Qu.: 63.25 3rd Qu.:258.8 3rd Qu.:11.500 3rd Qu.:85.00   
## Max. :168.00 Max. :334.0 Max. :20.700 Max. :97.00   
## NA's :37 NA's :7   
## Month Day   
## Min. :5.000 Min. : 1.0   
## 1st Qu.:6.000 1st Qu.: 8.0   
## Median :7.000 Median :16.0   
## Mean :6.993 Mean :15.8   
## 3rd Qu.:8.000 3rd Qu.:23.0   
## Max. :9.000 Max. :31.0   
##

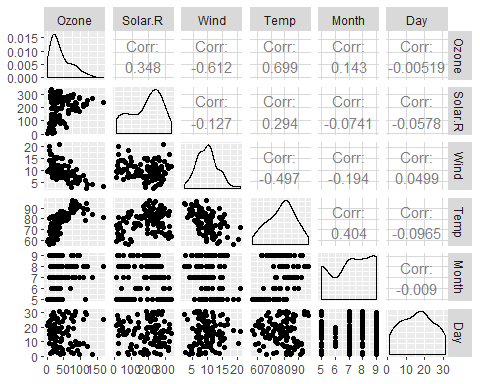
1. This dataset is daily air quality measurement sin New York, May to September 1973.
2. There are 153 observations and 6 variables.
3. Yes, there is missing data in the Ozone and Solar.R columns.
4. Ozone is likely to be the response variable.

air2 = air %>% filter(!is.na(Ozone)) %>% filter(!is.na(Solar.R))  
glimpse(air2)

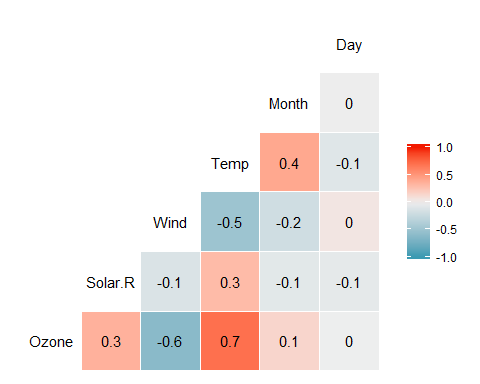
## Observations: 111  
## Variables: 6  
## $ Ozone <int> 41, 36, 12, 18, 23, 19, 8, 16, 11, 14, 18, 14, 34, 6, ...  
## $ Solar.R <int> 190, 118, 149, 313, 299, 99, 19, 256, 290, 274, 65, 33...  
## $ Wind <dbl> 7.4, 8.0, 12.6, 11.5, 8.6, 13.8, 20.1, 9.7, 9.2, 10.9,...  
## $ Temp <int> 67, 72, 74, 62, 65, 59, 61, 69, 66, 68, 58, 64, 66, 57...  
## $ Month <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ...  
## $ Day <int> 1, 2, 3, 4, 7, 8, 9, 12, 13, 14, 15, 16, 17, 18, 19, 2...

Now there are 111 observations and 6 variables left in the dataframe.

ggpairs(air2)

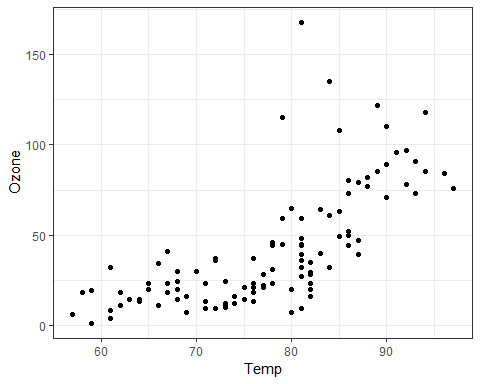


ggcorr(air2, label = TRUE)



1. Temp is mostly strongly correlated with the Ozone variable.
2. Day is least strongly correlated with the Ozone variable.

ggplot(air2,aes(x=Temp, y=Ozone)) + geom\_point() + theme\_bw()



It appears as if Temp and Ozone are closely correlated. The higher the Ozone gets, The more Temp rises.

model1 = lm(formula = Ozone ~ Temp, data = air2)  
summary(model1)

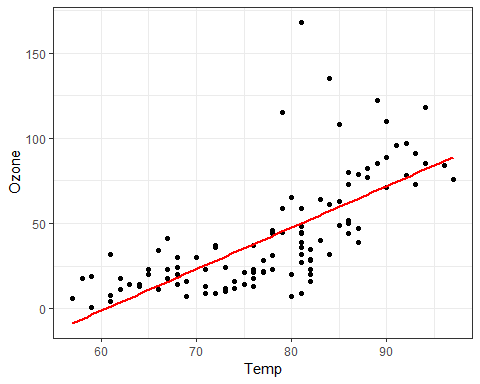
##   
## Call:  
## lm(formula = Ozone ~ Temp, data = air2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40.922 -17.459 -0.874 10.444 118.078   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -147.6461 18.7553 -7.872 2.76e-12 \*\*\*  
## Temp 2.4391 0.2393 10.192 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 23.92 on 109 degrees of freedom  
## Multiple R-squared: 0.488, Adjusted R-squared: 0.4833   
## F-statistic: 103.9 on 1 and 109 DF, p-value: < 2.2e-16

confint(model1)

## 2.5 % 97.5 %  
## (Intercept) -184.818372 -110.473773  
## Temp 1.964787 2.913433

a.This is a pretty good quality model. The p-value is less than 0.05 and the R square value is 0.4833. Both of these indicate a good model. b. 1.9 and 2.9

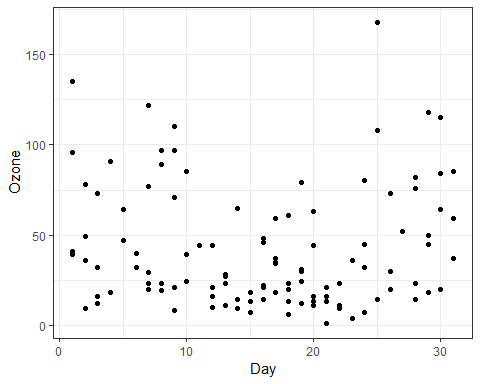
ggplot(air2,aes(x=Temp, y=Ozone)) + geom\_point() + geom\_smooth(method = "lm", color="red", se=FALSE) + theme\_bw()



testdata2 = data.frame(Temp = c(80))  
predict(model1, newdata = testdata2, interval = "predict")

## fit lwr upr  
## 1 47.48272 -0.1510188 95.11646

ggplot(air2,aes(x=Day, y=Ozone)) + geom\_point() + theme\_bw()



There is no coorelation between day and ozone.

model2 = lm(formula = Ozone ~ Day, data = air2)  
summary(model2)

##   
## Call:  
## lm(formula = Ozone ~ Day, data = air2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -41.00 -24.23 -11.04 19.96 126.08   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 42.41536 6.64353 6.384 4.32e-09 \*\*\*  
## Day -0.01983 0.36604 -0.054 0.957   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 33.43 on 109 degrees of freedom  
## Multiple R-squared: 2.693e-05, Adjusted R-squared: -0.009147   
## F-statistic: 0.002936 on 1 and 109 DF, p-value: 0.9569

confint(model2)

## 2.5 % 97.5 %  
## (Intercept) 29.248109 55.5826192  
## Day -0.745321 0.7056539

1. The quality of this model is horrible. The p-value is .95 (way above .05) and the R squared is 0. b.-.74 to .70

ggplot(air2,aes(x=Day, y=Ozone)) + geom\_point() + geom\_smooth(method = "lm", color="red", se=FALSE) + theme\_bw()

