library(tidyverse)

## -- 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

library(MASS)

##   
## Attaching package: 'MASS'

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

library(leaps)

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

library(caret)

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

## Loading required package: lattice

##   
## Attaching package: 'caret'

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

credit <- read\_csv("CreditData.csv")

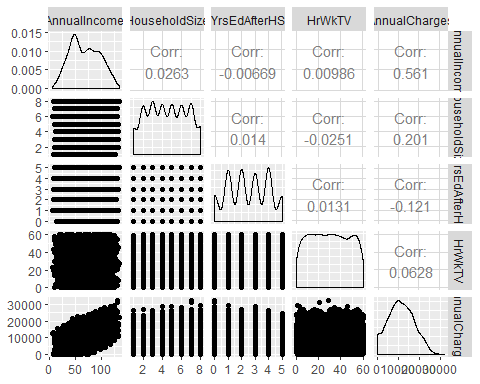
## Parsed with column specification:  
## cols(  
## AnnualIncome = col\_double(),  
## HouseholdSize = col\_integer(),  
## YrsEdAfterHS = col\_integer(),  
## HrWkTV = col\_integer(),  
## AnnualCharges = col\_double()  
## )

credit <- credit %>% drop\_na()  
str(credit)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 5000 obs. of 5 variables:  
## $ AnnualIncome : num 21.8 65.5 54.2 73.7 110.4 ...  
## $ HouseholdSize: int 4 7 3 6 7 8 5 8 1 3 ...  
## $ YrsEdAfterHS : int 5 3 2 0 5 3 4 5 4 1 ...  
## $ HrWkTV : int 29 46 18 44 39 39 40 27 15 3 ...  
## $ AnnualCharges: num 10024 11249 6115 9786 12634 ...

train.rows <- createDataPartition(y= credit$AnnualCharges, p= 0.7, list = FALSE)  
train <- credit[train.rows,]  
test <- credit[-train.rows,]

ggpairs(train)



mod1 <- lm(AnnualCharges ~ AnnualIncome, train)  
summary(mod1)

##   
## Call:  
## lm(formula = AnnualCharges ~ AnnualIncome, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12191.3 -3973.9 -76.1 3962.6 13199.1   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3066.765 224.268 13.68 <2e-16 \*\*\*  
## AnnualIncome 122.222 3.047 40.12 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5055 on 3498 degrees of freedom  
## Multiple R-squared: 0.3151, Adjusted R-squared: 0.3149   
## F-statistic: 1609 on 1 and 3498 DF, p-value: < 2.2e-16

test\_preds<- predict(mod1, newdata = test)

SSE <- sum((test$AnnualCharges - test\_preds)^2)  
SST <- sum((test$AnnualCharges - mean(test$AnnualCharges))^2)  
1 - SSE/SST

## [1] 0.3156217