Chapter1

Mohammed Ali

December 26, 2017

Contents

About

This document will represent the solutions for the applied part for *Intorduction to Statistical Learning* book.

Questions

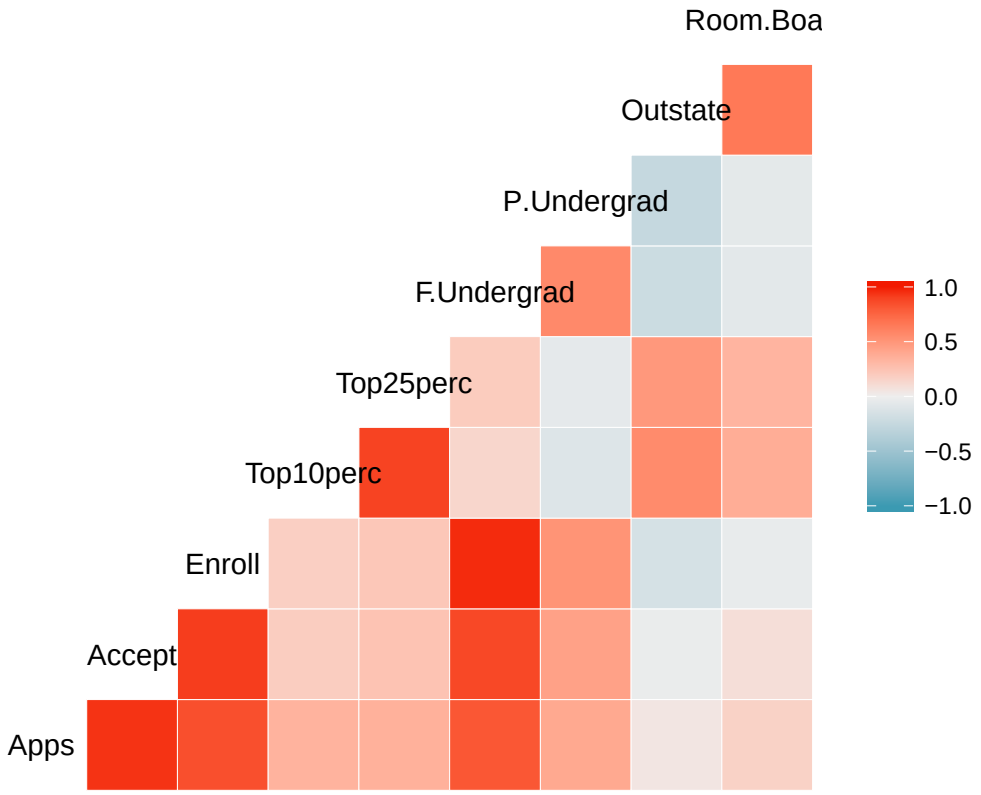
Q.8 Load college dataset and inspect it

```
data("College")
glimpse(College)
```

```
## Observations: 777
## Variables: 18
## $ Private     <fctr> Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes,...
## $ Apps        <dbl> 1660, 2186, 1428, 417, 193, 587, 353, 1899, 1038, ...
## $ Accept      <dbl> 1232, 1924, 1097, 349, 146, 479, 340, 1720, 839, 4...
## $ Enroll      <dbl> 721, 512, 336, 137, 55, 158, 103, 489, 227, 172, 4...
## $ Top10perc   <dbl> 23, 16, 22, 60, 16, 38, 17, 37, 30, 21, 37, 44, 38...
## $ Top25perc   <dbl> 52, 29, 50, 89, 44, 62, 45, 68, 63, 44, 75, 77, 64...
## $ F.Undergrad <dbl> 2885, 2683, 1036, 510, 249, 678, 416, 1594, 973, 7...
## $ P.Undergrad <dbl> 537, 1227, 99, 63, 869, 41, 230, 32, 306, 78, 110,...
## $ Outstate    <dbl> 7440, 12280, 11250, 12960, 7560, 13500, 13290, 138...
## $ Room.Board  <dbl> 3300, 6450, 3750, 5450, 4120, 3335, 5720, 4826, 44...
## $ Books       <dbl> 450, 750, 400, 450, 800, 500, 500, 450, 300, 660, ...
## $ Personal    <dbl> 2200, 1500, 1165, 875, 1500, 675, 1500, 850, 500, ...
## $ PhD         <dbl> 70, 29, 53, 92, 76, 67, 90, 89, 79, 40, 82, 73, 60...
## $ Terminal    <dbl> 78, 30, 66, 97, 72, 73, 93, 100, 84, 41, 88, 91, 8...
## $ S.F.Ratio   <dbl> 18.1, 12.2, 12.9, 7.7, 11.9, 9.4, 11.5, 13.7, 11.3...
## $ perc.alumni <dbl> 12, 16, 30, 37, 2, 11, 26, 37, 23, 15, 31, 41, 21,...
## $ Expend      <dbl> 7041, 10527, 8735, 19016, 10922, 9727, 8861, 11487...
## $ Grad.Rate   <dbl> 60, 56, 54, 59, 15, 55, 63, 73, 80, 52, 73, 76, 74...
```

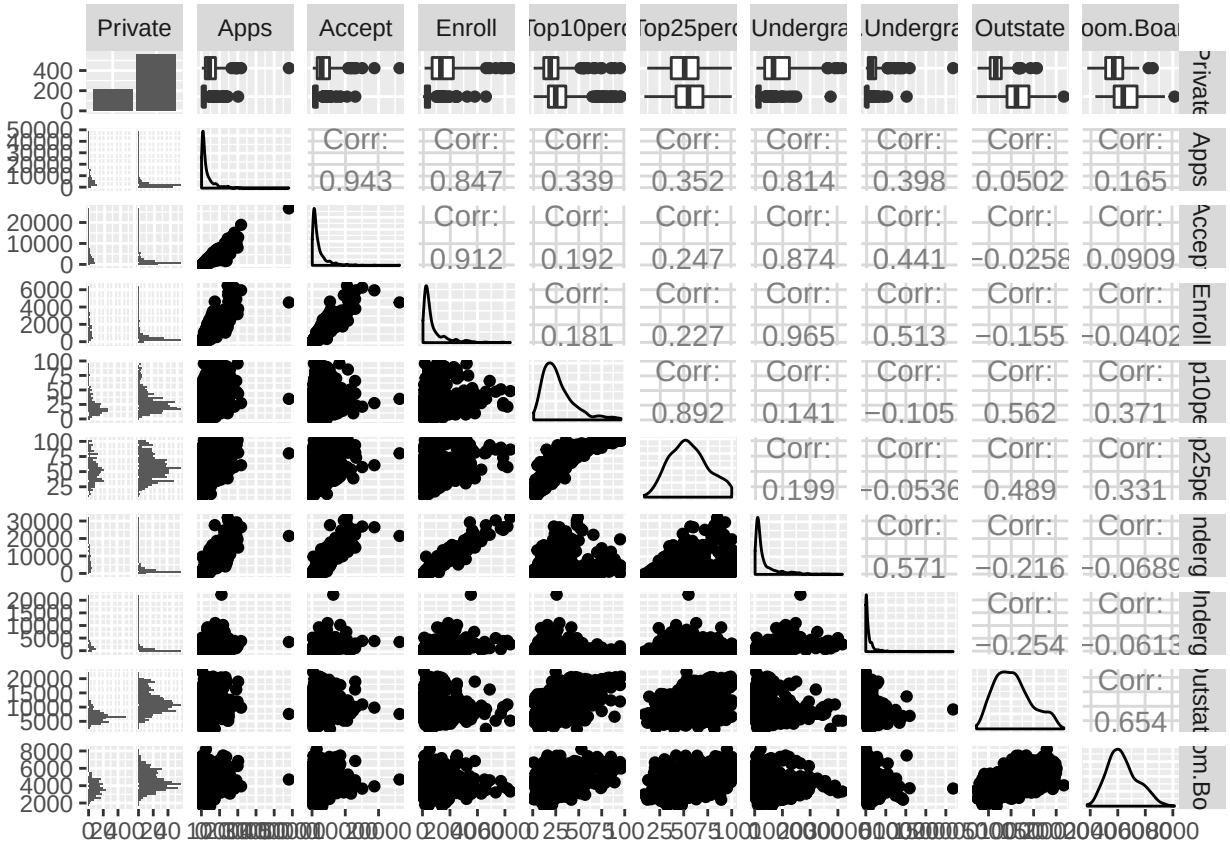
```
ggcorr(College[, 1:10])
```

```
## Warning in ggcorr(College[, 1:10]): data in column(s) 'Private' are not
## numeric and were ignored
```



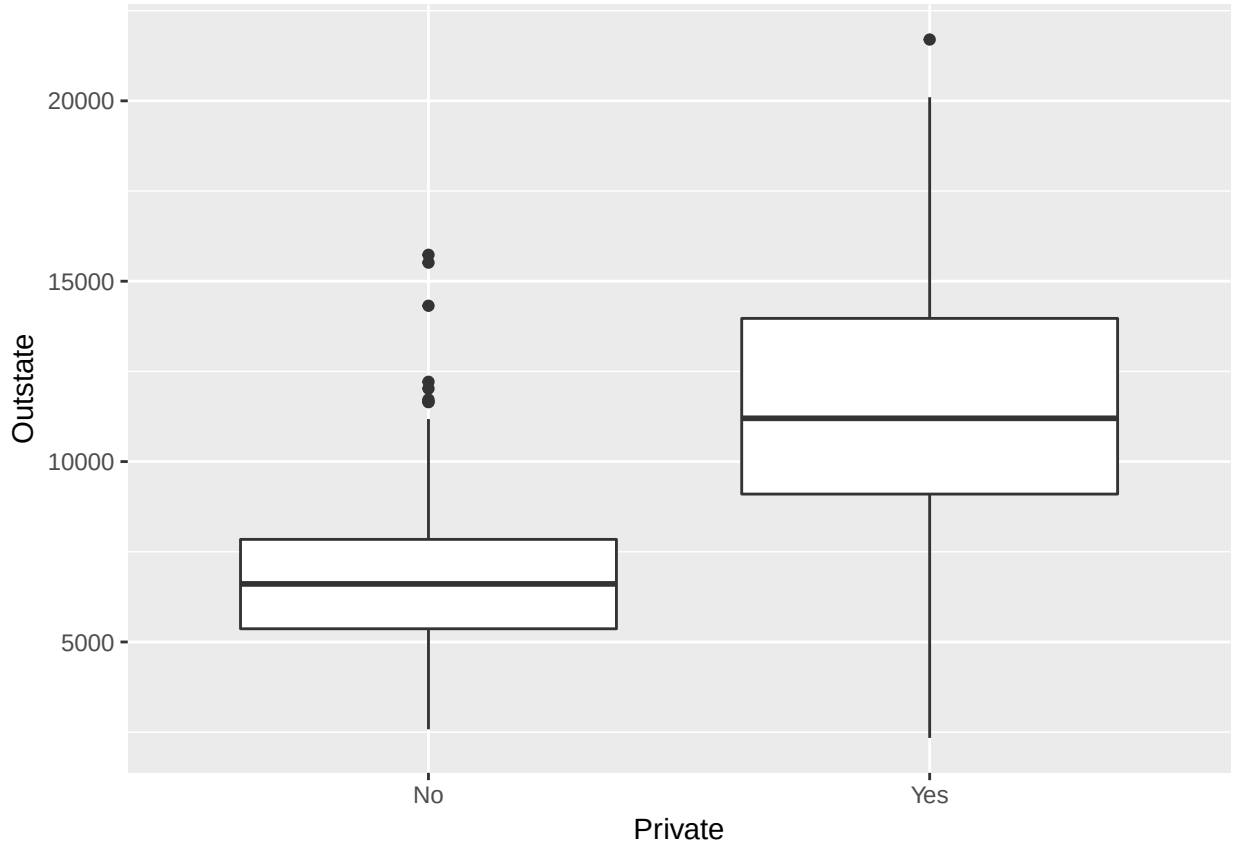

```
ggpairs(College[, 1:10])
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Create a new qualitative variable, called Elite to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

```
ggplot(College, aes(y=Outstate, x=Private)) +
  geom_boxplot()
```



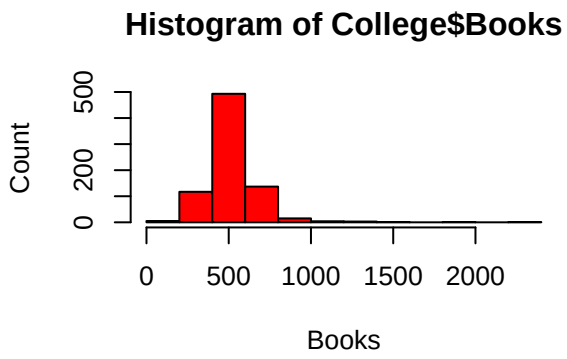

```
College <- College %>%
  mutate(Elite = ifelse(Top10perc > 50, "Yes", "No"))
```

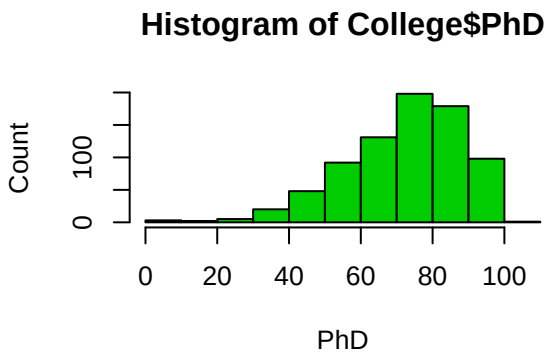
```
## Warning: package 'bindrcpp' was built under R version 3.4.2
```

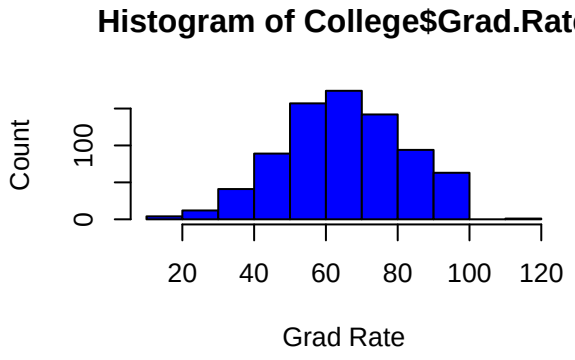
```
College$Elite <- as.factor(College$Elite)
table(College$Elite)
```

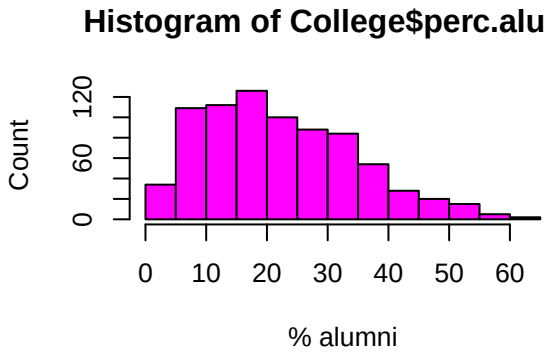
```
##
##  No Yes
## 699  78
```

```
par(mfrow = c(2,2))
hist(College$Books, col = 2, xlab = "Books", ylab = "Count")
hist(College$PhD, col = 3, xlab = "PhD", ylab = "Count")
hist(College$Grad.Rate, col = 4, xlab = "Grad Rate", ylab = "Count")
hist(College$perc.alumni, col = 6, xlab = "% alumni", ylab = "Count")
```










```
data("Auto")
glimpse(Auto)
```

```
## Observations: 392
## Variables: 9
## $ mpg          <dbl> 18, 15, 18, 16, 17, 15, 14, 14, 14, 15, 15, 14, 1...
## $ cylinders    <dbl> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 4, 6, 6...
## $ displacement <dbl> 307, 350, 318, 304, 302, 429, 454, 440, 455, 390,...
## $ horsepower   <dbl> 130, 165, 150, 150, 140, 198, 220, 215, 225, 190,...
## $ weight       <dbl> 3504, 3693, 3436, 3433, 3449, 4341, 4354, 4312, 4...
## $ acceleration <dbl> 12.0, 11.5, 11.0, 12.0, 10.5, 10.0, 9.0, 8.5, 10....
## $ year         <dbl> 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 7...
## $ origin       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1...
## $ name         <fctr> chevrolet chevelle malibu, buick skylark 320, pl...
```

```
summary(Auto[-9])
```

```
##       mpg          cylinders      displacement     horsepower
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0
##      weight      acceleration        year           origin
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
##  1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000
```

```
##  Median :2804   Median :15.50   Median :76.00   Median :1.000
##  Mean   :2978   Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000
```

```
map_dbl(Auto[-9], mean)
```

```
##          mpg    cylinders displacement   horsepower       weight
##    23.445918     5.471939   194.411990   104.469388  2977.584184
## acceleration         year       origin
##    15.541327    75.979592     1.576531
```

```
map_dbl(Auto[-9], sd)
```

```
##          mpg    cylinders displacement   horsepower       weight
##    7.8050075    1.7057832  104.6440039   38.4911599  849.4025600
## acceleration         year       origin
##    2.7588641    3.6837365    0.8055182
```

```
auto_subset <-Auto[- c(10:85), -9]

map_df(auto_subset, range)
```

```
## # A tibble: 2 x 8
##     mpg cylinders displacement horsepower weight acceleration  year origin
##   <dbl>     <dbl>        <dbl>      <dbl>  <dbl>        <dbl> <dbl>  <dbl>
## 1  11.0         3           68         46   1649          8.5    70      1
## 2  46.6         8          455        230   4997         24.8    82      3
```

```
map_dbl(auto_subset, mean)
```

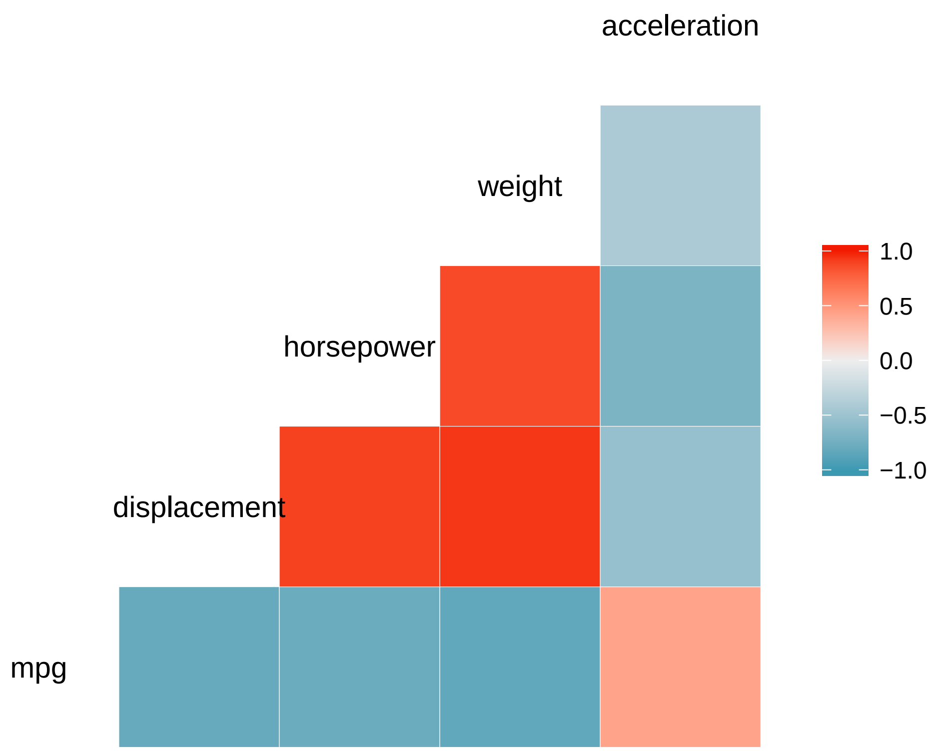
```
##          mpg    cylinders displacement   horsepower       weight
##    24.404430     5.373418   187.240506   100.721519  2935.971519
## acceleration         year       origin
##    15.726899    77.145570     1.601266
```

```
map_dbl(auto_subset, sd)
```

```
##          mpg    cylinders displacement   horsepower       weight
##     7.867283     1.654179    99.678367    35.708853   811.300208
## acceleration         year       origin
##     2.693721     3.106217     0.819910
```

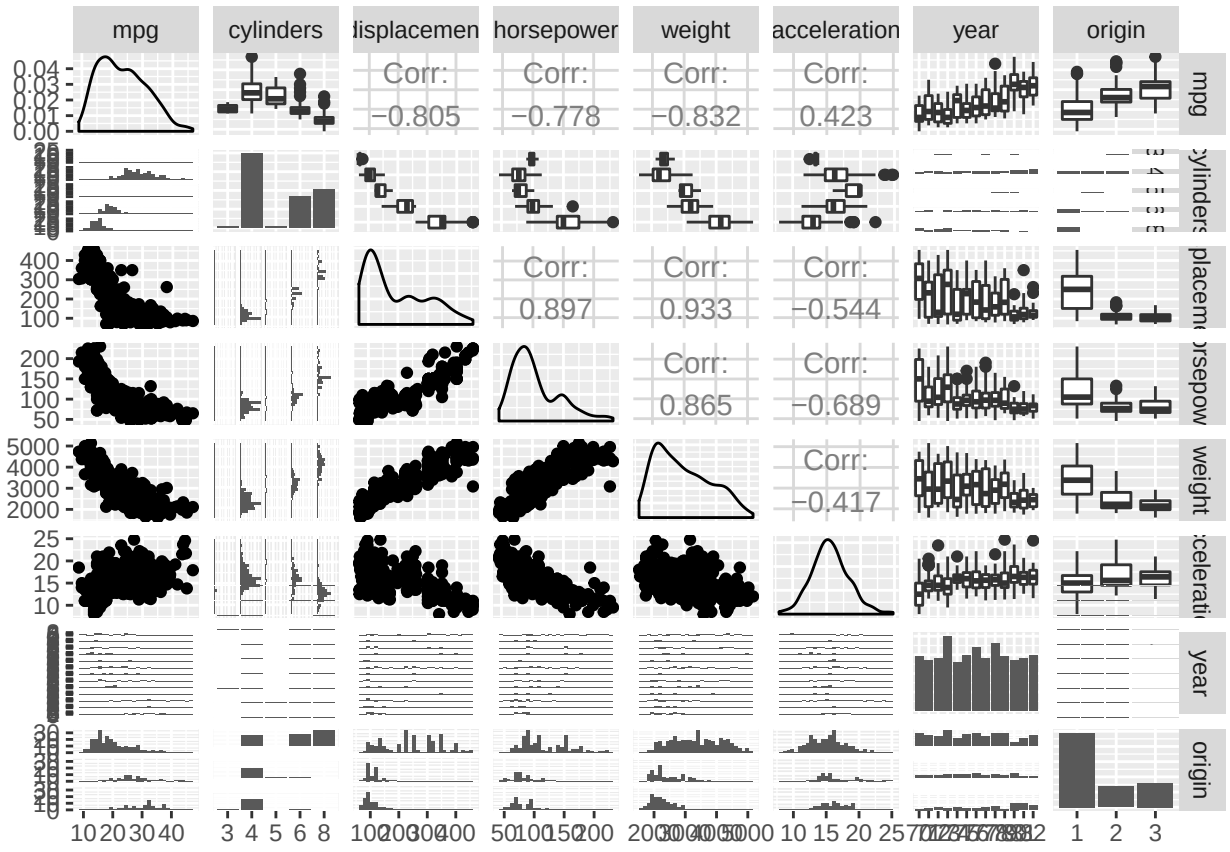
```
Auto$cylinders <- as.factor(Auto$cylinders)
Auto$year <- as.factor(Auto$year)
Auto$origin <- as.factor(Auto$origin)
ggcorr(Auto[-9])
```

```
## Warning in ggcorr(Auto[-9]): data in column(s) 'cylinders', 'year',
## 'origin' are not numeric and were ignored
```




```r
ggpairs(Auto[-9])
```

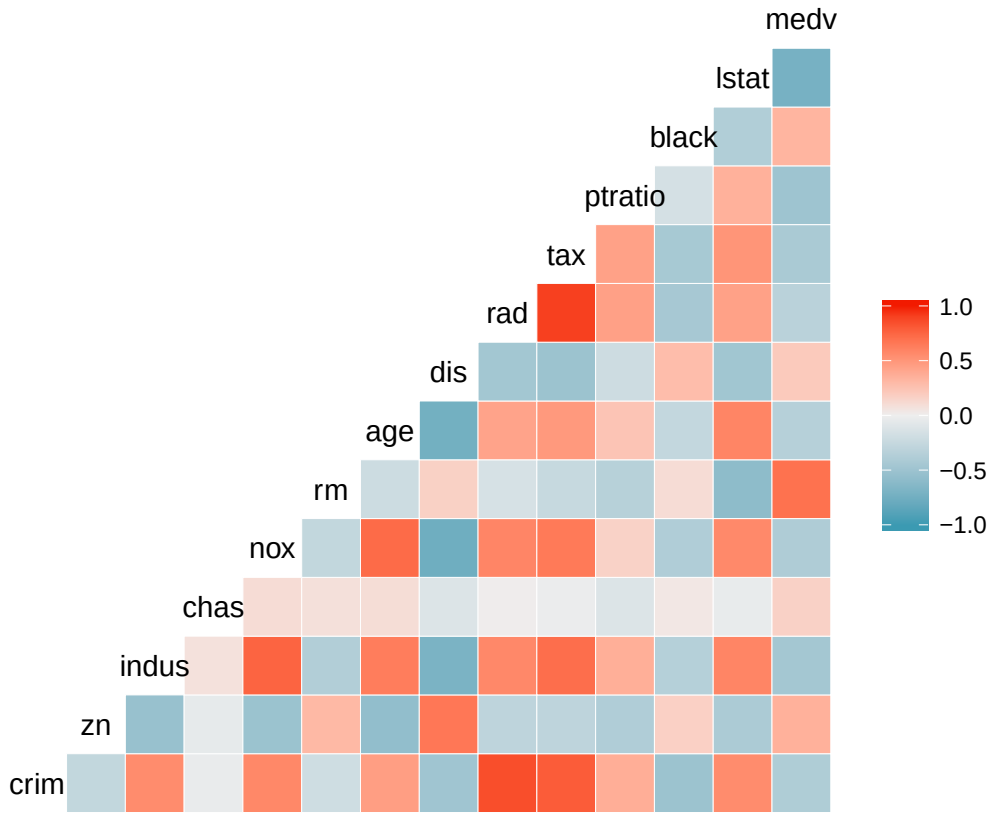
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



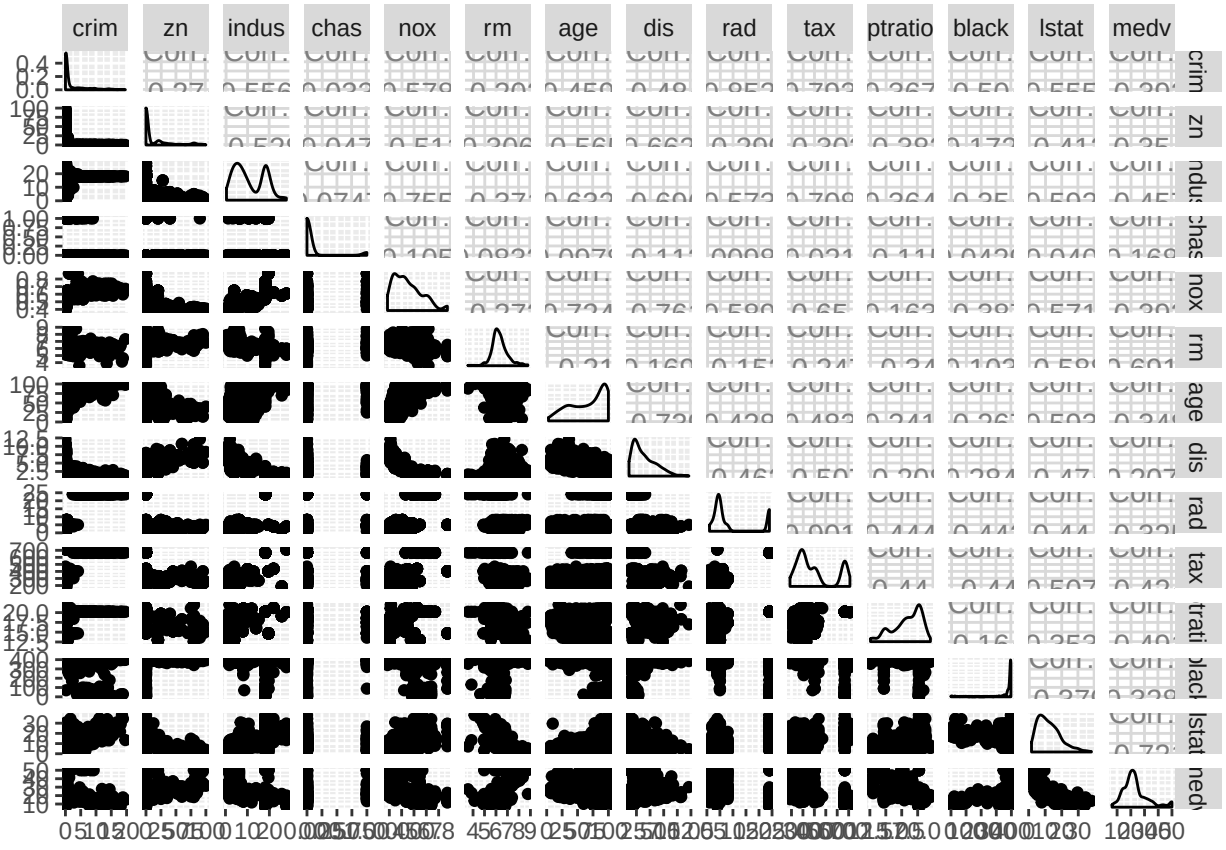

```
glimpse(Boston)
```

```
## Observations: 506
## Variables: 14
## $ crim    <dbl> 0.00632, 0.02731, 0.02729, 0.03237, 0.06905, 0.02985, ...
## $ zn      <dbl> 18.0, 0.0, 0.0, 0.0, 0.0, 0.0, 12.5, 12.5, 12.5, 12.5,...
## $ indus   <dbl> 2.31, 7.07, 7.07, 2.18, 2.18, 2.18, 7.87, 7.87, 7.87, ...
## $ chas    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ nox     <dbl> 0.538, 0.469, 0.469, 0.458, 0.458, 0.458, 0.524, 0.524...
## $ rm      <dbl> 6.575, 6.421, 7.185, 6.998, 7.147, 6.430, 6.012, 6.172...
## $ age     <dbl> 65.2, 78.9, 61.1, 45.8, 54.2, 58.7, 66.6, 96.1, 100.0,...
## $ dis     <dbl> 4.0900, 4.9671, 4.9671, 6.0622, 6.0622, 6.0622, 5.5605...
## $ rad     <int> 1, 2, 2, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4, 4, ...
## $ tax     <dbl> 296, 242, 242, 222, 222, 222, 311, 311, 311, 311, 311,...
## $ ptratio <dbl> 15.3, 17.8, 17.8, 18.7, 18.7, 18.7, 15.2, 15.2, 15.2, ...
## $ black   <dbl> 396.90, 396.90, 392.83, 394.63, 396.90, 394.12, 395.60...
## $ lstat   <dbl> 4.98, 9.14, 4.03, 2.94, 5.33, 5.21, 12.43, 19.15, 29.9...
## $ medv    <dbl> 24.0, 21.6, 34.7, 33.4, 36.2, 28.7, 22.9, 27.1, 16.5, ...
```

```
ggcorr(Boston[Boston$crim < 20, ])
```




```
ggpairs(Boston[Boston$crim < 20, ])
```


crim
zn
indus
chas
nox
rm
age
dis
rad
tax
ptratio
black
lstat
medv