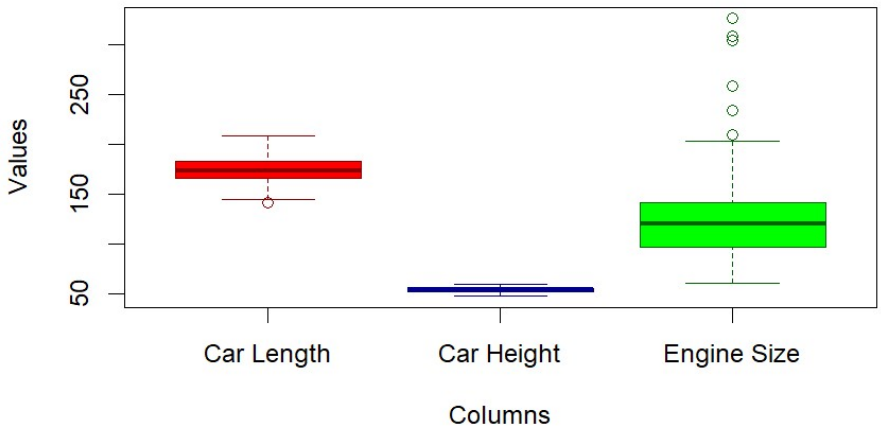
حنانه مبلغ توحید (۴۰۰۱۰۸۹۱۱)

- **بارگذاری و پیش‌پردازش دادگان:**

نمودار boxplot ستون‌های car height, car length, engine size رسم شده است.



مستطیل رنگی مربوط به هر کدام از داده‌ها، نشان‌دهنده‌ی ۵۰ درصد میانی داده‌ها است و خط مرکز آن میانه را نشان می‌دهد. خطوط بالا و پایین کشیدگی داده‌ها به راست یا چپ را نشان می‌دهند. خطوط بالا و پایین نشان‌دهنده typical range داده‌ها می‌باشند و به اندازه یک و نیم برابر چارک میانی ادامه پیدا می‌کنند.

با توجه به نمودار، داده‌های مربوط به قد ماشین‌ها واریانس کمتری دارد و متمرکزتر است. همچنین؛ خط پایین فاصله بیشتری از مستطیل مرکزی دارد پس می‌توان گفت که داده‌ها کشیدگی به سمت چپ دارند. همچنین هیچ داده‌ی پرتی وجود ندارد.

داده‌های مربوط به طول ماشین‌ها دارای واریانس نسبتاً بیشتری از قد ماشین‌هاست و نمودار آن تقریباً متقارن خواهد بود. یک داده پرت کمتر از مینیمم داریم.

داده‌های مربوط به اندازه موتور ماشین‌ها، کشیدگی به سمت راست دارد و واریانس آن از متغیرهای دیگر بیشتر است. همچنین تعدادی داده‌ی پرت بیشتر از ماکسیمم در نمودار مشاهده می‌شود، که با کشیدگی به سمت راست نمودار نیز همخوانی دارد.

در قسمت دوم داریم:

```
[1] "The column carbody contains missing values."
[1] "The column curbweight contains missing values."
[1] "The column cylindernumber contains missing values."
[1] "The column boreratio contains missing values."
```

برای رفع کردن مشکل داده‌های گم‌شده در ستون carbody می‌توان به نام ماشین و طول آن توجه کرد. اگر این دو یکسان باشند، آنگاه carbody نیز یکسان خواهد بود. برای تست کردن فرضیه گفته شده از کد زیر استفاده کردم:

```r
#creating a dataset with no missing values.
complete_cases <- data[complete.cases(rawData), ]

# Iterate over each row
for (i in 1:(num_rows - 1)) {
  # Get the current row's carbody, carname, and carlength
  current_carbody <- complete_cases[i, "carbody"]
  current_carname <- complete_cases[i, "CarName"]
  current_carlength <- complete_cases[i, "carlength"]

  # Iterate over the remaining rows
  for (j in (i + 1):num_rows) {
    # Get the comparison row's carbody, carname, and carlength
    comparison_carbody <- complete_cases[j, "carbody"]
    comparison_carname <- complete_cases[j, "CarName"]
    comparison_carlength <- complete_cases[j, "carlength"]

    # Check if carname and carlength match and carbody differs
    if (current_carname == comparison_carname && current_carlength == comparison_carlength &&
current_carbody != comparison_carbody) {
      print(paste("Cars with name", current_carname, "and length", current_carlength, "have different
car bodies."))
    }
  }
}
```

از آنجایی که هیچ دو خودرویی با نام و طول مشترک، بدنه متفاوتی ندارند، می‌توان هر بدنه خالی را با بدنه خودرویی که نام و طول یکسان دارد پر کرد.
با استفاده از این روش، تنها توانستم ۳ تا از داده‌های گمشده را پر کنم. تصمیم گرفتم با استفاده از کتابخانه mice تایپ بدنه خودرو را براساس طول و ارتفاع و عرض ماشین‌ها با استفاده از روش logistic regression تخمین بزنم.

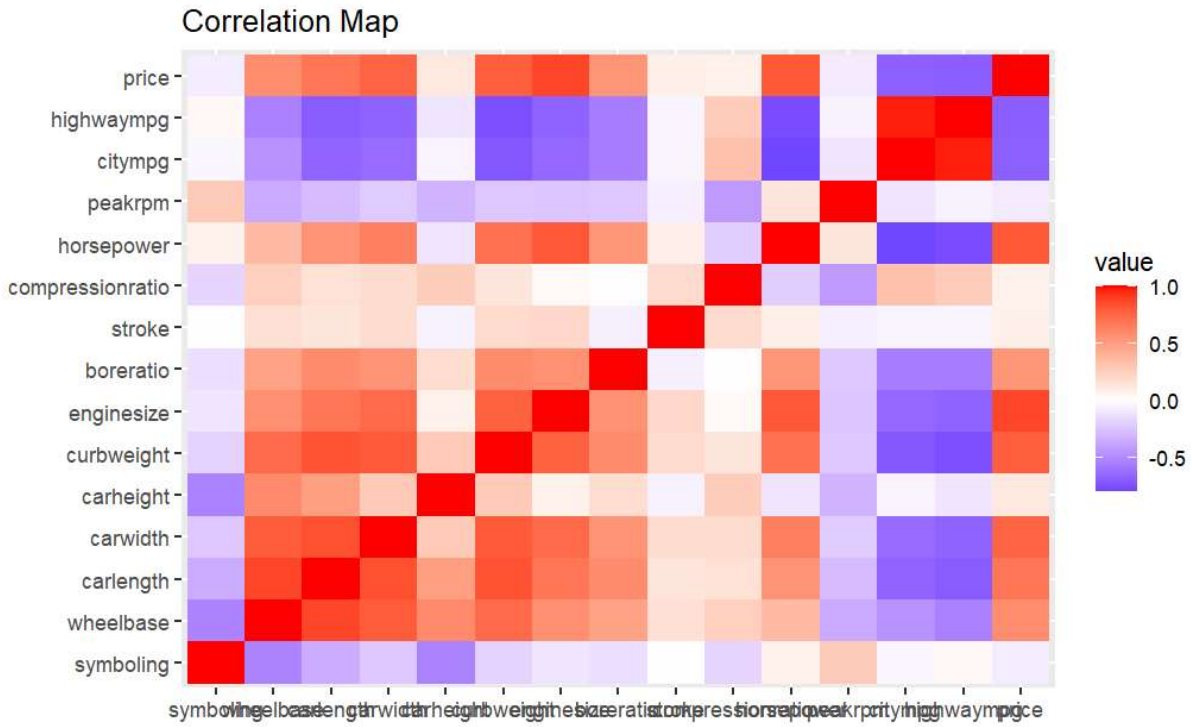
برای رفع مشکل داده‌های گم‌شده در ستون curbweight که وزن خودرو بدون بار یا مسافر را نشان می‌دهد، کوریلیشن بین وزن خودرو و اندازه موتور برابر ۰.۸۴۷۳۳۰۶ است که نشان‌دهنده کوریلیشن مثبت و بزرگی بین این دو می‌باشد. پس می‌توان براساس اندازه موتور تخمین زد.

تعداد سیلندرها را می‌توان براساس قدرت موتور (اسب بخار) تخمین زد. تعداد سیلندر می‌تواند مقادیر (۲، ۳، ۴، ۵، ۶، ۸ و ۱۲) را می‌تواند بگیرد. در هر کدام از تعداد سیلندرها، می‌توان mode قدرت موتور را پیدا کرد و داده‌های گمشده را با آن پر کرد. پس از استفاده از این روش متوجه شدم که هنوز ۲۰ داده گمشده وجود دارد و تصمیم گرفتم با استفاده از کتابخانه mice تعداد سیلندر خودرو را براساس اندازه موتور، قدرت موتور و تایپ موتور با استفاده از روش logistic regression تخمین بزنم.

با استفاده از کتابخانه mice boreratio را براساس *enginesize + horsepower + enginetype + compressionratio* با استفاده از روش logistic regression تخمین زدم.

سپس برند خودروها را از اسم آن‌ها جدا کردم. اما تعداد زیادی غلط املایی در این ستون از دیتاست وجود داشت که تعداد متغیرهای دامی را خیلی زیاد می‌کرد. پس تصمیم گرفتم این ستون را حذف کنم.

نقشه همبستگی:



- فرضیه اول: بین قیمت و اندازه موتور کوریلیشن مثبت قوی‌ای وجود دارد.

فرض اول را تست می‌کنیم:

```
        Pearson's product-moment correlation

data:  newData$price and newData$enginesize
t = 25.645, df = 203, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8374234 0.9030097
sample estimates:
      cor
0.8741448
```

آماره t نشان دهنده بزرگی تفاوت بین همبستگی مشاهده شده و همبستگی مورد انتظار تحت فرض صفر است (که هیچ همبستگی را فرض نمی کند). آماره t مثبت نشان دهنده همبستگی مثبت بین متغیرهای A و B است. هر چه قدر مطلق آماره t بزرگتر باشد، شواهد برای همبستگی قوی تر است.

مقدار p مرتبط با آزمون t نشان دهنده احتمال به دست آوردن همبستگی مشاهده شده (یا همبستگی شدیدتر) در صورت صحت فرضیه صفر (بدون همبستگی) است. اگر مقدار p بسیار کوچک باشد (معمولاً کمتر از سطح معناداری

انتخاب شده، مانند ۰.۰۵)، نشان می دهد که همبستگی مشاهده شده بعید است که به طور تصادفی رخ داده باشد. در این مورد، شواهدی در حمایت از فرضیه جایگزین ما مبنی بر همبستگی مثبت قوی بین متغیرهای *A* و *B* ارائه می کند. فرض صفر ما (عدم وجود همبستگی بین قیمت و اندازه موتور) رد می‌شود و نتیجه می‌گیریم که همبستگی مثبت قوی- ای بین قیمت و اندازه موتور وجود دارد.

- فرضیه دوم: بین وزن خودرو و highway mpg کوریلیشن منفی معناداری وجود دارد.

نتیجه *t-test*:

```
        Pearson's product-moment correlation

data:  newData$curbweight and newData$highwaympg
t = -16.489, df = 203, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8097246 -0.6912907
sample estimates:
       cor
-0.7566477
```

فرض صفر را رد می‌کنیم.

- فرضیه سوم: بین *boreratio* و *compressionratio* کوریلیشن معناداری وجود دارد.

نتیجه *t-test*:

```
        Pearson's product-moment correlation

data:  newData$boreratio and newData$compressionratio
t = 0.12438, df = 203, p-value = 0.9011
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1284594  0.1455901
sample estimates:
        cor
0.008729249
```

فرض صفر را قبول می‌کنیم.

- فرضیه چهارم: بین *horsepower* و *enginesize* کوریلیشن معناداری وجود ندارد.

نتیجه *t-test:*

```
        Pearson's product-moment correlation

data:  newData$horsepower and newData$enginesize
t = 19.663, df = 203, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7567024 0.8522341
sample estimates:
      cor
0.8097687
```

فرض صفر را قبول می‌کنیم. (کوریلیشن معناداری وجود دارد.)

از نسبت ۷۰-۳۰ برای قسمت *test\train* استفاده می‌کنیم.

با توجه به نقشه همبستگی و مقادیر بدست‌آمده پیش‌بینی کنید کدام ویژگی‌ها موثرتر و کدام موارد دارای اثر غیرموجه (اثر علّی کمتر روی متغیر پاسخ) هستند؟

با توجه به نقشه همبستگی متغیرهای wheelbase, carwidth, carlength, curbweight, borertio, enginesize,horsepower دارای کوریلیشن معنادار مثبت با متغیر پاسخ (قیمت) و متغیرهای highwaympg و citympg دارای کوریلیشن معنادار منفی با متغیر پاسخ هستند.

همچنین، متغیرهای carheight, stroke, compression ratio و peakrpm اثر کمی روی متغیر پاسخ دارند.

- **پردازش دادگان با مدل رگرسیون چندگانه:**

- [1] "the train RSS is: " "555320350.979994"
- [1] "the test RSS is: " "628950051.298672"

- [1] "the train TSS is: " "9245006164.13889"
- [1] "the test TSS is: " "3756173454.3241"

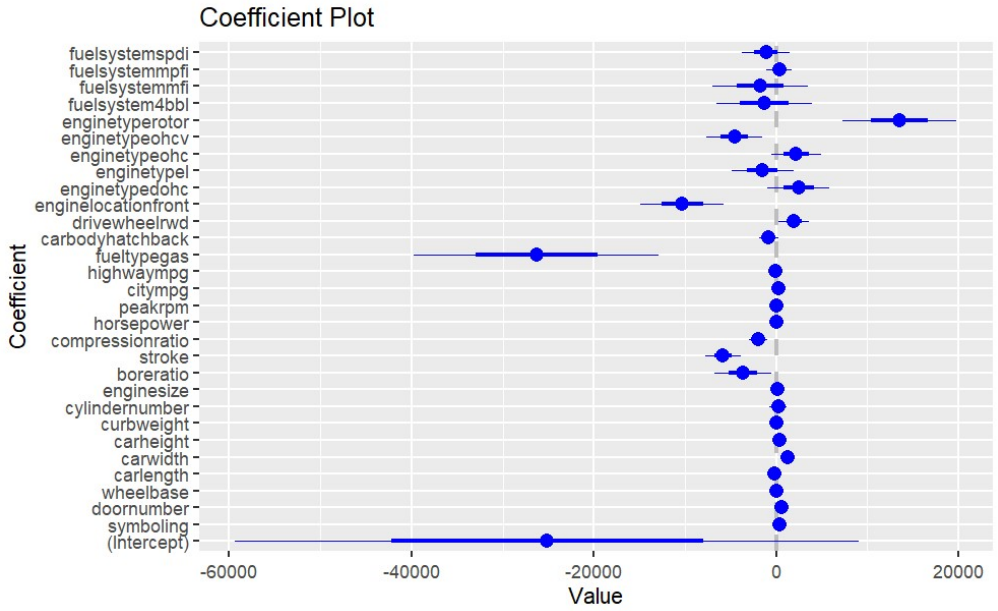
- [1] "the train MSE is: " "3856391.32624996"
- [1] "the test MSE is: " "10310656.5786668"

- [1] "the train R-squared is: " "0.93993293880819"
- [1] "the test R-squared is: " "0.83255564234537"

- [1] "the train R-adjusted is: " "0.923985931412134"
- [1] "the test R-adjusted is: " "0.66511128469074"

RSS- برابر جمع مجذور اختلاف‌های بین مقادیر مشاهده شده و مقادیر پیش بینی شده در یک مدل رگرسیونی است. این معیار برازش کلی مدل را با جمع تفاوت بین مقادیر مشاهده شده و پیش بینی شده برای ما قابل مشاهده می کند.
TSS- برابر جمع مجذور اختلاف‌های بین مقادیر مشاهده شده و میانگین متغیر پاسخ است. این متغیر تنوع (واریانس) کل متغیر پاسخ را نشان می دهد.
MSE – میانگین RSS در یک مدل رگرسیونی است. میانگین بزرگی باقیمانده ها را اندازه گیری می کند و تخمینی از دقت پیش بینی مدل ارائه می دهد.
R-squared یک معیار آماری است که نشان دهنده نسبت واریانس در متغیر وابسته است که توسط متغیرهای مستقل در یک مدل رگرسیونی توضیح داده می شود. از ۰ تا ۱ متغیر است، که ۰ نشان دهنده عدم وجود رابطه خطی بین متغیرها و ۱ نشان دهنده رابطه خطی کامل است.
Adjusted R-squared نسخه اصلاح شده R-squared است که تعداد پیش بینی کننده ها را در مدل تنظیم می‌کند و معیار قابل اعتمادتری از برازش خوب ارائه می‌کند. می‌تواند منفی نیز باشد.

نقشه مقایسه میزان ضرایب:



- بزرگ بودن ضریب یک متغیر، می‌تواند نشان‌دهنده تاثیر بالای آن باشد. با این حال، توجه به این نکته ضروری است که بزرگی یک ضریب لزوماً نشان دهنده اهمیت بالای آن نیست.

در یک مدل رگرسیونی، ضریب متغیر، میزان تغییر متغیر پاسخ با یک واحد تغییر در متغیر پیشگوی مربوطه با ثابت نگه داشتن سایر متغیرهاست. اهمیت یک ضریب به عوامل مختلفی از جمله مقیاس متغیرها و اهمیت آماری ضریب نیز بستگی دارد.

هنگامی که مقیاس متغیرها یکسان است، ضرایب نشان دهنده تغییر در متغیر وابسته مرتبط با تغییر یک واحدی در متغیر مستقل مربوطه است. در این مورد، بزرگی ضرایب را می‌توان به طور مستقیم برای ارزیابی اهمیت نسبی متغیرها مقایسه کرد.

- MSE داده‌های تست بسیار بیشتر از داده‌های ترین است و هر چند R-squared داده‌های تست و ترین به قابل قبول و نزدیک به هم هستند؛ adjusted R-squared داده‌های تست به طرز غیرقابل قبولی پایین‌تر از داده‌های ترین است. (کمتر از ۰.۷ است) می‌توان بررسی کرد که آیا ضرایب متغیرها از نظر آماری معنادار و با انتظارات ما همسو هستند؟ و آیا روابط بین متغیرها شهودی و منطقی است؟ همچنین؛ تعداد متغیرهای پیشگو بسیار زیاد است و می‌توان با حذف برخی متغیرهایی که با متغیر پاسخ رابطه معناداری ندارند به نتایج بهتری رسید. همچنین می‌توان از متغیرهایی که اثر مستقیم روی یکدیگر دارند یکی را حذف کرد و متغیرهای پیشگو را تا جای ممکن به داده‌هایی که استقلال خطی دارند، نزدیک کرد.

- **انتخاب ویژگی و تحلیل:**

ابتدا متغیرهایی که باعث ایجاد ضرایب NA می‌شدند را حذف کردم و سپس تمامی ضرایبی که p-value آن‌ها بیشتر از ۵ درصد بود را حذف کردم و در نهایت به ضرایب زیر رسیدم:

```
Call:
lm(formula = filtered_train_data$price ~ ., data = filtered_train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-6457.8 -1528.6  -274.1  1466.4  6656.7

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -5.706e+04  1.194e+04  -4.778 4.65e-06 ***
carlength          -9.938e+01  4.526e+01  -2.196 0.029840 *
carwidth            1.314e+03  2.292e+02   5.735 6.34e-08 ***
curbweight          2.915e+00  9.212e-01   3.164 0.001930 **
enginesize          1.301e+02  1.147e+01  11.339  < 2e-16 ***
boreratio          -3.184e+03  1.085e+03  -2.935 0.003933 **
stroke             -3.060e+03  7.971e+02  -3.839 0.000191 ***
peakrpm             1.296e+00  5.076e-01   2.552 0.011843 *
drivewheelrwd       1.568e+03  6.282e+02   2.496 0.013786 *
enginelocationfront -9.301e+03  2.152e+03  -4.323 3.01e-05 ***
enginetypeohcv      -5.142e+03  1.198e+03  -4.293 3.39e-05 ***
enginetyperotor      4.592e+03  1.557e+03   2.950 0.003761 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2525 on 132 degrees of freedom
Multiple R-squared:  0.9053,    Adjusted R-squared:  0.8974
F-statistic: 114.7 on 11 and 132 DF,  p-value: < 2.2e-16
```

حال داریم:

```
[1] "the filtered train RSS is: " "841582615.333699"
[1] "the filtered test RSS is: " "608129802.31699"

[1] "the filtered train TSS is: " "8883260079.47222"
[1] "the filtered test TSS is: " "4134992818.38784"

[1] "the filtered train MSE is: " "5844323.71759513"
[1] "the filtered test MSE is: " "9969341.02158999"

[1] "the filtered train R-squared is: " "0.905261963760527"
[1] "the filtered test R-squared is: " "0.852930868558536"

[1] "the filtered train R-adjusted is: " "0.898138803141018"
[1] "the filtered test R-adjusted is: " "0.823517042270244"
```

با توجه به متغیر R-adjusted می‌توان دید که به نتیجه بهتری رسیده‌ایم.

- Anova تست را ران می‌کنیم و ده ویژگی با بیشترین F-value را خروجی می‌دهیم.

```
[1] "wheelbase"          "carlength"          "carwidth"           "cylindernumber"
[5] "curbweight"         "enginesize"         "enginelocationfront" "stroke"
[9] "horsepower"         "carheight"
```

- در تحلیل رگرسیون خطی، می توان یک متغیر تعاملی بین متغیرهای مورد نظر برای ارزیابی اثر مشترک آنها گنجاند. اگر ضریب متغیر تعامل از نظر آماری معنی دار باشد، نشان دهنده وجود هم افزایی است.

```{r}
attach(train_data)
synergy_data <- cbind(train_data,
                      stroke_boreratio = stroke*boreratio,
                      compression_boreratio = compressionratio*boreratio,
                      wheelbase_carlength = wheelbase*carlength,
                      enginesize_curbweight = enginesize*curbweight,
                      enginesize_horsepower = enginesize*horsepower,
                      carheight_carwidth = carheight*carwidth,
                      citympg_horsepower = citympg*horsepower,
                      citympg_highwaympg = citympg*highwaympg,
                      carwidth_enginesize = carwidth*enginesize,
                      cylinder_horsepower = cylindernumber*horsepower)

synergy_model <- lm(price ~ . , data = synergy_data)
summary(synergy_model)
```

- مدل‌های دیگر، می‌توان مدلی پیشنهاد داد که ln(price) را برحسب متغیرهای پاسخ به دست بیاورد. برای اینکار باید از کل داده‌های ستون قیمت ln بگیریم و رگرسیون خطی را روی آن برازش کنیم. مدل‌های دیگری نیز می‌توان پیشنهاد داد. مثلا می‌توان از کل داده‌ها لگاریتم گرفت یا تمام آن‌ها را به توان ۲ رساند و تست کرد تا مشاهده کنیم کدام یک متغیر پاسخ را با دقت بالاتری پیش‌بینی می‌کند.

ابتدا از همه متغیرهای پاسخ ln می‌گیریم و سپس رگرسیون خطی را روی آن فیت می‌کنیم و داریم:

```
[1] "the train R-adjusted is: " "0.911629077950006"
[1] "the test R-adjusted is: " "0.723811511479459"
```