ISLR-R: Appendix D. Introduce ggplot2

2023-03-09

An Example

Check dataset summary:

```
Boston %>% summary()
```

```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm              age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00
```

Scatter plot with linear smoothing:

```
g <- Boston %>%
  dplyr::mutate(safe = ifelse(crim < 0.08, T, F)) %>%
  dplyr::mutate(new = ifelse(age < 45, T, F)) %>%
  ggplot(data = ., mapping = aes(x = lstat, y = medv)) +
  geom_point(
    mapping = aes(color = new %>% factor(), shape = safe %>% factor()),
```

```
    size = 2.0, alpha = 0.4,
  ) +
  scale_color_manual(values = c("#00AFBB", "#FC4E07", "#e7b800")) +
  scale_shape_manual(values = c(1, 19, 24)) +
  geom_smooth(
    method = "lm", formula = y ~ x, color = "firebrick",
    linewidth = 1.5,
  ) +
  labs(
    title = "Medium Value vs. Low Percentage", subtitle = "from boston dataset",
    caption = "figure 1. Linear Regression",
    x = "Low Percentage (%)", y = "Medium Value ($ 1,000)",
    color = "New", shape = "Safe"
  ) +
  theme_grey()

g + labs(tag = "original")
```

original

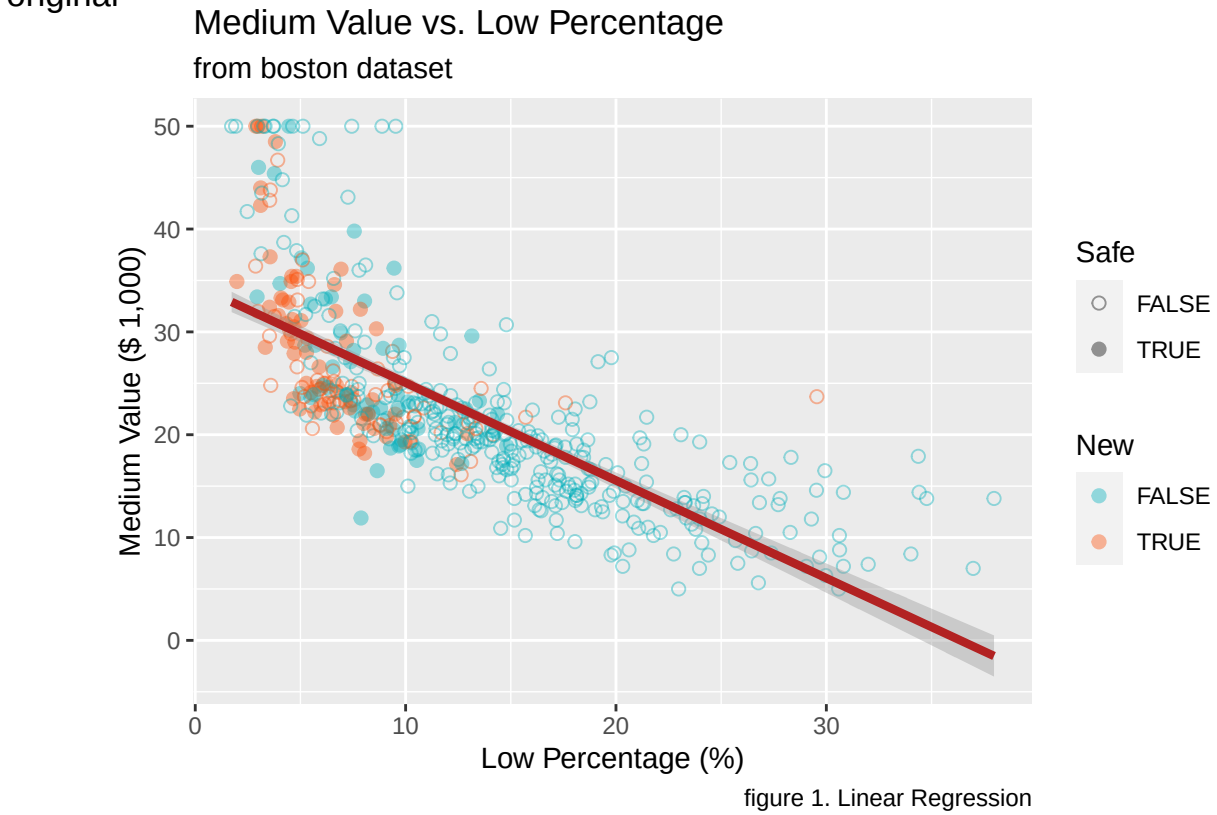


figure 1. Linear Regression

```
g + coord_cartesian(ylim = c(0, 30)) + labs(tag = "cropped")
```

cropped

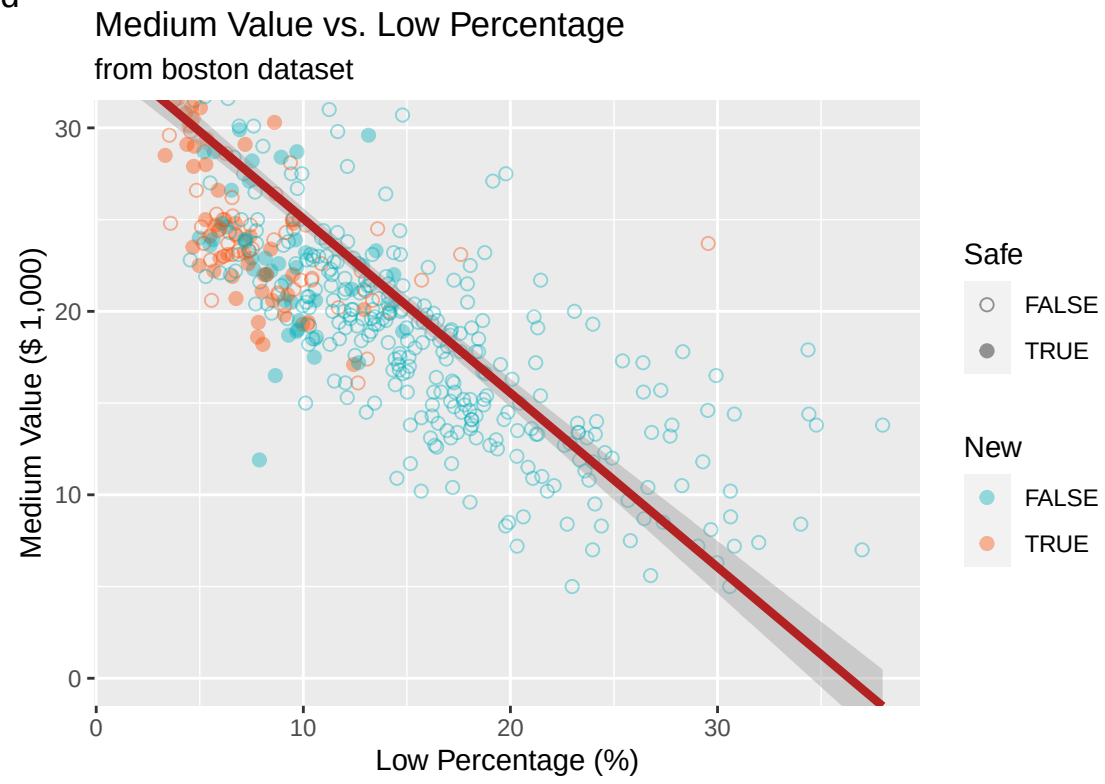
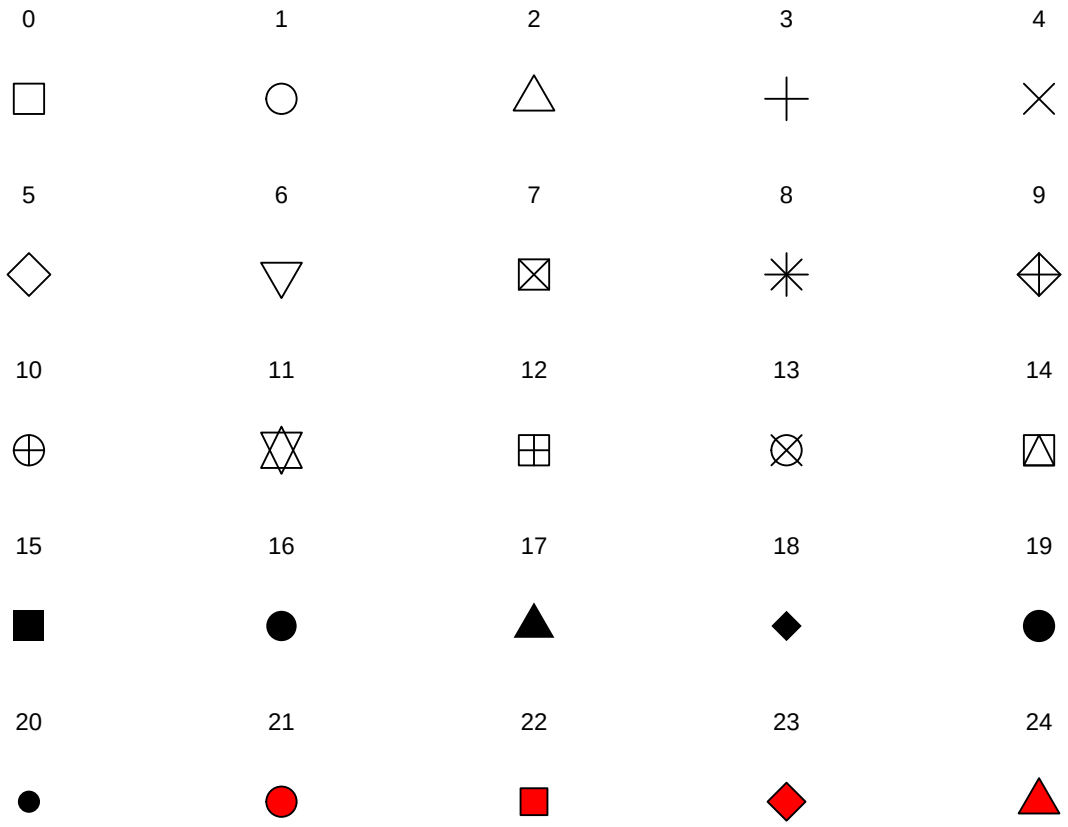


figure 1. Linear Regression

details

shapes

```
data.frame(shape = 0:24) %>% ggplot(data = ., mapping = aes(0, 0)) +
  geom_point(mapping = aes(shape = shape), size = 5, fill = "red") +
  facet_wrap(~shape) +
  scale_shape_identity() +
  theme_void()
```



Deja vu

```
Auto %>% summary()
```

```
##       mpg          cylinders      displacement     horsepower        weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
##   acceleration        year           origin          name
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   Length:392
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   Class :character
##  Median :15.50   Median :76.00   Median :1.000   Mode  :character
##  Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :24.80   Max.   :82.00   Max.   :3.000
```

Continous vs. Discrete

Box Plot:

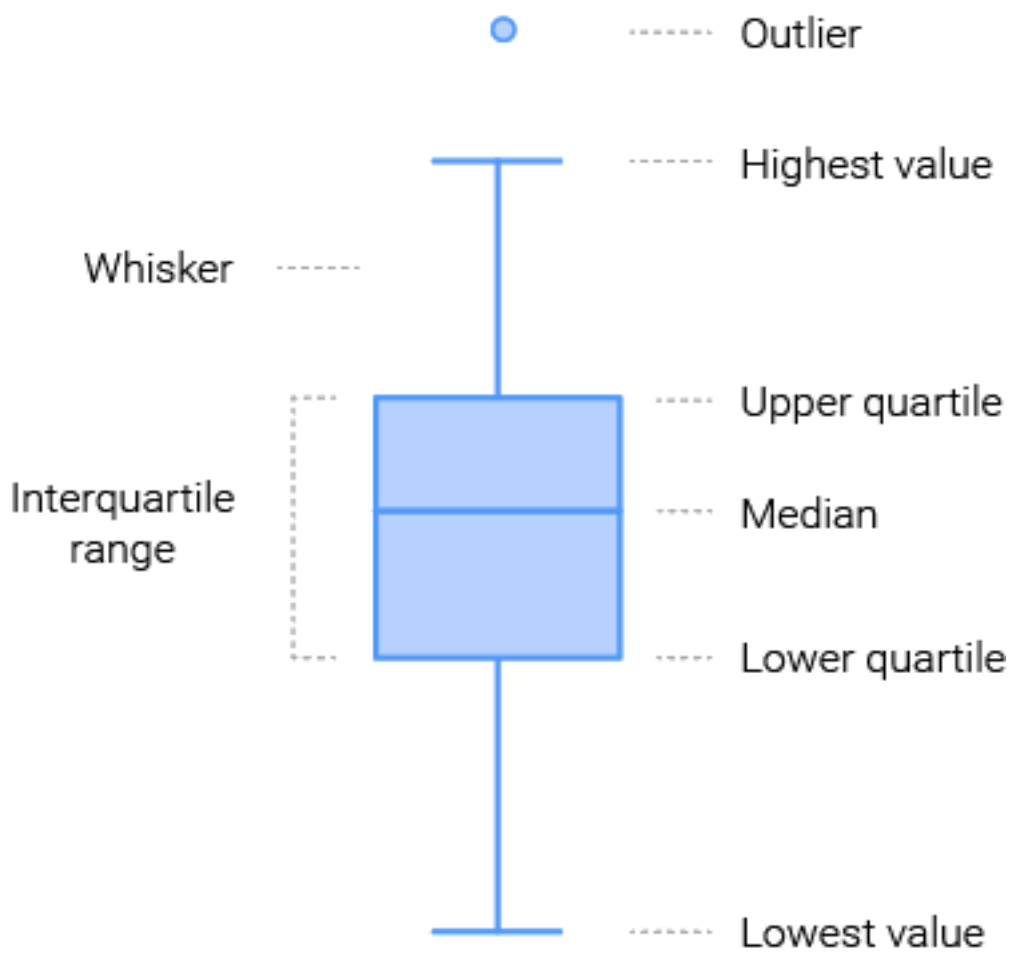


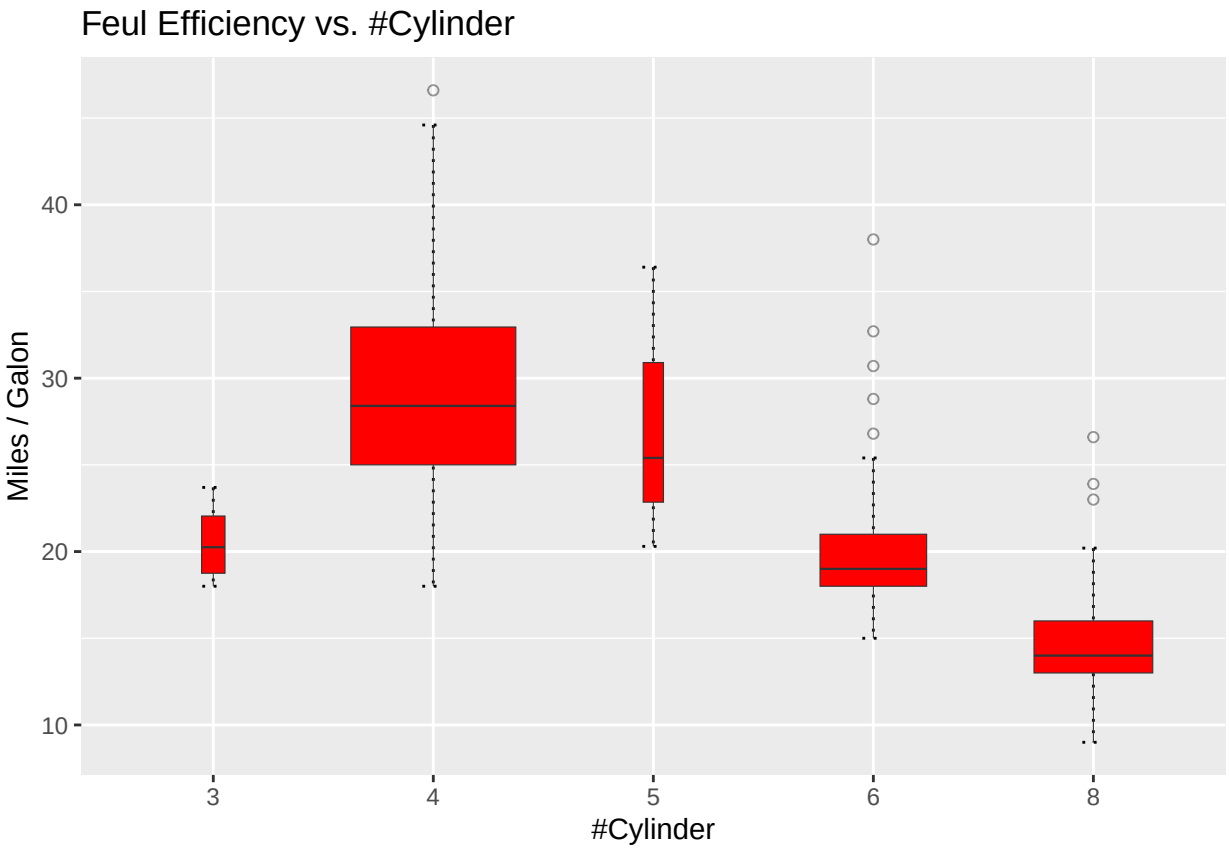
Figure 1: Fig 1. Boxplot Structure

TL;DR

$Q_1 = \mu - 0.675\sigma$

$Q_3 = \mu + 0.675\sigma$

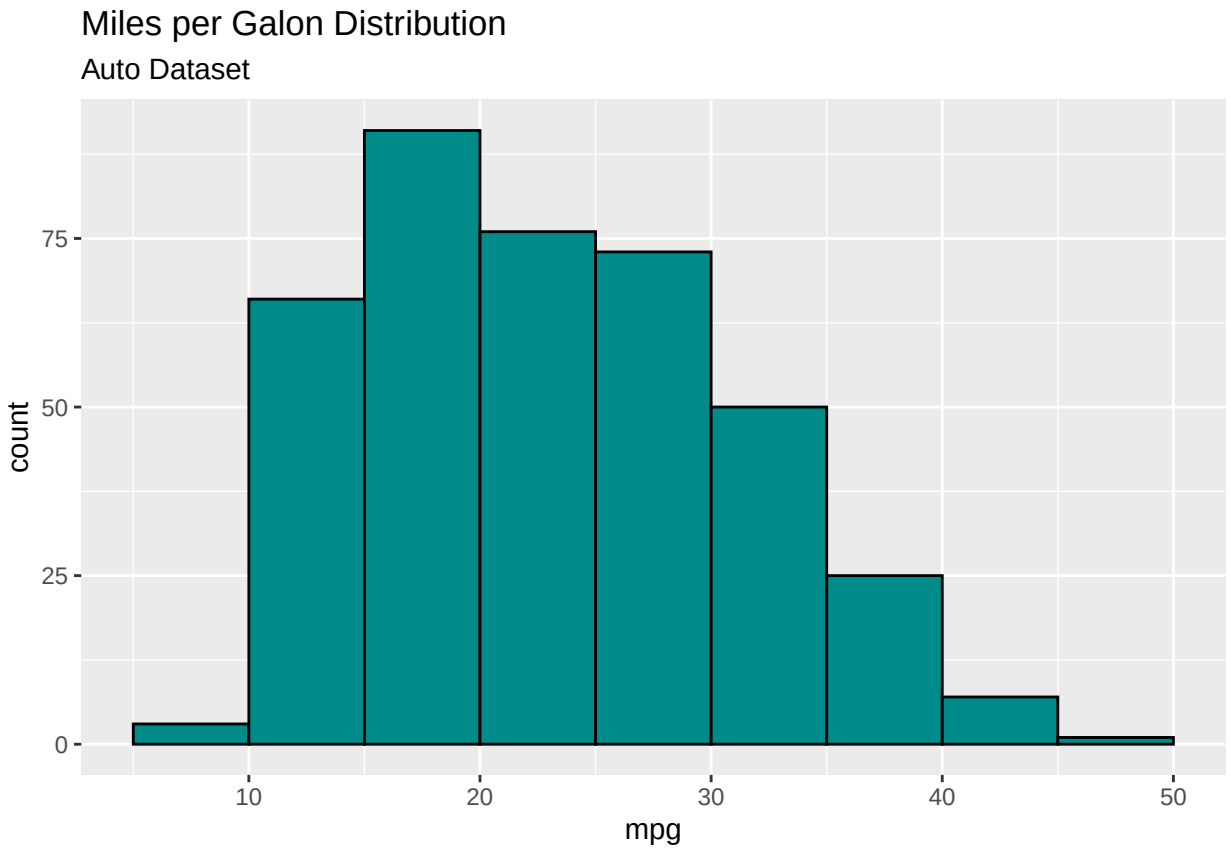
```
Auto %>%
  dplyr::mutate(new = ifelse(year > 79, T, F)) %>%
  ggplot(data = ., mapping = aes(x = cylinders %>% factor(), y = mpg)) +
  stat_boxplot(
    geom = "errorbar",
    # position = "dodge2",
    linetype = "dotted",
    width = 0.1,
    coef = 1.5,
  ) +
  geom_boxplot(
    linewidth = 0.2,
    shape = "dotted",
    fill = "red",
    outlier.stroke = 0.5,
    outlier.alpha = 0.5,
    outlier.shape = 1,
    varwidth = T, show.legend = T,
    # notch = F, notchwidth = 0.5,
    # coef = 1.5,
    # width = 0.8,
    # width.errorbar = 0.5,
  ) +
  theme_grey() +
  labs(
    title = "Feul Efficiency vs. #Cylinder", x = "#Cylinder",
    y = "Miles / Galon"
  )
```



One Continous

Histogram

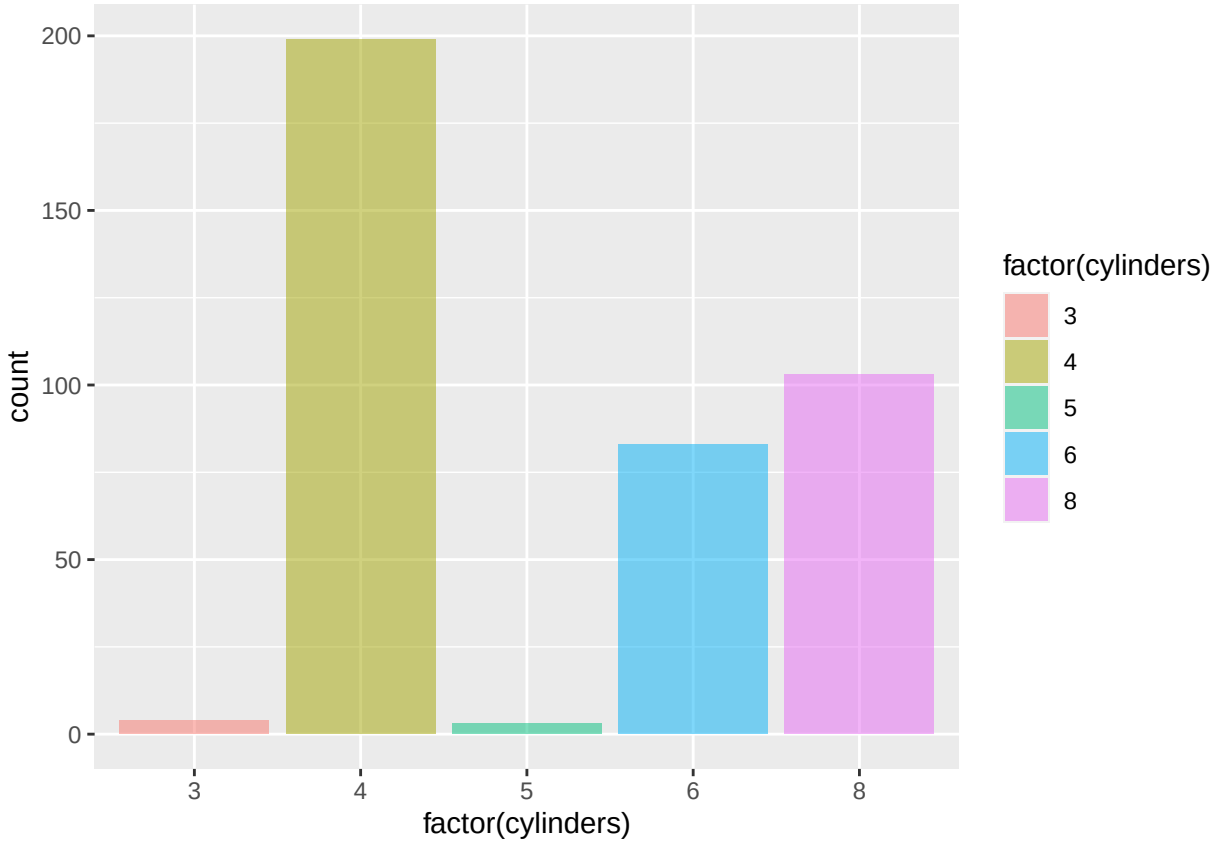
```
Auto %$% (ggplot(mapping = aes(x = mpg)) +
  geom_histogram(
    colour = 1,
    fill = "darkcyan",
    breaks = pretty(range(mpg),
      n = nclass.Sturges(mpg), min.n = 1
    )
  )) +
  labs(
    title = "Miles per Galon Distribution",
    subtitle = "Auto Dataset"
  )
```



One Discrete

Bar Chart

```
Auto %>% ggplot(data = .) +
  geom_bar(
    mapping = aes(
      x = factor(cylinders),
      fill = factor(cylinders)
    ),
    alpha = 0.5,
  ) +
  scale_color_brewer()
```



Linear Models

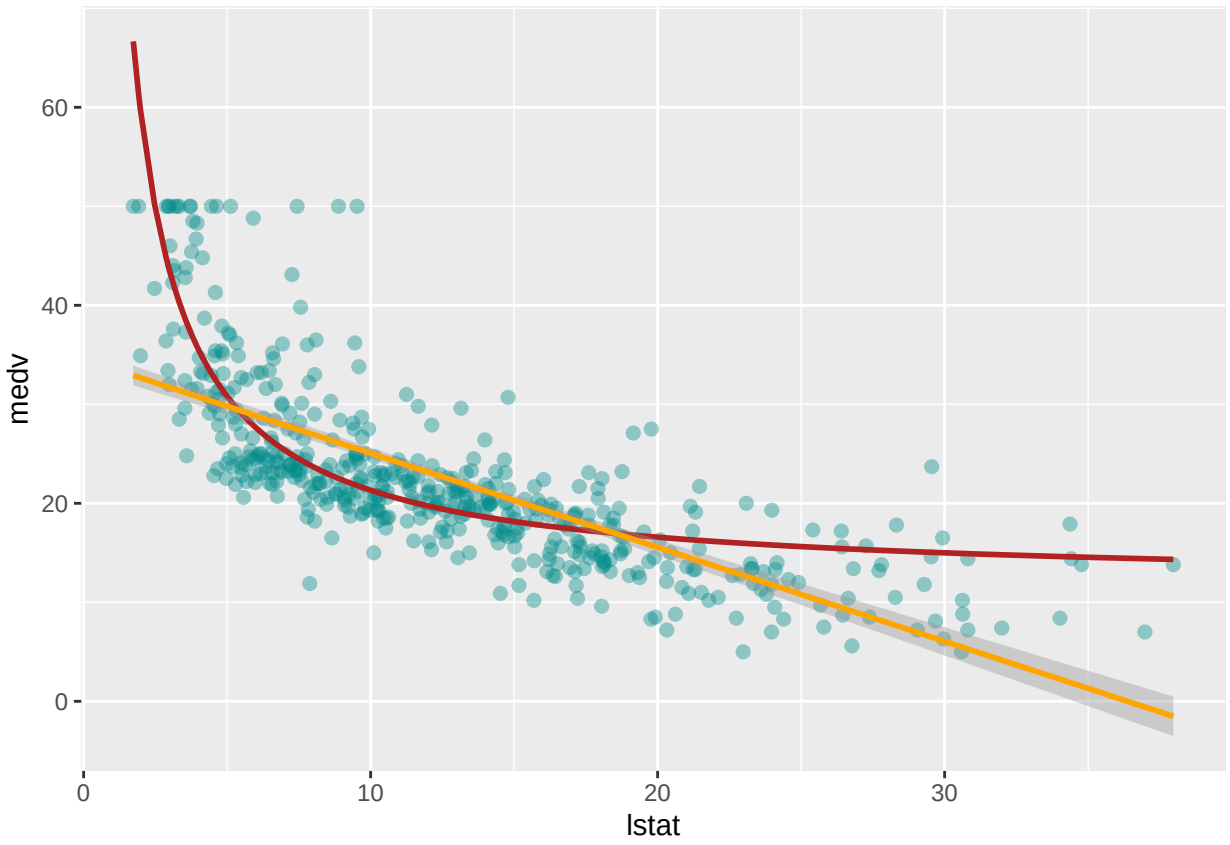
```
mdl <- Boston %$% lm(medv ~ I(1 / lstat))
```

Create predictions with the model above:

```
data <- Boston %>% dplyr::mutate(pred = predict(mdl))
```

Model Visualization:

```
data %>%
  ggplot(data = ., mapping = aes(x = lstat, y = medv)) +
  geom_point(
    mapping = aes(x = lstat, y = medv),
    size = 2.0, alpha = 0.4, color = "darkcyan"
  ) +
  geom_line(
    mapping = aes(x = lstat, y = pred),
    color = "firebrick", linewidth = 1.0,
  ) +
  geom_smooth(
    method = "lm", formula = y ~ x, color = "orange",
    linewidth = 1.0,
  )
```



More tricks:

```
data %>%
  dplyr::mutate(safe = ifelse(crim < 0.08, T, F)) %>%
  dplyr::mutate(new = ifelse(age < 45, T, F)) %>%
  ggplot(data = .) +
  geom_point(
    mapping = aes(
      x = lstat, y = medv, color = new %>% factor(),
      shape = safe %>% factor()
    ),
    size = 2.0, alpha = 0.4,
  ) +
  scale_color_manual(values = c("#00AFBB", "#FC4E07", "#e7b800")) +
  scale_shape_manual(values = c(1, 19, 24)) +
  geom_line(
    mapping = aes(x = lstat, y = pred), color = "firebrick",
    linewidth = 1.5,
  ) +
  labs(
    title = "Medium Value vs. Low Percentage",
    subtitle = "from boston dataset",
    caption = "figure 1. Linear Regression",
    x = "Low Percentage (%)", y = "Medium Value ($ 1,000)",
    color = "New", shape = "Safe"
  ) +
```

```
theme_grey()
```

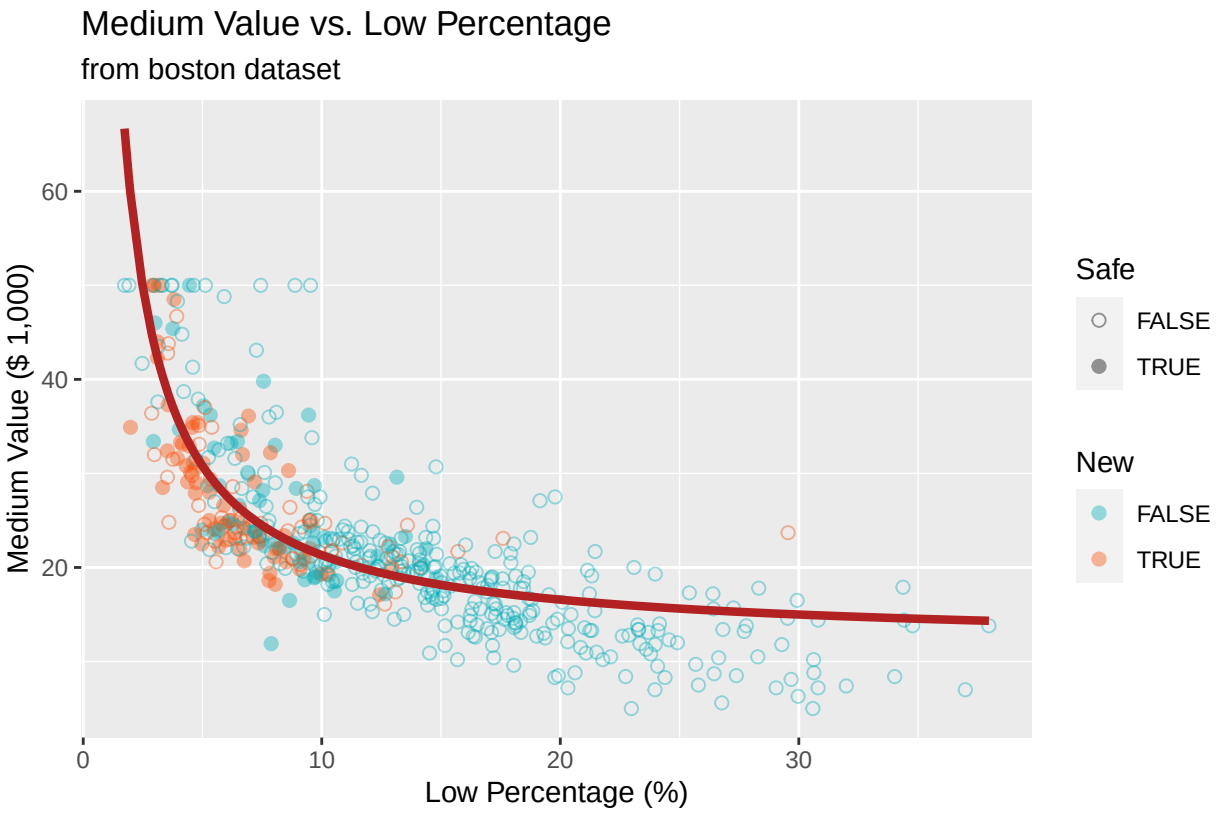


figure 1. Linear Regression

Reference

- http://r-statistics.co/Complete-Ggplot2-Tutorial-Part1-With-R-Code.html
- https://towardsdatascience.com/why-1-5-in-iqr-method-of-outlier-detection-5d07fdc82097