Exploratory Data Analysis (EDA) by ggplot2

Mr.Therarat Srisaswatakul

Project preparation before start analysis

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

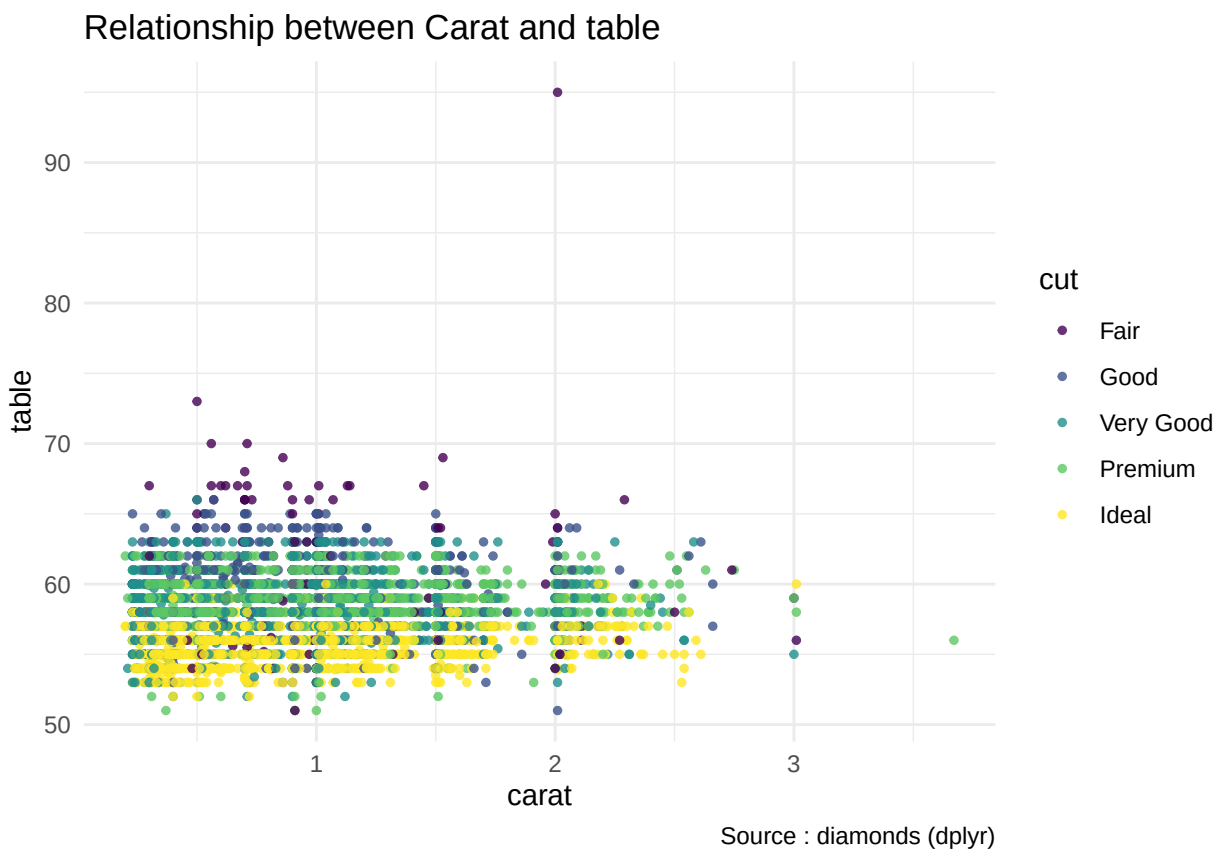
Data overview

```
glimpse(diamonds)
```

```
## Rows: 53,940
## Columns: 10
## $ carat   <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, 0.23, 0.~
## $ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very Good, Ver~
## $ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, J, J, I,~
## $ clarity <ord> SI2, SI1, VS1, VS2, SI2, VVS2, VVS1, SI1, VS2, VS1, SI1, VS1, ~
## $ depth   <dbl> 61.5, 59.8, 56.9, 62.4, 63.3, 62.8, 62.3, 61.9, 65.1, 59.4, 64~
## $ table   <dbl> 55, 61, 65, 58, 58, 57, 57, 55, 61, 61, 55, 56, 61, 54, 62, 58~
## $ price   <int> 326, 326, 327, 334, 335, 336, 336, 337, 337, 338, 339, 340, 34~
## $ x       <dbl> 3.95, 3.89, 4.05, 4.20, 4.34, 3.94, 3.95, 4.07, 3.87, 4.00, 4.~
## $ y       <dbl> 3.98, 3.84, 4.07, 4.23, 4.35, 3.96, 3.98, 4.11, 3.78, 4.05, 4.~
## $ z       <dbl> 2.43, 2.31, 2.31, 2.63, 2.75, 2.48, 2.47, 2.53, 2.49, 2.39, 2.~
```

1. Compare the Carat and Table

```
set.seed(40)
ggplot(sample_n(diamonds, 10000),
       aes(carat,
           table,
           col = cut)) +
  geom_point(size = 1,
             alpha = 0.8) +
  labs(title = "Relationship between Carat and table",
       caption = "Source : diamonds (dplyr)") +
  theme_minimal()
```

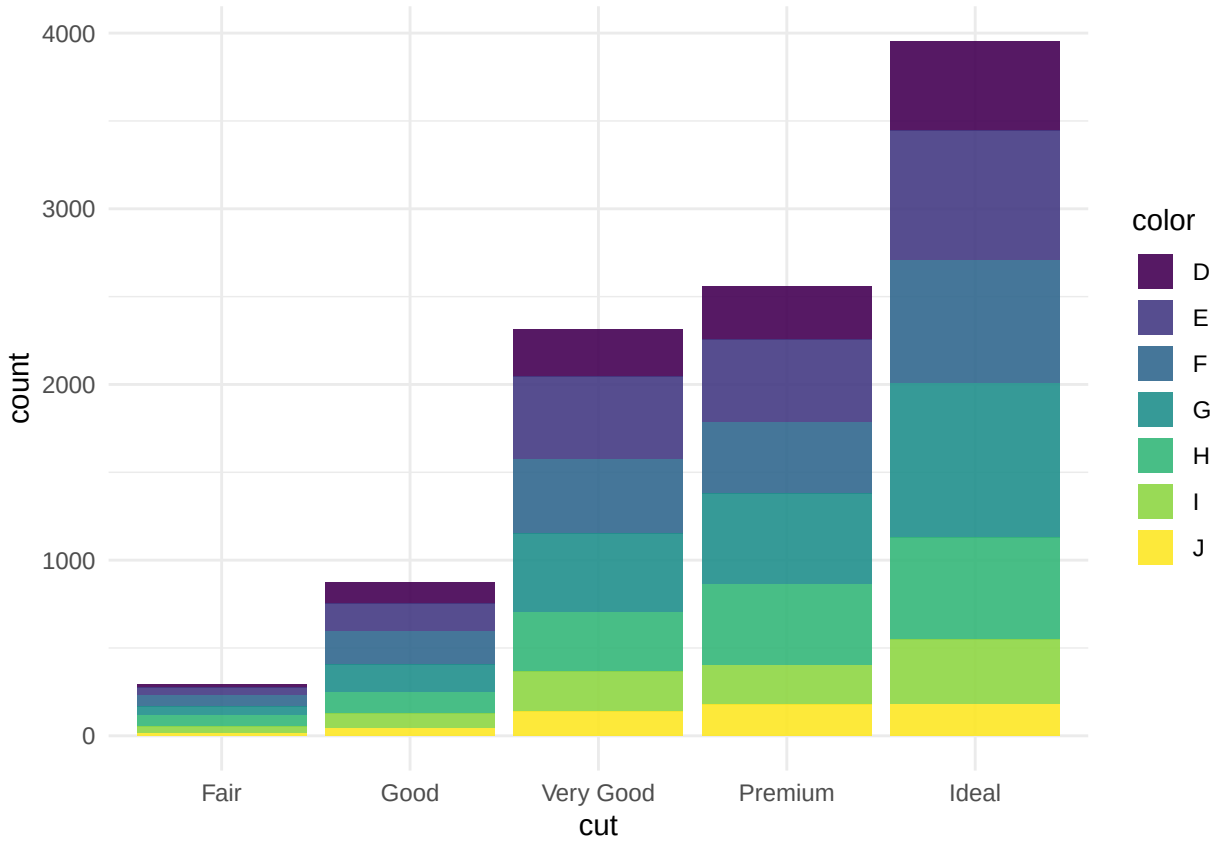


The higher of table cause the low quality of cut compare by the same carat

---

2.Count the diamond seperated by cut (Group by color)

```
set.seed(99)
sample_n(diamonds, 10000) %>%
ggplot(aes(cut, fill = color)) +
  geom_bar(size = 2, alpha = 0.9) +
  theme_minimal()
```



The ideal cut got highest quantity by sequencially from ideal to fair cut

Data Overview 2

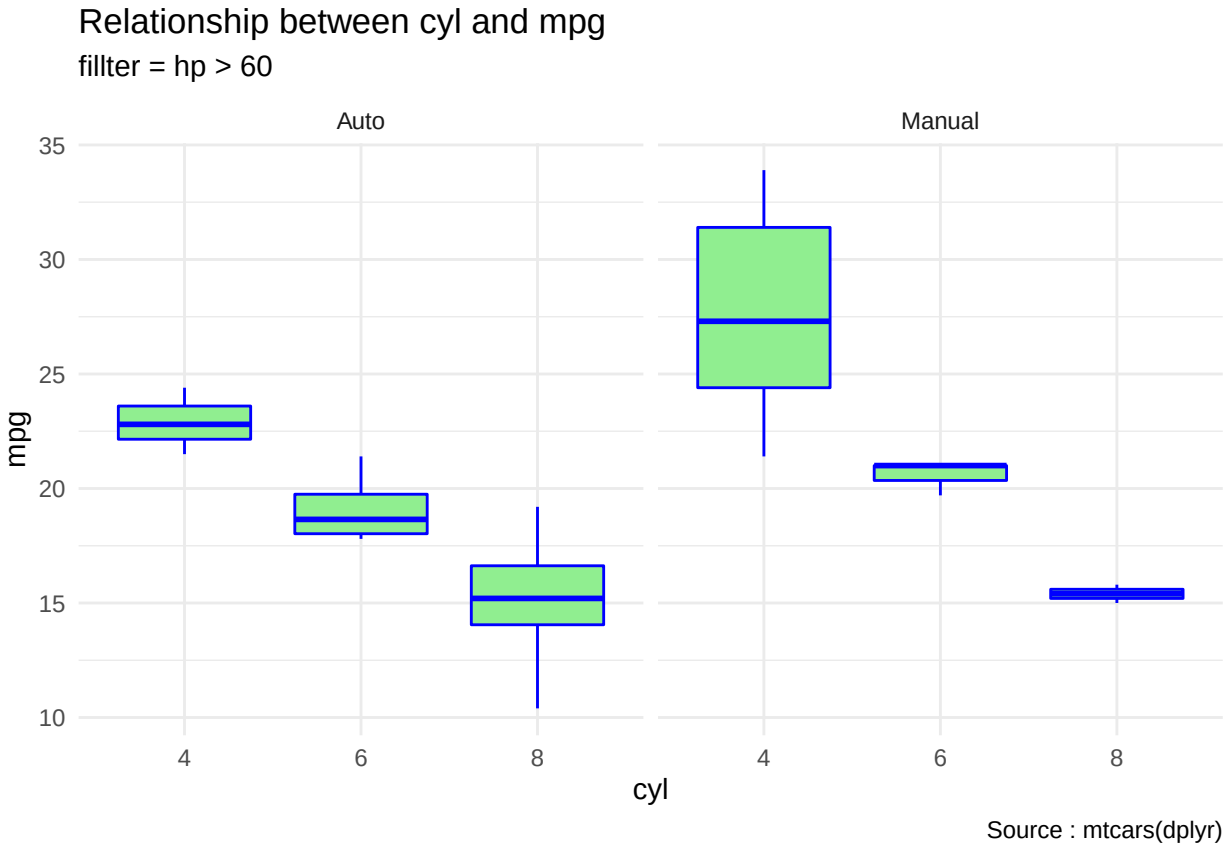
```
glimpse(mtcars)
```

```
## Rows: 32
## Columns: 11
## $ mpg  <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2, 17.8,~
## $ cyl  <dbl> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4, 4, 4, 8,~
## $ disp <dbl> 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 146.7, 140.8, 16~
## $ hp   <dbl> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, 180, 180~
## $ drat <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3.92, 3.92,~
## $ wt   <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3.150, 3.~
## $ qsec <dbl> 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 20.00, 22.90, 18~
## $ vs   <dbl> 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0,~
## $ am   <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0,~
## $ gear <dbl> 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3, 3,~
## $ carb <dbl> 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2, 1, 1, 2,~
```

3.The relation between the fuel consumption and cylinder

condition : Standard Engine in market (hp > 60)

```
mtcars %>%
  tibble %>%
  filter(hp > 60) %>%
  mutate(cyl = factor(cyl,  ## change to factor type
                      levels = c("2", "4", "6", "8", "10"),
                      labels = c("2", "4", "6", "8", "10"))) %>%
  mutate(am = factor(am,    ## change to factor type
                     levels = c("0", "1"),
                     labels = c("Auto", "Manual"))) %>%
  ggplot(aes(cyl, mpg)) +
  geom_boxplot(col = "blue", fill = "light green") +
  labs(title = "Relationship between cyl and mpg",
       caption = "Source : mtcars(dplyr)",
       subtitle = "fillter = hp > 60") +
  facet_wrap(~ am) +
  theme_minimal()
```



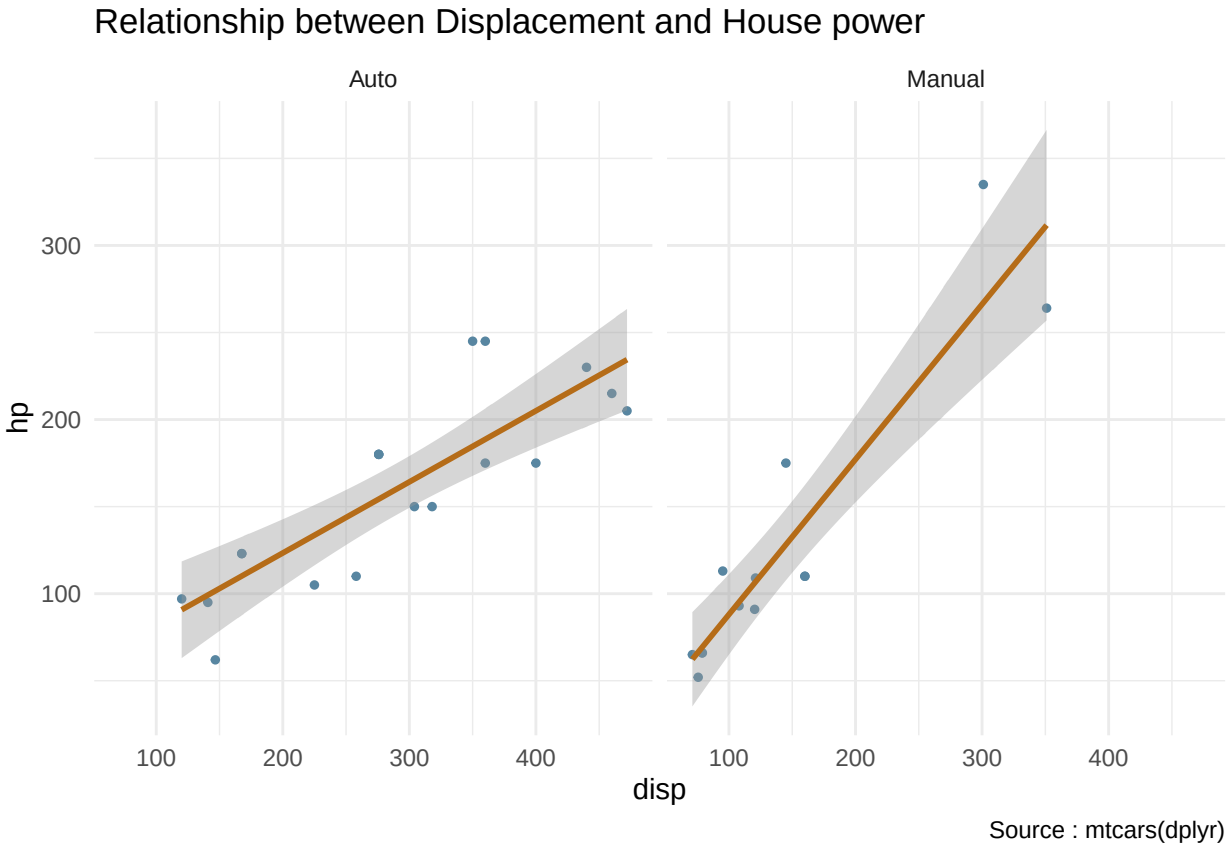
MT transmission has more miles per gallon compare with AT transmission

The more cylinder will cause less of miles per gallon

4.The relationship between disp and hp

```
mtcars %>%
  mutate(am = factor(am,
                     levels = c("0", "1"),
                     labels = c("Auto", "Manual"))) %>%
ggplot(aes(disp, hp)) +
  geom_point(size = 1, col = "#5886a1") + #hexcolor
  geom_smooth(col = "#b56c18", se = T, method = "lm") +
  labs(title = "Relationship between Displacement and House power",
       caption = "Source : mtcars(dplyr)") +
  facet_wrap(~ am) +
  theme_minimal()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The more of displacement will cause the more of house power

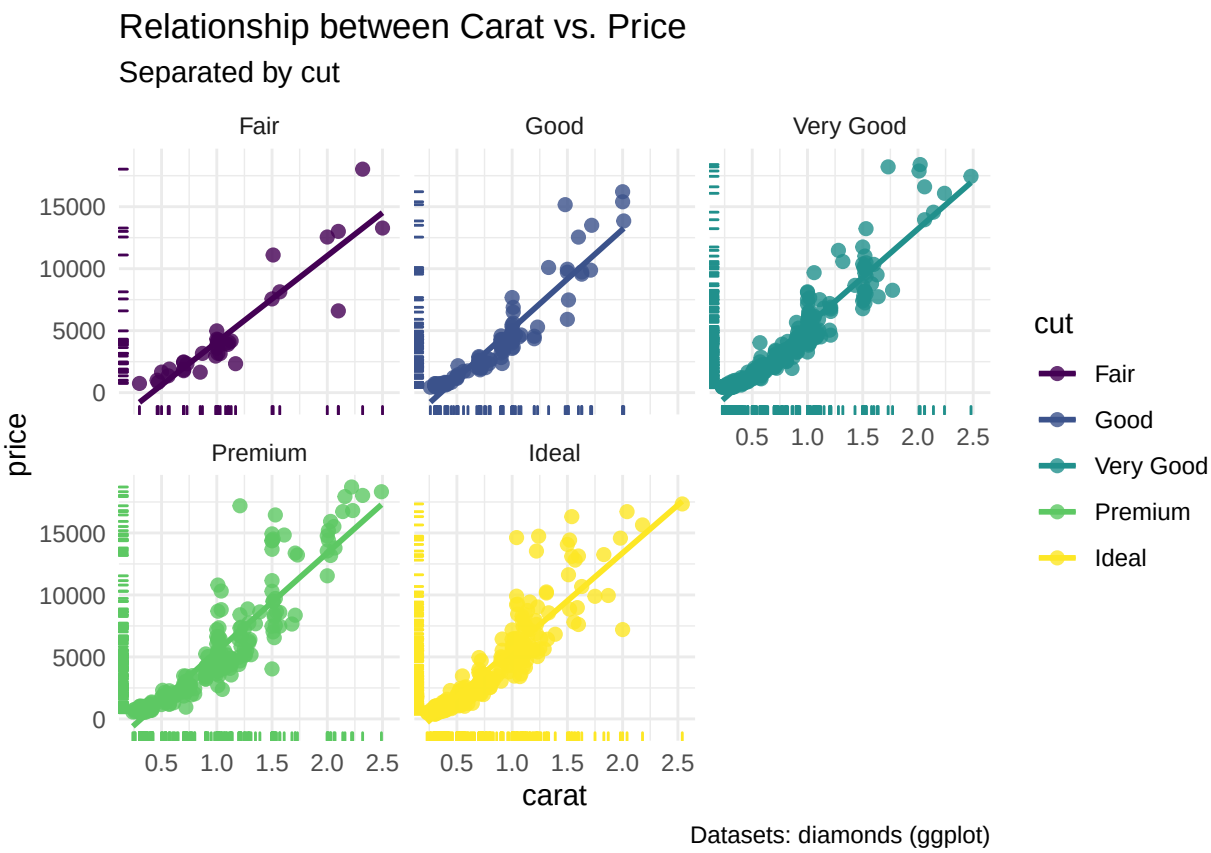
The slope between displacement and housepower of Manual transmission is greater than automatic transimission

---

5.The relationship between Carat vs. Price

```
ggplot(data = sample_n(diamonds,1000),
       aes(carat, price, color=cut)) +
  geom_smooth(method="lm",
              se = FALSE) +
  geom_point(size=2, alpha = 0.8) +
  geom_rug() +
  facet_wrap(~cut, ncol = 3) +
  theme_minimal() +
  labs( title = "Relationship between Carat vs. Price",
        subtitle = "Separated by cut",
        caption = "Datasets: diamonds (ggplot)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



##The more of carat is cause the more of price