Exploratory Data Analysis using ggplot

Warisara Yuwansiri

Dataset: Diamonds

Using Package

```
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   1.2.0      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()
```

```
library(patchwork)
library(dplyr)
library(ggplot2)
library(RColorBrewer)
```

Data Overview

```
head(diamonds)
```

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

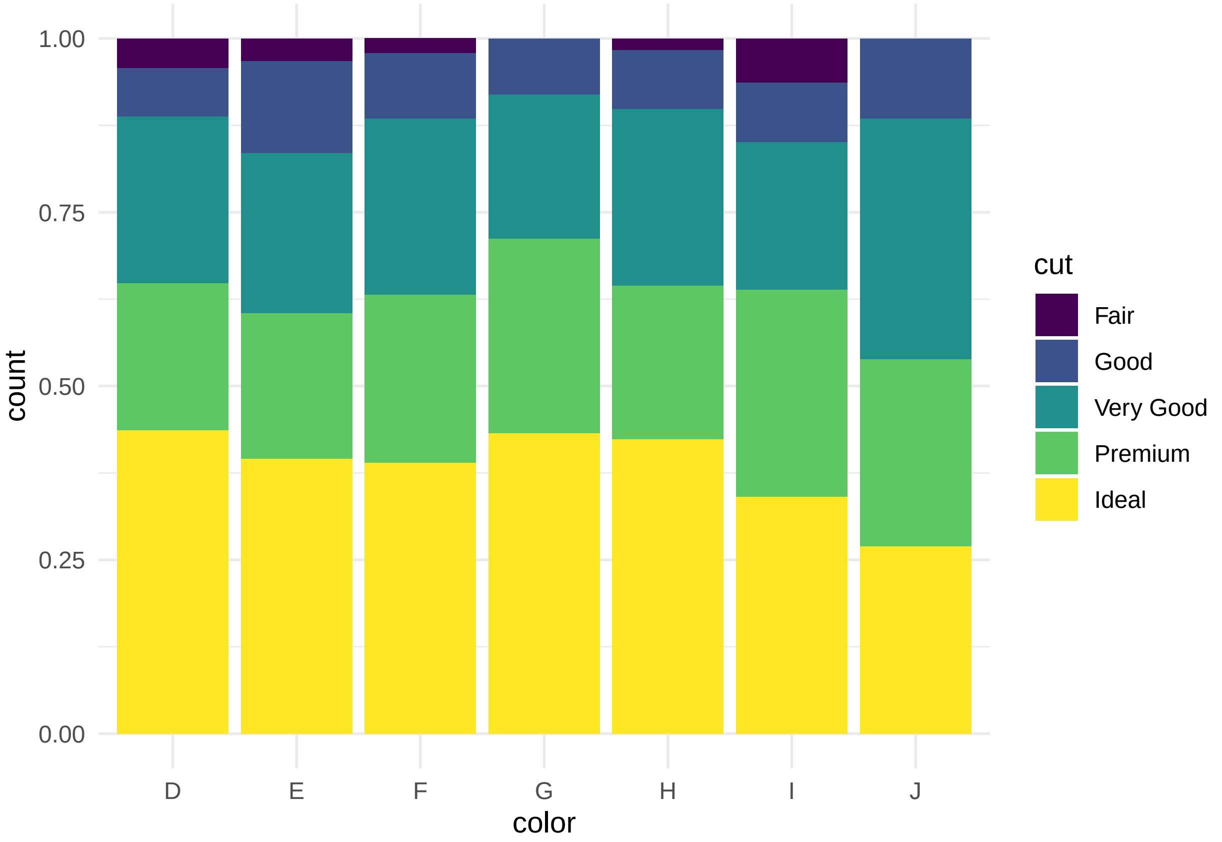
```
glimpse(diamonds)
```

```
## 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.~
```

Create New Visualization

This is a stacked bar chart to show color distribution.

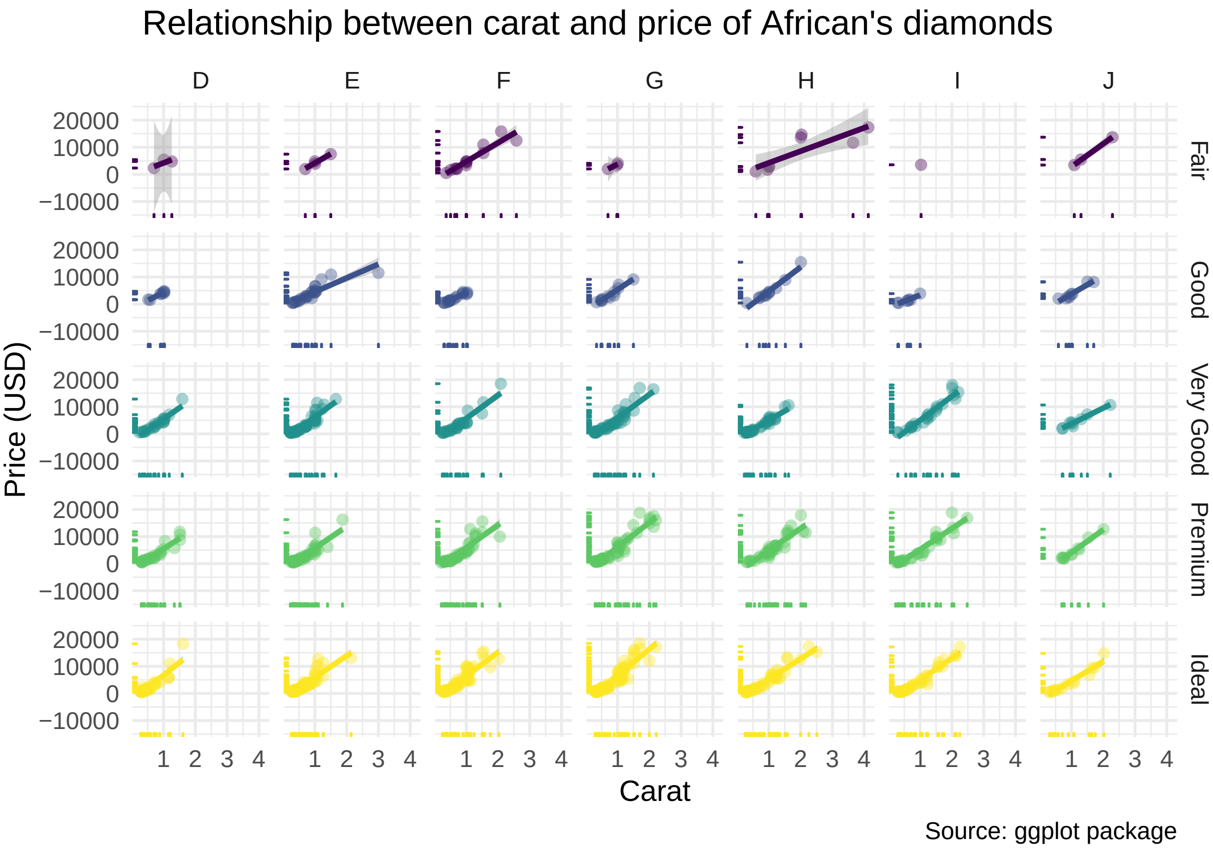
```
ggplot(sample_n(diamonds,500),aes(color , fill = cut)) +
  geom_bar(position = "fill")+
  theme_minimal()
```



This graph show about between carat and price of African's diamonds by using color for separate.

```
set.seed(45)
ggplot(sample_n(diamonds, 1000), aes(carat, price, color = cut)) +
  geom_point(alpha = 0.4)+
  geom_smooth(method = "lm")+
  theme_minimal()+
  facet_grid(cut ~ color)+
  labs(
    title = " Relationship between carat and price of African's diamonds",
    x = "Carat",
    y = "Price (USD)",
    caption = "Source: ggplot package"
  )+
  theme(legend.position = 'none')+
  geom_rug()
```

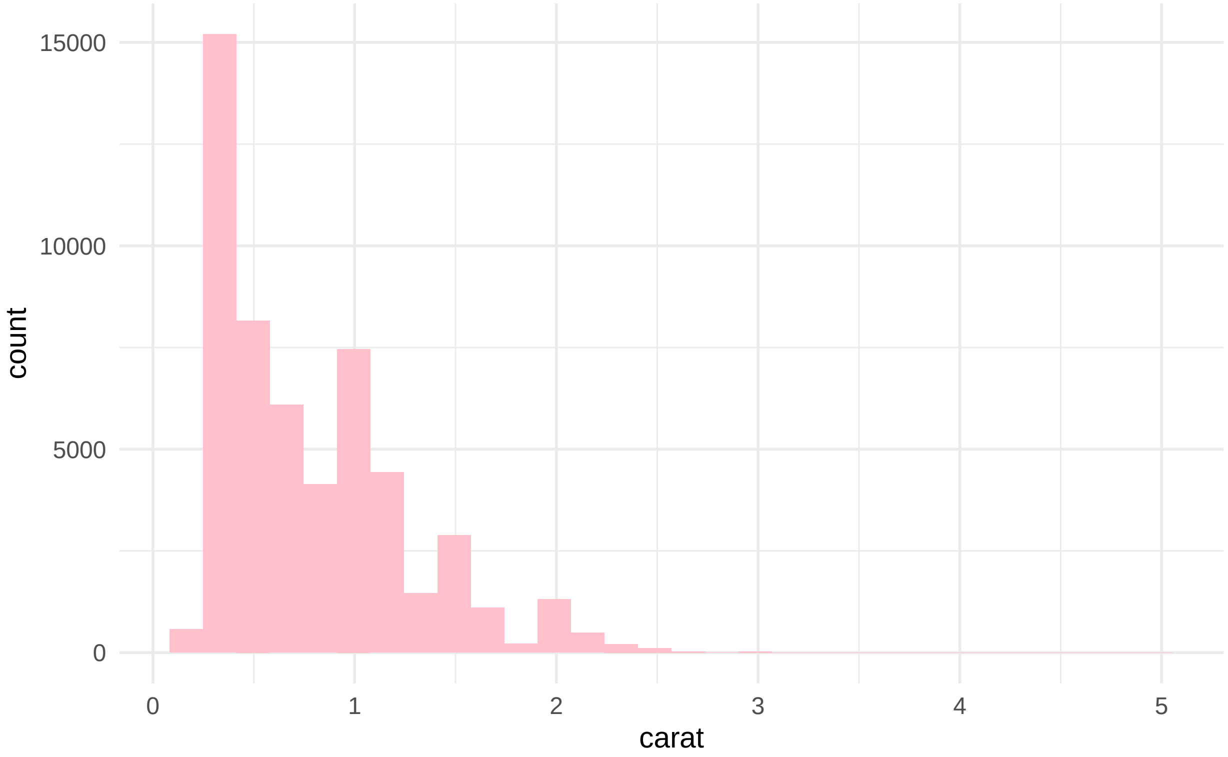
```
## `geom_smooth()` using formula 'y ~ x'
```



Histogram graph for look number of carat using single continuous variable.

```
ggplot(diamonds, mapping= aes(carat))+
  geom_histogram(bins = 30, fill = "pink")+
  theme_minimal()+
  labs(
    title = "Amount of carat",
    caption = "Datasets: diamonds from ggplot"
  )
```

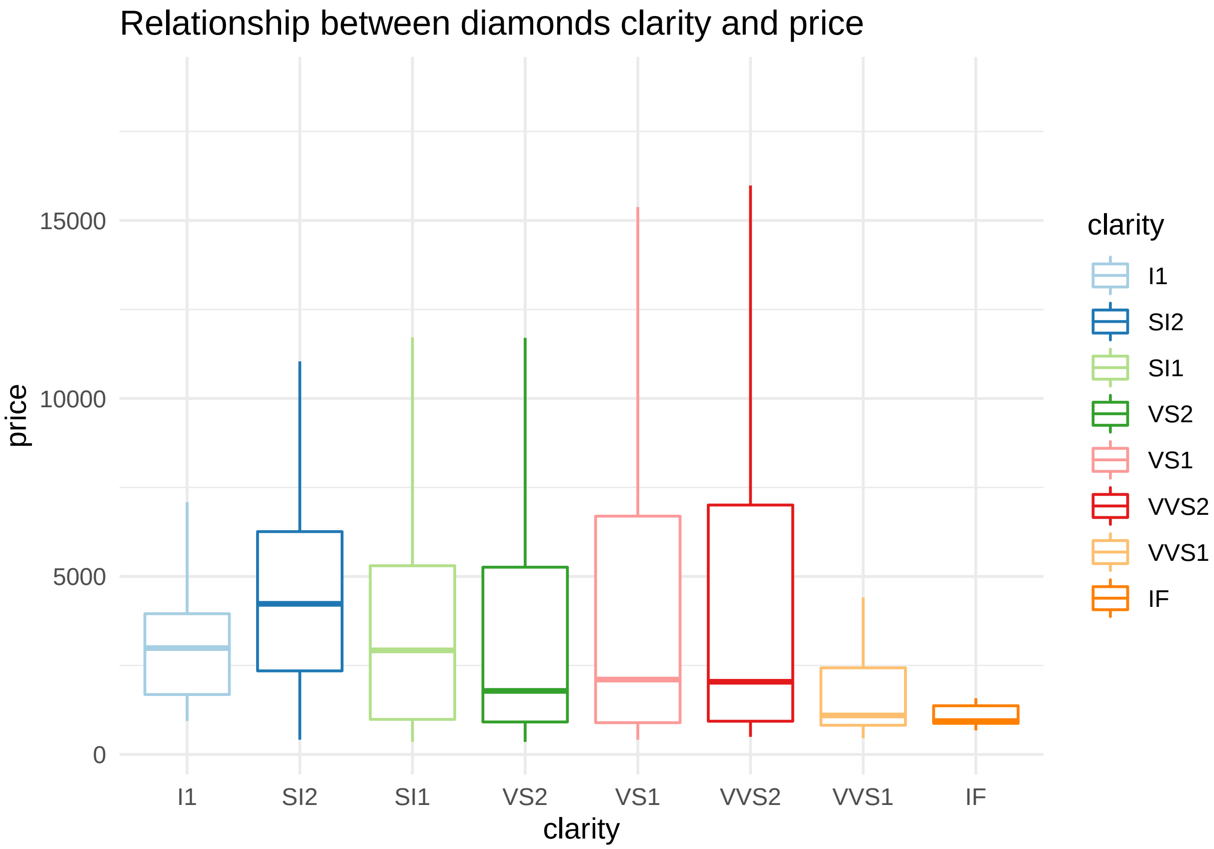
Amount of carat



Datasets: diamonds from ggplot

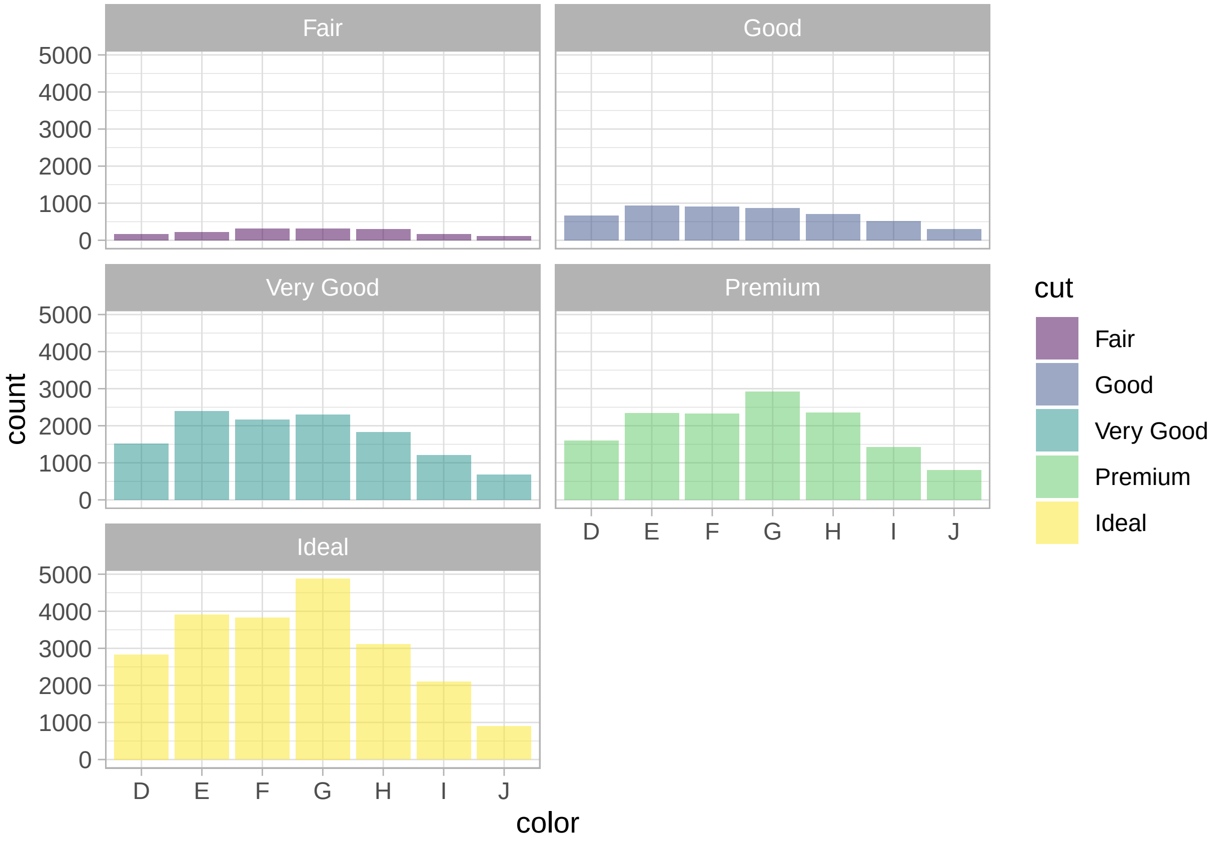
This is a boxplot graph to show relationship between diamonds clarity and price.

```
set.seed(50)
ggplot(sample_n(diamonds,1000), aes(clarity, price, color = clarity))+
  geom_boxplot(outlier.shape = NA)+
  labs(
    title = "Relationship between diamonds clarity and price",
    captiion = "Source: Diamonds dataset"
  ) +
  theme_minimal() +
  scale_color_brewer(type = "qual", palette = 3)
```



This is bar chart to show color frequency count in each cut level

```
ggplot(diamonds, aes(color, fill = cut)) +
  geom_bar(alpha = 0.5)+
  theme_light()+
  facet_wrap(~cut, ncol = 2)
```

Fair
Good
Very Good
Premium
Ideal
5000
4000
3000
2000
1000
0
count
D
E
F
G
H
I
J
color
cut
Fair
Good
Very Good
Premium
Ideal