Diamonds Exploratory Data Analysis

Diamonds EDA

The goal of this document is to explore the diamonds dataset. First, we begin by loading the important packages

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyverse
```

```
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------ tidyverse_conflicts()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Checking out the first 10 rows of the dataset

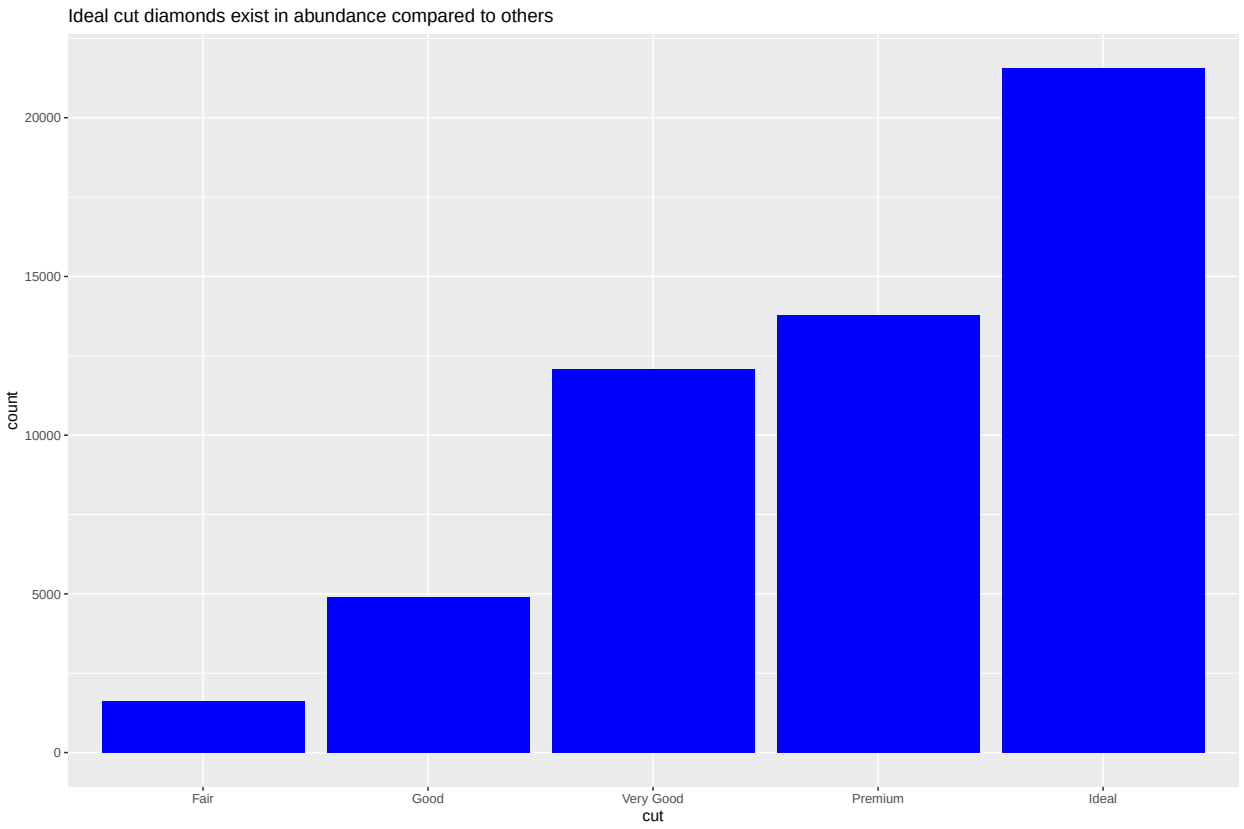
```
head(diamonds, 10)
```

```
## # A tibble: 10 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 0.290 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
##  7 0.24  Very Good I     VVS1     62.3    57   336  3.95  3.98  2.47
##  8 0.26  Very Good H     SI1      61.9    55   337  4.07  4.11  2.53
##  9 0.22  Fair      E     VS2      65.1    61   337  3.87  3.78  2.49
## 10 0.23  Very Good H     VS1      59.4    61   338  4     4.05  2.39
```

```
df <- diamonds
```

A bar chart for the diamond's cut

```
bar_cut <- ggplot(df) +
  geom_bar(aes(x = cut), fill = "blue") +
  labs(title = "Ideal cut diamonds exist in abundance compared to others", x = "cut", y = "count")
bar_cut
```



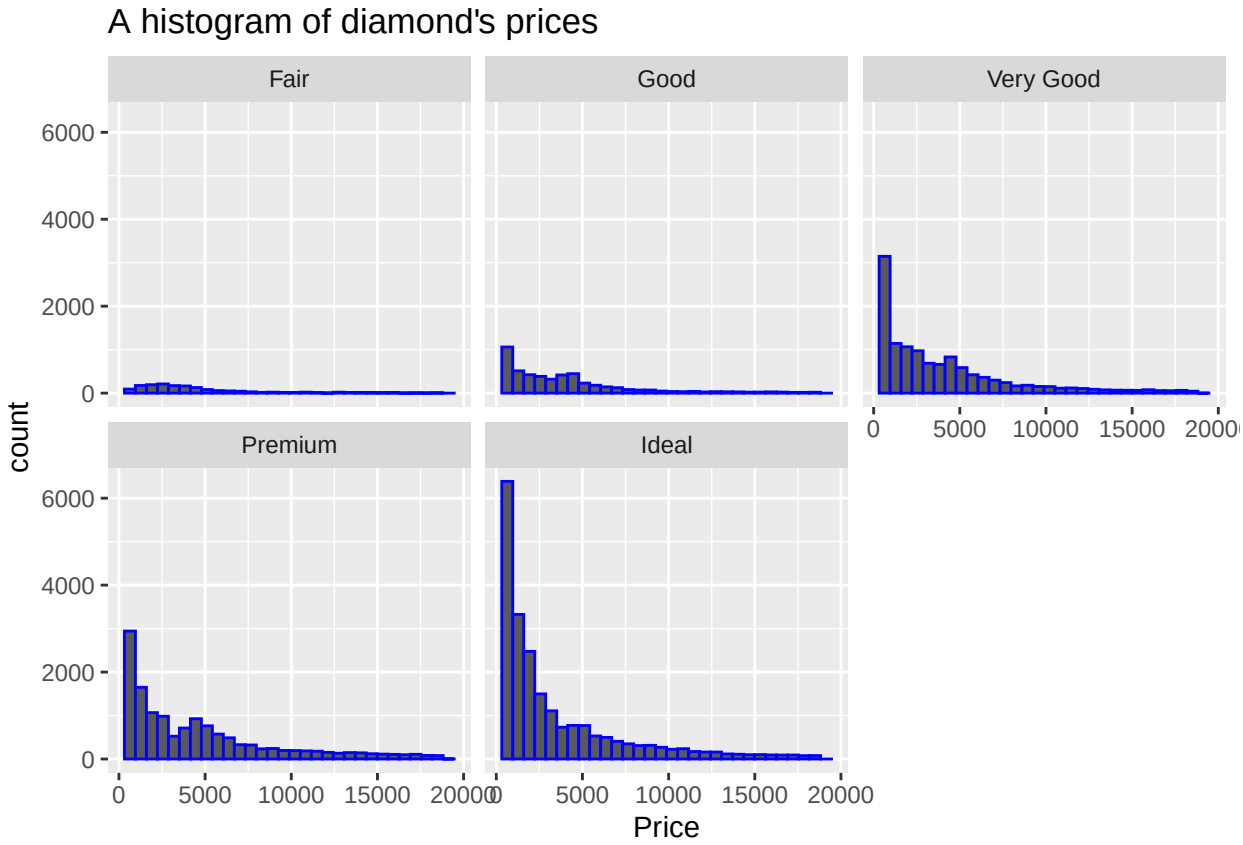

```
ggsave("bar_cut.png", bar_cut)
```

```
## Saving 12 x 8 in image
```

A histogram of diamond's price

```
ggplot(df) +
  geom_histogram(aes(x = price), color = "blue") +
  facet_wrap(~cut, nrow = 2) +
  labs(title = "A histogram of diamond's prices", x = "Price")
```

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

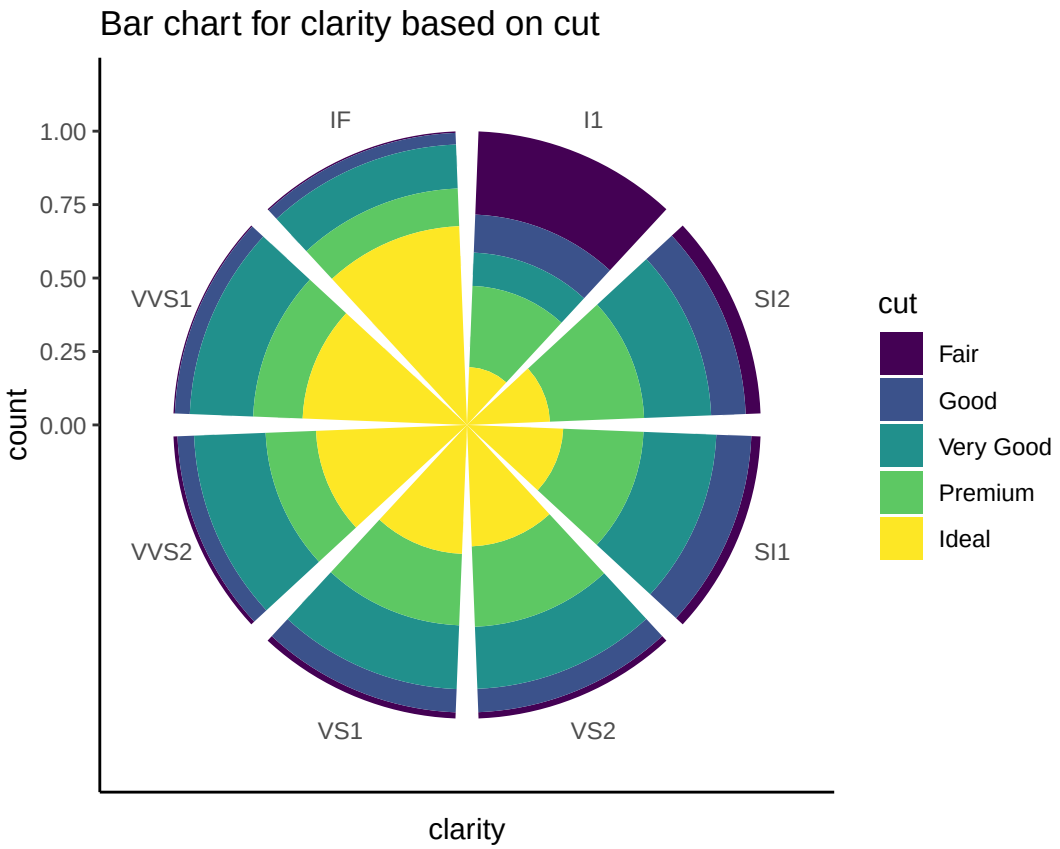



```
max(df$price)
```

```
## [1] 18823
```

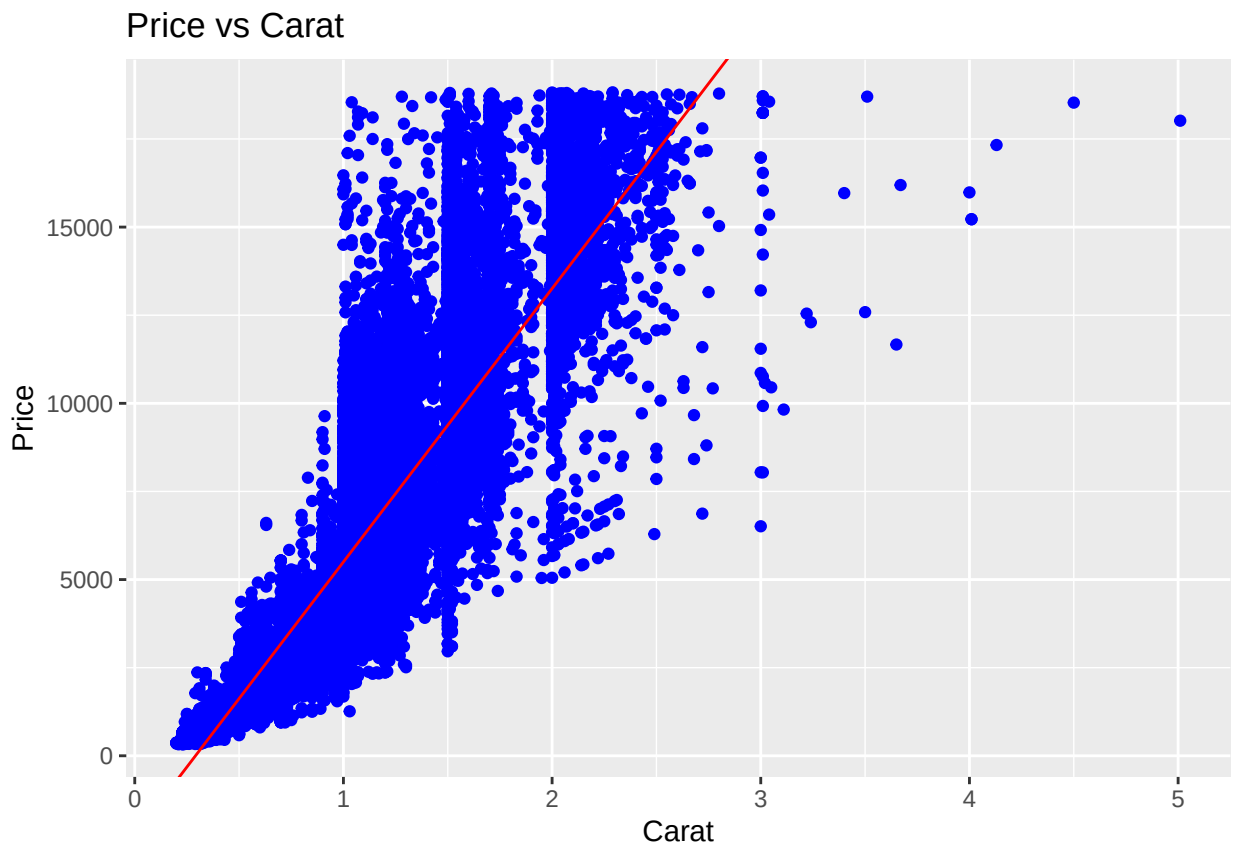
Bar chart for clarity based on cut

```
ggplot(df) +
  geom_bar(aes(x = clarity, fill = cut), position = "fill") +
  labs(title = "Bar chart for clarity based on cut", x = "clarity") +
  coord_polar() +
  theme_classic()
```



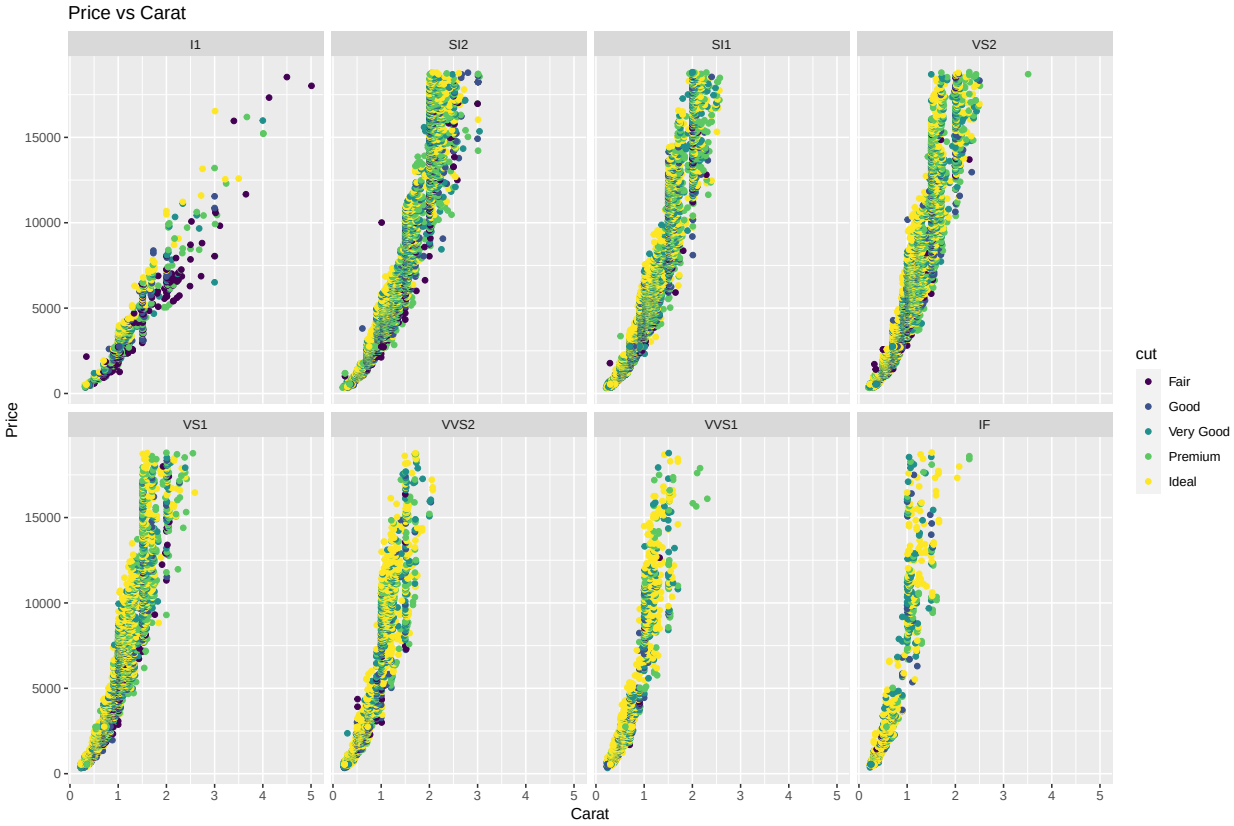
Price vs Carat

```
model <- lm(price ~ carat, data = df)
coeff <- coef(model)
ggplot(df) +
  geom_point(aes(x = carat, y = price), color = "blue") +
  geom_abline(intercept = coeff[1], slope = coeff[2], color = "red") +
  labs(title = "Price vs Carat", x = "Carat", y = "Price")
```



Price vs Carat based on cut

```
ggplot(df) +
  geom_point(aes(x = carat, y = price, color = cut)) +
  facet_wrap(~clarity, nrow = 2) +
  labs(title = "Price vs Carat", x = "Carat", y = "Price")
```


Price vs Carat
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
Price
Carat
0
5000
10000
15000
0
1
2
3
4
5
cut
Fair
Good
Very Good
Premium
Ideal