Prashanth.S 19MID0020

Univariate Analysis

```
library(plotly)
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

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

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:plotly':
##
##     arrange, mutate, rename, summarise
```

Diamond Data-Set

```
df1 = diamonds
head(df1)
```

carat <dbl>	cut <ord>	color <ord>	clarity <ord>	depth <dbl>	table <dbl>	price <int>	x <dbl>	y <dbl>	z <dbl>
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48

6 rows

PreProcessing

```
cat("Number of instances  : ",nrow(df1))
```

```
## Number of instances  :  53940
```

```
cat("\nNumber of attributes : ",ncol(df1))
```

```
##
## Number of attributes :  10
```

```
str(df1)
```

```
## tibble [53,940 × 10] (S3: tbl_df/tbl/data.frame)
##  $ carat  : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
##  $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
##  $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
##  $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
##  $ depth  : num [1:53940] 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
##  $ table  : num [1:53940] 55 61 65 58 58 57 57 55 61 61 ...
##  $ price  : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...
##  $ x      : num [1:53940] 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
##  $ y      : num [1:53940] 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
##  $ z      : num [1:53940] 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

```
summary(df1)
```

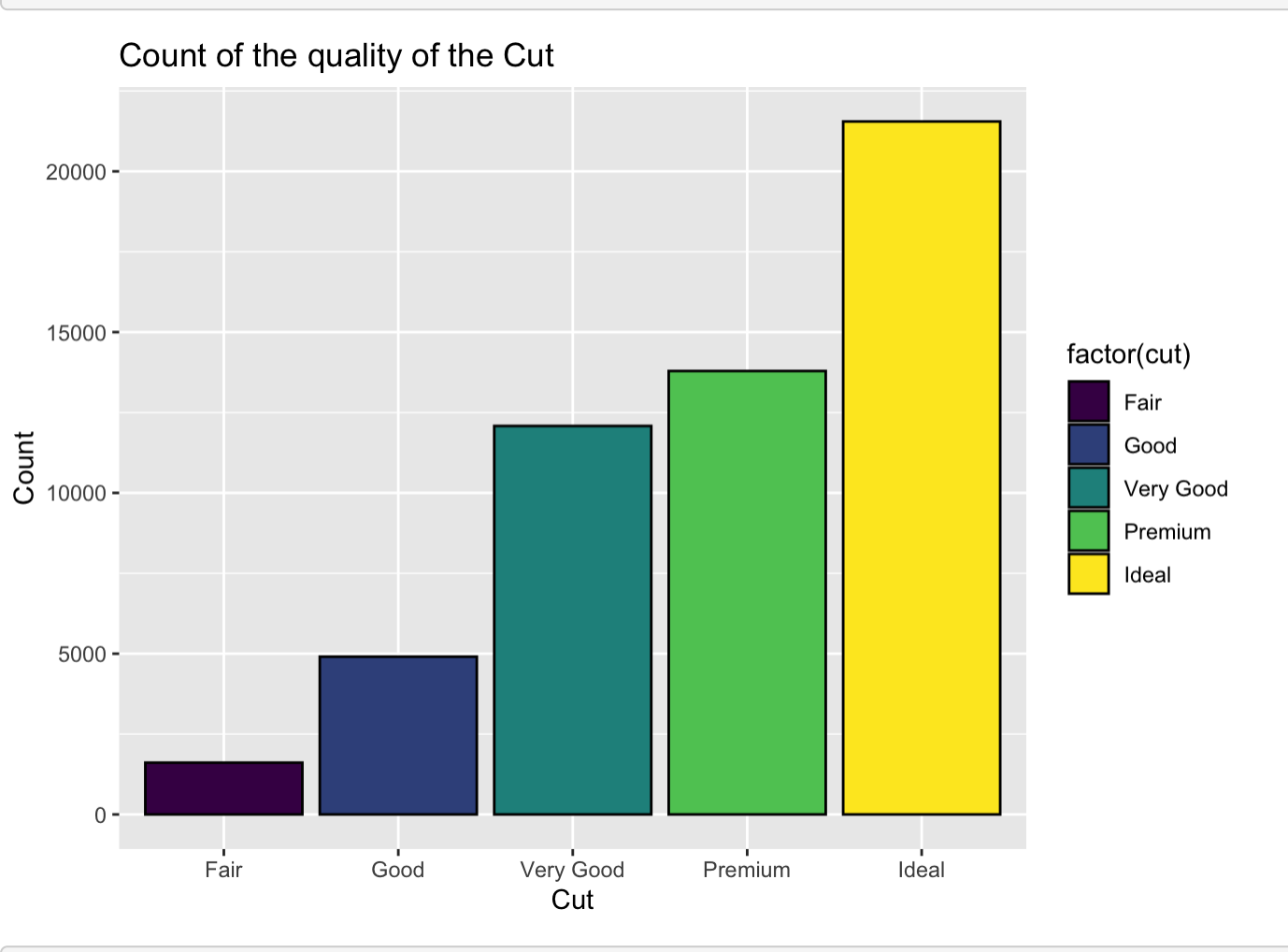
```
##      carat               cut        color        clarity          depth
##  Min.   :0.2000   Fair     : 1610   D: 6775   SI1    :13065   Min.   :43.00
##  1st Qu.:0.4000   Good     : 4906   E: 9797   VS2    :12258   1st Qu.:61.00
##  Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194   Median :61.80
##  Mean   :0.7979   Premium  :13791   G:11292   VS1    : 8171   Mean   :61.75
##  3rd Qu.:1.0400   Ideal    :21551   H: 8304   VVS2   : 5066   3rd Qu.:62.50
##  Max.   :5.0100                     I: 5422   VVS1   : 3655   Max.   :79.00
##                                     J: 2808   (Other): 2531
##      table           price             x                y
##  Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720
##  Median :57.00   Median : 2401   Median : 5.700   Median : 5.710
##  Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735
##  3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540
##  Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900
##
##        z
##  Min.   : 0.000
##  1st Qu.: 2.910
##  Median : 3.530
##  Mean   : 3.539
##  3rd Qu.: 4.040
##  Max.   :31.800
##
```

Bar-Chart

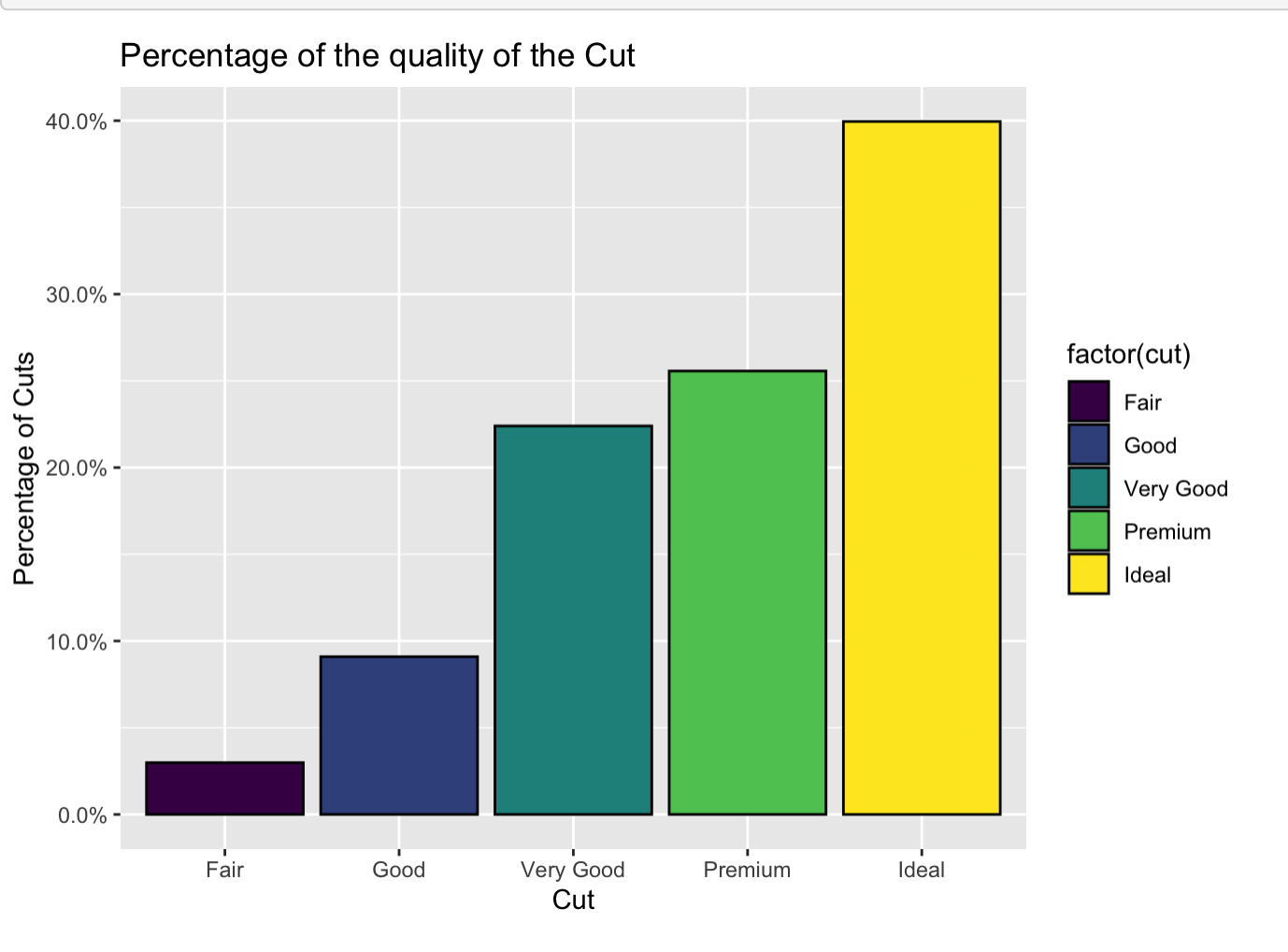
```
table(df1$cut)
```

```
##
##      Fair      Good Very Good   Premium     Ideal
##      1610      4906     12082     13791     21551
```

```
ggplot(df1, aes(x=factor(cut),fill = factor(cut))) +
        geom_bar(color='black') +
        labs(x = "Cut", y = "Count", title  = "Count of the quality of the Cut")
```



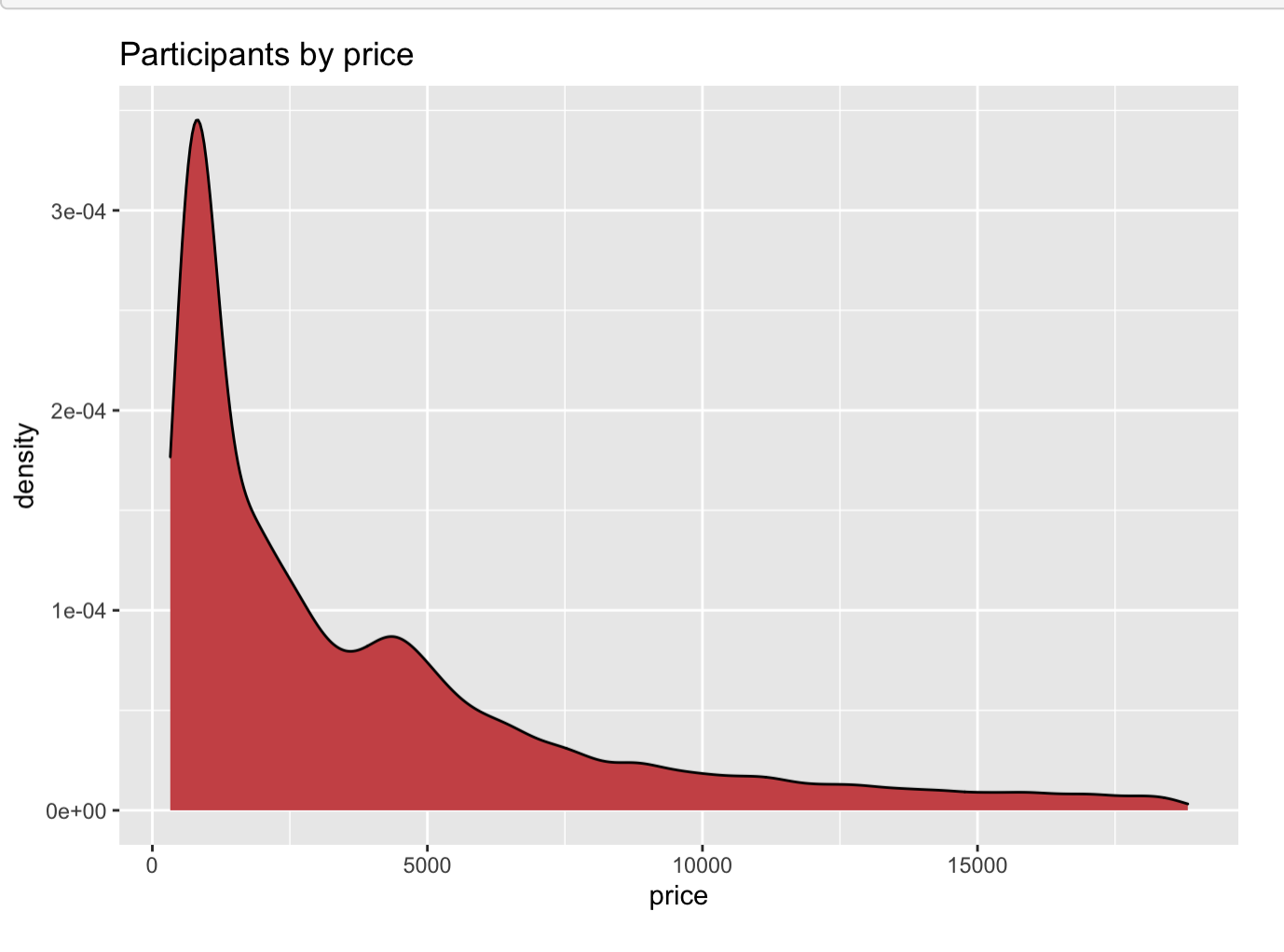

```
ggplot(df1,
       aes(x = factor(cut), y = ..count.. / sum(..count..),fill = factor(cut))) +
       geom_bar(color='black') +
       labs(x = "Cut", y = "Percentage of Cuts", title  = "Percentage of the quality of the Cut") +
       scale_y_continuous(labels = scales::percent)
```



Pie-Chart

Kernel Density Plot

```
ggplot(df1, aes(x = price)) +
       geom_density(fill = "indianred3") +
       labs(title = "Participants by price")
```



MPG data-set

```
df2 = mpg
head(df2)
```

manufacturer <chr>	model <chr>	displ <dbl>	year <int>	cyl <int>	trans <chr>	drv <chr>	cty <int>	hwy <int>	fl <chr>
audi	a4	1.8	1999	4	auto(l5)	f	18	29	p
audi	a4	1.8	1999	4	manual(m5)	f	21	29	p
audi	a4	2.0	2008	4	manual(m6)	f	20	31	p
audi	a4	2.0	2008	4	auto(av)	f	21	30	p
audi	a4	2.8	1999	6	auto(l5)	f	16	26	p
audi	a4	2.8	1999	6	manual(m5)	f	18	26	p

6 rows | 1-10 of 11 columns

PreProcessing

```
cat("Number of instances  : ",nrow(df2))
```

```
## Number of instances  :  234
```

```
cat("\nNumber of attributes : ",ncol(df2))
```

```
##
## Number of attributes :  11
```

```
str(df2)
```

```
## tibble [234 × 11] (S3: tbl_df/tbl/data.frame)
##  $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
##  $ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
##  $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr [1:234] "f" "f" "f" "f" ...
##  $ cty         : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr [1:234] "p" "p" "p" "p" ...
##  $ class       : chr [1:234] "compact" "compact" "compact" "compact" ...
```

```
summary(df2)
```

```
##  manufacturer          model               displ            year
##  Length:234         Length:234         Min.   :1.600   Min.   :1999
##  Class :character   Class :character   1st Qu.:2.400   1st Qu.:1999
##  Mode  :character   Mode  :character   Median :3.300   Median :2004
##                                        Mean   :3.472   Mean   :2004
##                                        3rd Qu.:4.600   3rd Qu.:2008
##                                        Max.   :7.000   Max.   :2008
##       cyl           trans               drv                 cty
##  Min.   :4.000   Length:234         Length:234         Min.   : 9.00
##  1st Qu.:4.000   Class :character   Class :character   1st Qu.:14.00
##  Median :6.000   Mode  :character   Mode  :character   Median :17.00
##  Mean   :5.889                                         Mean   :16.86
##  3rd Qu.:8.000                                         3rd Qu.:19.00
##  Max.   :8.000                                         Max.   :35.00
##       hwy             fl               class
##  Min.   :12.00   Length:234         Length:234
##  1st Qu.:18.00   Class :character   Class :character
##  Median :24.00   Mode  :character   Mode  :character
##  Mean   :23.44
##  3rd Qu.:27.00
##  Max.   :44.00
```

```
df2 %>%
    summarize(variable = "cty",
              q0.2 = quantile(cty, 0.2), ## 20% city values
              q0.4 = quantile(cty, 0.4), ## 40% cty values
              q0.6 = quantile(cty, 0.6), ## 60% cty values
              q0.8 = quantile(cty, 0.8)) ## 80% cty values
```

variable <chr>	q0.2 <dbl>	q0.4 <dbl>	q0.6 <dbl>	q0.8 <dbl>
cty	13	15	18	20

1 row

```
df2 %>%
    summarize(
      variable = "cty",
      mean = mean(cty),
      variance = var(cty),
      standard_deviation = sd(cty),
      Inter_Quartile_Range = IQR(cty)
)
```

variable <chr>	mean <dbl>	variance <dbl>	standard_deviation <dbl>	Inter_Quartile_Range <dbl>
cty	16.85897	18.11307	4.255946	5

1 row

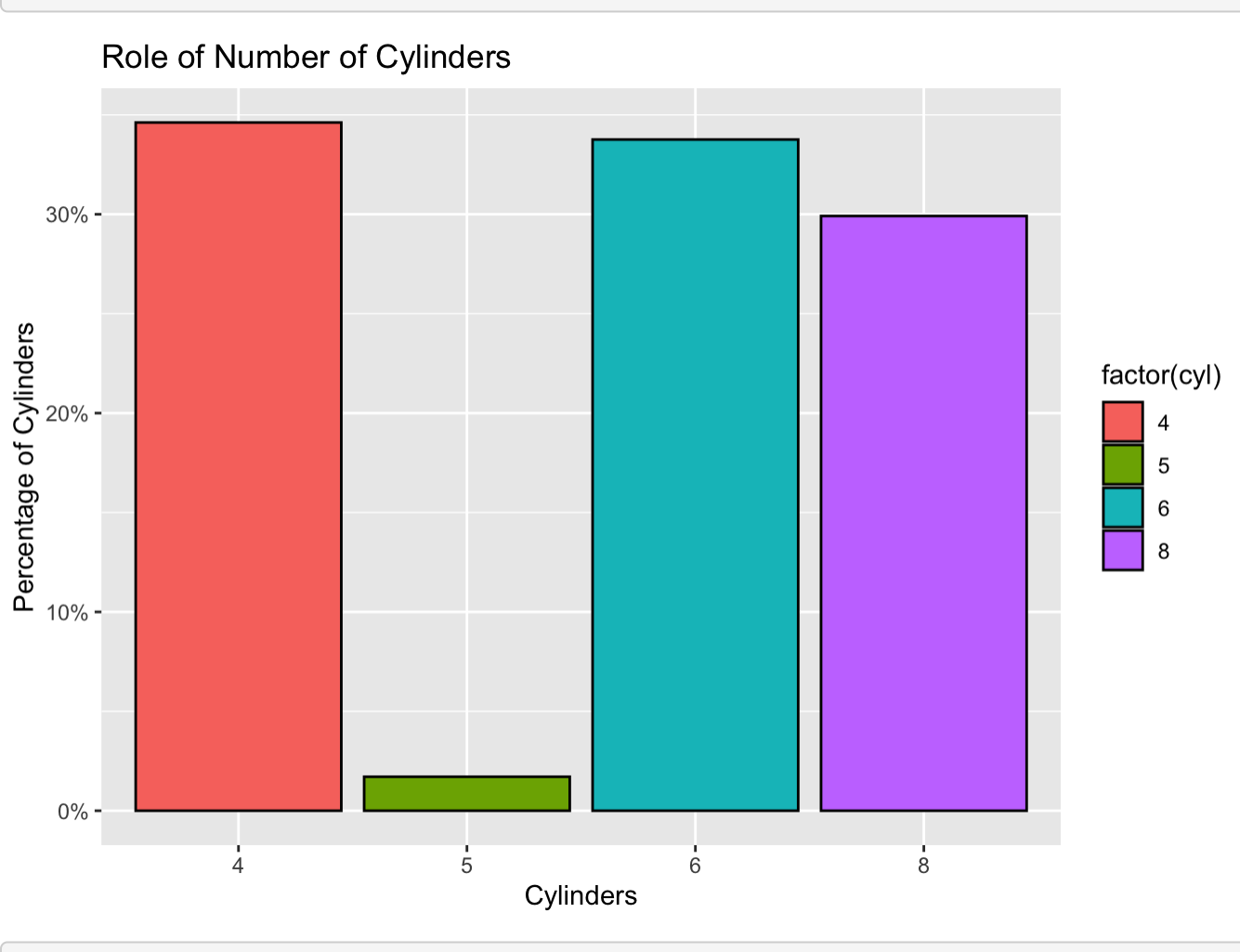
```
df2 %>%
    group_by(drv) %>%
    summarize(mean_cty = mean(cty), Standard_Deviation_cty = sd(cty))
```

mean_cty <dbl>	Standard_Deviation_cty <dbl>
16.85897	4.255946

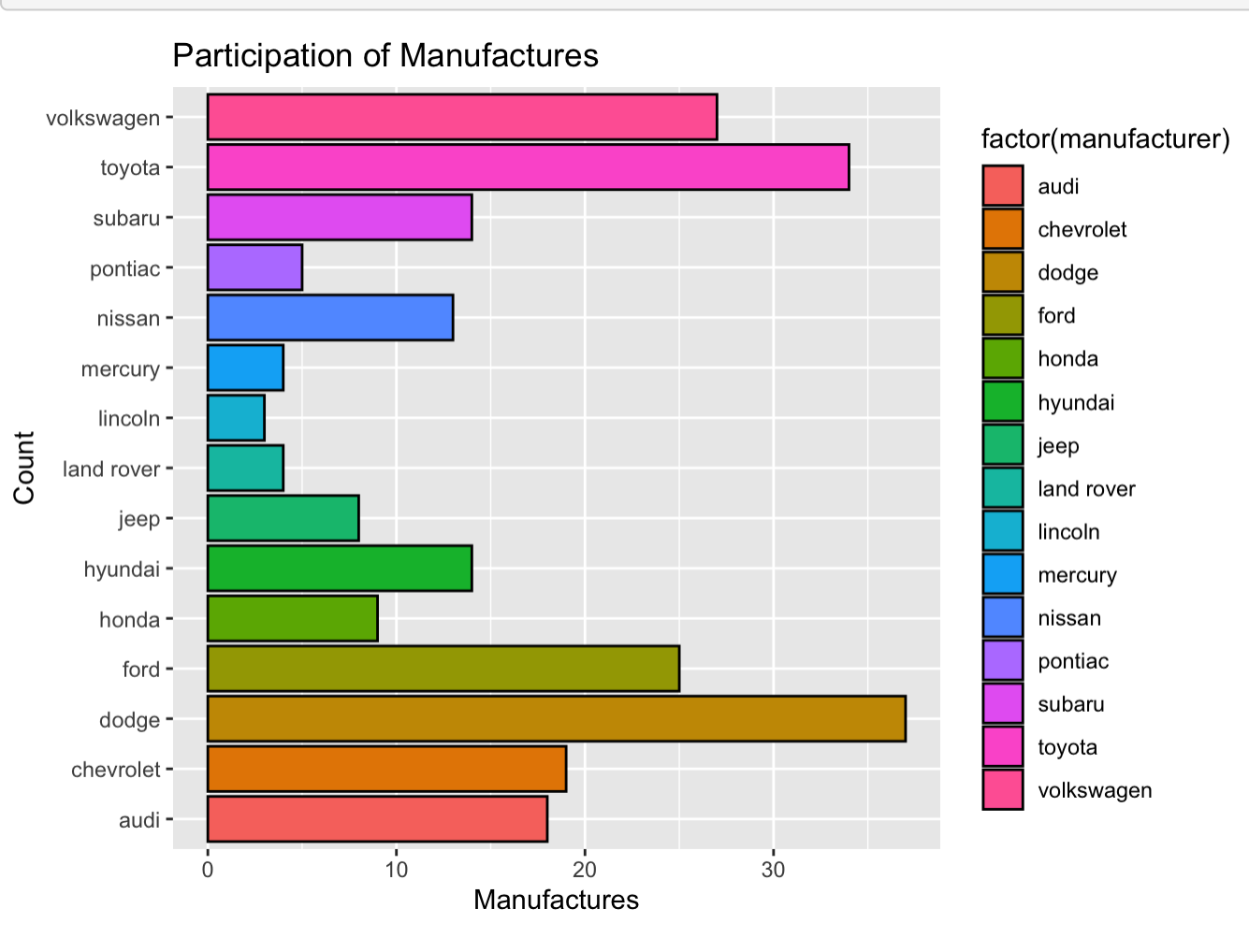
1 row

Barchart

```
ggplot(df2,
       aes(x = factor(cyl), y = ..count.. / sum(..count..),fill = factor(cyl))) +
       geom_bar(color='black') +
       labs(x = "Cylinders", y = "Percentage of Cylinders", title  = "Role of Number of Cylinders") +
       scale_y_continuous(labels = scales::percent)
```

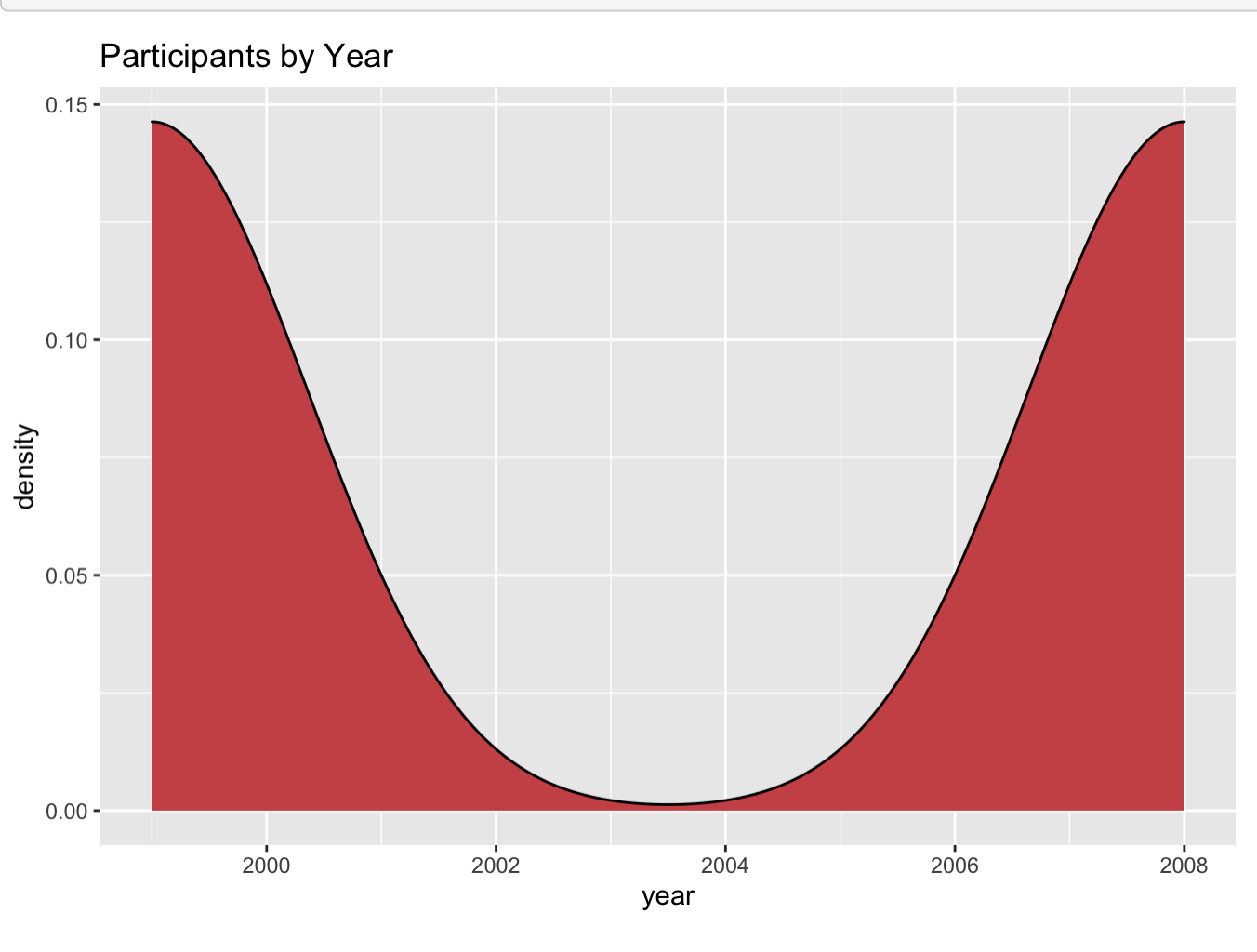



```
ggplot(df2, aes(factor(manufacturer),fill = factor(manufacturer))) +
      geom_bar(color='black') +
      labs(x = "Count", y = "Manufactures", title  = "Participation of Manufactures") +
      coord_flip()
```



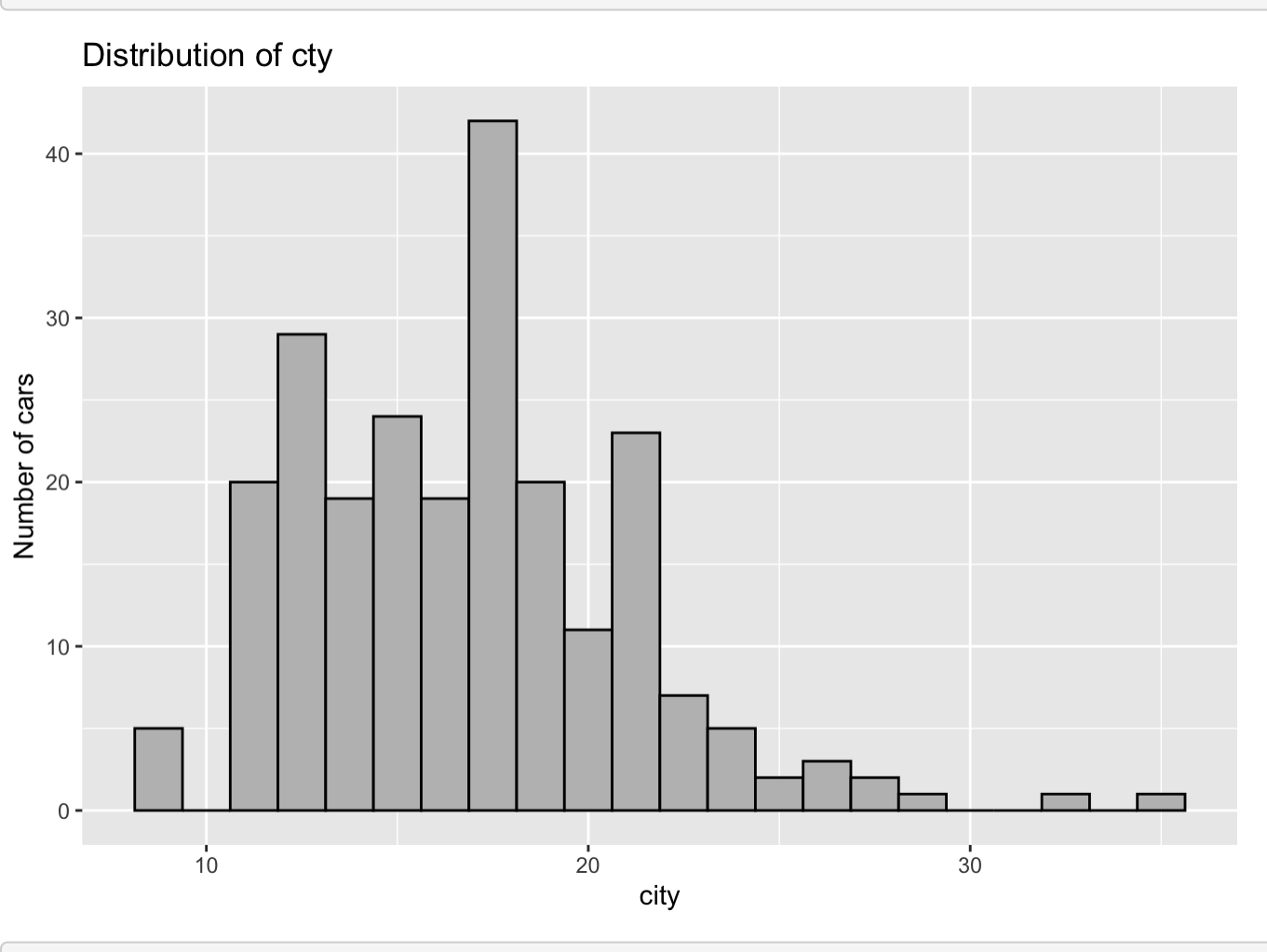
Kernel Density Plot

```
ggplot(df2, aes(x = year)) +
       geom_density(fill = "indianred3") +
       labs(title = "Participants by Year")
```



bivariate analysis

```
df2 %>%
    ggplot(aes(cty)) +
    geom_histogram(binwidth = 1.25, color = "black",fill = "grey") +
    labs(
        title = "Distribution of cty",
         x = "city",
         y = "Number of cars"
    )
```




```
df2 %>%
    ggplot(aes(drv,cty)) +
    geom_boxplot() +
    labs(title = "Distribution of cty relative to Driving Mode",
         x = "Driving Mode",
         y = "cty-value")
```