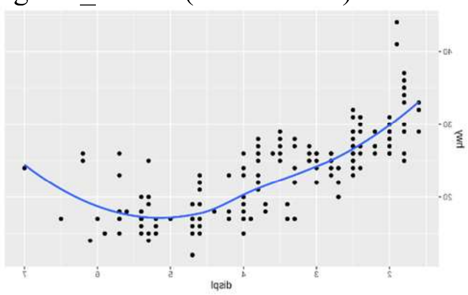
DATASCIENCE ASSIGNMENT-1

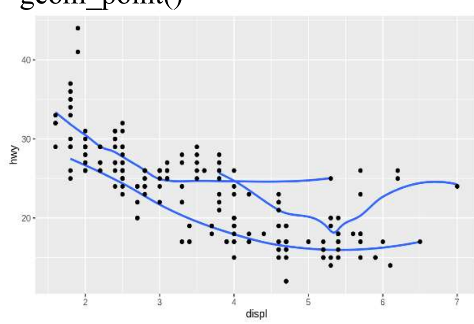
NAME:T.SRIPOORNIMA
CLASS:III CSE-A
ROLLNO:19BCS027
1. Re-create the R code necessary to generate the following graphs using mtcars dataset.

Solution:

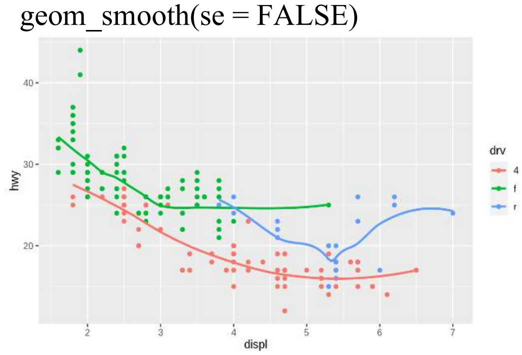
a). ggplot(mpg, aes(x = displ, y = hwy)) +
 geom_point() +
 geom_smooth(se = FALSE)



b). ggplot(mpg, aes(x = displ, y = hwy)) +
 geom_smooth(mapping = aes(group = drv), se = FALSE) +
 geom_point()

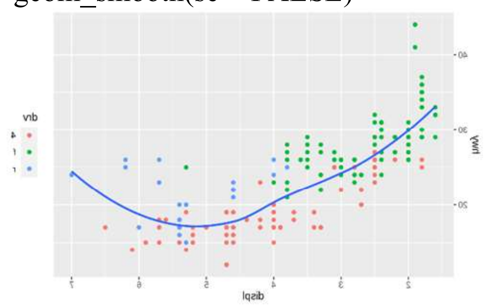


c). ggplot(mpg, aes(x = displ, y = hwy, colour = drv)) +
 geom_point() +
 geom_smooth(se = FALSE)

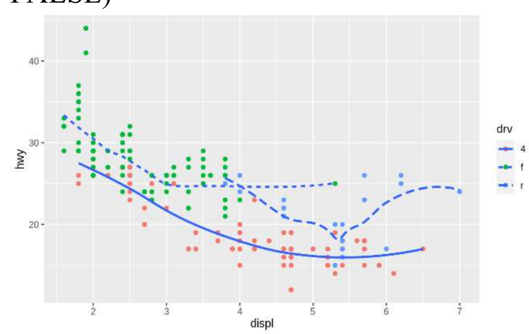


d). ggplot(mpg, aes(x = displ, y = hwy)) +

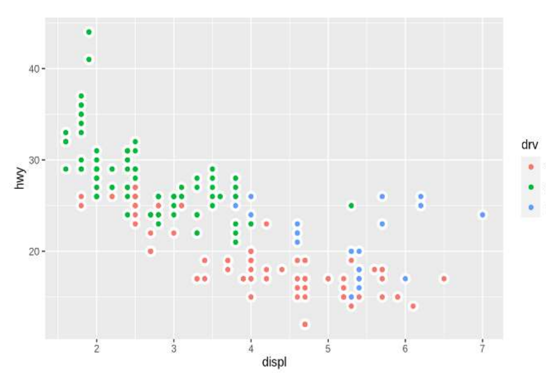
 geom_point(aes(colour = drv)) +
 geom_smooth(se = FALSE)



e). ggplot(mpg, aes(x = displ, y = hwy)) +
 geom_point(aes(colour = drv)) +
 geom_smooth(aes(linetype = drv), se = FALSE)



f). ggplot(mpg, aes(x = displ, y = hwy)) +
 geom_point(size = 4, color = "white")
+geom_point(aes(colour = drv))



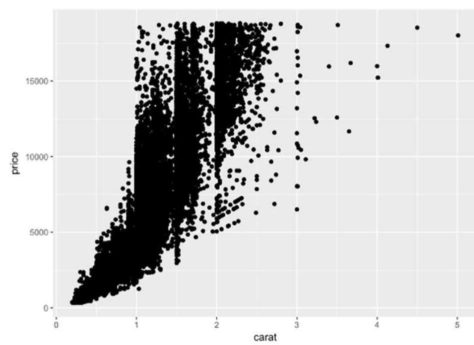
DATASCIENCE ASSIGNMENT-1

NAME:T.SRIPOORNIMA
CLASS:III CSE-A
ROLLNO:19BCS027

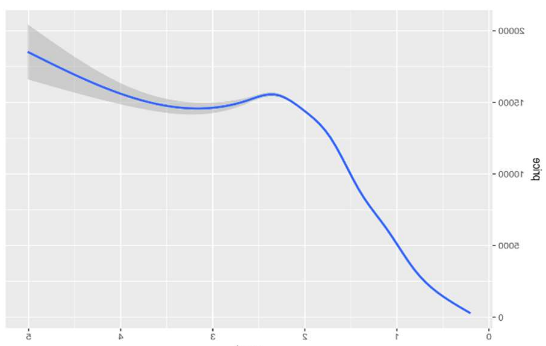
2). Use diamonds dataset and explore using 5 different plots What variable in the diamond's dataset is most important for predicting the price of a diamond?

Solution:

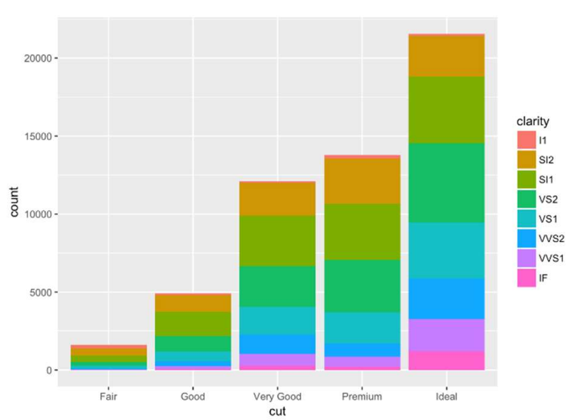
library(ggplot2)
library(dplyr)
View(diamonds)
a). ggplot(diamonds, aes(x=carat, y=price)) + geom_point()



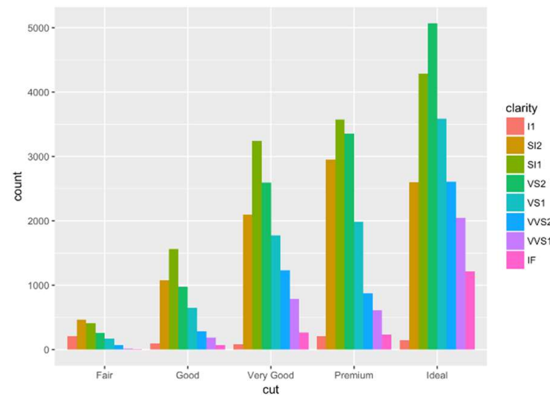
b). ggplot(diamonds, aes(x=carat, y=price)) + geom_smooth()



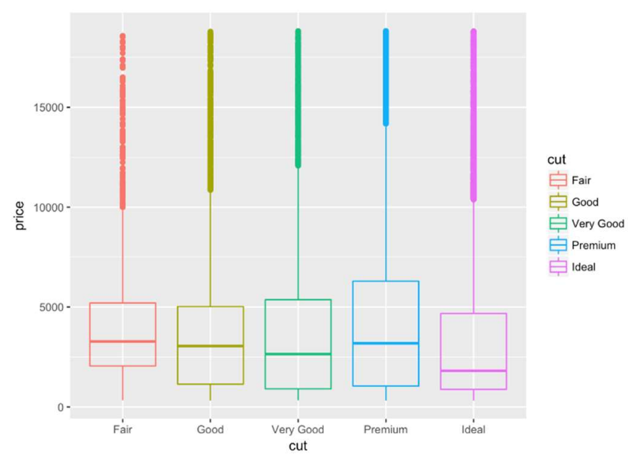
c). ggplot(diamonds, aes(cut)) + geom_bar(aes(fill = clarity))



d). ggplot(diamonds, aes(cut)) + geom_bar(aes(fill = clarity), position = "dodge")



e). ggplot(diamonds, aes(cut, price)) + geom_boxplot(aes(color=cut), fill=NA)



DATASCIENCE ASSIGNMENT-1

NAME:T.SRIPOORNIMA
CLASS:III CSE-A
ROLLNO:19BCS027

2). Write R code to do the following using flights dataset,

1. Sort flights to find the most delayed and the fastest flights. Find the flights that left earliest.

>library(nycflights13)
> library(tidyverse)
> arrange(flights, dep_delay)

```
# A tibble: 336,776 x 19
   year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay
carrier flight tailnum origin dest  air_time distance  hour minute time_hour
  <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl> <chr>    <int>
<chr>   <chr>  <chr>    <dbl>    <dbl> <dbl>  <dbl> <dttm>
 1  2013    12     7     2040           2123       -43       40           2352        48 B6          97 N592JB
JFK    DEN       265     1626    21     23 2013-12-07 21:00:00
 2  2013     2     3     2022           2055       -33     2240           2338       -58 DL        1715
N612DL  LGA    MSY       162     1183    20     55 2013-02-03 20:00:00
 3  2013    11    10     1408           1440       -32     1549           1559       -10 EV        5713
N825AS  LGA    IAD        52      229    14     40 2013-11-10 14:00:00
 4  2013     1    11     1900           1930       -30     2233           2243       -10 DL        1435
N934DL  LGA    TPA       139     1010    19     30 2013-01-11 19:00:00
 5  2013     1    29     1703           1730       -27     1947           1957       -10 F9         837
N208FR  LGA    DEN       250     1620    17     30 2013-01-29 17:00:00
```

> arrange(flights, air_time)

```
# A tibble: 336,776 x 19
   year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay
carrier flight tailnum origin dest  air_time distance  hour minute time_hour
  <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl> <chr>    <int>
<chr>   <chr>  <chr>    <dbl>    <dbl> <dbl>  <dbl> <dttm>
 1  2013     1    16     1355           1315        40     1442           1411        31 EV        4368
N16911  EWR    BDL        20      116    13     15 2013-01-16 13:00:00
 2  2013     4    13      537            527        10      622            628        -6 EV        4631 N12167
EWR    BDL        20      116     5     27 2013-04-13 05:00:00
 3  2013    12     6      922            851        31     1021            954        27 EV        4276 N27200
EWR    BDL        21      116     8     51 2013-12-06 08:00:00
 4  2013     2     3     2153           2129        24     2247           2224        23 EV        4619
N13913  EWR    PHL        21       80    21     29 2013-02-03 21:00:00
 5  2013     2     5     1303           1315       -12     1342           1411       -29 EV        4368
N13955  EWR    BDL        21      116    13     15 2013-02-05 13:00:00
```

2. Find the 10 most delayed flights using a ranking function.

> flights %>%
+   top_n(10, dep_delay)

```
# A tibble: 10 x 19
   year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay
carrier flight tailnum origin dest  air_time distance  hour minute time_hour
  <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl> <chr>    <int>
<chr>   <chr>  <chr>    <dbl>    <dbl> <dbl>  <dbl> <dttm>
```

<u>**DATASCIENCE ASSIGNMENT-1**</u>

NAME:T.SRIPOORNIMA
CLASS:III CSE-A
ROLLNO:19BCS027

```
 1  2013     1     9      641           900      1301     1242          1530      1272 HA          51
N384HA  JFK    HNL        640     4983     9      0 2013-01-09 09:00:00
 2  2013     1    10     1121          1635      1126     1239          1810      1109 MQ        3695
N517MQ  EWR    ORD        111      719    16     35 2013-01-10 16:00:00
 3  2013    12     5      756          1700       896     1058          2020       878 AA         172
N5DMAA  EWR    MIA        149     1085    17      0 2013-12-05 17:00:00
 4  2013     3    17     2321           810       911      135          1020       915 DL        2119
N927DA  LGA    MSP        167     1020     8     10 2013-03-17 08:00:00
 5  2013     4    10     1100          1900       960     1342          2211       931 DL        2391
N959DL  JFK    TPA        139     1005    19      0 2013-04-10 19:00:00
```

3. Which carrier has the worst delays?

```
> filter(flights,arr_delay>1000, dep_delay>1000)
# A tibble: 4 x 19
   year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay
carrier flight tailnum origin dest  air_time distance  hour minute
  <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl> <chr>    <int>
<chr>   <chr>  <chr>    <dbl>    <dbl> <dbl>  <dbl>
1  2013     1     9      641           900      1301     1242          1530      1272 HA          51
N384HA  JFK    HNL        640     4983     9      0
2  2013     1    10     1121          1635      1126     1239          1810      1109 MQ        3695
N517MQ  EWR    ORD        111      719    16     35
3  2013     6    15     1432          1935      1137     1607          2120      1127 MQ        3535
N504MQ  JFK    CMH         74      483    19     35
4  2013     9    20     1139          1845      1014     1457          2210      1007 AA         177
N338AA  JFK    SFO        354     2586    18    454.
```

4. Which plane (tailnum) has the worst on-time record?

```
> flights %>%
+     filter(!is.na(arr_delay)) %>%
+     group_by(tailnum) %>%
+     summarise(prop_time = sum(arr_delay <= 30)/n(),
+               mean_arr = mean(arr_delay, na.rm = T),
+               fl = n()) %>%
+     arrange(desc(prop_time))
# A tibble: 4,037 x 4
   tailnum prop_time mean_arr    fl
   <chr>       <dbl>    <dbl> <int>
 1 N103US          1    -6.93    46
 2 N1200K          1    -9.38    21
 3 N121DE          1    15        2
 4 N137DL          1    -5        1
 5 N143DA          1    24        1
```

5. Find all destinations that are flown by at least two carriers. Use that information to rank the carriers.

```
> flights %>%
+     group_by(dest) %>%
+     filter(n_distinct(carrier) > 2) %>%
```

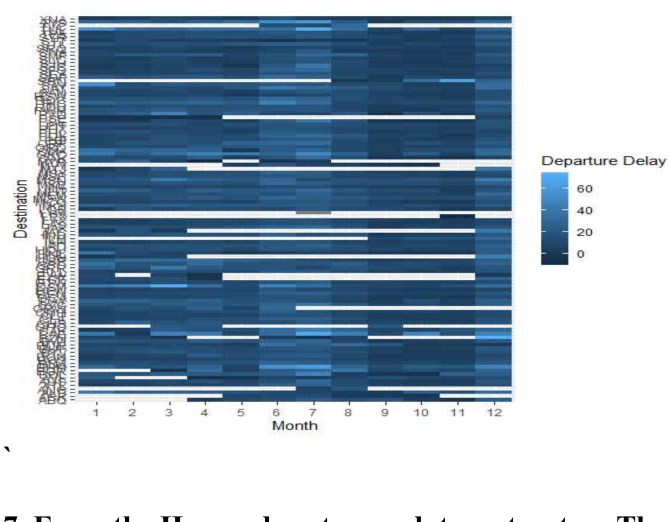
DATASCIENCE ASSIGNMENT-1

NAME:T.SRIPOORNIMA
CLASS:III CSE-A
ROLLNO:19BCS027

```
+    group_by(carrier) %>%
+    summarise(n = n_distinct(dest)) %>%
+    arrange(-n)
# A tibble: 15 x 2
   carrier     n
   <chr>   <int>
 1 DL         37
 2 EV         36
 3 UA         36
 4 9E         35
 5 B6         30
 6 AA         17
 7 MQ         17
 8 WN          9
 9 OO          5
10 US          5
11 VX          3
12 YV          3
13 FL          2
14 AS          1
15 F9          1
```

6. Use geom_tile() together with dplyr to explore how average flight delays vary by destination and month of year.

```
> flights %>%
+    group_by(month, dest) %>%
+    summarise(dep_delay = mean(dep_delay, na.rm = TRUE)) %>%
+    ggplot(aes(x = factor(month), y = dest, fill = dep_delay)) +
+    geom_tile() +
+    labs(x = "Month", y = "Destination", fill = "Departure Delay")
```



`

7. From the Harvard sentences data, extract: a. The first word from each sentence. b. All words ending in ing. c. All plurals.

```
color <- c("red", "orange", "yellow", "green", "blue", "purple")
```

<u>**DATASCIENCE ASSIGNMENT-1**</u>

NAME:T.SRIPOORNIMA
CLASS:III CSE-A
ROLLNO:19BCS027

color_match <- str_c(color, collapse = "|")

str_extract(sentences, "^[a-zA-Z]+")

```
[1] "The"     "Glue"    "It"      "These"   "Rice"    "The"     "The"     "The"
"Four"    "Large"   "The"
 [12] "A"      "The"     "Kick"    "Help"    "A"       "Smoky"   "The"     "The"
"The"     "The"     "The"
 [23] "Press"  "The"     "The"     "Two"     "Her"     "The"     "It"      "Read"
"Hoist"   "Take"    "Note"
```

str_extract_all(sentences, "[a-zA-Z]+ing")

```
[[429]]
[1] "hous"   "robins"

[[430]]
[1] "mats"

[[431]]
[1] "This"  "hors"  "finis"

[[432]]
[1] "protects"
```

str_extract_all(sentences, "[a-zA-Z]{3,}s")

```
[[716]]
[1] "grass"  "bushes"

[[717]]
[1] "coins"

[[718]]
character(0)

[[719]]
[1] "times"

[[720]]
character(0)
```