Homework assignment #2

2017100057 / 이영노

October 11, 2022

Question Number 1

```
library(gapminder)
library(dplyr)
```

```
##
## 다음의 패키지를 부착합니다: 'dplyr'
```

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

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

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v stringr 1.4.0
## v tidyr   1.2.0     v forcats 0.5.2
## v readr   2.1.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(gridExtra)
```

```
##
## 다음의 패키지를 부착합니다: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
data(gapminder)
```

Data exploration

```
str(gapminder)
```

```
## tibble [1,704 x 6] (S3: tbl_df/tbl/data.frame)
##  $ country  : Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ continent: Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ year     : int [1:1704] 1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ lifeExp  : num [1:1704] 28.8 30.3 32 34 36.1 ...
##  $ pop      : int [1:1704] 8425333 9240934 10267083 11537966 13079460 14880372 12881816 138
##  $ gdpPercap: num [1:1704] 779 821 853 836 740 ...
```

```
gapminder%>%head()
```

```
## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1952    28.8  8425333      779.
## 2 Afghanistan Asia       1957    30.3  9240934      821.
## 3 Afghanistan Asia       1962    32.0 10267083      853.
## 4 Afghanistan Asia       1967    34.0 11537966      836.
## 5 Afghanistan Asia       1972    36.1 13079460      740.
## 6 Afghanistan Asia       1977    38.4 14880372      786.
```

a.

How many unique countries?

```
gapminder %>% group_by(continent) %>% summarize(num_of_countries=n_distinct(country))
```

```
## # A tibble: 5 x 2
##   continent num_of_countries
##   <fct>                <int>
## 1 Africa                  52
## 2 Americas                25
## 3 Asia                    33
## 4 Europe                  30
## 5 Oceania                  2
```

Answer: shown on the code result

b.

Which European nation had the lowest GDP per capita in 1997 and 2007?

```
gapminder %>% filter(continent=="Europe" & year %in% c(2007)) %>%
  arrange(gdpPercap) %>% .$country %>% head(1)
```

```
## [1] Albania
## 142 Levels: Afghanistan Albania Algeria Angola Argentina Australia ... Zimbabwe
```

```
gapminder %>% filter(continent=="Europe" & year %in% c(1997)) %>%
  arrange(gdpPercap) %>% .$country %>% head(1)
```

```
## [1] Albania
## 142 Levels: Afghanistan Albania Algeria Angola Argentina Australia ... Zimbabwe
```

Answer: both year, Albania

c.

What is the average life expectancy within each continent in 1970s?

```
gapminder %>% filter(year %in% 1970:1979) %>% group_by(continent) %>%
  summarize(avg_lifeExp=mean(lifeExp))
```

```
## # A tibble: 5 x 2
##   continent avg_lifeExp
##   <fct>           <dbl>
## 1 Africa           48.5
## 2 Americas         63.4
## 3 Asia             58.5
## 4 Europe           71.4
## 5 Oceania          72.4
```

Answer: shown on the code result

d.

What 5 countries have the highest total GDP over all years combined?

```
gapminder %>% mutate(gdp=pop*gdpPercap)%>%group_by(country)%>%
  summarize(total=sum(gdp))%>%arrange(-total)%>%head(5)
```

```
## # A tibble: 5 x 2
##   country          total
##   <fct>            <dbl>
## 1 United States  7.68e13
## 2 Japan          2.54e13
## 3 China          2.04e13
## 4 Germany        1.95e13
## 5 United Kingdom 1.33e13
```

Answer: US,Japan,China,Germany,UK

e.

What countries and years had life expectancies of at least 82 years? (columns only)

```
gapminder%>%filter(lifeExp>=82)%>%select(country,lifeExp,year)%>%
  dplyr::rename(c(Country=country,Life_expectancy=lifeExp,Year=year))
```

```
## # A tibble: 3 x 3
##   Country          Life_expectancy  Year
##   <fct>                      <dbl> <int>
## 1 Hong Kong, China            82.2  2007
```

```
## 2 Japan                          82    2002
## 3 Japan                          82.6  2007
```

Answer: Shown on the code result

f.

Which combinations of continent and year have the highest average population across all countries?

```
gapminder%>%filter(continent !="Europe")%>%group_by(continent,year)%>%
  summarize(avg_pop=mean(pop))%>%arrange(-avg_pop)
```

```
## `summarise()` has grouped output by 'continent'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 48 x 3
## # Groups:   continent [4]
##    continent  year    avg_pop
##    <fct>     <int>      <dbl>
##  1 Asia       2007 115513752.
##  2 Asia       2002 109145521.
##  3 Asia       1997 102523803.
##  4 Asia       1992  94948248.
##  5 Asia       1987  87006690.
##  6 Asia       1982  79095018.
##  7 Asia       1977  72257987.
##  8 Asia       1972  65180977.
##  9 Asia       1967  57747361.
## 10 Asia       1962  51404763.
## # ... with 38 more rows
```

Answer: shown on the code result

Question Number 2

```
library(nycflights13)
data(flights)
data(planes)
data(weather)
```

a.

What month ahd the highest proportion of cancelled flights?

What month ahd the lowest? Interpret any seasonal patterns.

Answer: Define term **"cancelled"** as situation that plane did not leave. Thereby, if condition `is.na(dep_delay)` is `TRUE`, then the flight is cancelled.

```
cancelprop=flights%>%mutate(cancel=case_when(
  is.na(dep_delay) ~ 1, TRUE ~ 0))%>%group_by(month)%>%
```

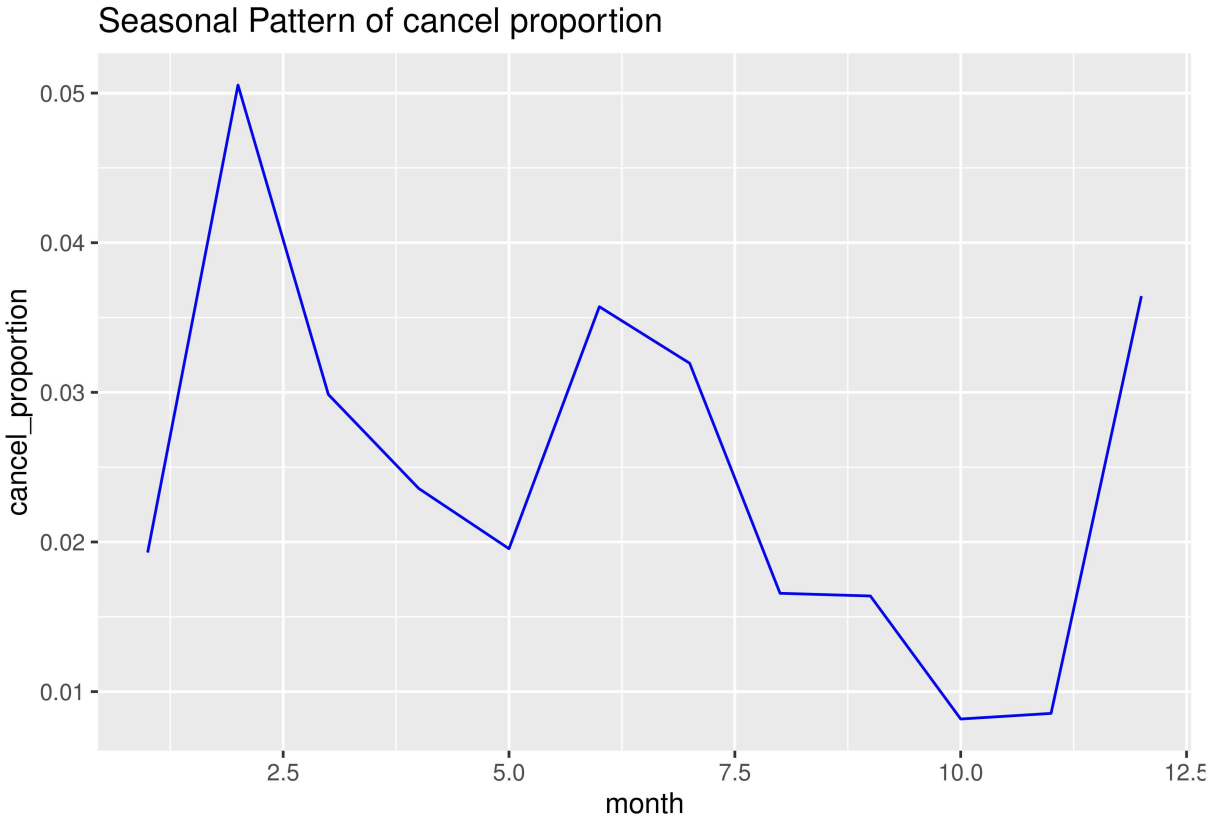
```
  summarize(cancel_proportion=sum(cancel)/length(cancel))
cancelprop%>%filter(cancel_proportion==max(cancel_proportion)|cancel_proportion==min(cancel_pr
```

```
## # A tibble: 2 x 2
##   month cancel_proportion
##   <int>             <dbl>
## 1     2           0.0505
## 2    10           0.00817
```

Then, February(2) had highest cancelled proportion while October(10) had lowest cancelled proportion.

We can interpret the seasonal patterns by plot.

```
cancelprop%>%ggplot(aes(month,cancel_proportion))+geom_line(color="blue")+
  ggtitle("Seasonal Pattern of cancel proportion")
```



By plot above, there seem to be increase in summer(6,7)& winter(12,1,2), and decrease in fall(9,10,11)

b.

What plane traveled most times from NYC airports in 2013?

```
flights%>%group_by(tailnum)%>%
  summarize(time=sum(air_time))%>%arrange(-time)%>%head(10)
```

```
## # A tibble: 10 x 2
##    tailnum  time
##    <chr>   <dbl>
##  1 N502UA  97320
##  2 N512UA  95943
##  3 N505UA  95591
##  4 N557UA  87371
##  5 N518UA  80772
##  6 N508UA  79998
##  7 N555UA  77736
##  8 N597UA  59006
##  9 N598JB  55225
## 10 N591JB  52137
```

Answer: Plane with tailnum "N502UA" traveld most time.

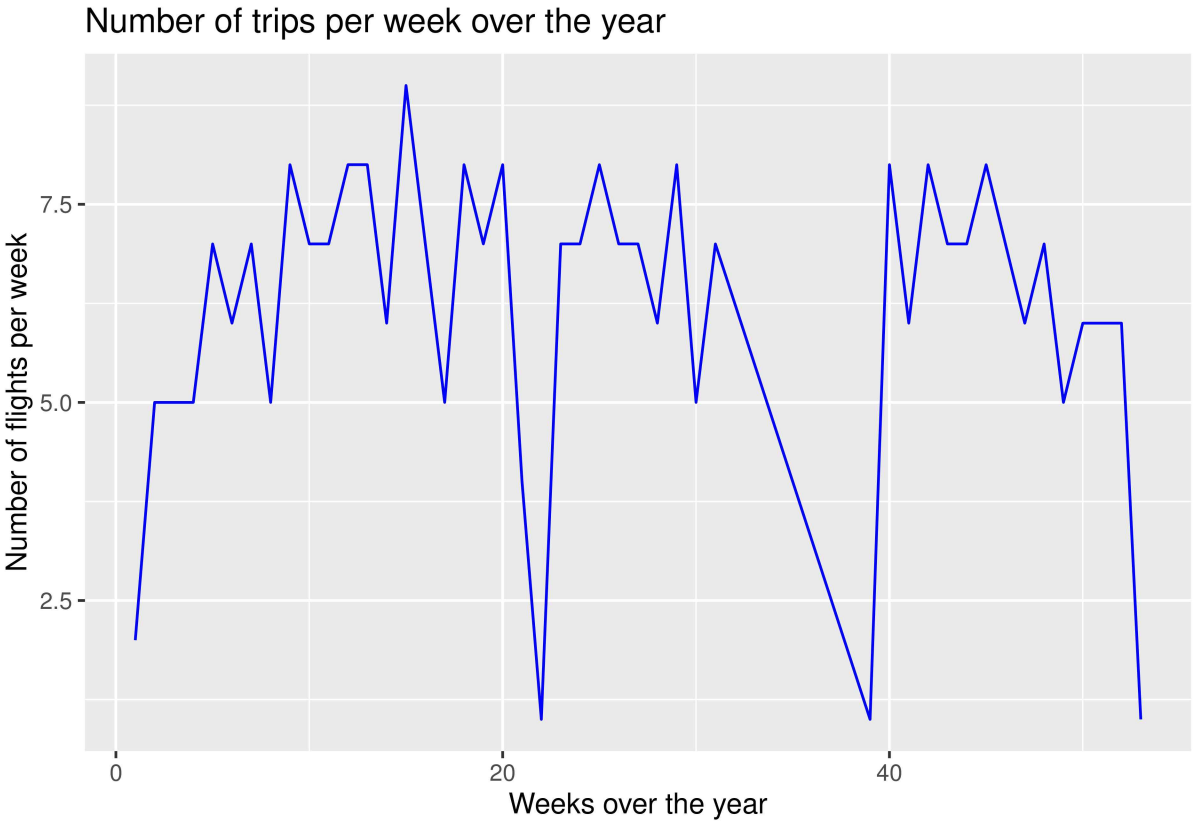
Plot the number of trips per week over the year

```
library(lubridate)
```

```
##
## 다음의 패키지를 부착합니다: 'lubridate'
```

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

```
flights%>%filter(tailnum=="N502UA")%>%
  mutate(date=paste0(year,"-",month,"-",day))%>%
  mutate(nweek=week(date))%>%dplyr::select(tailnum, nweek,date)%>%
  group_by(nweek)%>%summarize(nflight=n())%>%
  ggplot(aes(x=nweek,y=nflight))+geom_line(color="blue")+
  xlab("Weeks over the year")+
  ylab("Number of flights per week")+
  ggtitle("Number of trips per week over the year")
```



c.

What is the oldest plane?

Answer: By explanation in metadata of `planes`, `planes$year` tells the year manufactured. Thereby,"N381AA" is the oldest plane.

```
planes%>%group_by(year)%>%summarize(count=n())%>%head(5) # oldest plane manufactured in 1956.
```

```
## # A tibble: 5 x 2
##    year count
##   <int> <int>
## 1  1956     1
## 2  1959     2
## 3  1963     2
## 4  1965     1
## 5  1967     1
```

```
planes%>%dplyr::select(year,tailnum)%>%filter(year==1956)
```

```
## # A tibble: 1 x 2
##    year tailnum
##   <int> <chr>
## 1  1956 N381AA
```

How many airplanes that flew from NYC are included in planes table?

```
# number of unique values in tailnum in planes table
list=planes%>%group_by(tailnum)%>%summarize(unique=n_distinct(tailnum))%>%
  dplyr::select(tailnum)%>%pull(tailnum)

# using %in%
flights%>%filter(!is.na(dep_delay))%>%group_by(tailnum)%>%
  summarize(unique=n_distinct(tailnum))%>%
  dplyr::select(tailnum)%>%pull(tailnum)%in%list%>%sum()
```

```
## [1] 3316
```

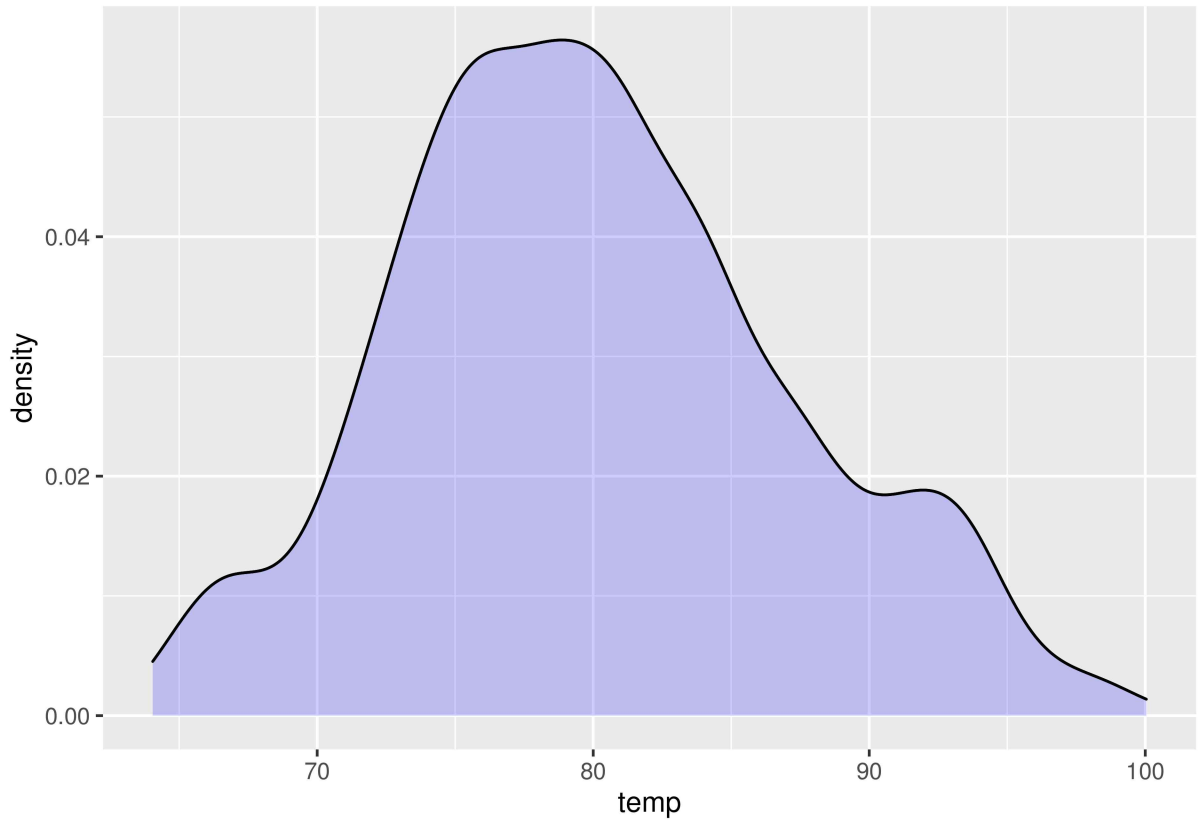
Answer: 3316 planes that actually 'flew'(not cancelled) are included in `planes` table.

d.

What is the distribution of temperature in July 2013?

Answer: shown on the code below.

```
weather%>%filter(year==2013 & month==7)%>%dplyr::select(temp)%>%
  ggplot(aes(temp))+geom_density(fill="blue",alpha=.2)+geom_line(stat='density')
```

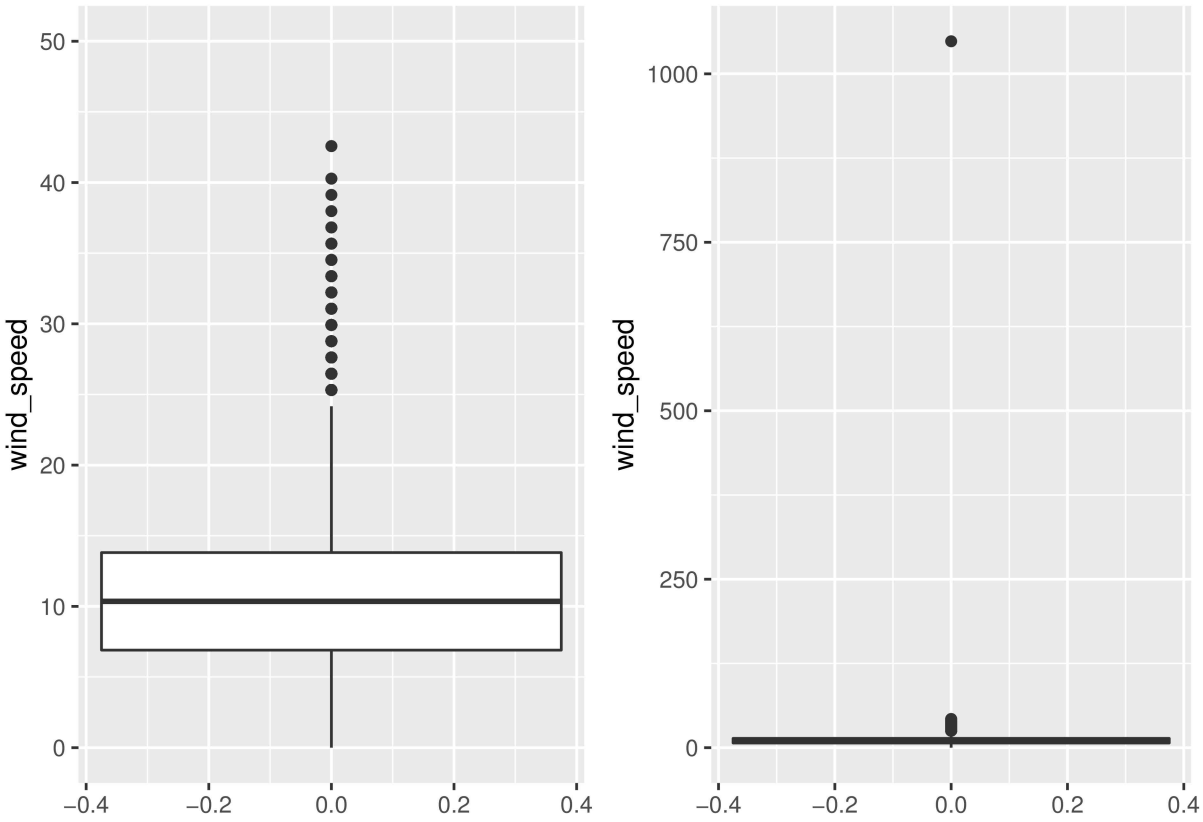


Identify any important ouliers in terms of the `wind_speed` variable.

Answer: `wind_speed`was too skewed, with outlier that has value more than 1000, which is important

outlier that might affect entire data statistics.

```
p1=weather%>%ggplot(aes(y=wind_speed))+geom_boxplot()+
  scale_y_continuous(lim=c(0,50))
p2=weather%>%ggplot(aes(y=wind_speed))+geom_boxplot()
grid.arrange(p1,p2,ncol=2)
```

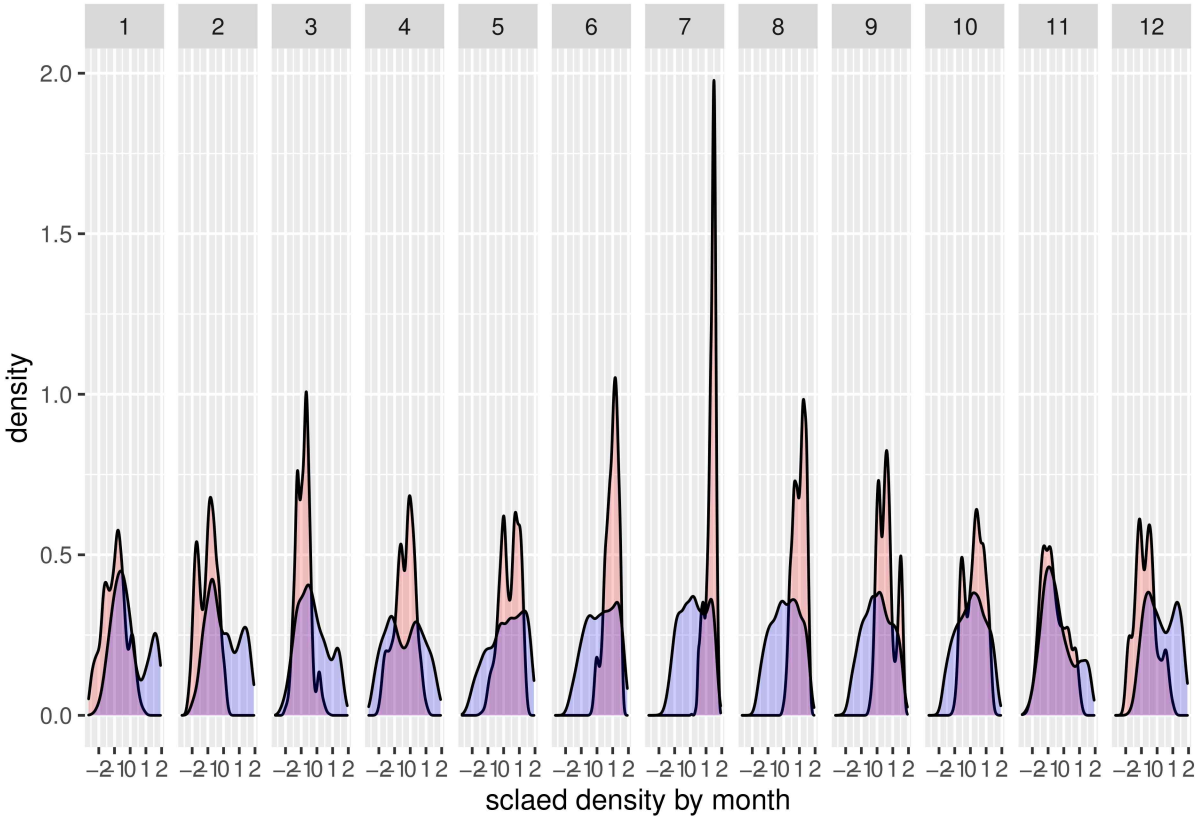


What is the relationship between `dewp` and `humid`?

```
r1=weather%>%filter(!is.na(dewp) & !is.na(humid))
cor(r1$dewp,r1$humid)
```

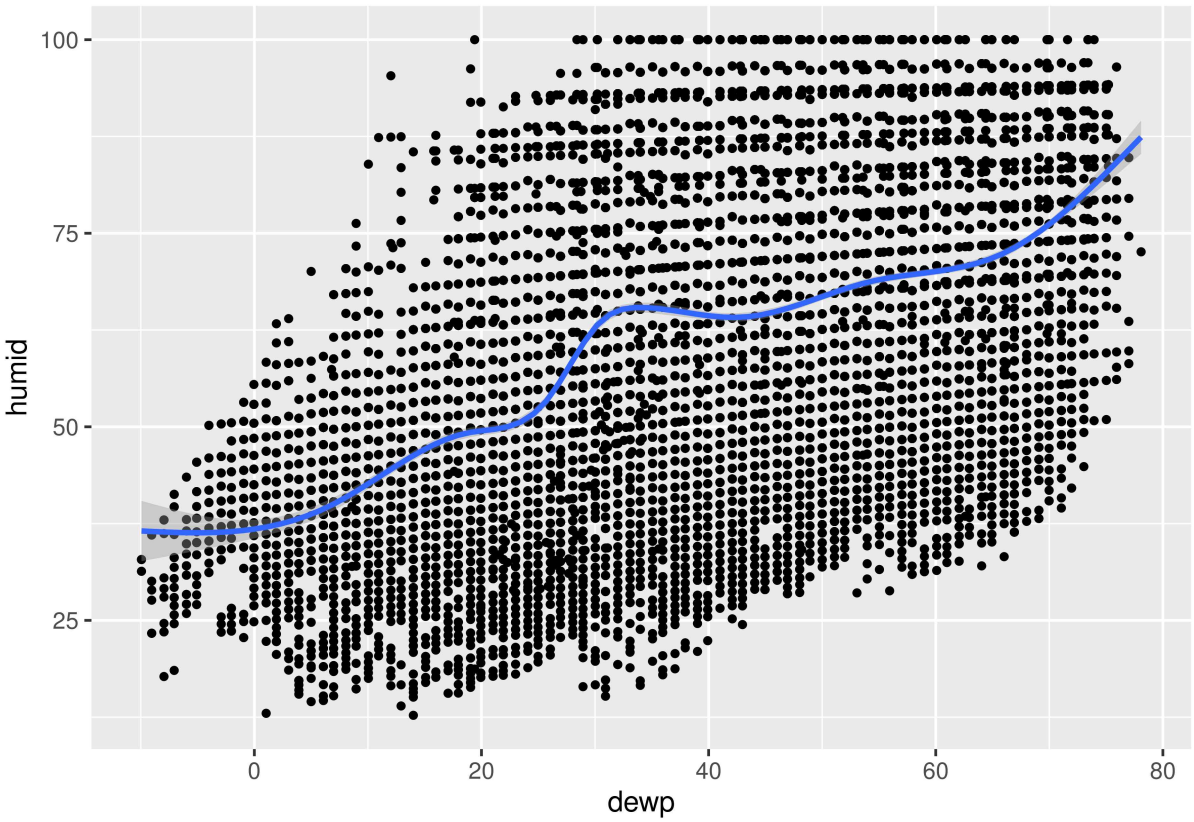
```
## [1] 0.5121952
```

```
weather%>%ggplot()+geom_density(aes(scale(dewp)),fill='red',alpha=.2)+
  geom_density(aes(scale(humid)),fill='blue',alpha=.2)+
  facet_grid(.~ month)+
  xlab("sclaed density by month")
```




```
weather%>%
  ggplot(aes(x = dewp, y = humid)) +
  geom_point(size = 1) +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Answer: distribution is simillar when weather gets cold, and distribution is different when weather gets hot. And correlation is 0.5 positive. Which means that as dewp goes up, humid tends to go up too.

What is the relationship between `precip` and `visib`?

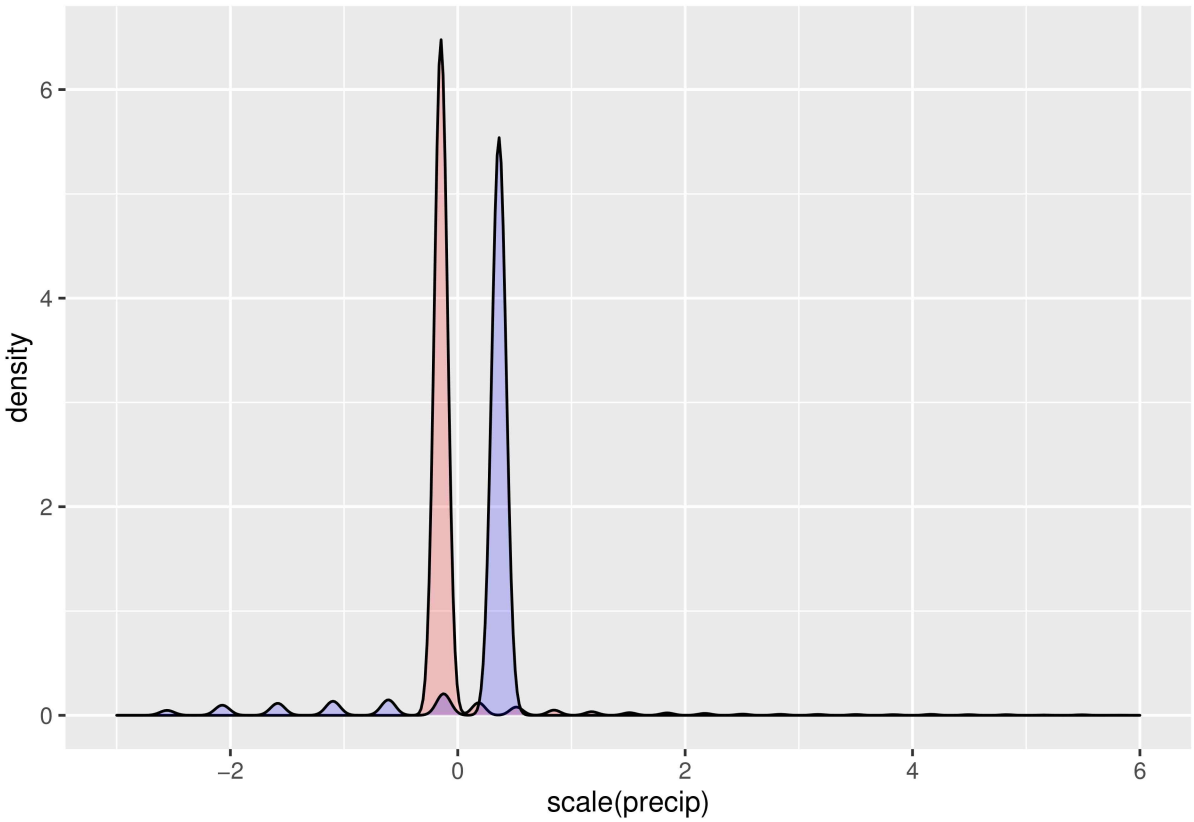
```
r2=weather%>%filter(!is.na(precip) & !is.na(visib))
cor(r2$precip,r2$visib)
```

```
## [1] -0.3199118
```

```
weather%>%ggplot()+geom_density(aes(scale(precip)),fill='red',alpha=.2)+
  geom_density(aes(scale(visib)),fill='blue',alpha=.2)+
  xlim(-3,6)
```

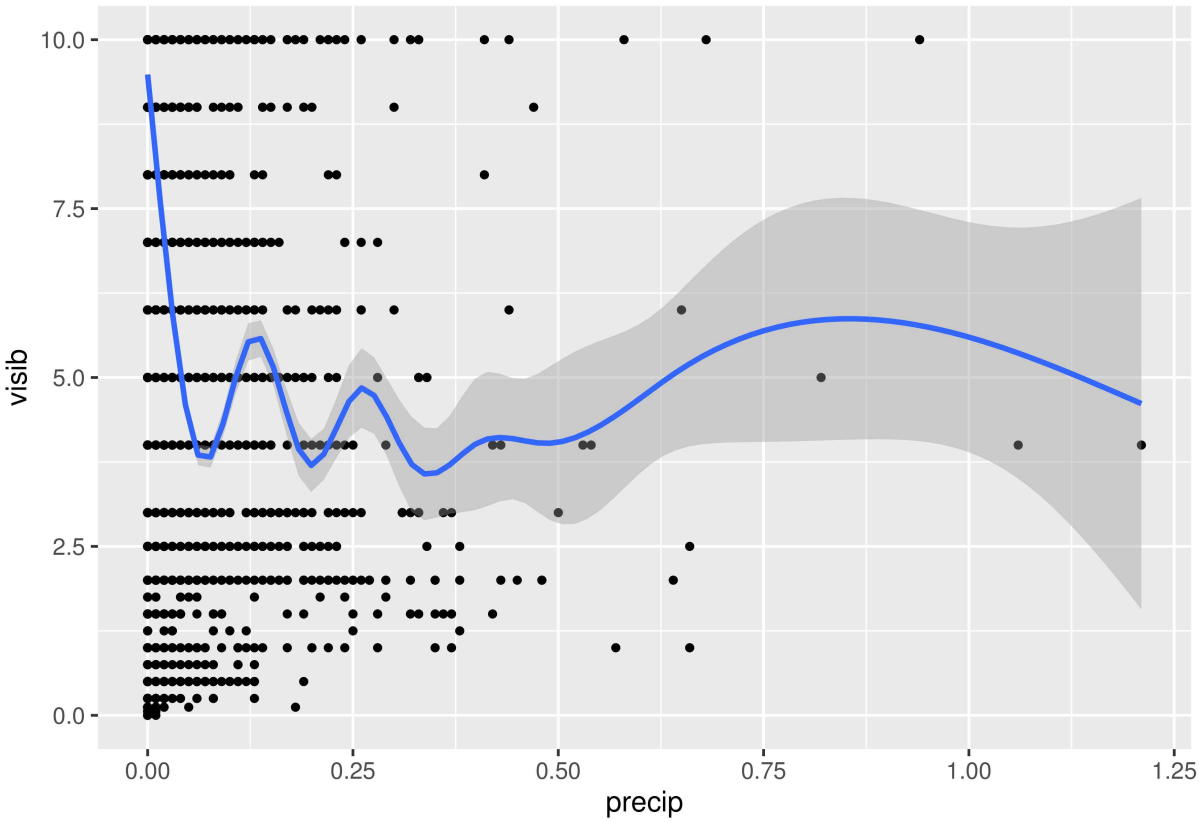
```
## Warning: Removed 155 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 1322 rows containing non-finite values (stat_density).
```




```
weather%>%
  ggplot(aes(x = precip, y = visib)) +
  geom_point(size = 1) +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Answer: Distribution looks simillar in variance but different in mean.

They have -0.3 negative correlation. Which means that as precip goes up,

humid tends to go down.

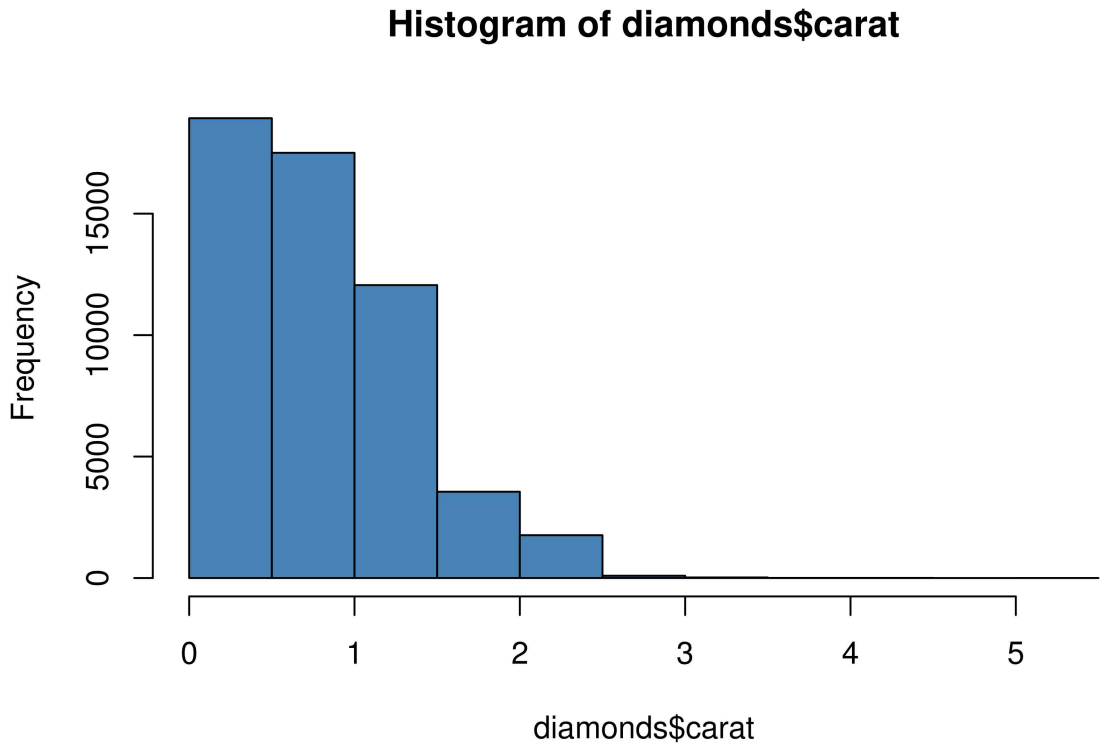
Question Number 3

```
library(ggplot2)
data(diamonds)
```

a.

Use `hist`function to create histogram of carat with bars colored steelblue.

```
hist(diamonds$carat,col="steelblue")
```

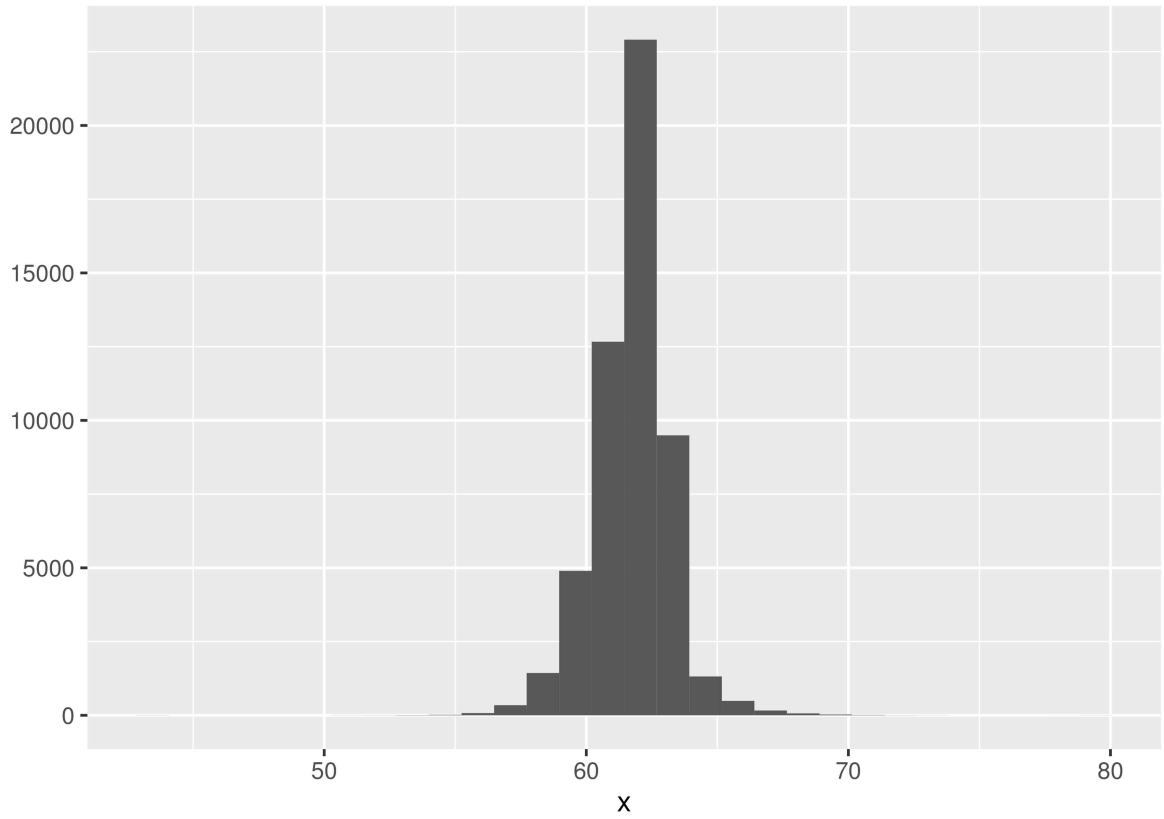


b.

Use `qplot` function to create histogram of `depth`

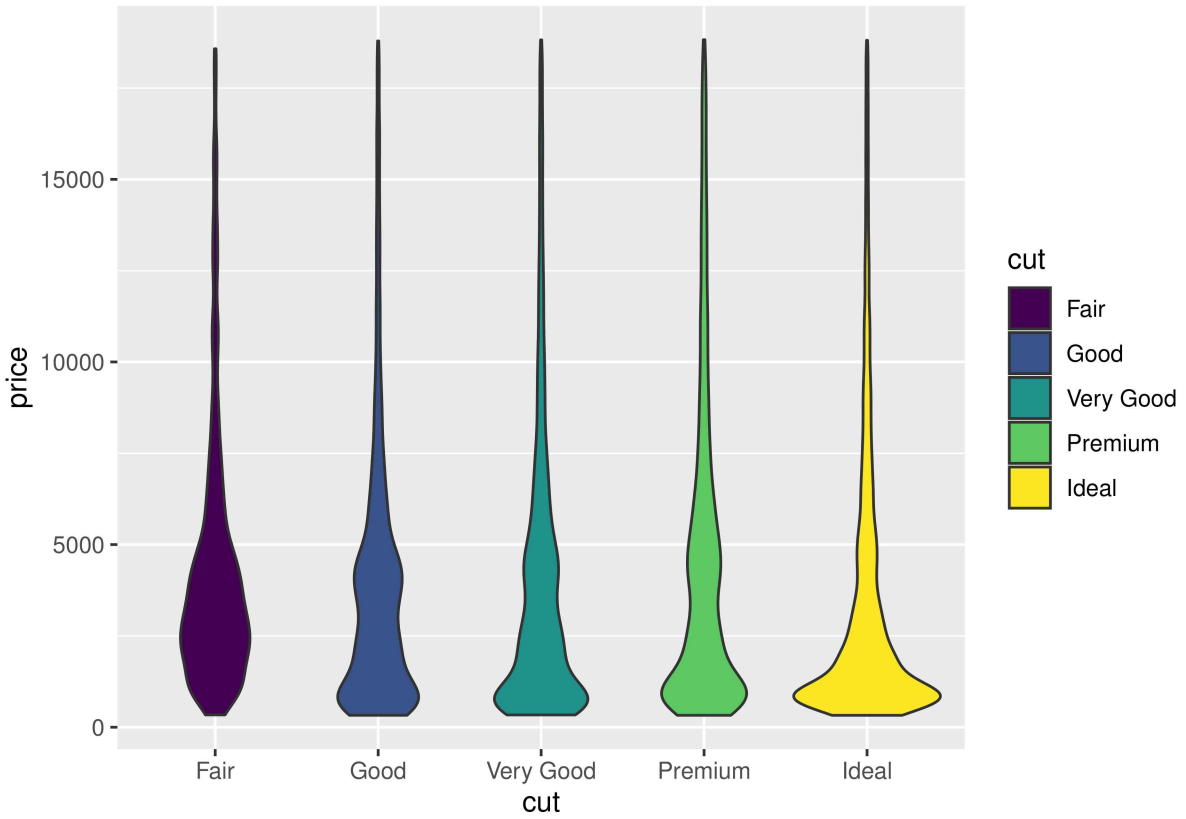
```
x<-diamonds%>%pull(depth)
qplot(x)
```

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



c.

```
qplot(cut,price,data=diamonds,geom='violin',fill=cut)
```



Question Number 4

```
library(MASS)
```

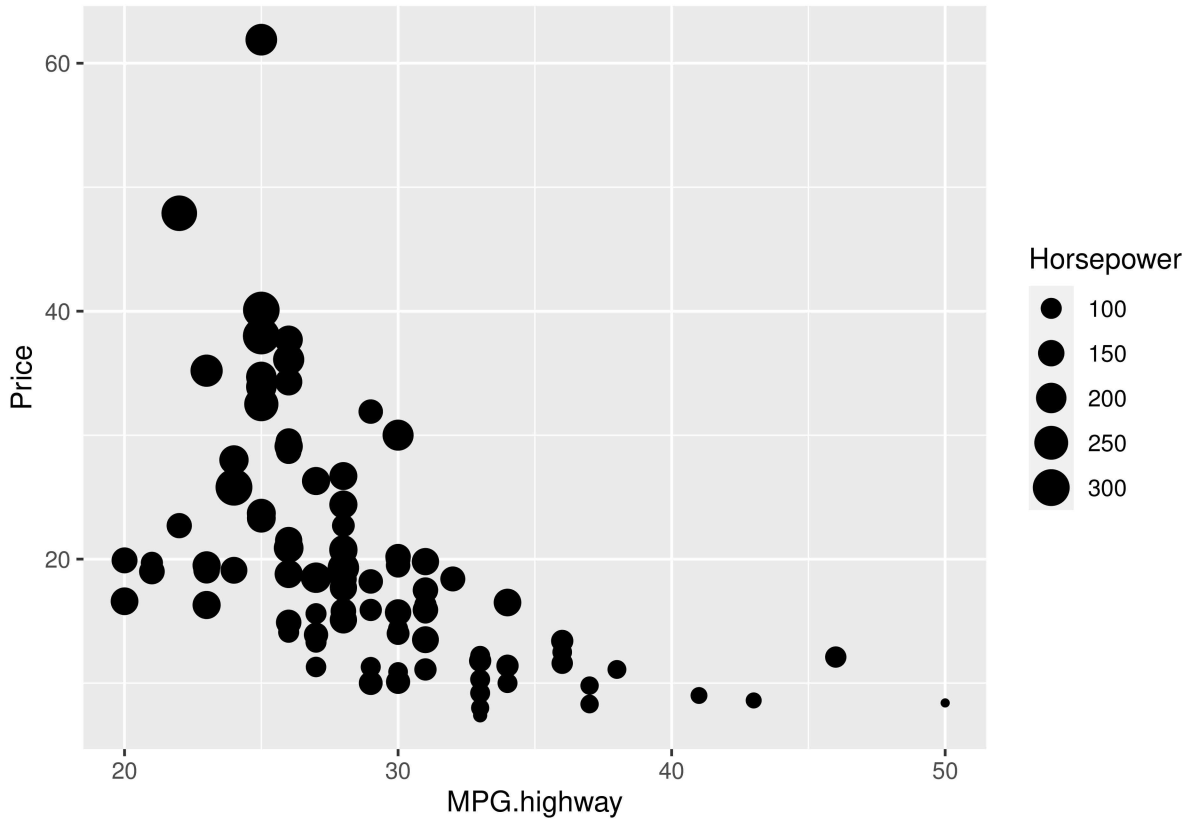
```
##
## 다음의 패키지를 부착합니다: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(tidyverse)
as_tibble(Cars93)
```

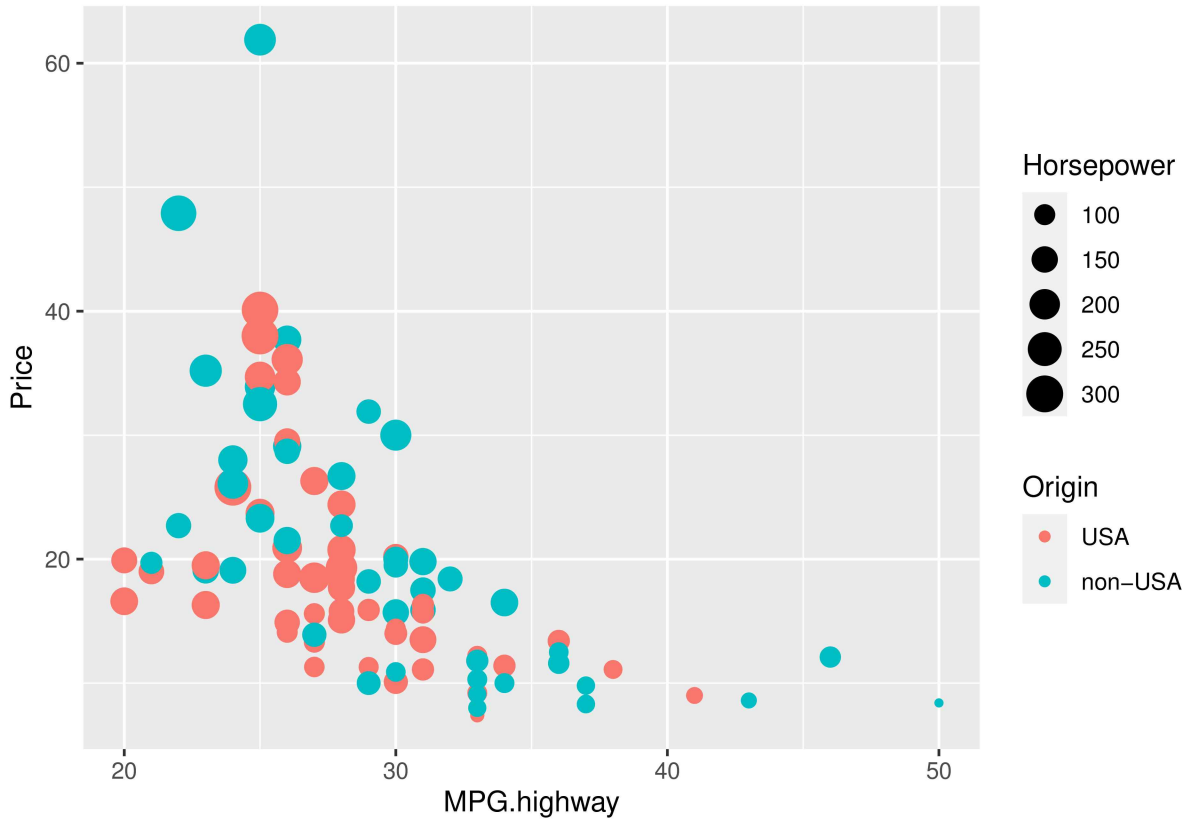
a.

```
Cars93%>%ggplot(aes(x=MPG.highway,y=Price,size=Horsepower))+geom_point()
```



b.

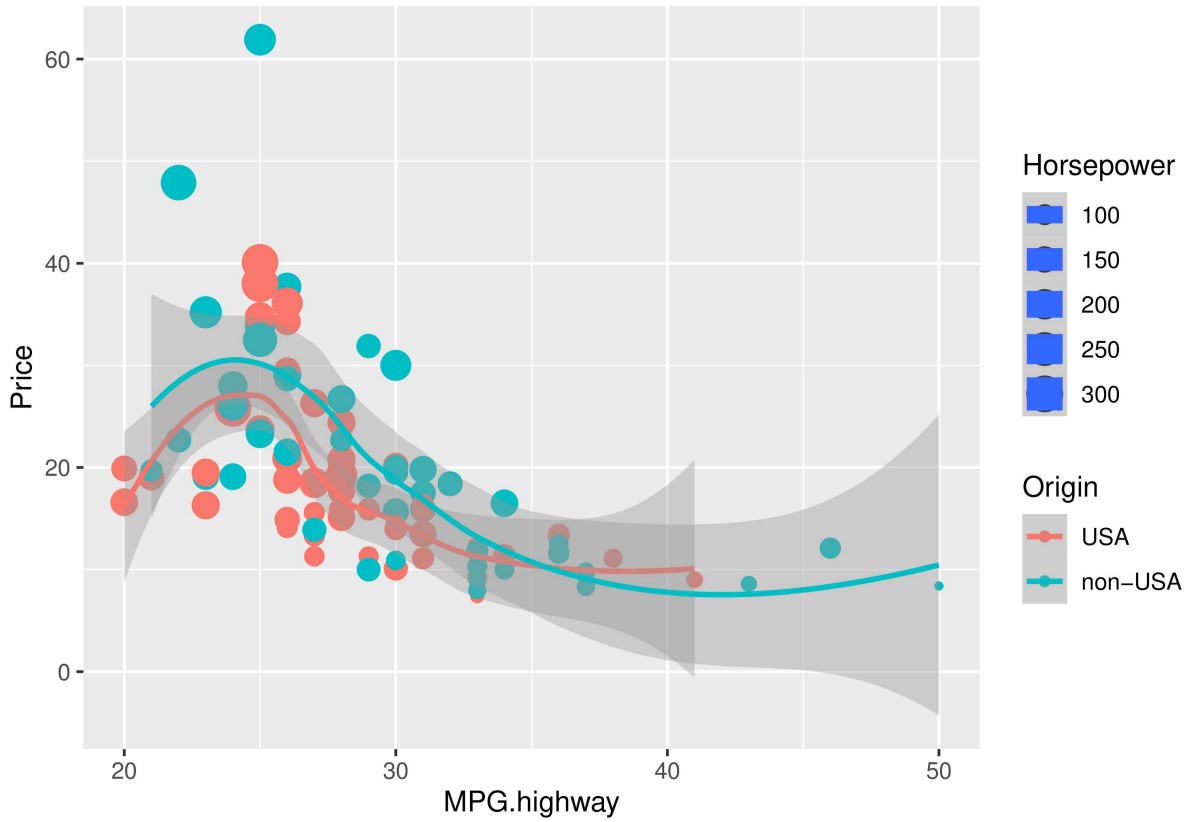
```
Cars93%>%ggplot(aes(x=MPG.highway,y=Price,color=Origin,size=Horsepower))+geom_point()
```



c.

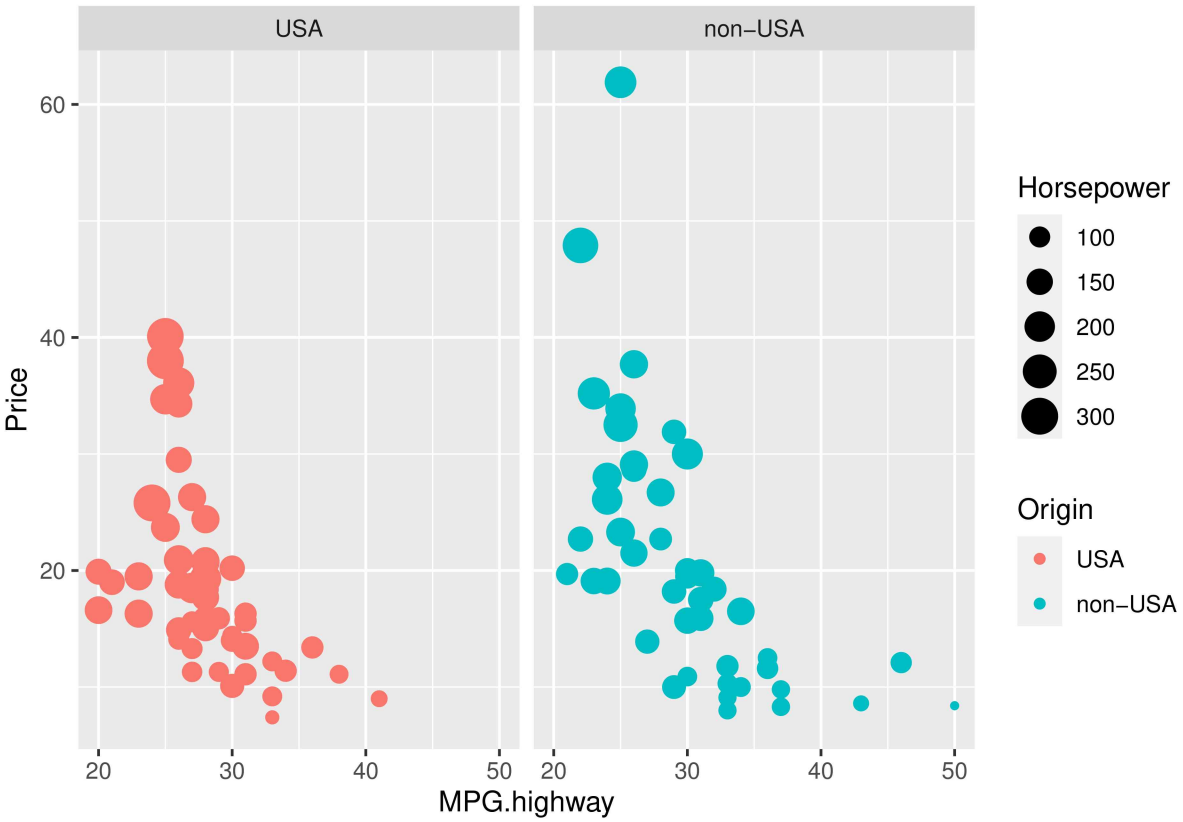
```
Cars93%>%ggplot(aes(x=MPG.highway,y=Price,color=Origin,size=Horsepower))+geom_point()+
  stat_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



d.

```
Cars93%>%ggplot(aes(x=MPG.highway,y=Price,color=Origin,size=Horsepower))+geom_point()+
  facet_grid(.~Origin)
```


USA
non−USA
60
40
20
Price
20
30
40
50
MPG.highway
Horsepower
100
150
200
250
300
Origin
USA
non−USA