STAT346_hw2_2019150445

2019150445/ShinBaekRok

2020 10 2

1.

```
library(gapminder)
data(gapminder)
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
```

a)

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

```
## `summarise()` ungrouping output (override with `.groups` argument)

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

b)

```
gapminder %>% filter(continent=='Europe',year==1997) %>%
  arrange(gdpPercap) %>% head(n=1) %>% select(country)
```

```
## # A tibble: 1 x 1
##   country
##   <fct>
## 1 Albania
```

```
gapminder %>% filter(continent=='Europe',year==2007) %>%
  arrange(gdpPercap) %>% head(n=1) %>% select(country)
```

```
## # A tibble: 1 x 1
##   country
##   <fct>
## 1 Albania
```

c)

```
gapminder %>% filter(year%in% 1980:1989) %>%
  group_by(continent) %>% summarize(mean(lifeExp))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 5 x 2
##   continent `mean(lifeExp)`
##   <fct>                <dbl>
## 1 Africa                52.5
## 2 Americas              67.2
## 3 Asia                  63.7
## 4 Europe                73.2
## 5 Oceania               74.8
```

d)

```
gapminder %>% mutate(GDP=gdpPercap*pop) %>%
  group_by(country) %>% summarize(totalGDP=sum(GDP))%>%
  arrange(-totalGDP) %>% head(n=5)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

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

e)

```
gapminder %>% filter(lifeExp>=80) %>% select(country,lifeExp,year)
```

```
## # A tibble: 22 x 3
##    country          lifeExp  year
##    <fct>              <dbl> <int>
##  1 Australia           80.4  2002
##  2 Australia           81.2  2007
##  3 Canada              80.7  2007
##  4 France              80.7  2007
##  5 Hong Kong, China    80    1997
##  6 Hong Kong, China    81.5  2002
##  7 Hong Kong, China    82.2  2007
##  8 Iceland             80.5  2002
##  9 Iceland             81.8  2007
## 10 Israel              80.7  2007
## # ... with 12 more rows
```

f)

```
gapminder %>% group_by(country) %>% summarize(corr=cor(lifeExp,gdpPercap))%>%
  arrange(-abs(corr)) %>% head(10)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 10 x 2
##    country         corr
##    <fct>          <dbl>
##  1 France         0.996
##  2 Austria        0.993
##  3 Belgium        0.993
##  4 Norway         0.992
##  5 Oman           0.991
##  6 United Kingdom 0.990
##  7 Italy          0.990
##  8 Israel         0.988
##  9 Denmark        0.987
## 10 Australia      0.986
```

g)

```
gapminder %>% filter(continent!='Asia') %>% group_by(continent, year) %>%
  summarize(avg=mean(pop)) %>% arrange(-avg)
```

```
## `summarise()` regrouping output by 'continent' (override with `.groups` argument)
```

```
## # A tibble: 48 x 3
## # Groups:   continent [4]
##    continent  year       avg
##    <fct>     <int>     <dbl>
##  1 Americas   2007 35954847.
```

```
##  2 Americas   2002 33990910.
##  3 Americas   1997 31876016.
##  4 Americas   1992 29570964.
##  5 Americas   1987 27310159.
##  6 Americas   1982 25211637.
##  7 Americas   1977 23122708.
##  8 Americas   1972 21175368.
##  9 Europe     2007 19536618.
## 10 Europe     2002 19274129.
## # ... with 38 more rows
```

h)

```
gapminder %>% group_by(country) %>%
  summarize(sd=sd(pop)) %>% arrange(sd) %>% head(3)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 2
##   country                   sd
##   <fct>                  <dbl>
## 1 Sao Tome and Principe 45906.
## 2 Iceland               48542.
## 3 Montenegro            99738.
```

2.

```
library(nycflights13)
library(ggplot2) #To use ggplot package
data(flights)
data(planes)
data(weather)
```

a)

```
#dep_time=NA means that plane does not depart(Plane has cancelled).
x<-flights %>% group_by(month) %>%
  summarize(total=n(),frequency=sum(is.na(dep_time)),proportion=frequency/total)
```

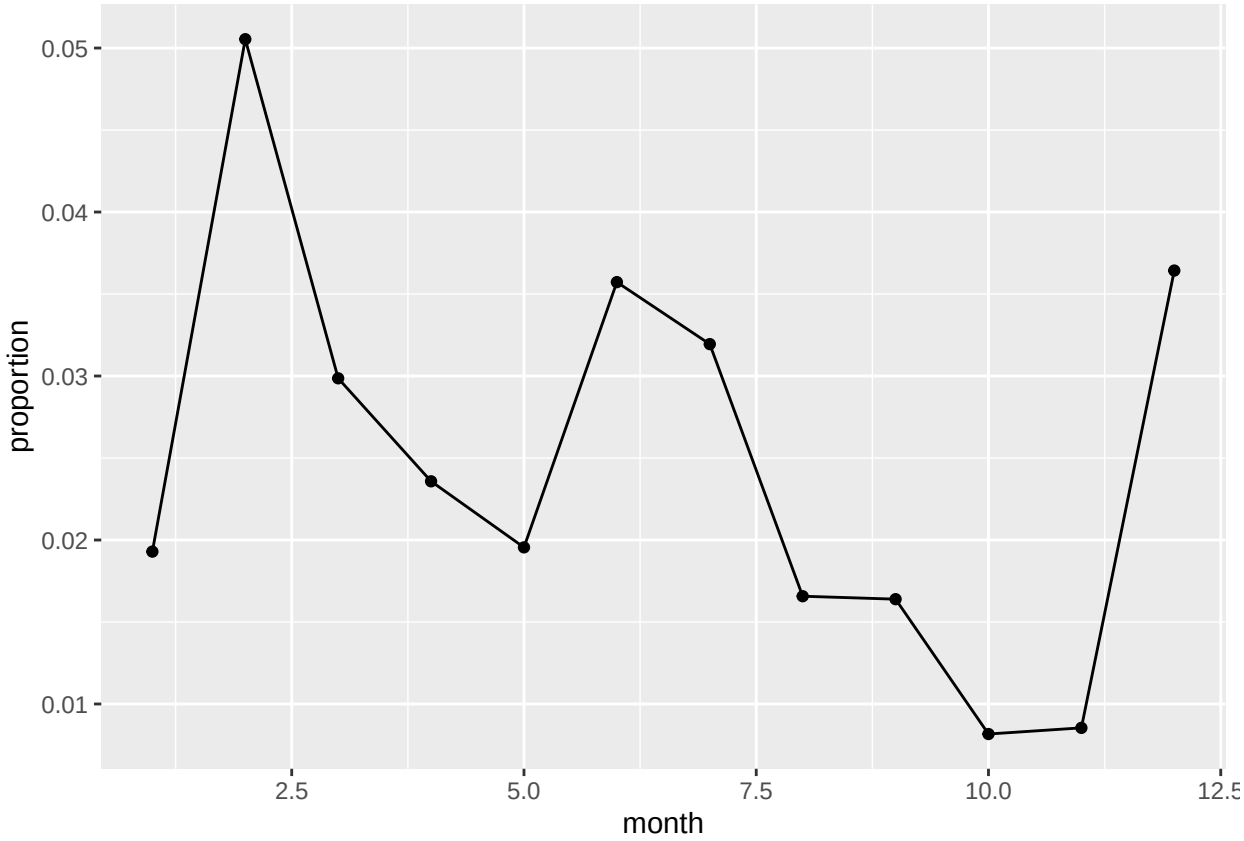
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
x %>% head()
```

```
## # A tibble: 6 x 4
##   month total frequency proportion
##   <int> <int>     <int>      <dbl>
## 1     1 27004       521     0.0193
## 2     2 24951      1261     0.0505
```

```
## 3     3 28834       861     0.0299
## 4     4 28330       668     0.0236
## 5     5 28796       563     0.0196
## 6     6 28243      1009     0.0357
```

```
x %>% ggplot(aes(x=month,y=proportion))+geom_point()+geom_line()
```



Month 2(i.e.Febuary) is the highest. Month 10(i.e. October) is the lowest. As we can see, Summer & Winter has high proportion of cancelled flights. We can interpret this as an impact of snow and storm.

b)

```
flights %>% filter(year==2013, !is.na(tailnum)) %>%
group_by(tailnum) %>%summarize(freq=n()) %>%
arrange(-freq)%>%head(1)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 1 x 2
##   tailnum  freq
##   <chr>   <int>
## 1 N725MQ    575
```

```
library(lubridate) #To calculate each week.
```

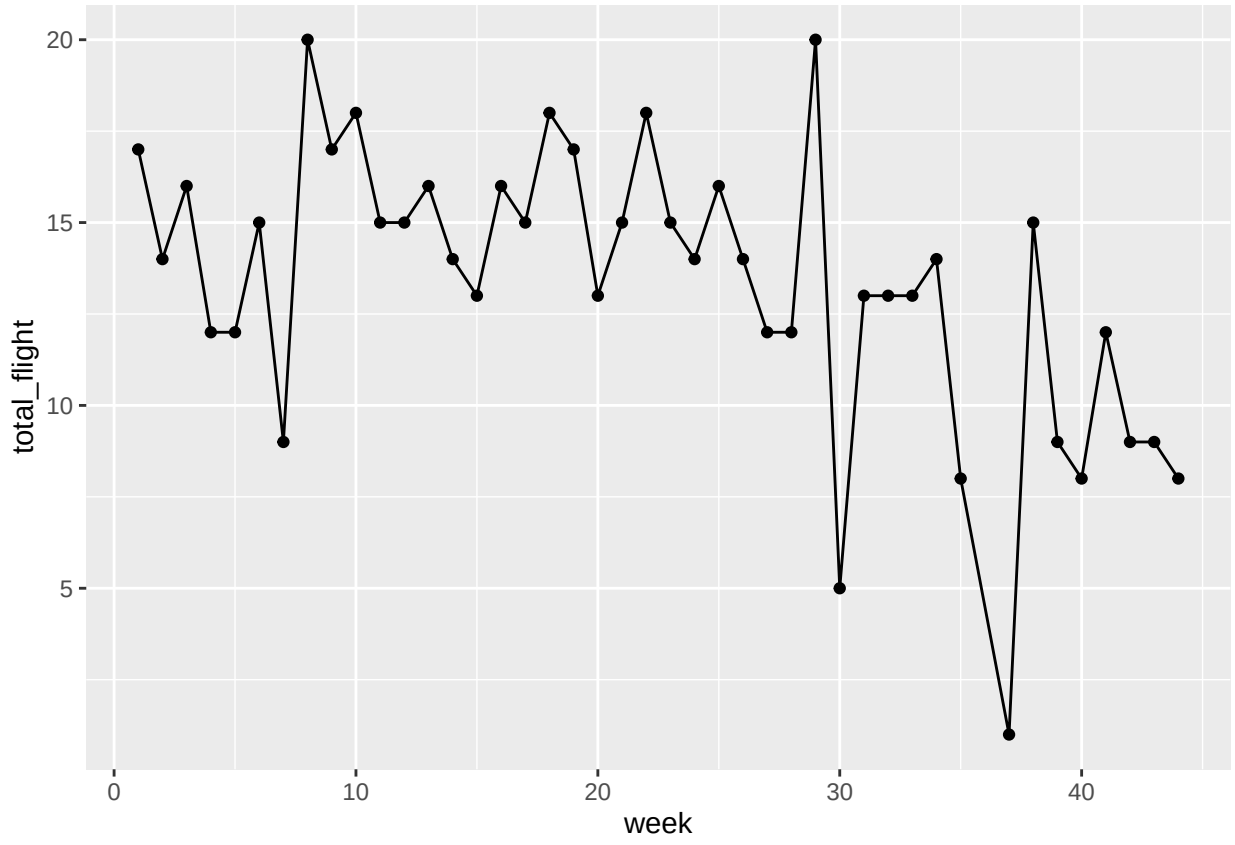
```
## Warning: package 'lubridate' was built under R version 4.0.3
```

```
##
## Attaching package: 'lubridate'
```

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

```
flights$week<-make_date(flights$year,flights$month,flights$day) %>% week()#Add week column to flights .

#Next, filter tailnum='N725MQ' that traveled the most times from NYC in 2013
#and plot the number of trips per week over the year 2013.
flights %>% filter(year==2013 & tailnum=='N725MQ') %>% group_by(week) %>%
  summarize(total_flight=n()) %>%
  ggplot(aes(x=week,y=total_flight))+geom_point()+geom_line()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



c)

```
planes %>% select(tailnum, old=year) %>%
  inner_join(flights,by='tailnum') %>% arrange(old) %>% head(1)
```

```
## # A tibble: 1 x 21
##   tailnum   old  year month   day dep_time sched_dep_time dep_delay arr_time
##   <chr>   <int> <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1 N381AA   1956  2013     1    30      741            745        -4     1059
## # ... with 12 more variables: sched_arr_time <int>, arr_delay <dbl>,
## #   carrier <chr>, flight <int>, origin <chr>, dest <chr>, air_time <dbl>,
## #   distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>, week <dbl>
```

```
sum(unique(flights$tailnum) %in% unique(planes$tailnum))
```

```
## [1] 3322
```

d)

```
#planes$year indicates the year that plane has manufactured.
planes$year %>% is.na() %>% sum()
```

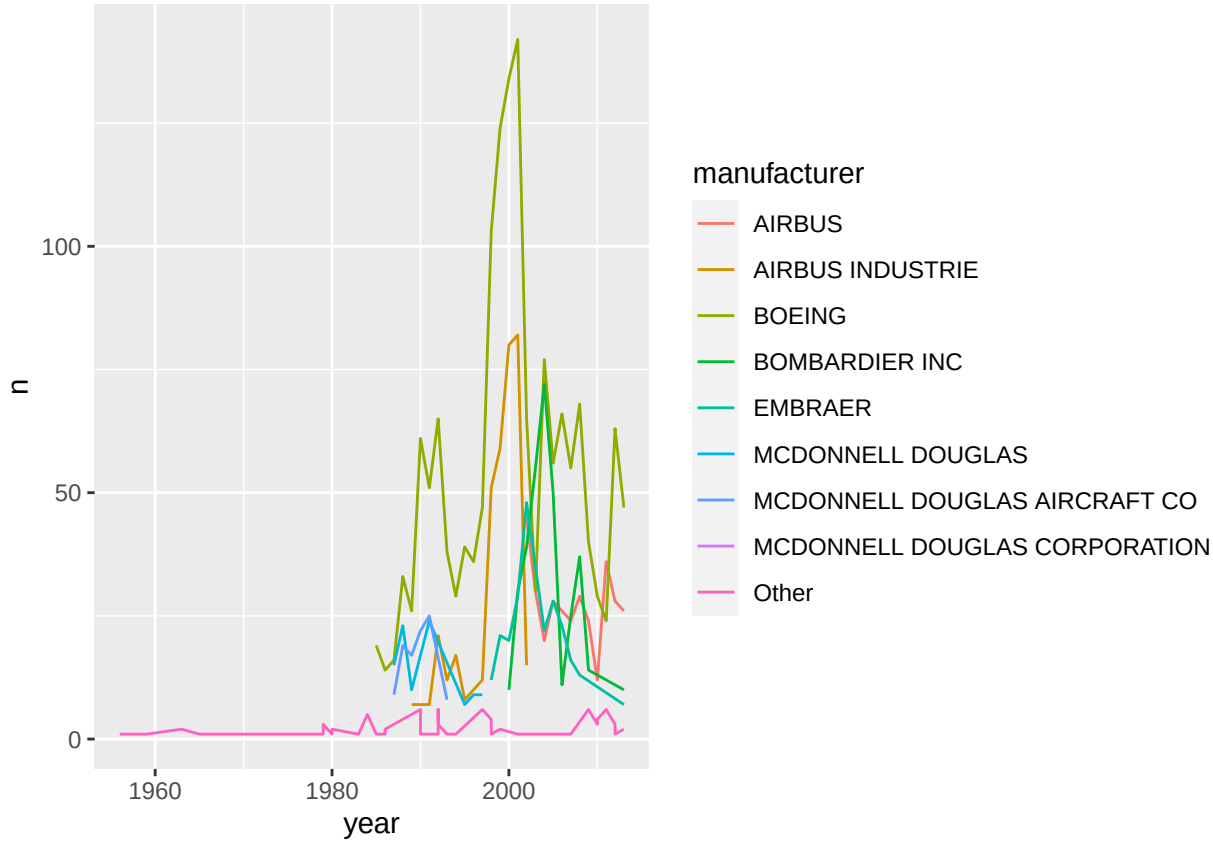
```
## [1] 70
```

```
table(planes$manufacturer) %>% data.frame() %>% arrange(-Freq) %>% head(5)
```

```
##                Var1 Freq
## 1           BOEING 1630
## 2 AIRBUS INDUSTRIE  400
## 3   BOMBARDIER INC  368
## 4           AIRBUS  336
## 5          EMBRAER  299
```

```
planes<-planes %>% count(manufacturer,year) %>%
  mutate(manufacturer=ifelse(n<7,'Other',manufacturer))
# When n<7, manufacturer is classfied as Other.


planes %>% ggplot(aes(x=year,y=n,col=manufacturer)) + geom_line()
```

```
## Warning: Removed 14 row(s) containing missing values (geom_path).
```

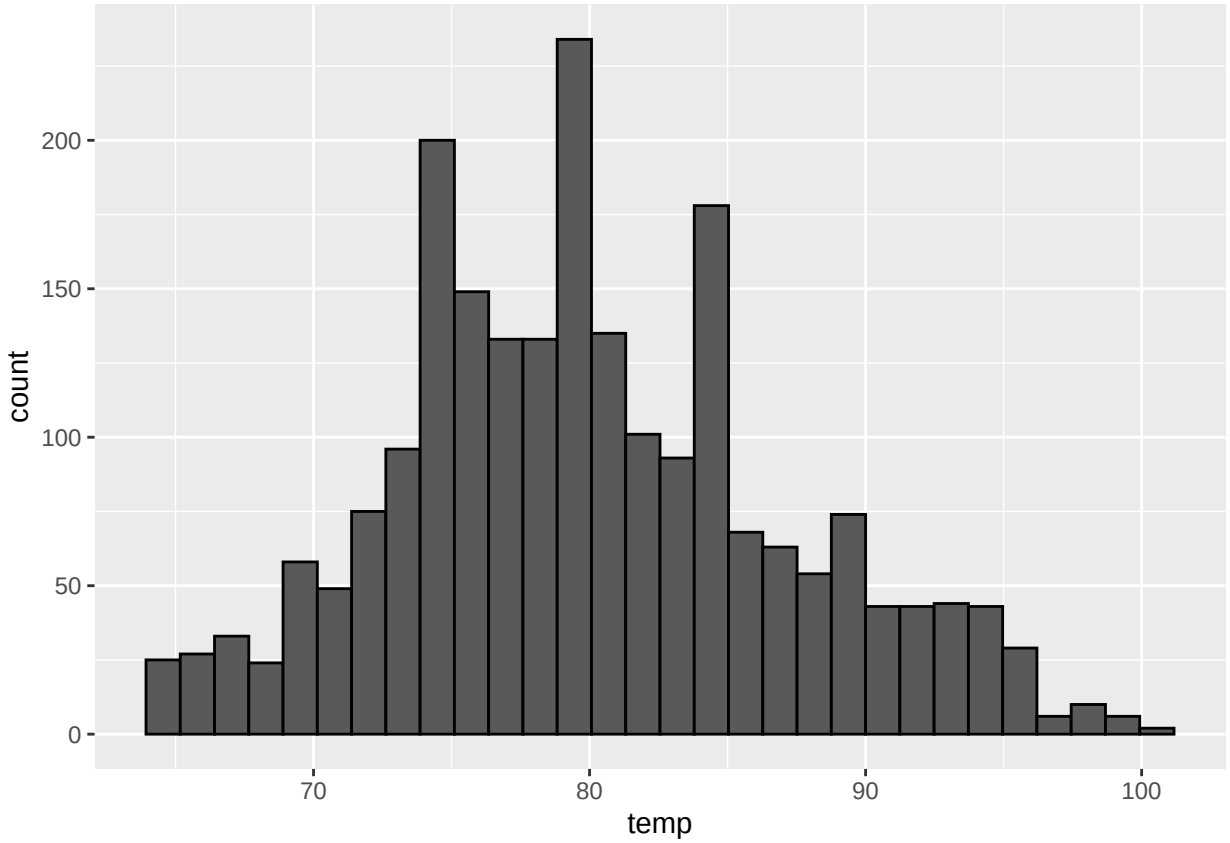


Since 1980's, Boeing is higher in almost all years than any other company.

e)

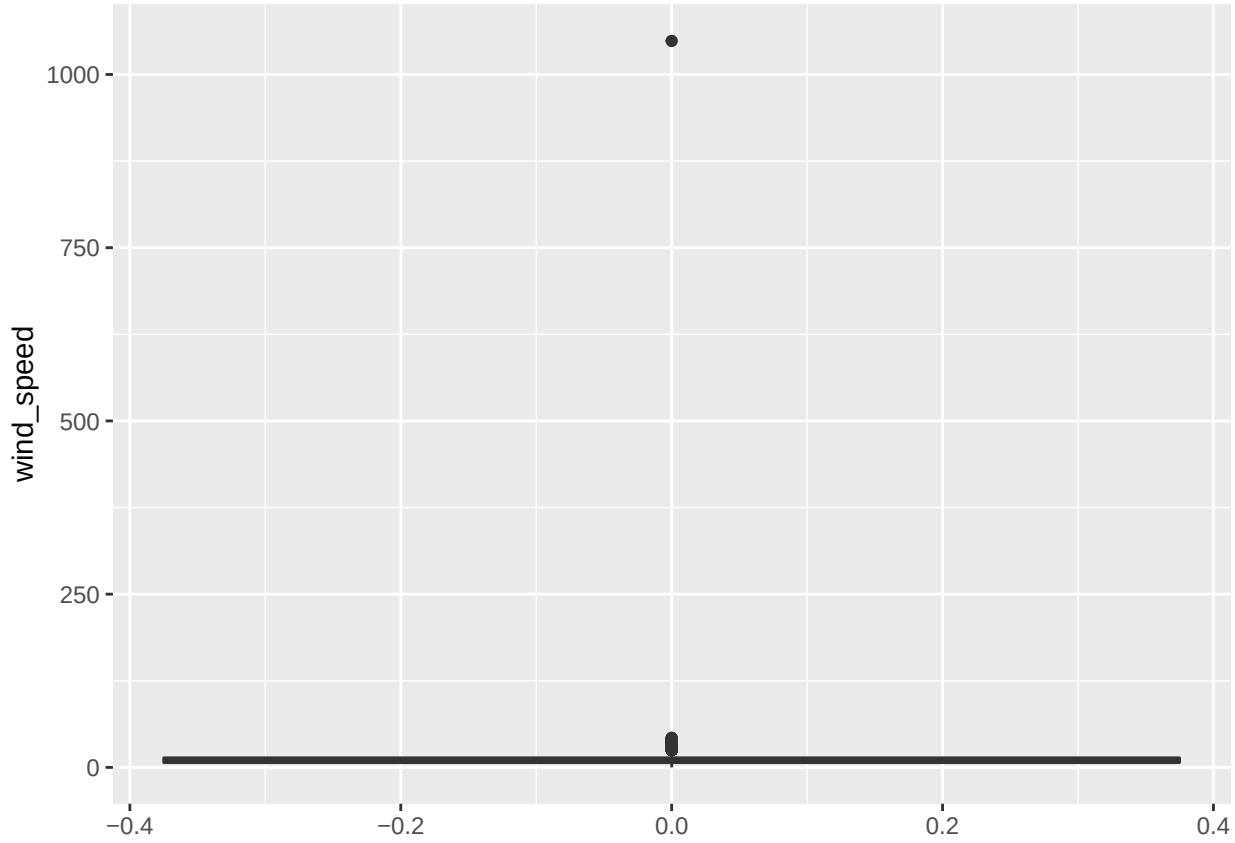
```
weather %>% filter(year==2013 & month==7) %>% ggplot(aes(temp))+geom_histogram(col='black')
```

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




```
#Temp distribution is Slightly skewed to the right.

weather %>% ggplot(aes(y=wind_speed)) + geom_boxplot()
```

```
## Warning: Removed 4 rows containing non-finite values (stat_boxplot).
```

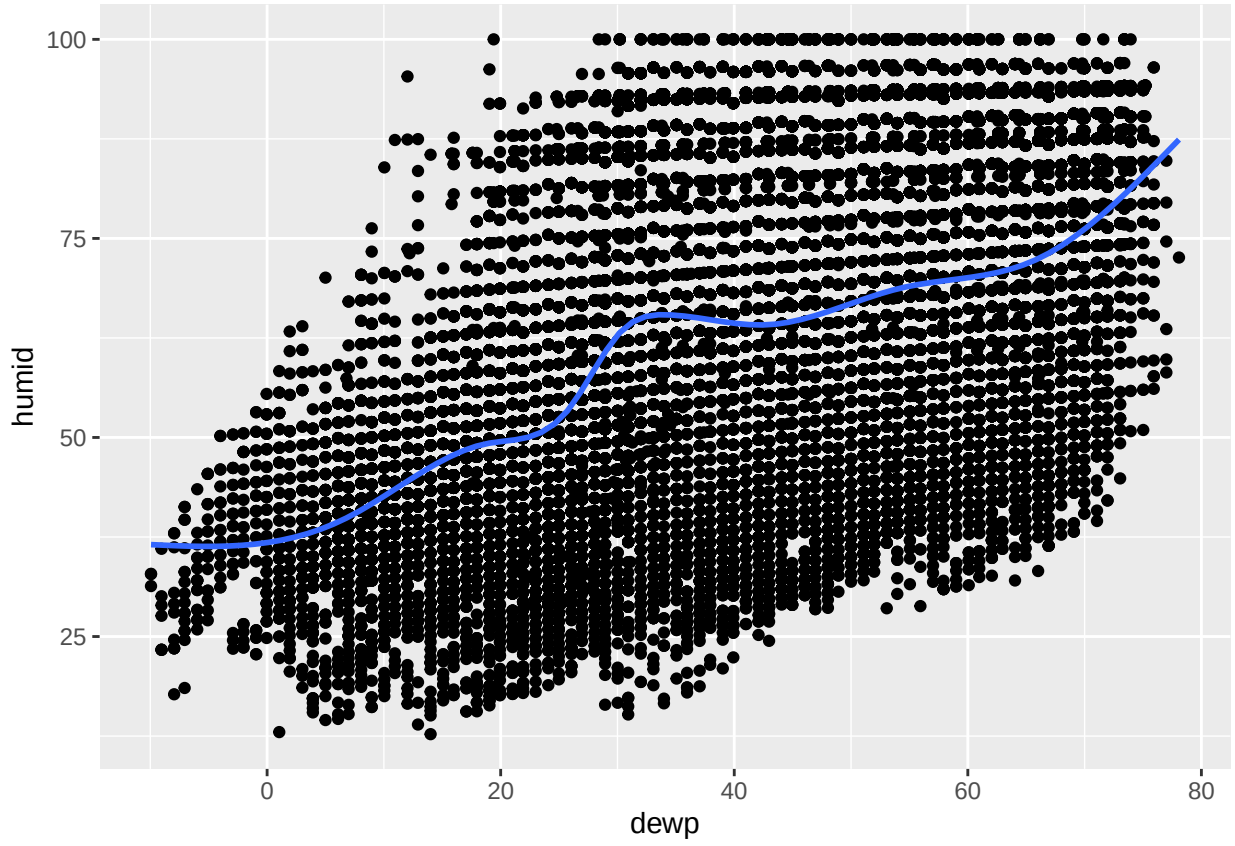



```
# There are some outliers, let's see when wind_speed>1000.

weather %>% filter(wind_speed>1000)
```

```
## # A tibble: 1 x 15
##   origin  year month   day  hour  temp  dewp humid wind_dir wind_speed wind_gust
##   <chr>  <int> <int> <int> <int> <dbl> <dbl> <dbl>    <dbl>      <dbl>     <dbl>
## 1 EWR     2013     2    12     3  39.0  27.0  61.6      260      1048.        NA
## # ... with 4 more variables: precip <dbl>, pressure <dbl>, visib <dbl>,
## #   time_hour <dttm>
```

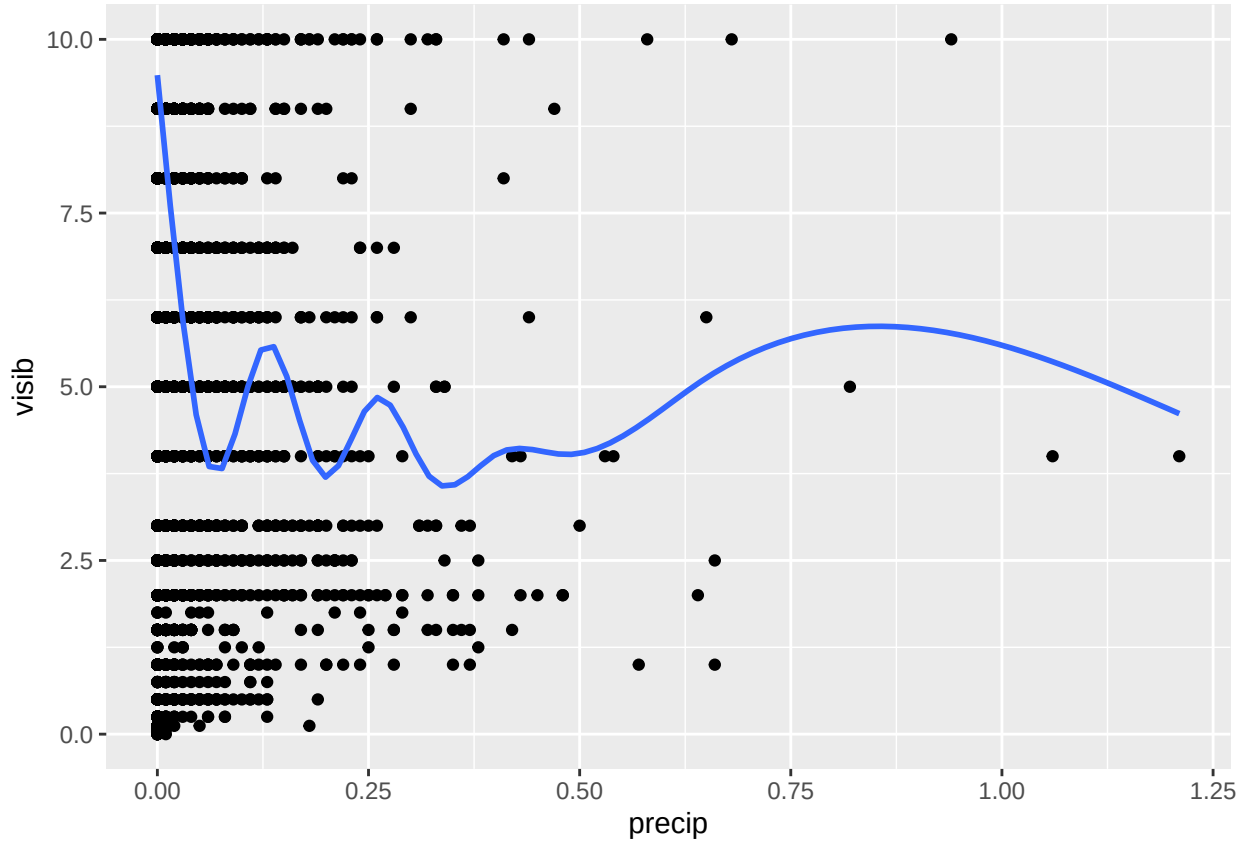
```
weather %>% ggplot(aes(dewp, humid)) +geom_point()+geom_smooth(fill=NA)
```

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

## Warning: Removed 1 rows containing non-finite values (stat_smooth).

## Warning: Removed 1 rows containing missing values (geom_point).
```




```
#There seem to be a positive relationship between dewp & humid.

weather %>% ggplot(aes(precip,visib)) +geom_point()+geom_smooth(fill=NA)
```

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




```
#There seem to be there isn't any relationship between precip & visib
```

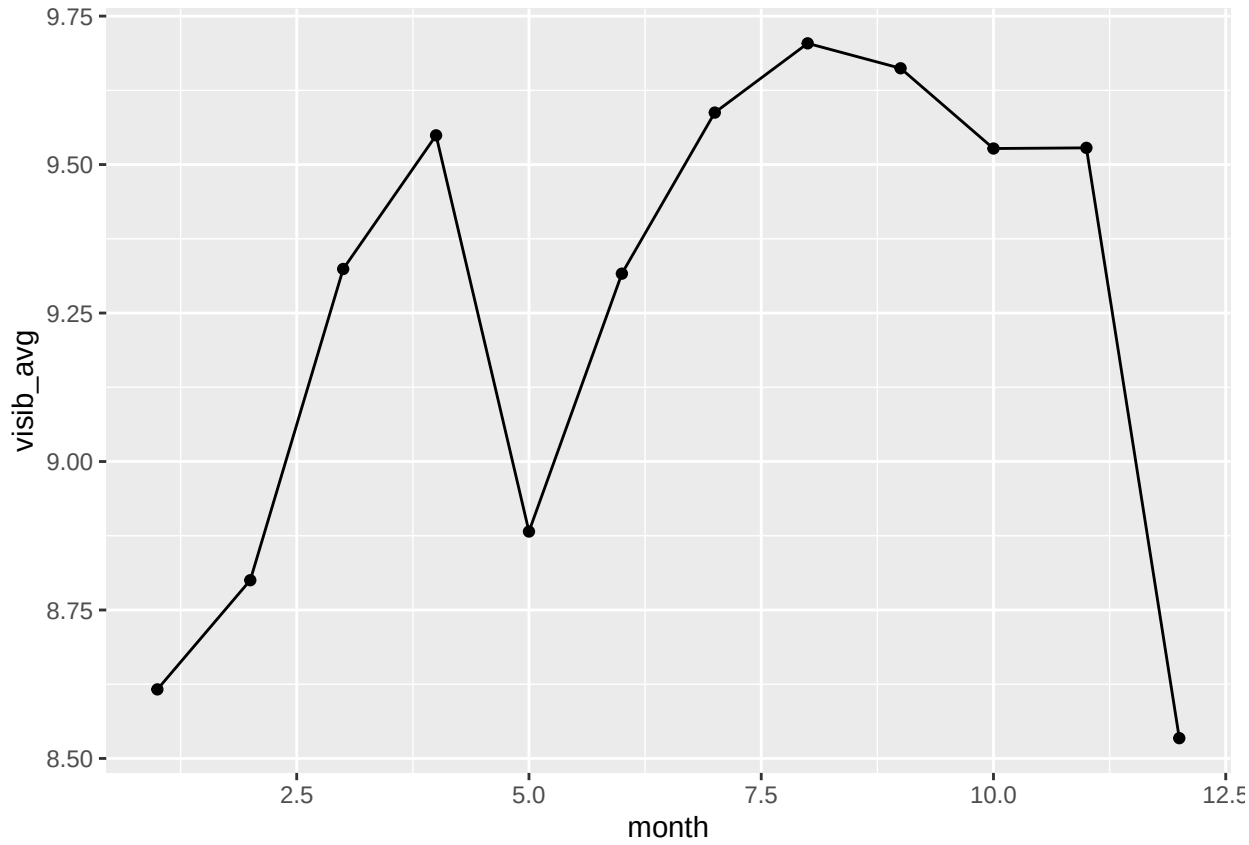
f)

```
weather1<-weather %>% filter(year==2013 & precip>0)
lubridate::make_date(weather1$year, weather1$month, weather1$day) %>% n_distinct()
```

```
## [1] 141
```

```
weather %>% group_by(month) %>% summarize(visib_avg=mean(visib)) %>%
  ggplot(aes(month, visib_avg))+geom_line()+geom_point()
```

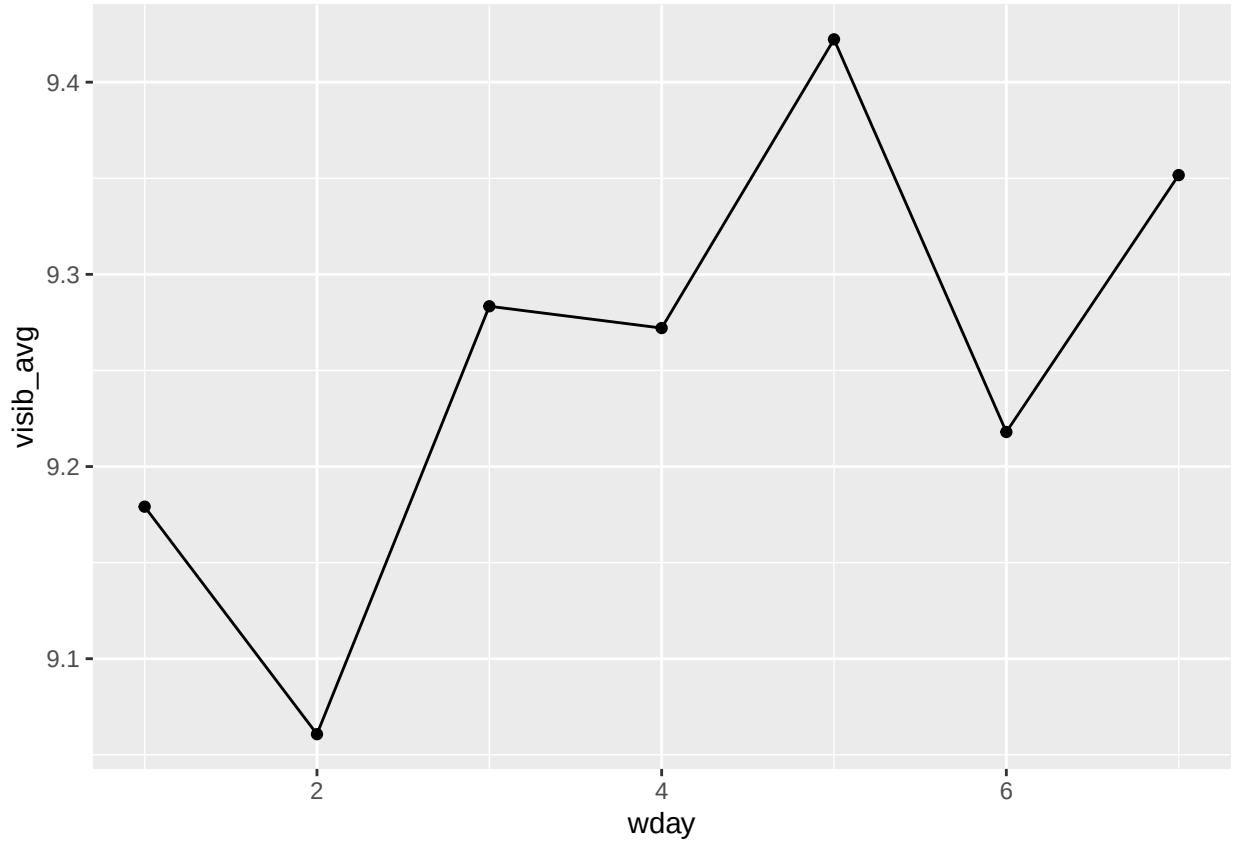
```
## `summarise()` ungrouping output (override with `.groups` argument)
```



Month 12, Month 1,2 are lowest. That means in winter season, visibility is low.

```
weather<-weather %>% mutate(date=make_date(weather$year,weather$month,weather$day))

weather$wday<-weather$date %>% wday()
weather %>% group_by(wday) %>%
  summarize(visib_avg=mean(visib)) %>% ggplot(aes(wday,visib_avg))+geom_line()+geom_point()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



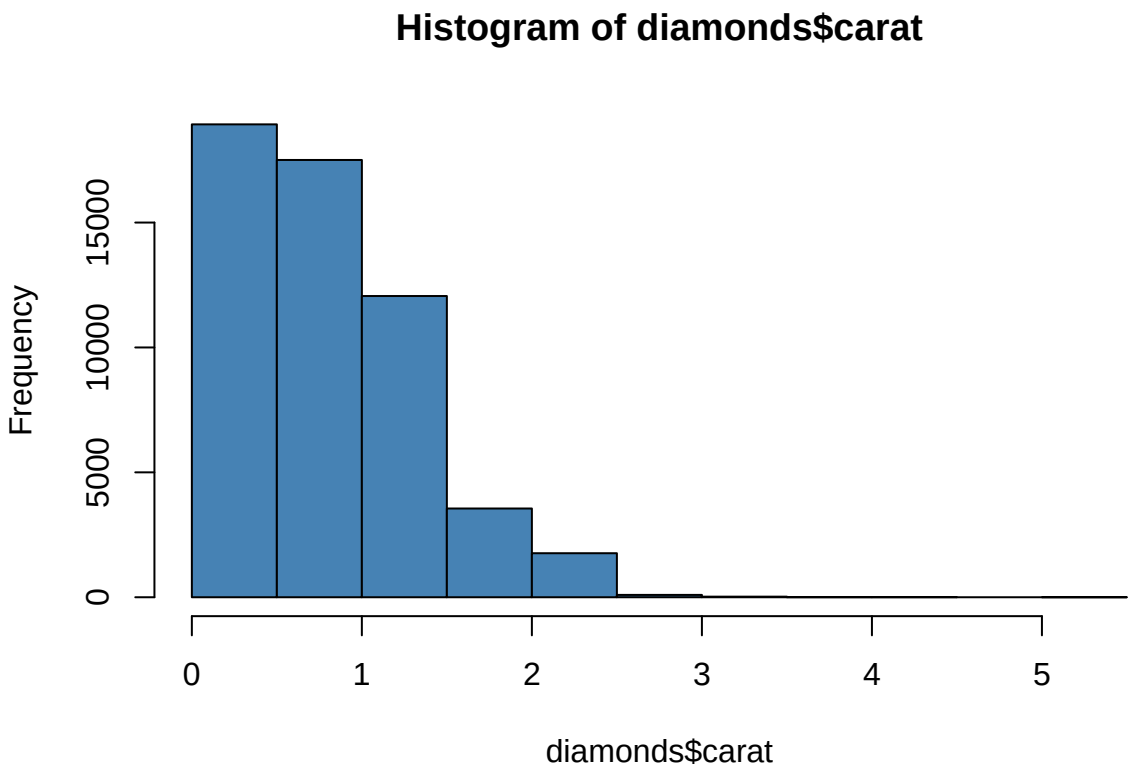
It seems like there is some difference between the day of week, but its lowest average is 9.15 & its highest average is 9.49.(Not that differ)

3.

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

a)

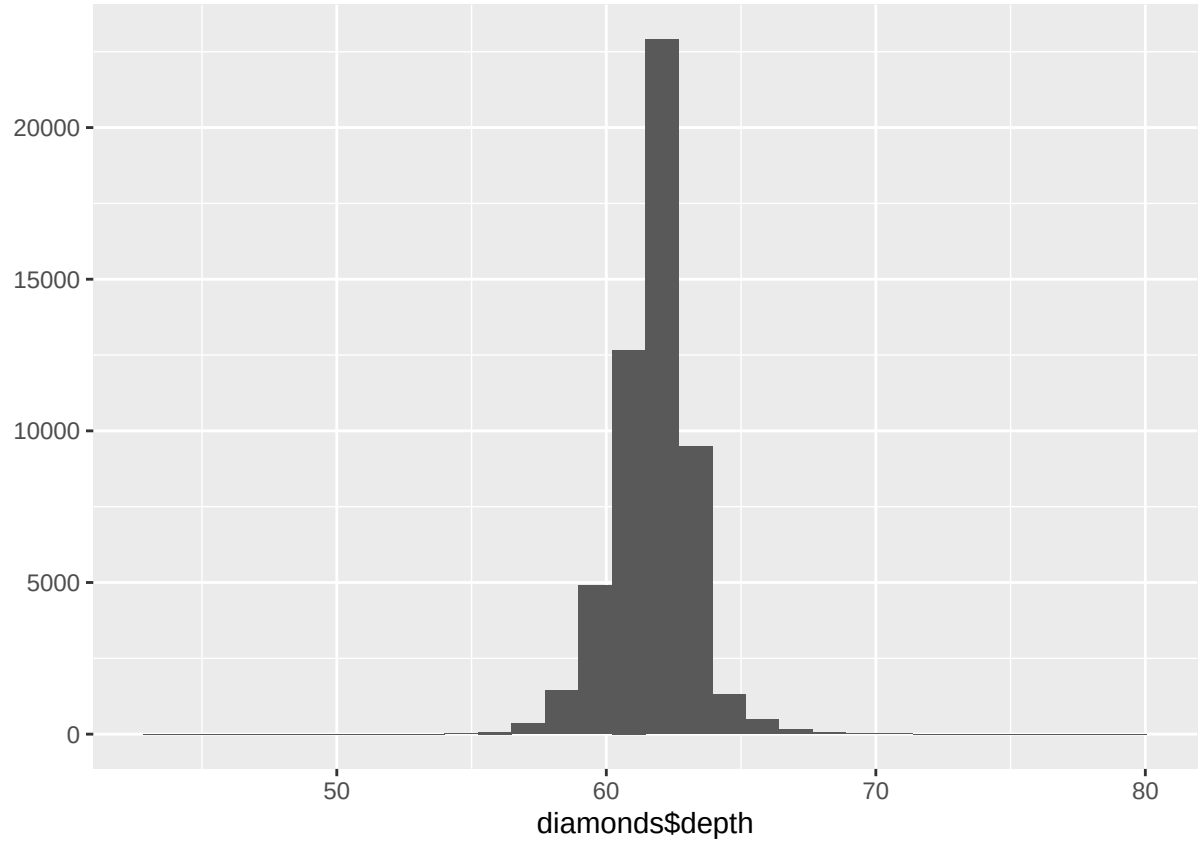
```
hist(diamonds$carat,col='steelblue')
```



b)

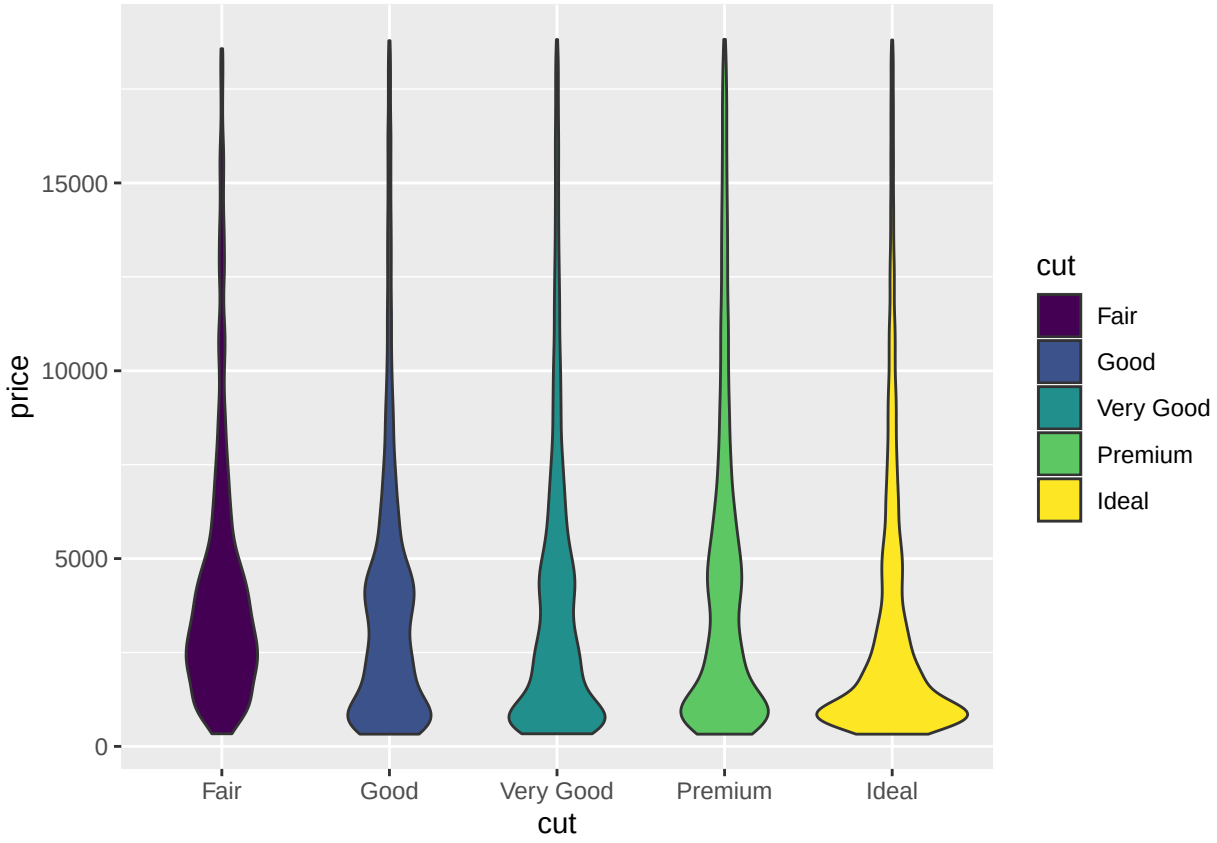
```
qplot(diamonds$depth)
```

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



c)

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



4.

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.3
```

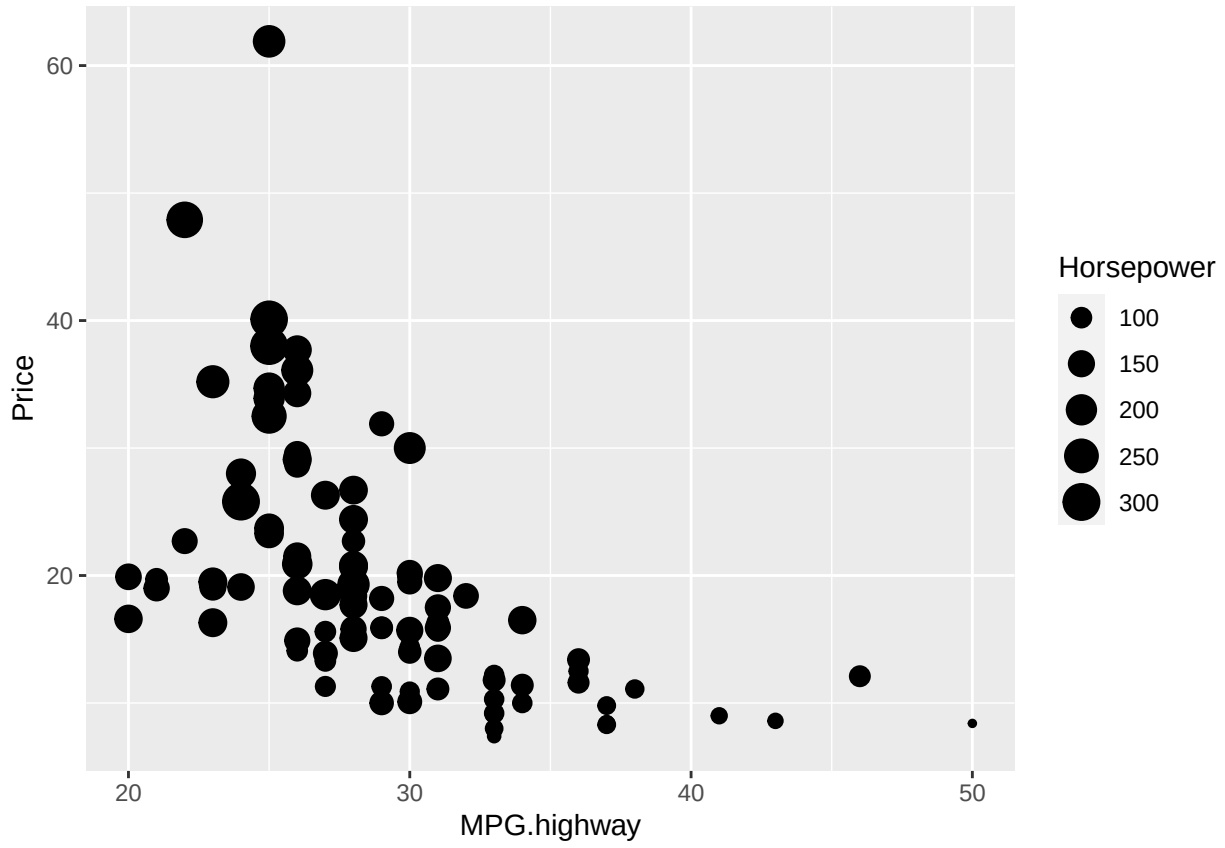
```
##
## Attaching package: 'MASS'
```

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

```
Cars93<-as_tibble(Cars93)
```

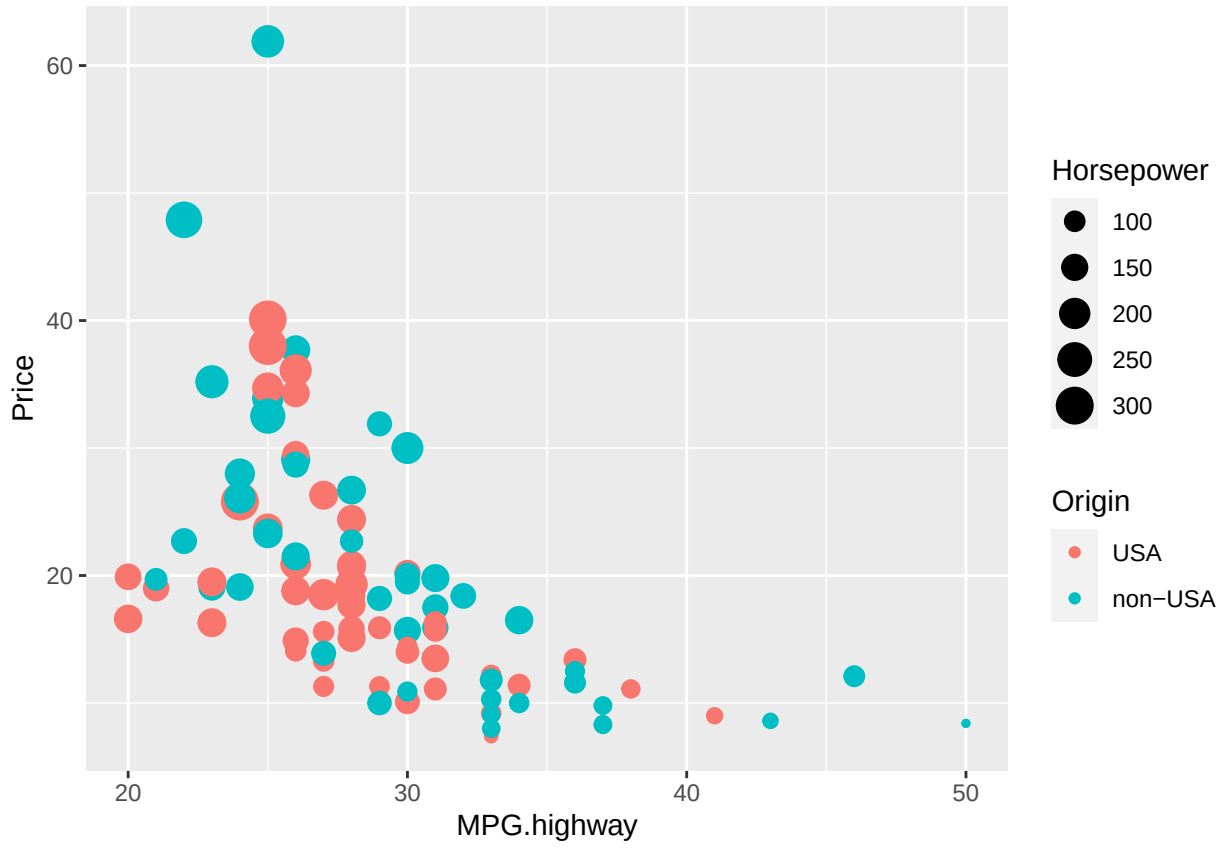
a)

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



b)

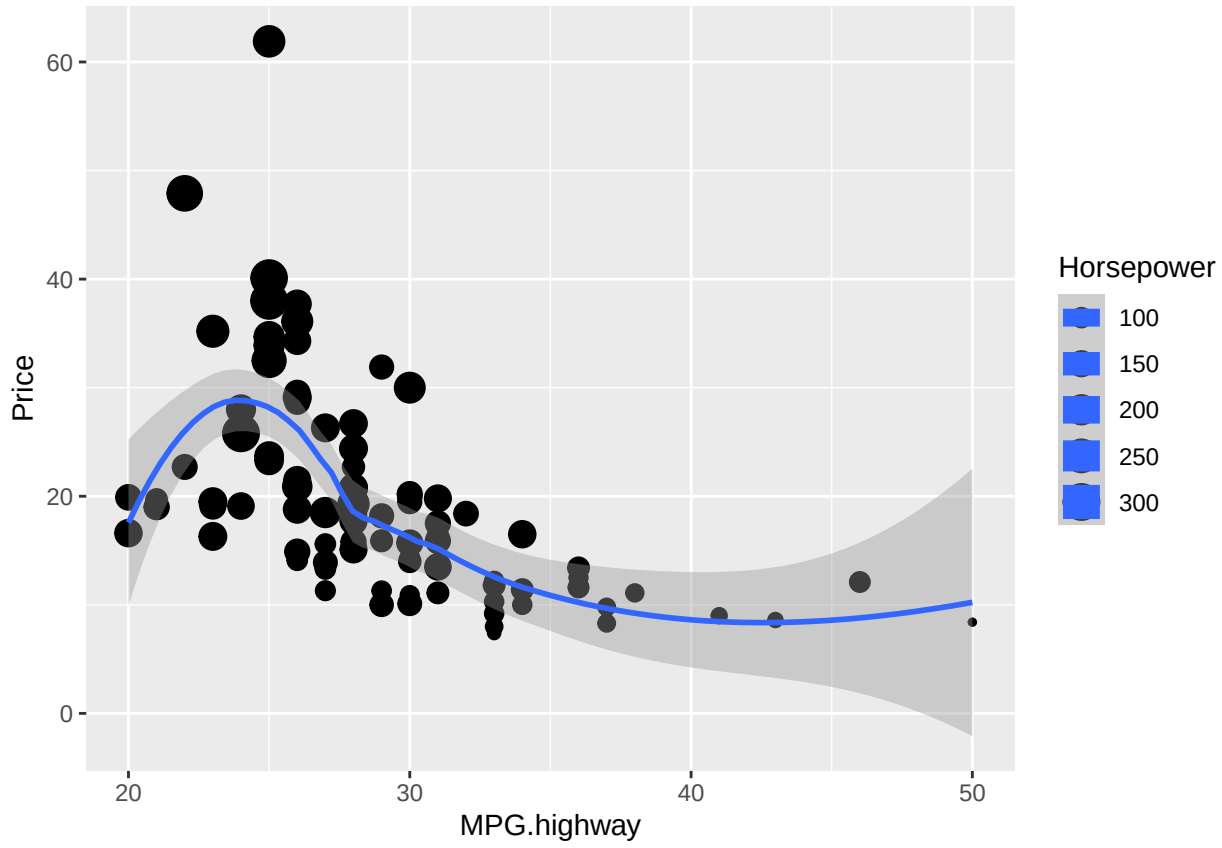
```
Cars93 %>% ggplot(aes(MPG.highway,Price,size=Horsepower,col=Origin))+
  geom_point()
```



c)

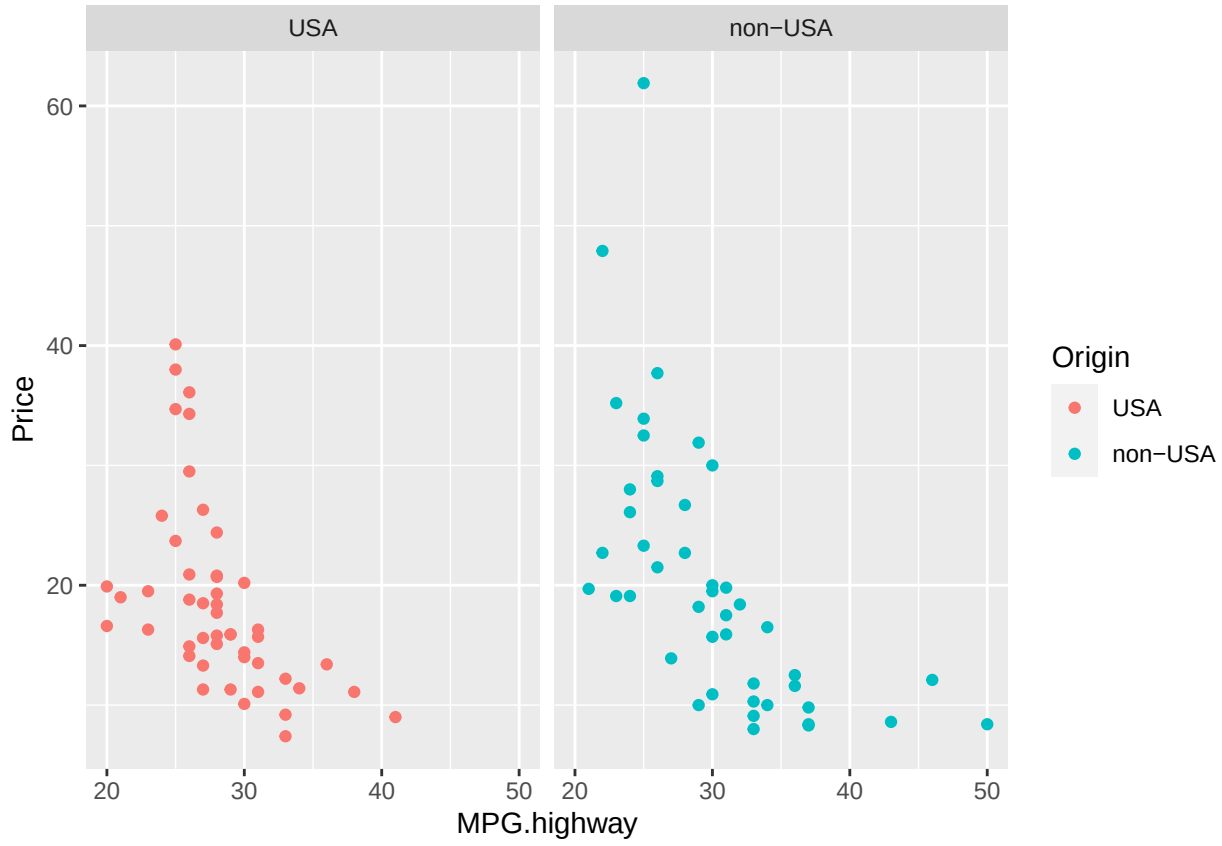
```
Cars93 %>% ggplot(aes(MPG.highway,Price,size=Horsepower))+
  geom_point()+stat_smooth()
```

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



d)

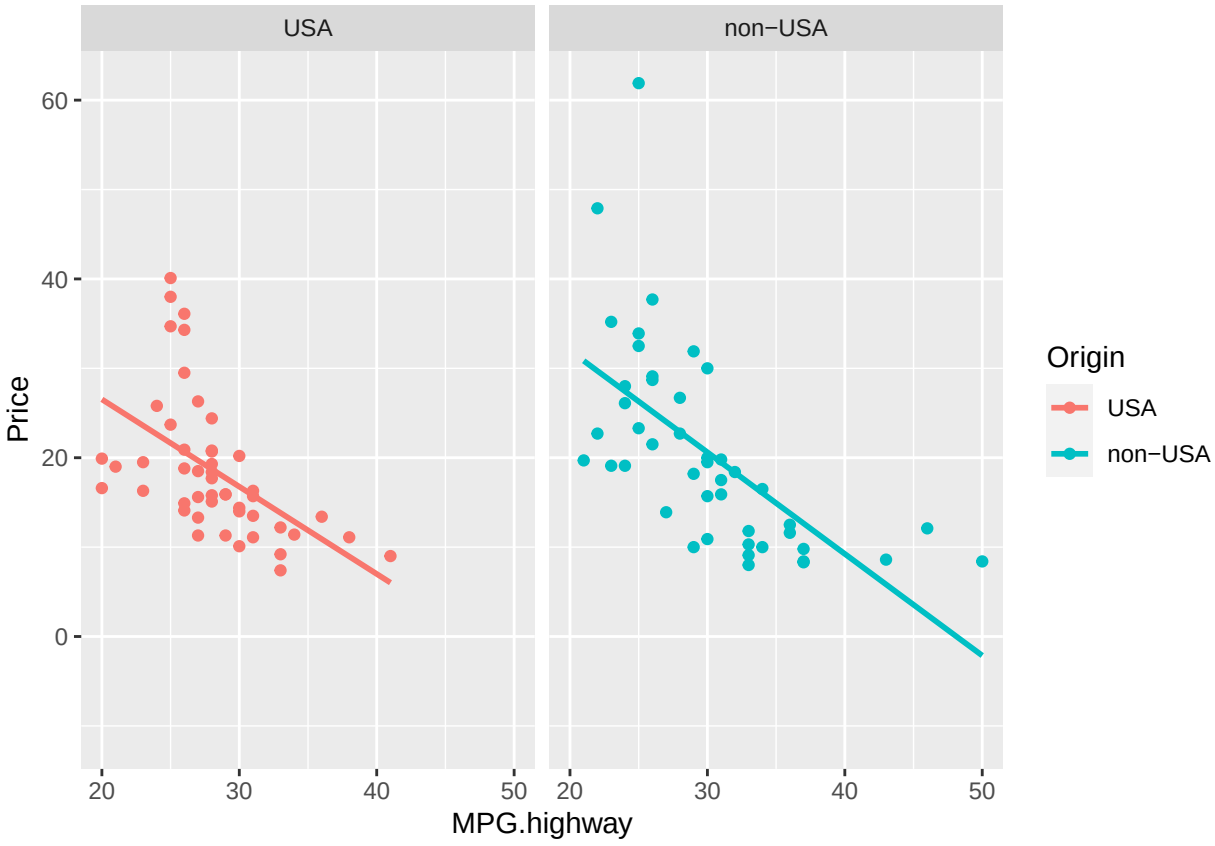
```
Cars93 %>% ggplot(aes(MPG.highway,Price,col=Origin))+
  geom_point()+facet_grid(.~Origin)
```



e)

```
Cars93 %>% ggplot(aes(MPG.highway,Price, col=Origin))+
  geom_point()+geom_smooth(method='lm',fill=NA)+facet_grid(.~Origin)
```

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


USA
non−USA
60
40
20
0
Price
20
30
40
50
MPG.highway
Origin
USA
non−USA