Homework #2

2019150432 임효진

September 30, 2020

1

(a)

```
library("gapminder")
library(dplyr)
gapminder%>%group_by(continent)%>%
    summarize(country_num=n_distinct(country))
```

```
## # A tibble: 5 x 2
##   continent country_num
##   <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)
```

```
## # A tibble: 1 x 6
##   country continent  year lifeExp     pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>   <int>     <dbl>
## 1 Albania Europe     1997    73.0 3428038     3193.
```

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

```
## # A tibble: 1 x 6
##   country continent  year lifeExp     pop gdpPercap
```

```
##   <fct>   <fct>     <int>   <dbl>   <int>     <dbl>
## 1 Albania Europe     2007    76.4 3600523     5937.
```

(c)

```
gapminder%>%filter(between(year, 1980, 1989))%>%
    group_by(continent)%>%
    summarize(avg_life_expectancy=mean(lifeExp))
```

```
## # A tibble: 5 x 2
##   continent avg_life_expectancy
##   <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(GDP=sum(GDP))%>%
  arrange(desc(GDP))%>%
  head(5)
```

```
## # A tibble: 5 x 2
##   country            GDP
##   <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%>%select(country, lifeExp, year)%>%
  filter(lifeExp>=80)
```

```
## # 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(cor=cor(lifeExp, gdpPercap))%>%
  arrange(desc(abs(cor)))%>%
  head(10)
```

```
## # A tibble: 10 x 2
##    country          cor
##    <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(average_population=mean(pop))%>%
  arrange(desc(average_population))%>%
  head(1)
```

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

(h)

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

```
  arrange(sd)%>%
  head(3)
```

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

2

(a)

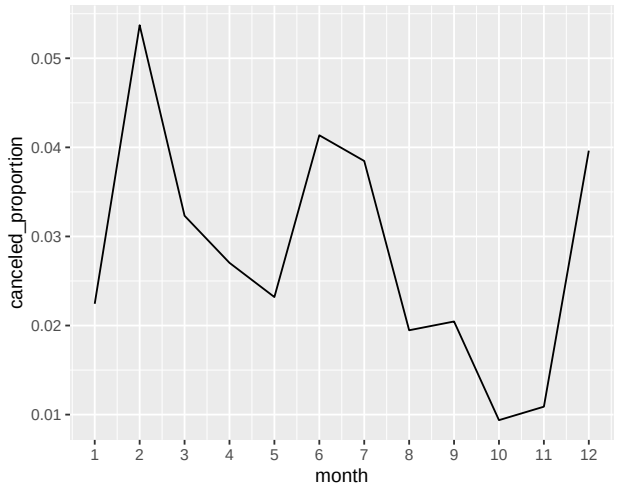
```
library("nycflights13")
data("flights")
data(planes)
data(weather)

flights%>%mutate(canceled=ifelse(is.na(air_time), 1, 0))%>%
  group_by(month)%>%
  summarize(canceled_proportion=sum(canceled)/n())%>%
  arrange(canceled_proportion)%>%
  head(1)
```

```
## # A tibble: 1 x 2
##   month canceled_proportion
##   <int>               <dbl>
## 1    10             0.00938
```

```
flights%>%mutate(canceled=ifelse(is.na(air_time), 1, 0))%>%
  group_by(month)%>%
  summarize(canceled_proportion=sum(canceled)/n())%>%
  arrange(desc(canceled_proportion))%>%
  head(1)
```

```
## # A tibble: 1 x 2
##   month canceled_proportion
##   <int>               <dbl>
## 1     2              0.0537
```

```
library(ggplot2)
canceled_data=flights%>%
  mutate(canceled=ifelse(is.na(air_time), 1, 0))%>%
  group_by(month)%>%
  summarize(canceled_proportion=sum(canceled)/n())
canceled_data%>%ggplot(aes(month, canceled_proportion))+
  geom_line()+scale_x_continuous(breaks = seq(1, 12))
```



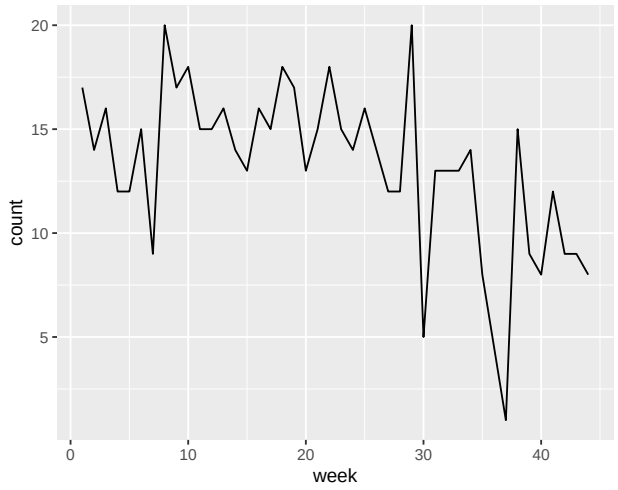
We can interpret that winter and summer has a high cancel proprotion compared to spring and autumn.

(b)

```
flights%>%filter(!is.na(tailnum))%>%
  group_by(tailnum)%>%
  summarize(count=n())%>%
  arrange(desc(count))%>%
  head(1)
```

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

```
library(lubridate)

flights%>%filter(tailnum=="N725MQ")%>%
  mutate(week=week(time_hour))%>%
  group_by(week)%>%
  summarize(count=n())%>%
  ggplot(aes(week, count))+
  geom_line()
```



(c)

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

```
## # A tibble: 1 x 2
## # Groups:   tailnum [1]
##   tailnum year.y
##   <chr>    <int>
## 1 N10156    2013
```

```
planes%>%inner_join(flights, by='tailnum')%>%
  select(tailnum)%>%
  n_distinct()
```

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

(d)

```
planes%>%group_by(tailnum)%>%
  filter(is.na(manufacturer)==T)
```

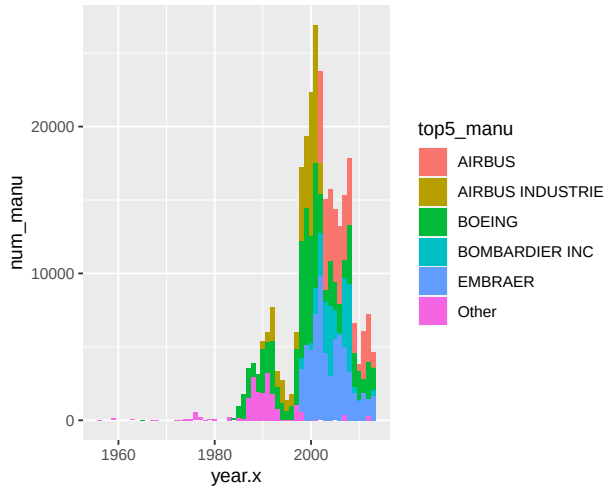
```
## # A tibble: 0 x 9
## # Groups:   tailnum [0]
## # ... with 9 variables: tailnum <chr>, year <int>, type <chr>,
## #   manufacturer <chr>, model <chr>, engines <int>, seats <int>, speed <int>,
## #   engine <chr>
```

```
planes%>%group_by(manufacturer)%>%
  summarize(count=n())%>%
```

```
  arrange(desc(count))%>%
  head(5)
```

```
## # A tibble: 5 x 2
##   manufacturer     count
##   <chr>            <int>
## 1 BOEING            1630
## 2 AIRBUS INDUSTRIE   400
## 3 BOMBARDIER INC     368
## 4 AIRBUS             336
## 5 EMBRAER            299
```

```
planes%>%inner_join(flights, by='tailnum')%>%
  mutate(top5_manu=
                 ifelse(manufacturer%in%c("BOEING",
                                          "AIRBUS INDUSTRIE",
                                          "BOMBARDIER INC",
                                          "AIRBUS",
                                          "EMBRAER"),
                        manufacturer, "Other"))%>%
  group_by(year.x, top5_manu)%>%
  summarize(num_manu=n())%>%
  ggplot(aes(year.x, num_manu, fill=top5_manu))+
  geom_bar(stat = "identity")
```

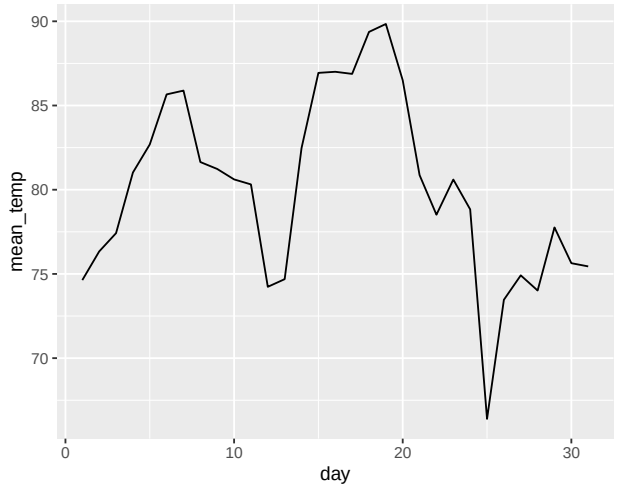


The minor manufacturer companies are no seldom seen after the year of 2000. Therfore we can conclude that the distribution of the manufacturers changed as time went by.

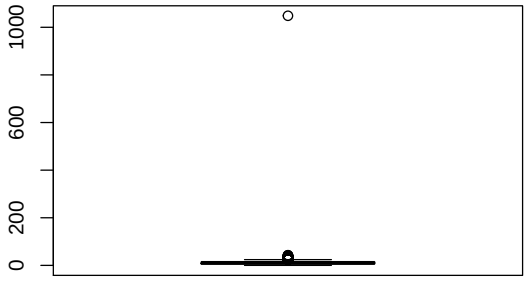
(e)

```
weather%>%filter(month==7)%>%
  group_by(day)%>%
  summarize(mean_temp=mean(temp))%>%
```

```
  ggplot(aes(day, mean_temp))+
  geom_line()
```




```
boxplot(weather$wind_speed)
```




```
boxplot.stats(weather$wind_speed)$out%>%head(10)
```

```
##  [1] 25.31716 26.46794 25.31716 28.76950 25.31716 25.31716 31.07106 27.61872
##  [9] 40.27730 42.57886
```

```
# There seems to be a lot of outliers.
which.max(weather$wind_speed)
```

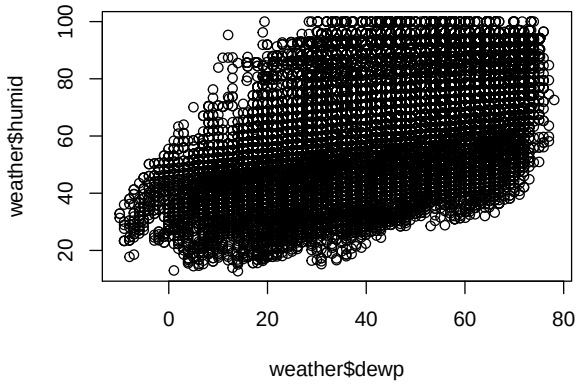
```
## [1] 1010
```

```
weather[1010, ]%>%select(wind_speed)
```

```
## # A tibble: 1 x 1
##   wind_speed
##        <dbl>
## 1      1048.
```

```
# The most extreme outlier is in row 1010, with value 1048.361
```

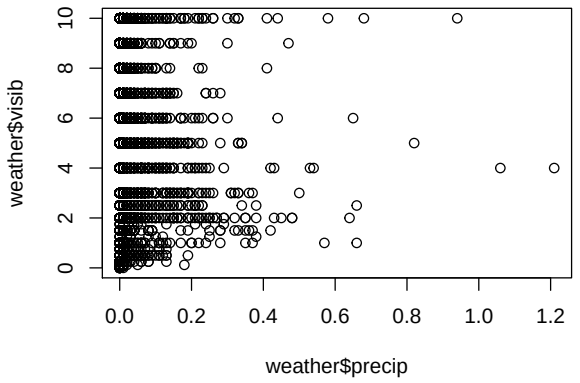
```
plot(weather$dewp, weather$humid)
```




```
cor(weather$dewp, weather$humid,  use="complete.obs")
```

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

```
plot(weather$precip, weather$visib)
```




```
cor(weather$precip, weather$visib,  use="complete.obs")
```

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

The former seems to have a positive correlation while the latter seems to have little correlation.

(f)

```
weather%>%group_by(month, day)%>%
  summarize(count=sum(precip))%>%
```

```
  filter(!count==0)%>%
  nrow()
```

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

```
weather%>%mutate(weekday=weekdays.POSIXt(time_hour))%>%
  group_by(weekday)%>%
  select(visib, weekday)%>%
  summarize(avg_visib=mean(visib))
```

```
## # A tibble: 7 x 2
##   weekday avg_visib
##   <chr>       <dbl>
## 1          9.22
## 2          9.42
## 3          9.27
## 4          9.06
## 5          9.18
## 6          9.35
## 7          9.28
```

```
weather%>%group_by(month)%>%
  summarize(avg_visib=mean(visib))
```

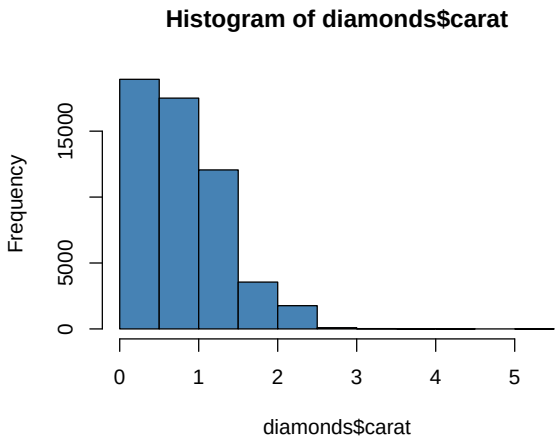
```
## # A tibble: 12 x 2
##    month avg_visib
##    <int>     <dbl>
##  1     1      8.62
##  2     2      8.80
##  3     3      9.32
##  4     4      9.55
##  5     5      8.88
##  6     6      9.32
##  7     7      9.59
##  8     8      9.70
##  9     9      9.66
## 10    10      9.53
## 11    11      9.53
## 12    12      8.53
```

The diffrence between the days of the week and month is insignificant.

3

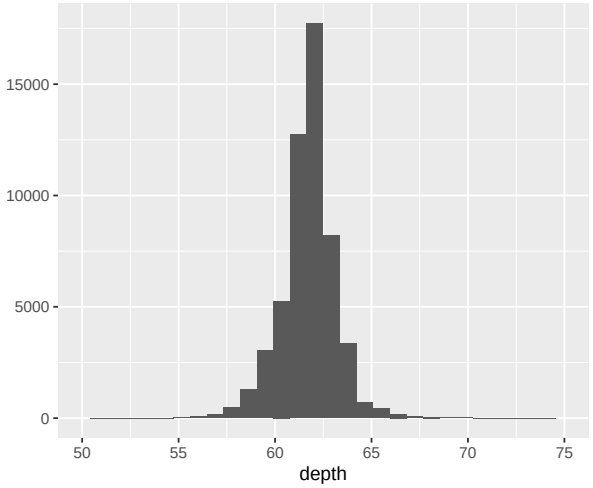
(a)

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



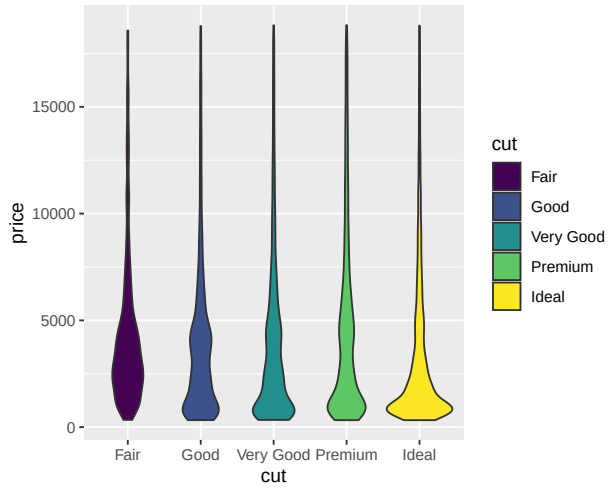
(b)

```
qplot(depth, data=diamonds)+xlim(50, 75)
```



(c)

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



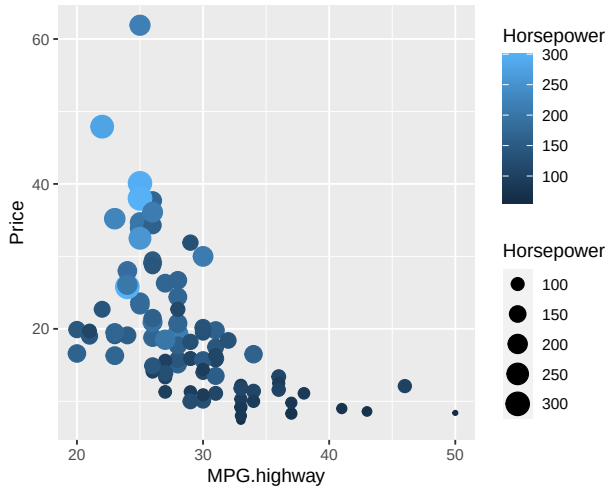
4

(a)

```
library(MASS)
data("Cars93")
as_tibble(Cars93)
```

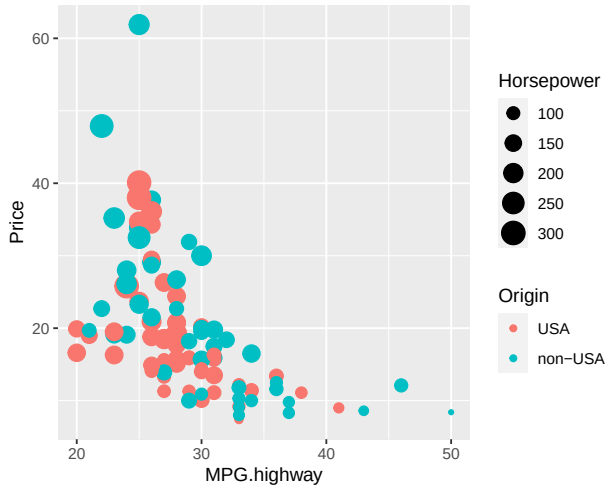
```
## # A tibble: 93 x 27
##    Manufacturer Model Type  Min.Price Price Max.Price MPG.city MPG.highway
##    <fct>        <fct> <fct>     <dbl> <dbl>     <dbl>    <int>       <int>
##  1 Acura        Inte~ Small      12.9  15.9      18.8       25          31
##  2 Acura        Lege~ Mids~      29.2  33.9      38.7       18          25
##  3 Audi         90    Comp~      25.9  29.1      32.3       20          26
##  4 Audi         100   Mids~      30.8  37.7      44.6       19          26
##  5 BMW          535i  Mids~      23.7  30        36.2       22          30
##  6 Buick        Cent~ Mids~      14.2  15.7      17.3       22          31
##  7 Buick        LeSa~ Large      19.9  20.8      21.7       19          28
##  8 Buick        Road~ Large      22.6  23.7      24.9       16          25
##  9 Buick        Rivi~ Mids~      26.3  26.3      26.3       19          27
## 10 Cadillac     DeVi~ Large      33    34.7      36.3       16          25
## # ... with 83 more rows, and 19 more variables: AirBags <fct>,
## #   DriveTrain <fct>, Cylinders <fct>, EngineSize <dbl>, Horsepower <int>,
## #   RPM <int>, Rev.per.mile <int>, Man.trans.avail <fct>,
## #   Fuel.tank.capacity <dbl>, Passengers <int>, Length <int>, Wheelbase <int>,
## #   Width <int>, Turn.circle <int>, Rear.seat.room <dbl>, Luggage.room <int>,
## #   Weight <int>, Origin <fct>, Make <fct>
```

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



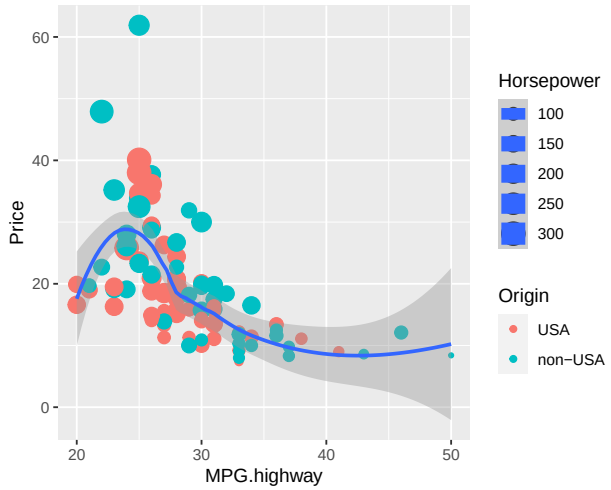
(b)

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



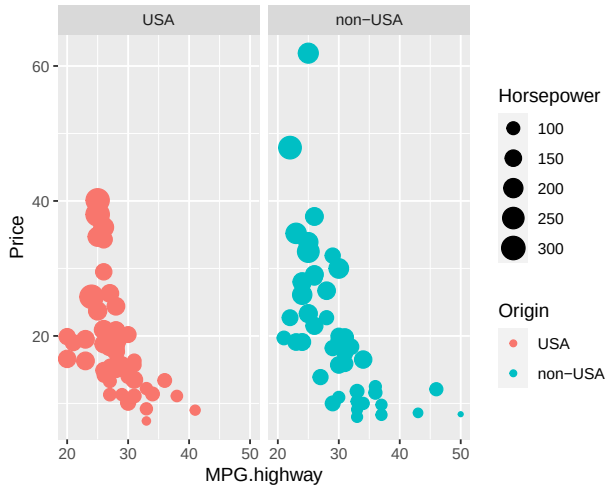
(c)

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



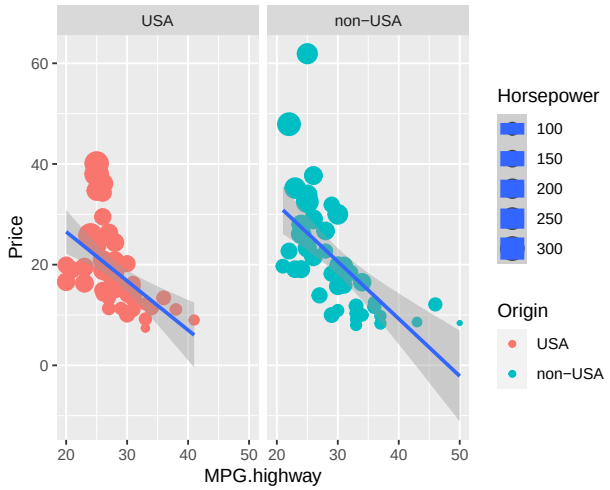
(d)

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



(e)

```
Cars93%>%ggplot(aes(MPG.highway, Price, size=Horsepower))+
  geom_point(aes(color=Origin))+
  facet_grid(~Origin)+
  geom_smooth(method="lm")
```


USA
non−USA
60
40
20
0
Price
20
30
40
50
MPG.highway
Horsepower
100
150
200
250
300
Origin
USA
non−USA