Homework assignment #2

2021320322 / Minseo Yoon

October 11, 2022

1

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

(a)

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

```
## # 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)
```

```
## # 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.
```

Answer: Albania

```
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.
```

Answer: Albania

(c)

```
gapminder %>%
  filter(year >= 1970 & year < 1980) %>%
  group_by(continent) %>%
  summarize(avg = mean(lifeExp))
```

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

(d)

```
gapminder %>%
  mutate(gdp = gdpPercap * pop) %>%
  group_by(country) %>%
  summarize(totalgdp = sum(gdp)) %>%
  arrange(desc(totalgdp)) %>%
  head(n=5)
```

```
## # 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 %>%
  select(country, lifeExp, year) %>%
  filter(lifeExp >= 82)
```

```
## # A tibble: 3 x 3
##   country          lifeExp  year
##   <fct>              <dbl> <int>
## 1 Hong Kong, China    82.2  2007
## 2 Japan              82    2002
## 3 Japan              82.6  2007
```

(f)

```
gapminder %>%
  filter(continent != 'Europe') %>%
  group_by(continent, year) %>%
  summarize(meanPop = mean(pop)) %>%
  arrange(desc(meanPop))
```

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

```
## # A tibble: 48 x 3
## # Groups:   continent [4]
##    continent  year    meanPop
##    <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: Asia, 2007

2

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

(a)

```
flights %>%
  group_by(month) %>%
  summarize(cancelled = sum(is.na(dep_time)), total = n(),
            prop = cancelled / total) %>%
  arrange(prop)
```

```
## # A tibble: 12 x 4
##    month cancelled total    prop
##    <int>     <int> <int>   <dbl>
##  1    10       236 28889 0.00817
##  2    11       233 27268 0.00854
##  3     9       452 27574 0.0164
##  4     8       486 29327 0.0166
##  5     1       521 27004 0.0193
##  6     5       563 28796 0.0196
##  7     4       668 28330 0.0236
##  8     3       861 28834 0.0299
##  9     7       940 29425 0.0319
## 10     6      1009 28243 0.0357
## 11    12      1025 28135 0.0364
## 12     2      1261 24951 0.0505
```

Highest: February, Lowest: October

It is estimated that flights are often cancelled in summer and winter due to bad weather (rain, snow, and so on).

(b)

```
library(ggplot2)
```

```
flights %>%
  filter(year == 2013 & !is.na(tailnum)) %>%
  group_by(tailnum) %>%
  summarize(num = n()) %>%
  arrange(desc(num)) %>%
  head(n=1)
```

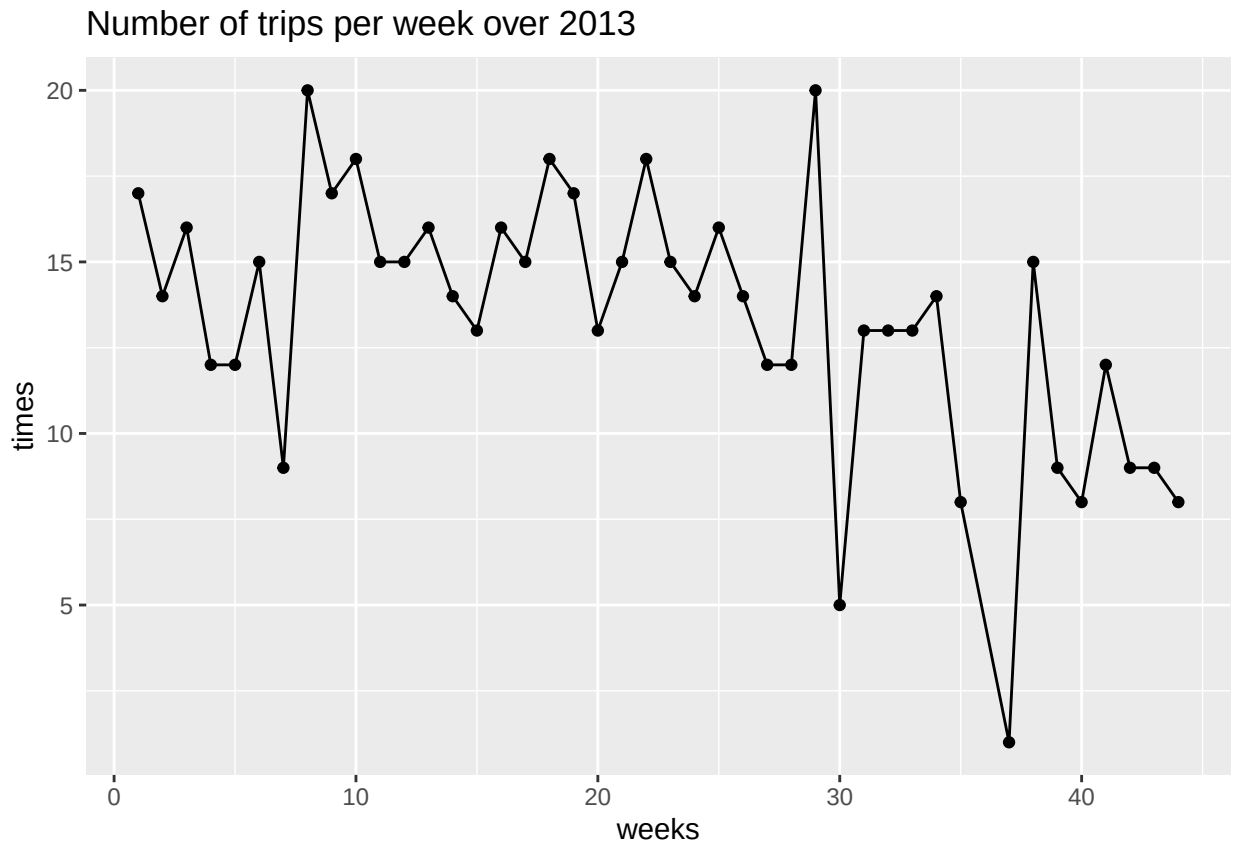
```
## # A tibble: 1 x 2
##   tailnum   num
##   <chr>   <int>
## 1 N725MQ    575
```

```
library(lubridate)
```

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

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

```
flights %>%
  filter(tailnum == 'N725MQ') %>%
  mutate(date = paste(sprintf('%04d-%02d-%02d', year, month, day)),
         weeks = week(date)) %>%
  group_by(weeks) %>%
  summarize(times = n()) %>%
  ggplot(aes(x = weeks, y = times)) +
  geom_point() +
  geom_line() +
  ggtitle('Number of trips per week over 2013')
```



(c)

```
planes %>%
  select(tailnum, year) %>%
  arrange(year) %>%
  head(n=1)
```

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

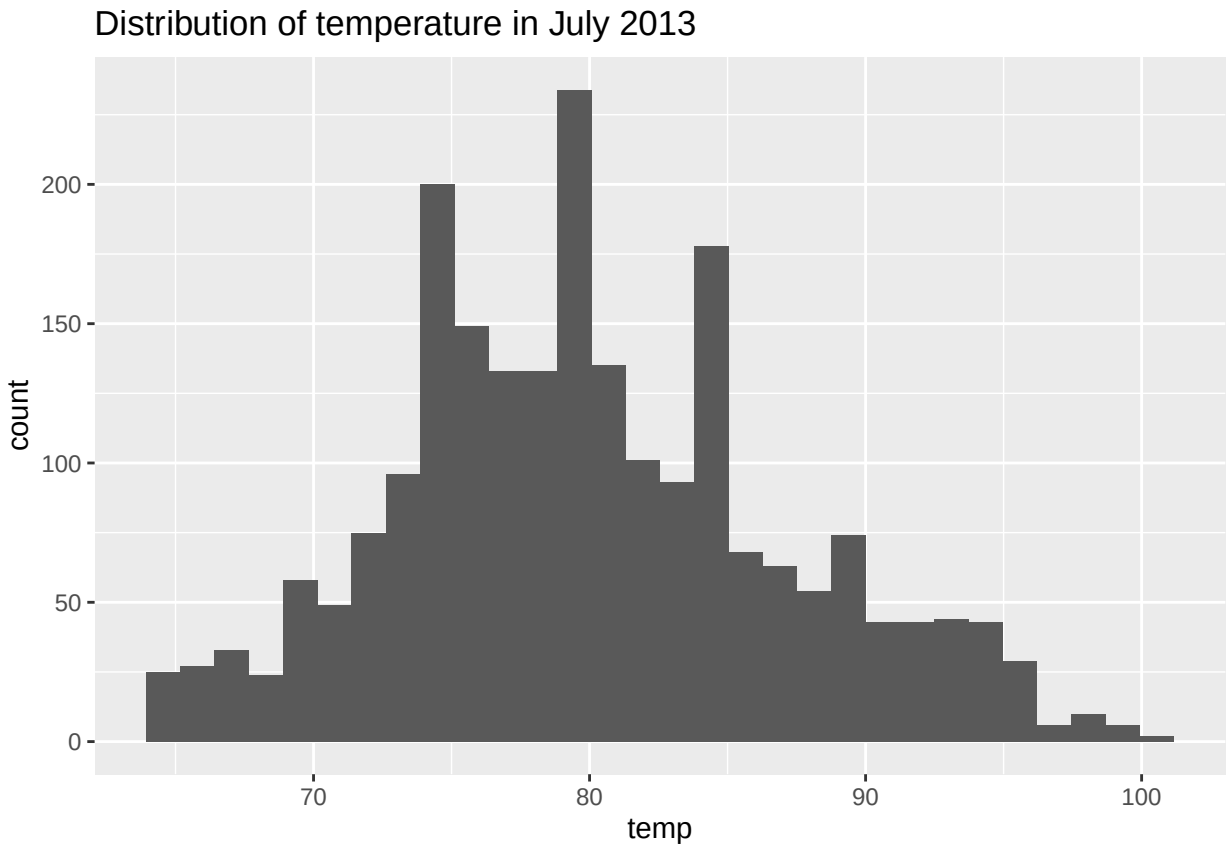
```
planes %>%
  summarize(plane.num = n_distinct(tailnum))
```

```
## # A tibble: 1 x 1
##   plane.num
##       <int>
## 1      3322
```

(d)

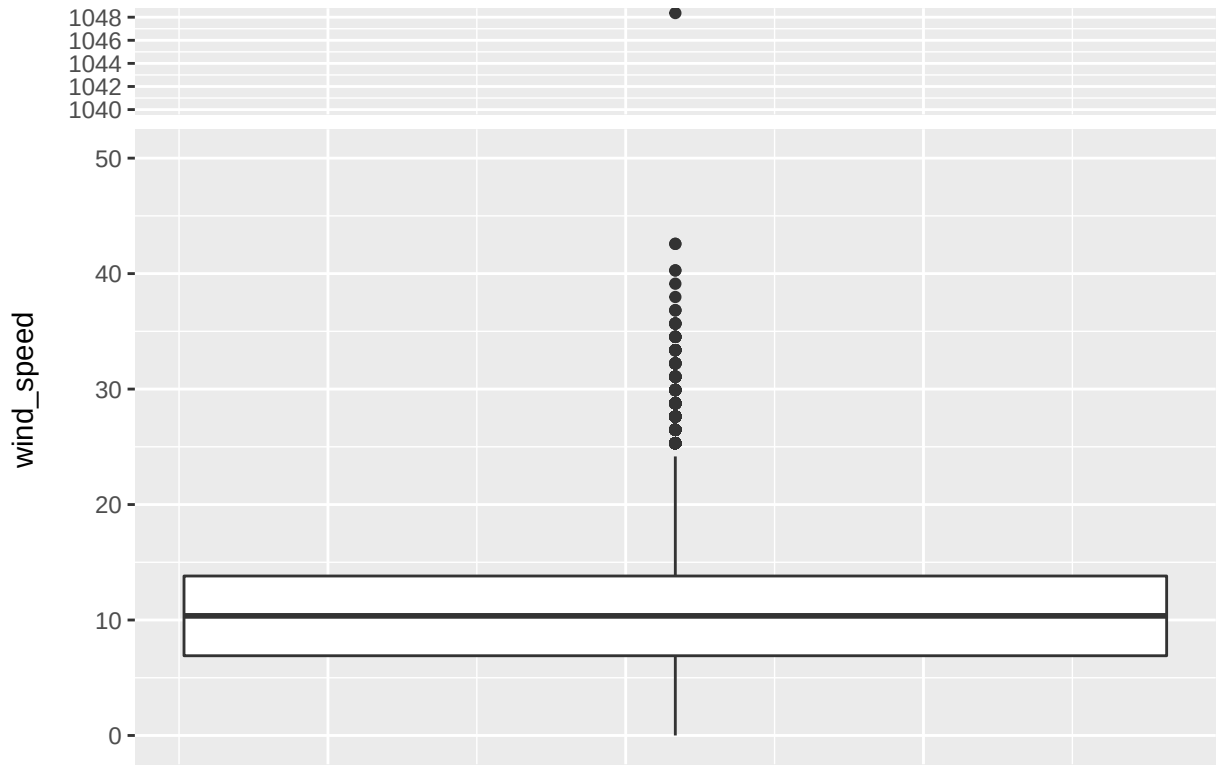
```
weather %>%
  filter(month == 7) %>%
  ggplot(aes(x = temp)) +
  geom_histogram() +
  ggtitle('Distribution of temperature in July 2013')
```

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




```
library(ggbreak)
```

```
weather %>%
  group_by(month) %>%
  ggplot(aes(x = month, y = wind_speed)) +
  geom_boxplot() +
  xlab('') +
  theme(axis.ticks.x = element_line(NA),
        axis.text.x = element_blank()) +
  scale_y_break(c(50, 1040))
```



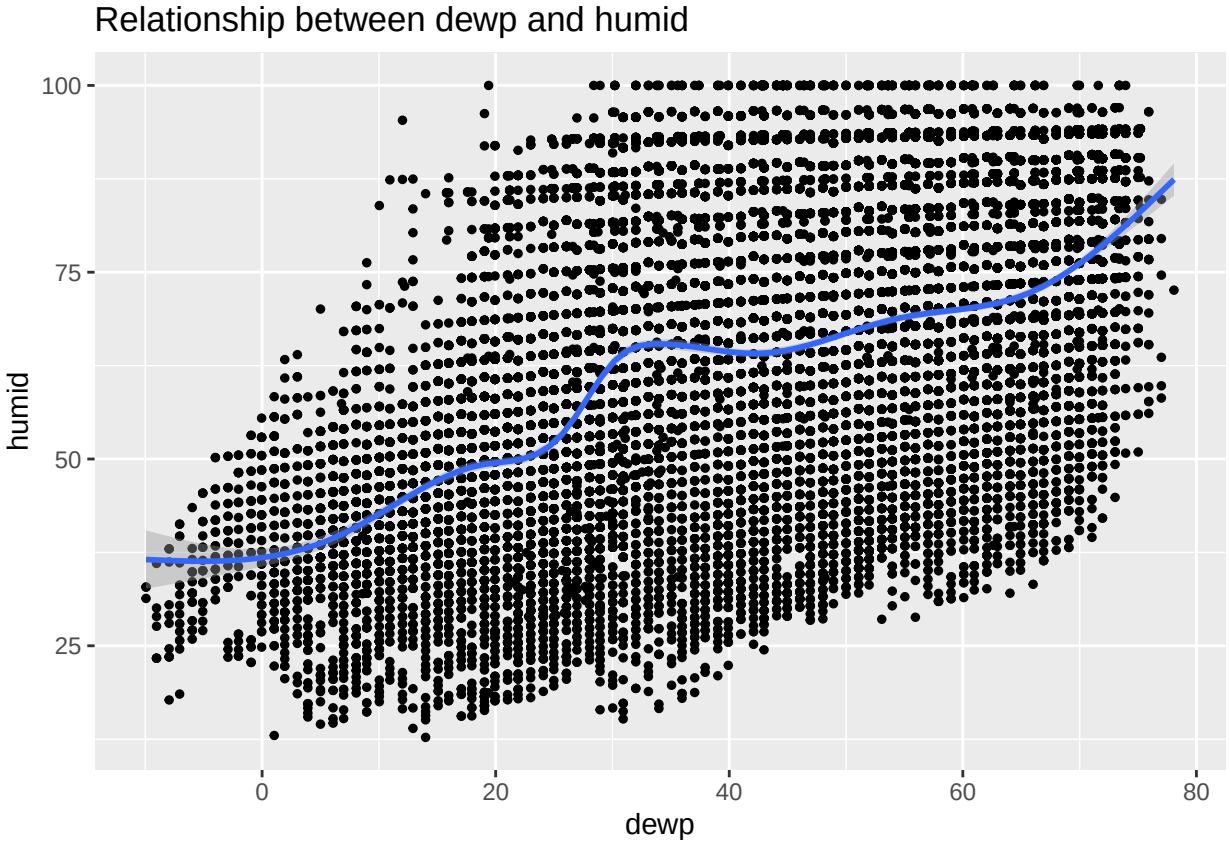

```
quan <- quantile(weather$wind_speed, na.rm=TRUE)
iqr <- quan[4] - quan[2]

weather %>%
  filter(wind_speed >= quan[4] + 1.5 * iqr) %>%
  arrange(desc(wind_speed))
```

```
## # A tibble: 580 x 15
##    origin  year month   day  hour  temp  dewp humid wind_dir wind_speed
##    <chr>  <int> <int> <int> <int> <dbl> <dbl> <dbl>    <dbl>      <dbl>
##  1 EWR     2013     2    12     3  39.0 27.0   61.6      260     1048.
##  2 EWR     2013     1    31     6  57.2 53.6   87.7      270       42.6
##  3 JFK     2013     1    31     4  53.6 53.1  100        200       42.6
##  4 EWR     2013     1    31     4  60.8 59     93.8      230       40.3
##  5 LGA     2013     1    31     4  59   55.4   93.7      230       40.3
##  6 EWR     2013     1    31     8  46.0 30.0   53.3      270       39.1
##  7 JFK     2013     3     6    14  41   28.9   61.9       50       38.0
##  8 JFK     2013     1    31     3  53.1 52.0  100        180       36.8
##  9 JFK     2013     1    31     7  51.8 46.4   81.7      270       36.8
## 10 JFK     2013    11    24    10  28.0 -0.04  29.2      310       36.8
## # ... with 570 more rows, and 5 more variables: wind_gust <dbl>, precip <dbl>,
## #   pressure <dbl>, visib <dbl>, time_hour <dttm>
```

Important outlier: 1048.361

```
weather %>%
  filter(dewp != is.na(dewp) & humid != is.na(humid)) %>%
  ggplot(aes(dewp, humid)) +
  geom_point(size = 1) +
  geom_smooth() +
  ggtitle('Relationship between dewp and humid')
```

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



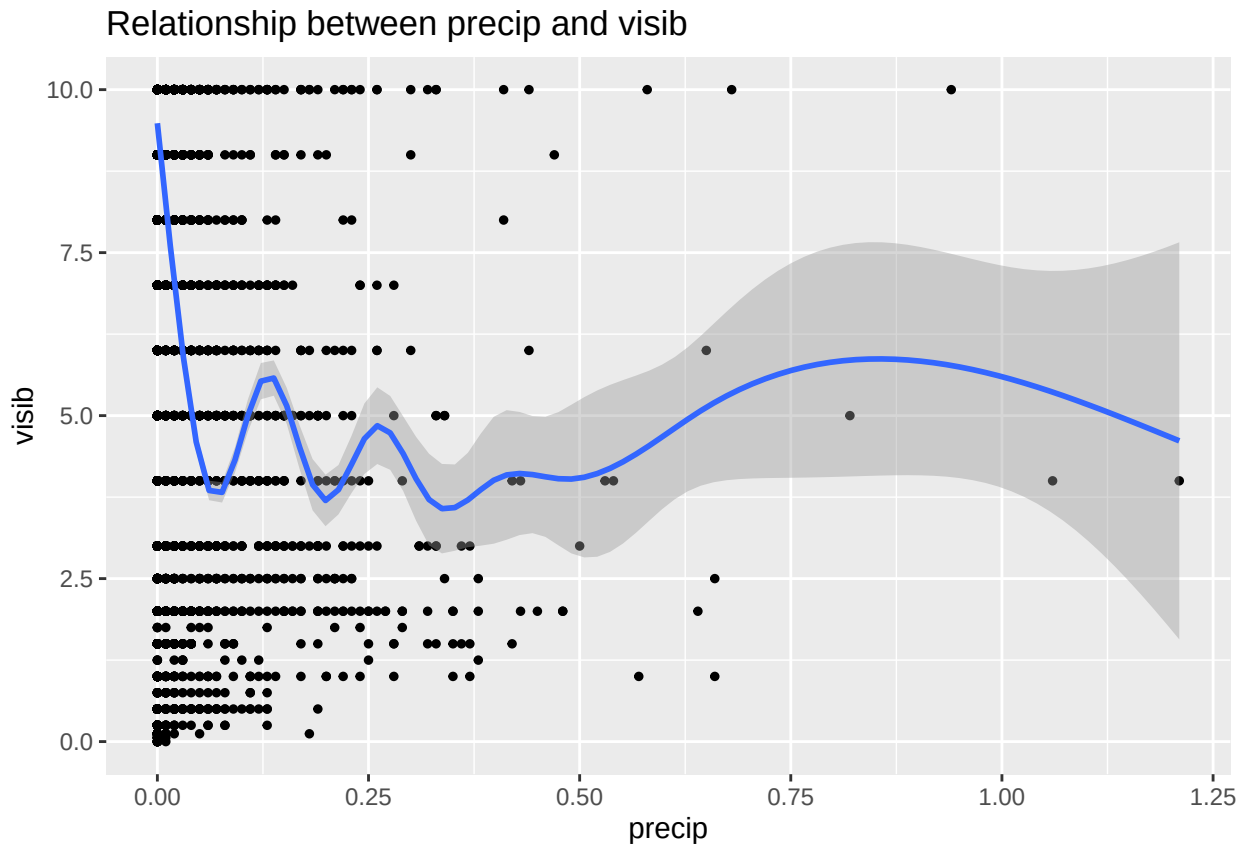

```
weather %>%
  summarize(cor(dewp, humid, use = 'complete.obs'))
```

```
## # A tibble: 1 x 1
##   `cor(dewp, humid, use = "complete.obs")`
##                                      <dbl>
## 1                                    0.512
```

Answer: As the dewp increases, so does the humid. - a positive relationship

```
weather %>%
  ggplot(aes(precip, visib)) +
  geom_point(size = 1) +
  geom_smooth() +
  ggtitle('Relationship between precip and visib')
```

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




```
weather %>%
  summarize(cor(precip, visib, use = 'complete.obs'))
```

```
## # A tibble: 1 x 1
##   `cor(precip, visib, use = "complete.obs")`
##                                        <dbl>
## 1                                     -0.320
```

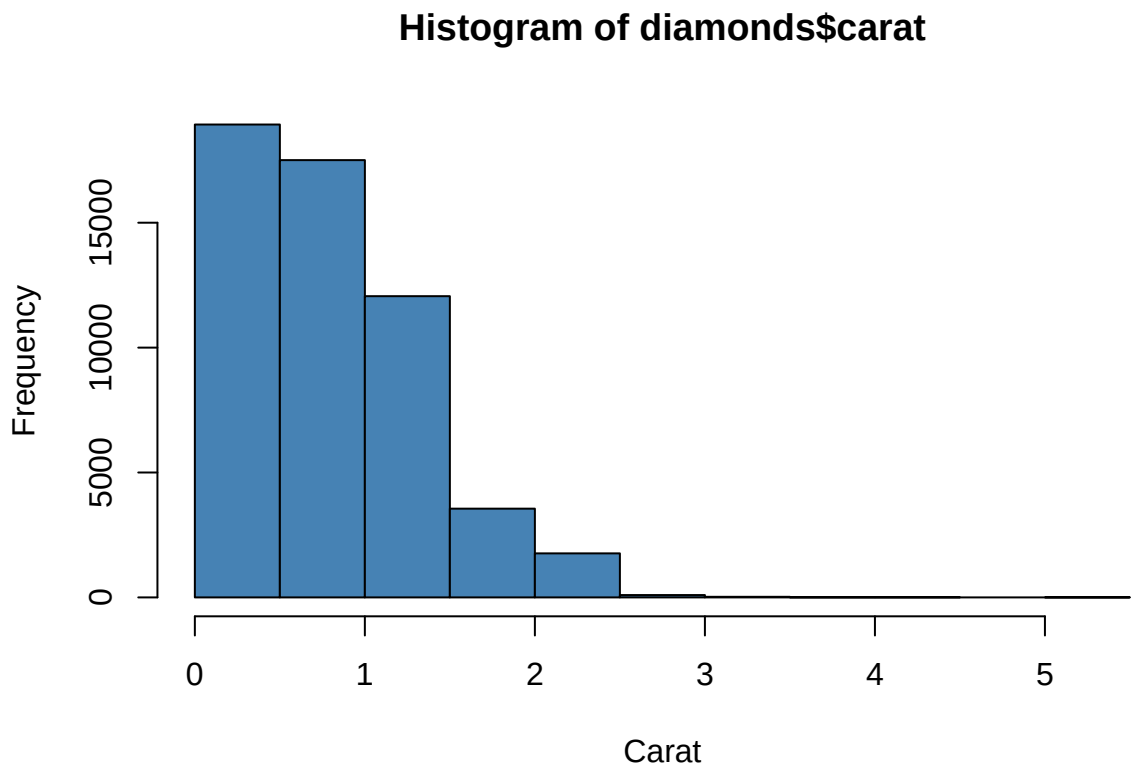
Answer: Although not apparent, visib tends to decrease as the precip increases. - a weak, negative relationship

3

```
?diamonds
```

(a)

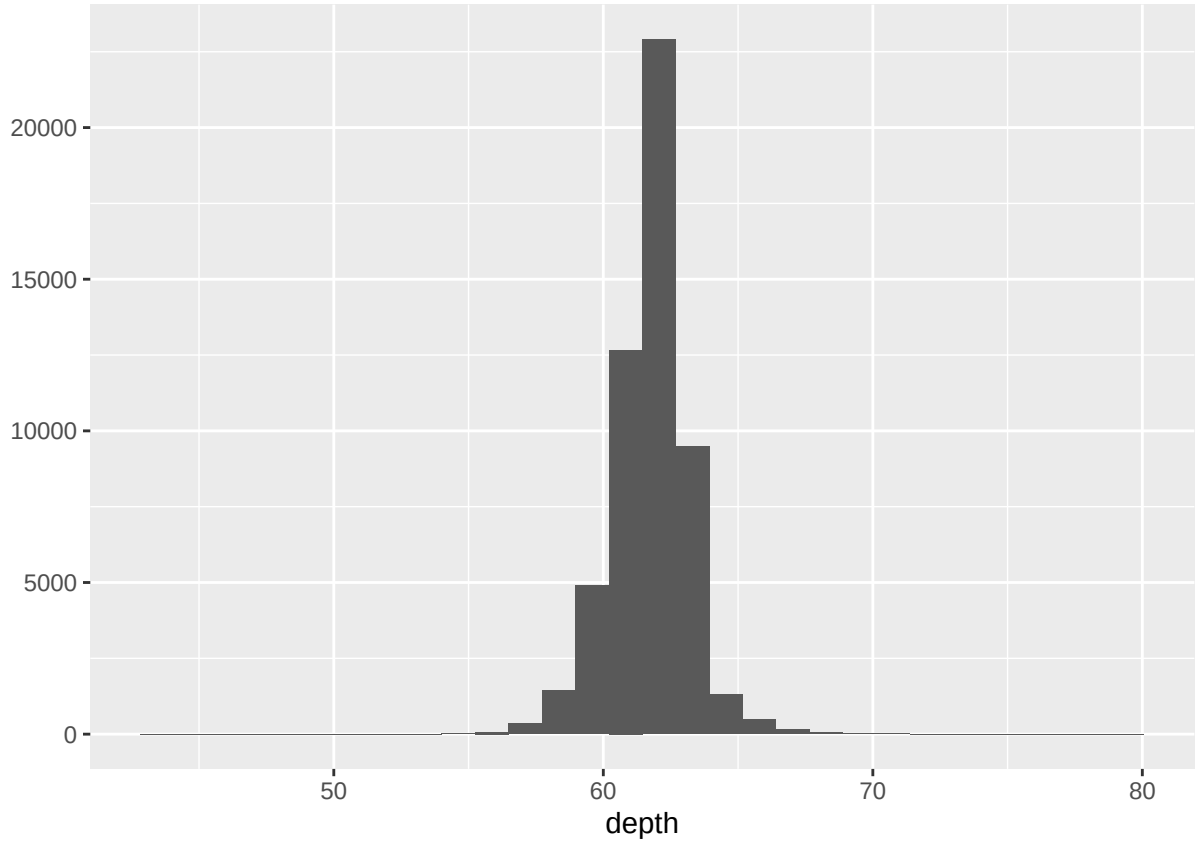
```
hist(x = diamonds$carat, col = 'steelblue', xlab = 'Carat')
```



(b)

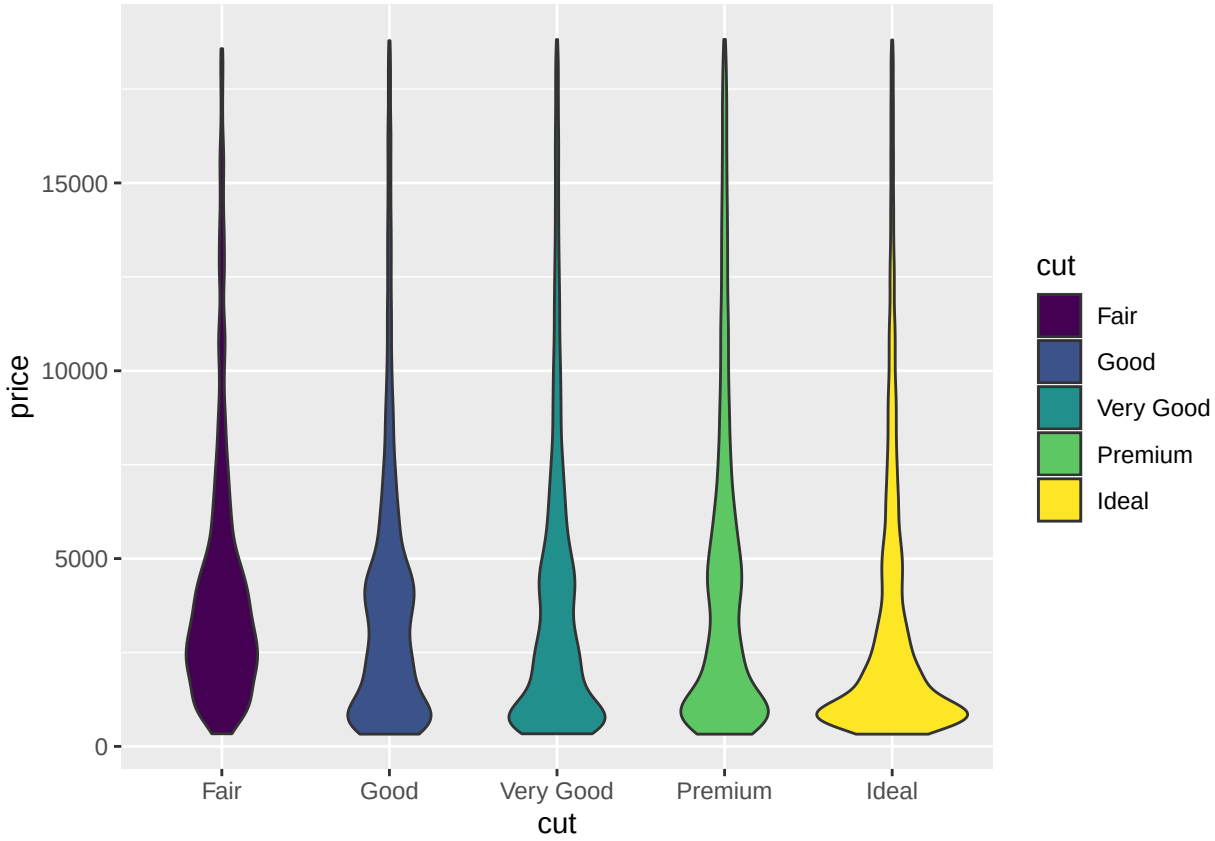
```
qplot(x = depth, data = diamonds)
```

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



(c)

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



4

```
library(MASS)
library(tidyverse)
```

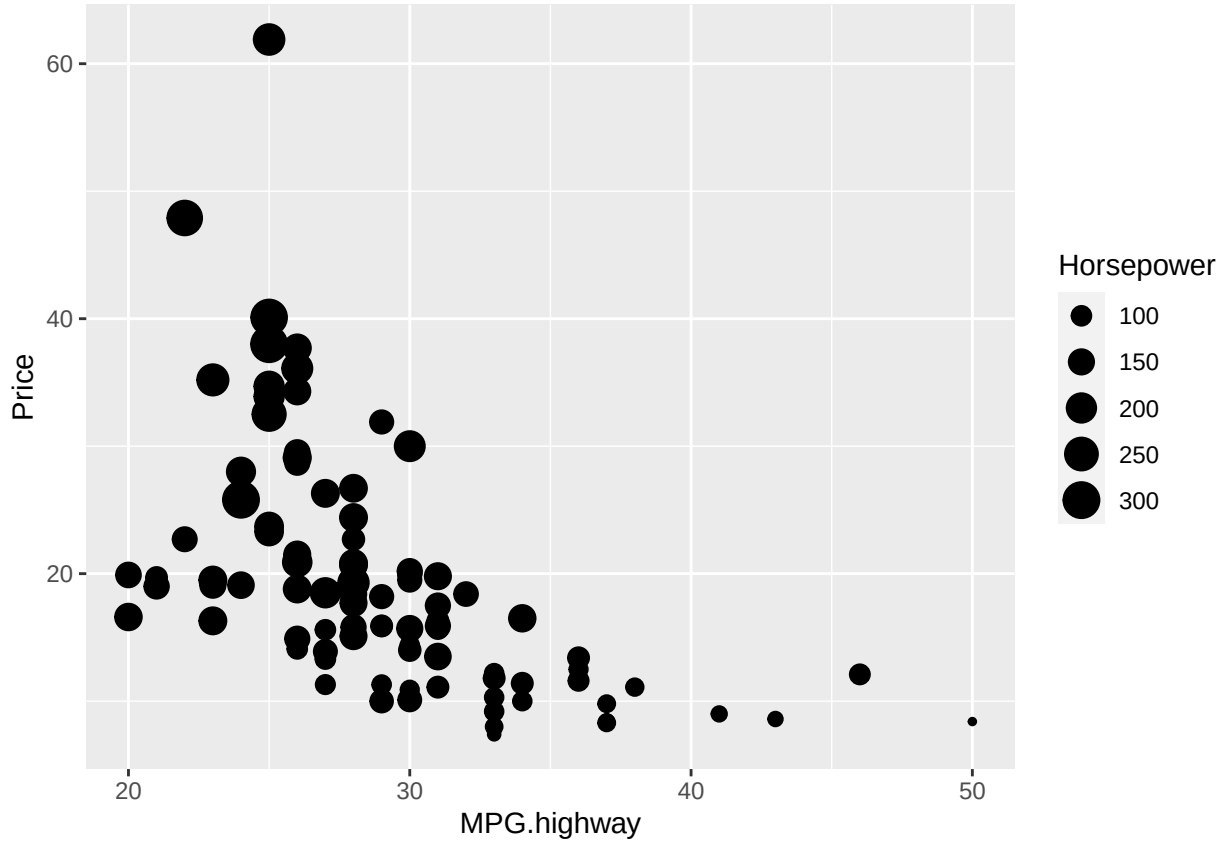
```
as_tibble(Cars93)
```

```
## # 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        Integra    Small        12.9  15.9      18.8       25          31
##  2 Acura        Legend     Midsize      29.2  33.9      38.7       18          25
##  3 Audi         90         Compact      25.9  29.1      32.3       20          26
##  4 Audi         100        Midsize      30.8  37.7      44.6       19          26
##  5 BMW          535i       Midsize      23.7  30        36.2       22          30
##  6 Buick        Century    Midsize      14.2  15.7      17.3       22          31
##  7 Buick        LeSabre    Large        19.9  20.8      21.7       19          28
##  8 Buick        Roadmaster Large        22.6  23.7      24.9       16          25
##  9 Buick        Riviera    Midsize      26.3  26.3      26.3       19          27
## 10 Cadillac     DeVille    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>
```

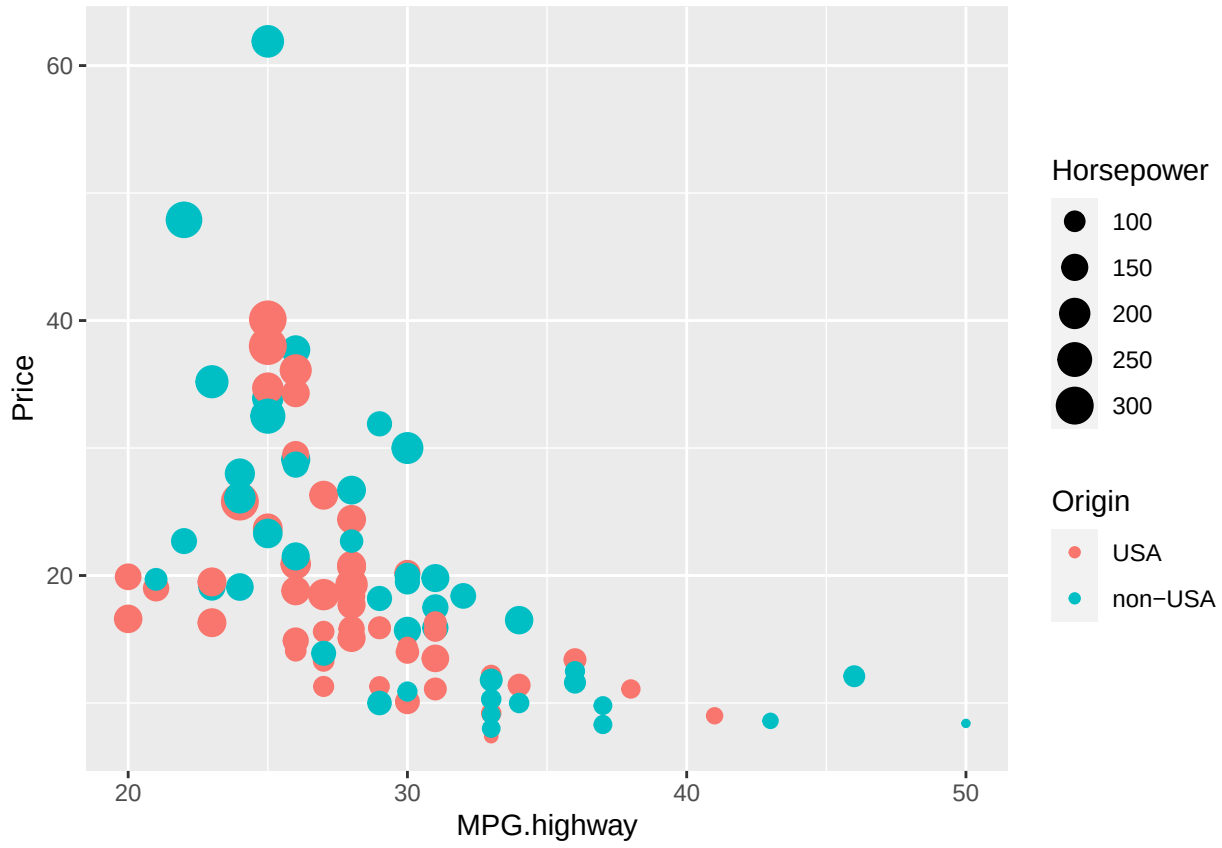
(a)

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



(b)

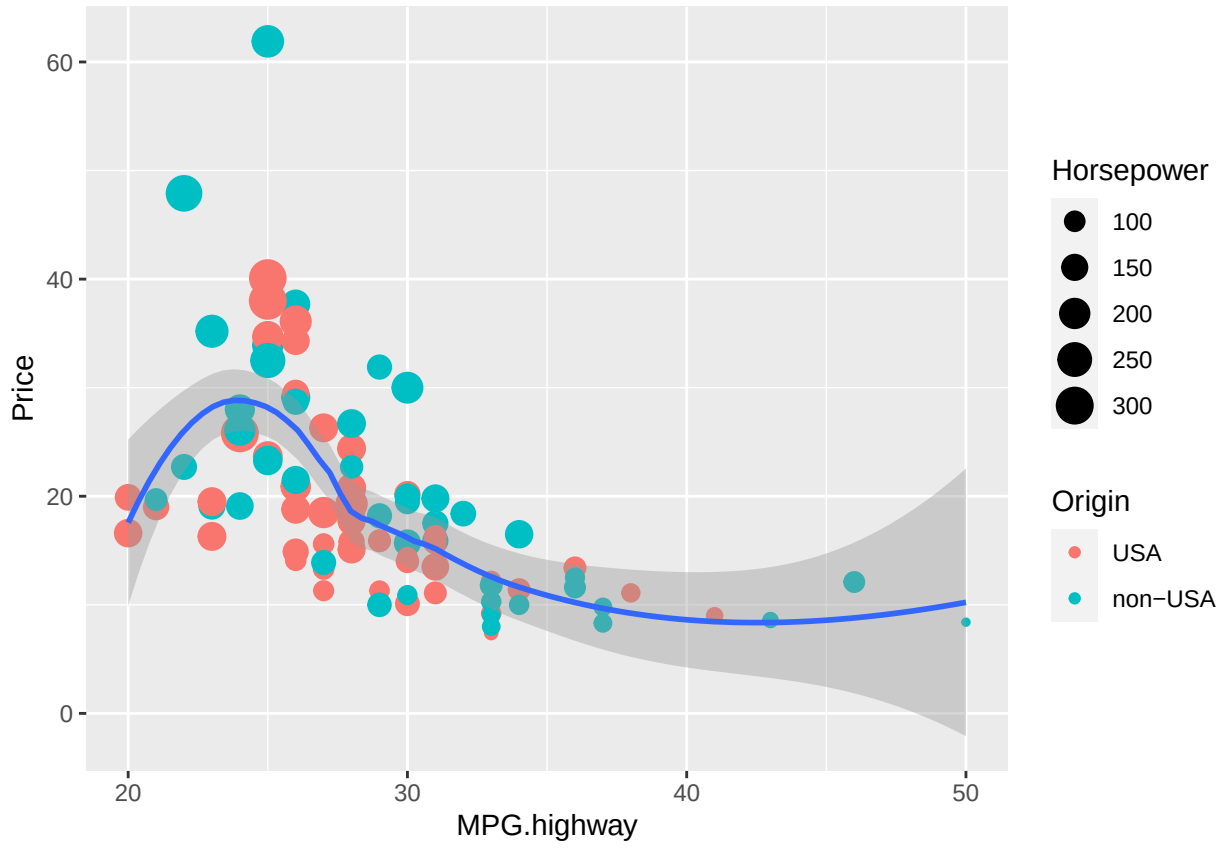
```
ggplot(data = Cars93, aes(x = MPG.highway, y = Price)) +
  geom_point(aes(size = Horsepower, color = Origin))
```



(c)

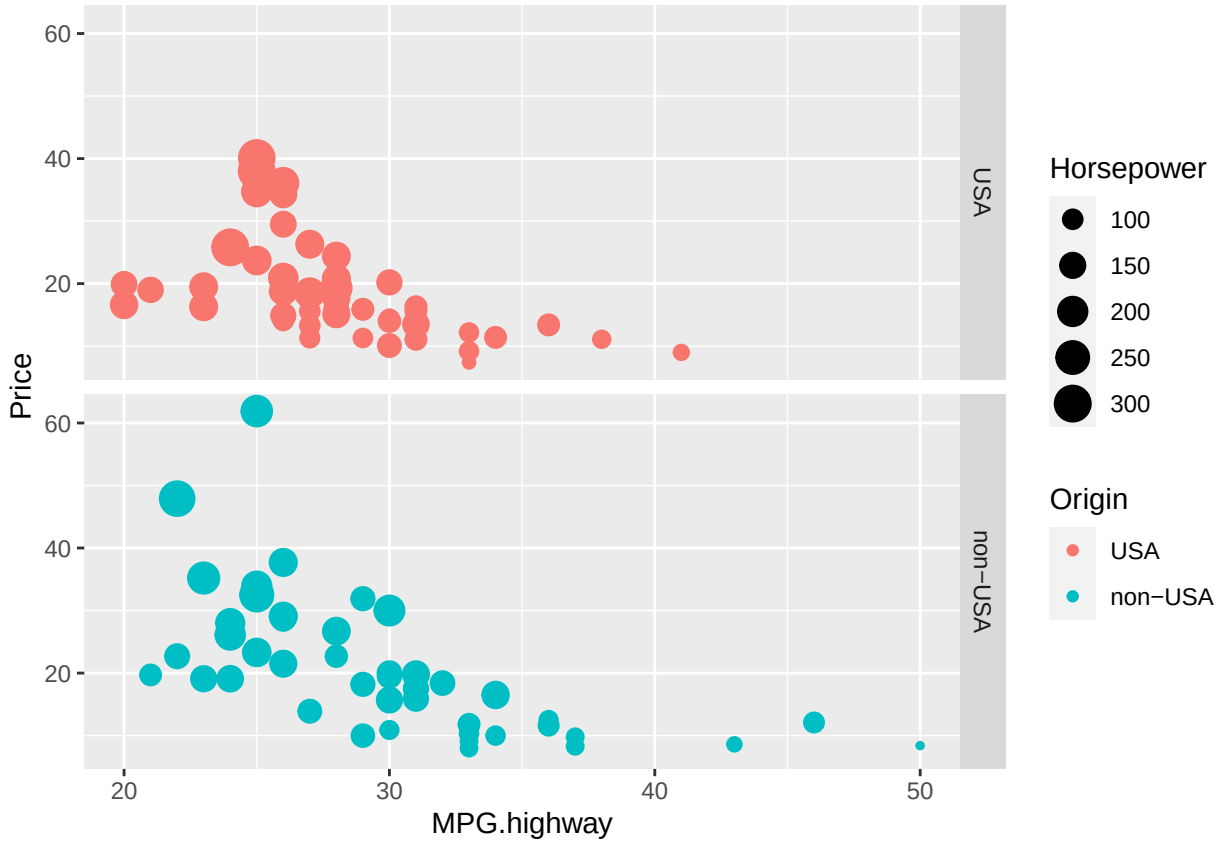
```
ggplot(data = Cars93, aes(x = MPG.highway, y = Price)) +
  geom_point(aes(size = Horsepower, color = Origin)) +
  stat_smooth()
```

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



(d)

```
ggplot(data = Cars93, aes(x = MPG.highway, y = Price)) +
  geom_point(aes(size = Horsepower, color = Origin)) +
  facet_grid(facets = Cars93$Origin)
```


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