SIMPLE CASE OF STUDY

Introducción a la Ciencia de Datos

DATA

- Engine car data from 2015. *Based on an old Sharp Sight tutorial*
- Data available in car_example.xls
- We are going to use tidyverse, dplyr and ggplot2 graphics
- Libraries than we are going to need

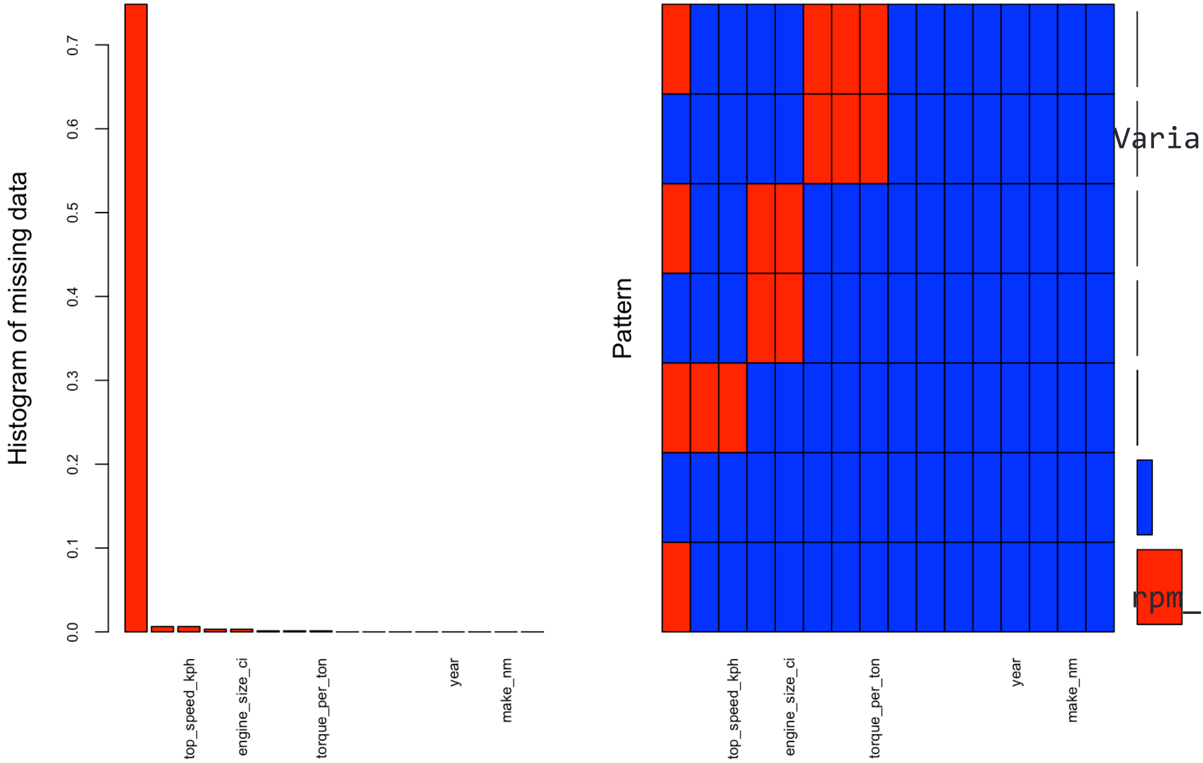
```
library(tidyverse)
library(dplyr)
# for working of %>%
library(magrittr)
library(ggplot2)
library(VIM)
```

Data inspection

- How many variables do you have?
- Which type are they?
- Did R imported all variables with the class that you consider the right one?
- If not change it
- Do you have missing values?

Missing values

```
library(VIM)
aggr_plot <- aggr(df.car_spec_data, col=c('blue','red'), numbers=TRUE,
                  sortVars=TRUE, labels=names(df.car_spec_data), cex.axis=.7, gap=3,
                  ylab=c("Histogram of missing data","Pattern"))
```



```
Variables sorted by number of missings:
                   Variable       Count
       car_0_60_time_seconds 0.748415716
               top_speed_mph 0.006337136
               top_speed_kph 0.006337136
              engine_size_cc 0.003168568
              engine_size_ci 0.003168568
                torque_lb_ft 0.001267427
     rpm_torque_measure_point 0.001267427
              torque_per_ton 0.001267427
                 car_full_nm 0.000000000
              horsepower_bhp 0.000000000
 rpm_horsepower_measure_point 0.000000000
      horsepower_per_ton_bhp 0.000000000
                        year 0.000000000
                      decade 0.000000000
                     make_nm 0.000000000
             car_weight_tons 0.000000000
```

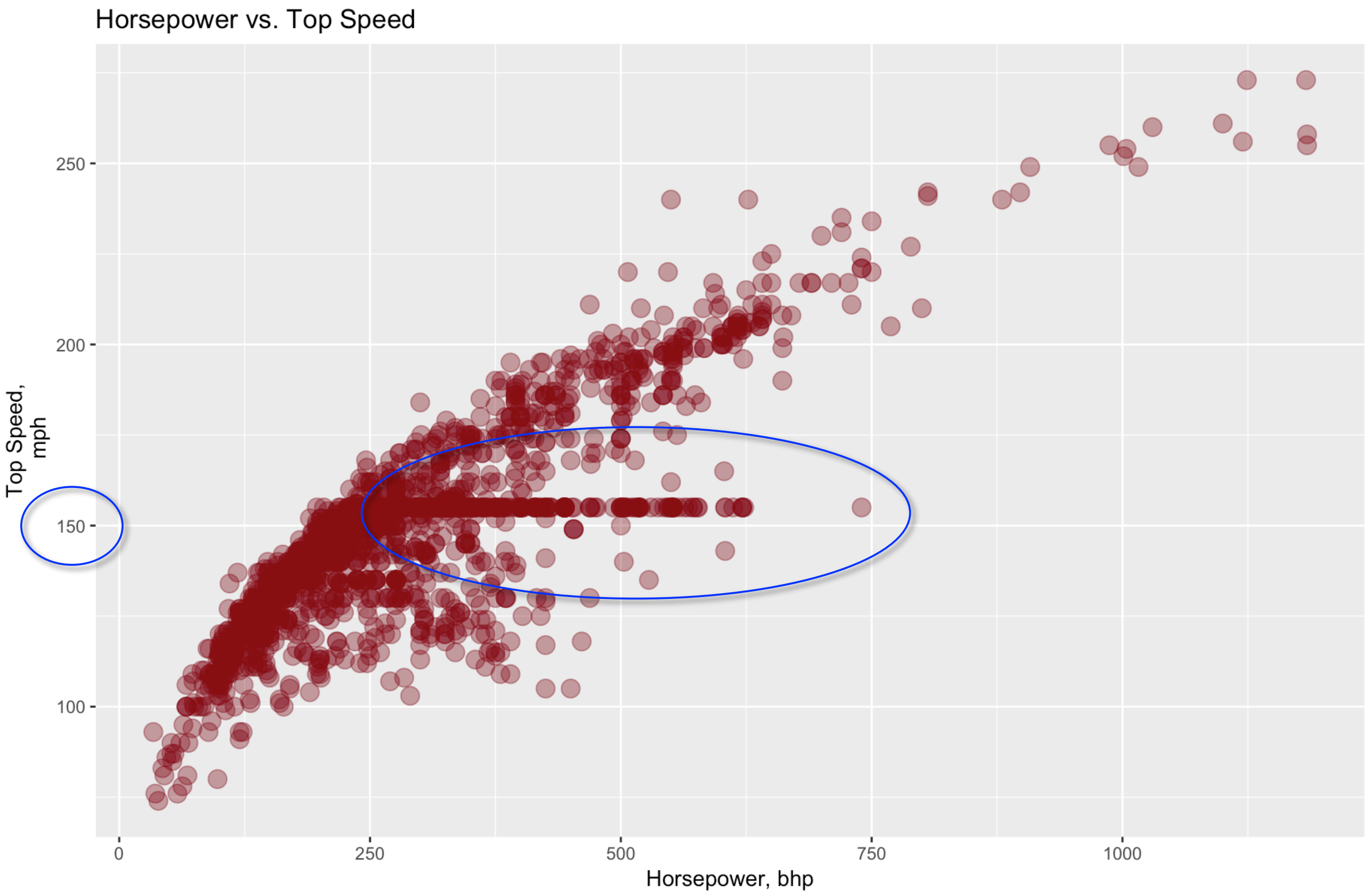
Compare Horsepower vs. Top Speed

- Hypothesis: greater Horsepower higher speed

```
ggplot(data=df.car_spec_data, aes(x=horsepower_bhp, y=top_speed_mph)) +
        geom_point(alpha=.4, size=4, color="#880011") +
        ggtitle("Horsepower vs. Top Speed") +
        labs(x="Horsepower, bhp", y="Top Speed,\n mph")
```

Compare Horsepower vs. Top Speed

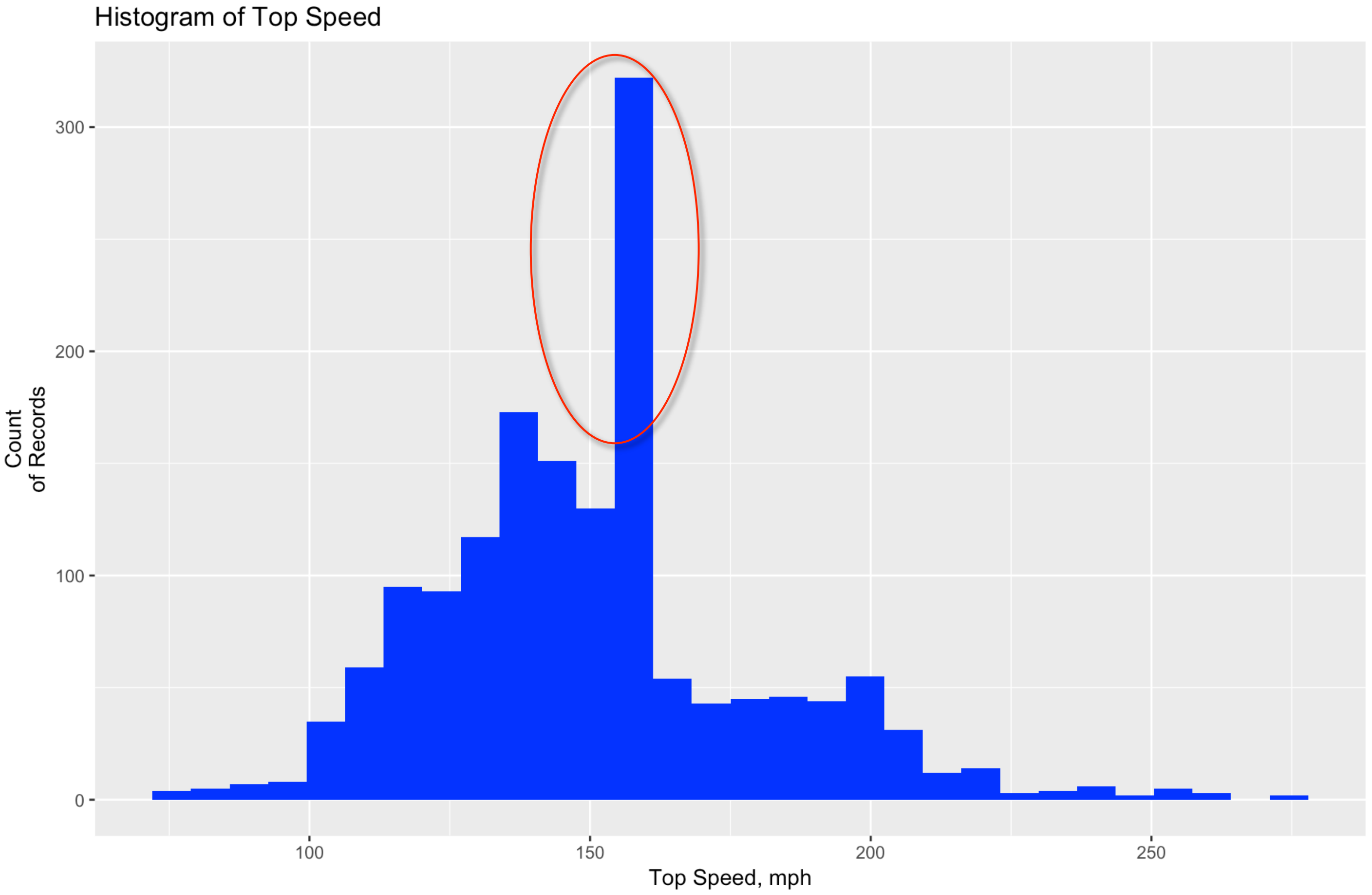
- Hypothesis: greater Horsepower higher speed



Histogram of Top Speed

```
ggplot(data=df.car_spec_data, aes(x=top_speed_mph)) +
        geom_histogram(fill="blue") +
        ggtitle("Histogram of Top Speed") +
        labs(x="Top Speed, mph", y="Count\nof Records")
```

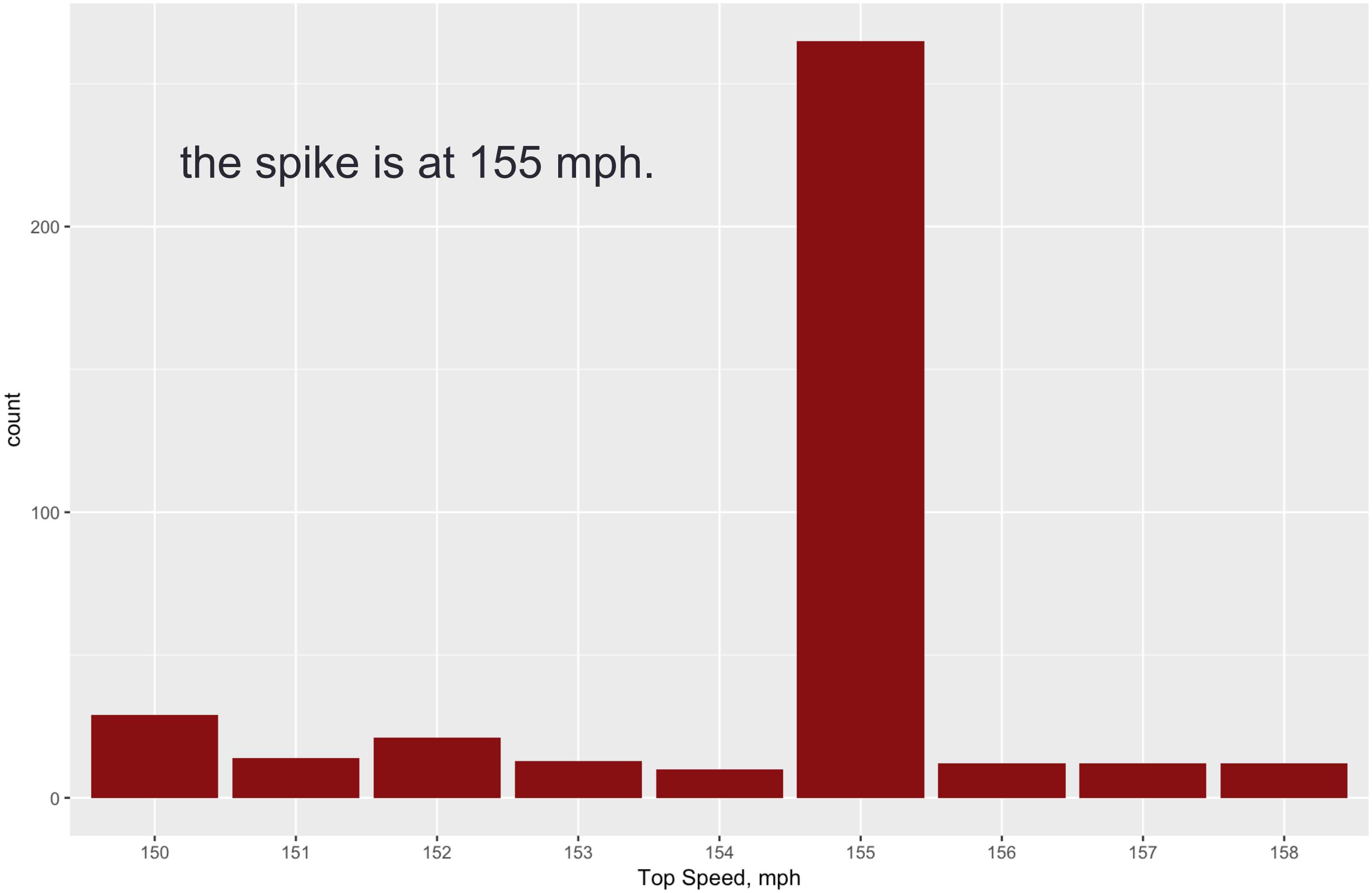
Histogram of Top Speed



Speed between 149 and 159

- Subset the dataset with speed between 149 and 150 using dplyr
- Make a barchart of the results
- What do you see?

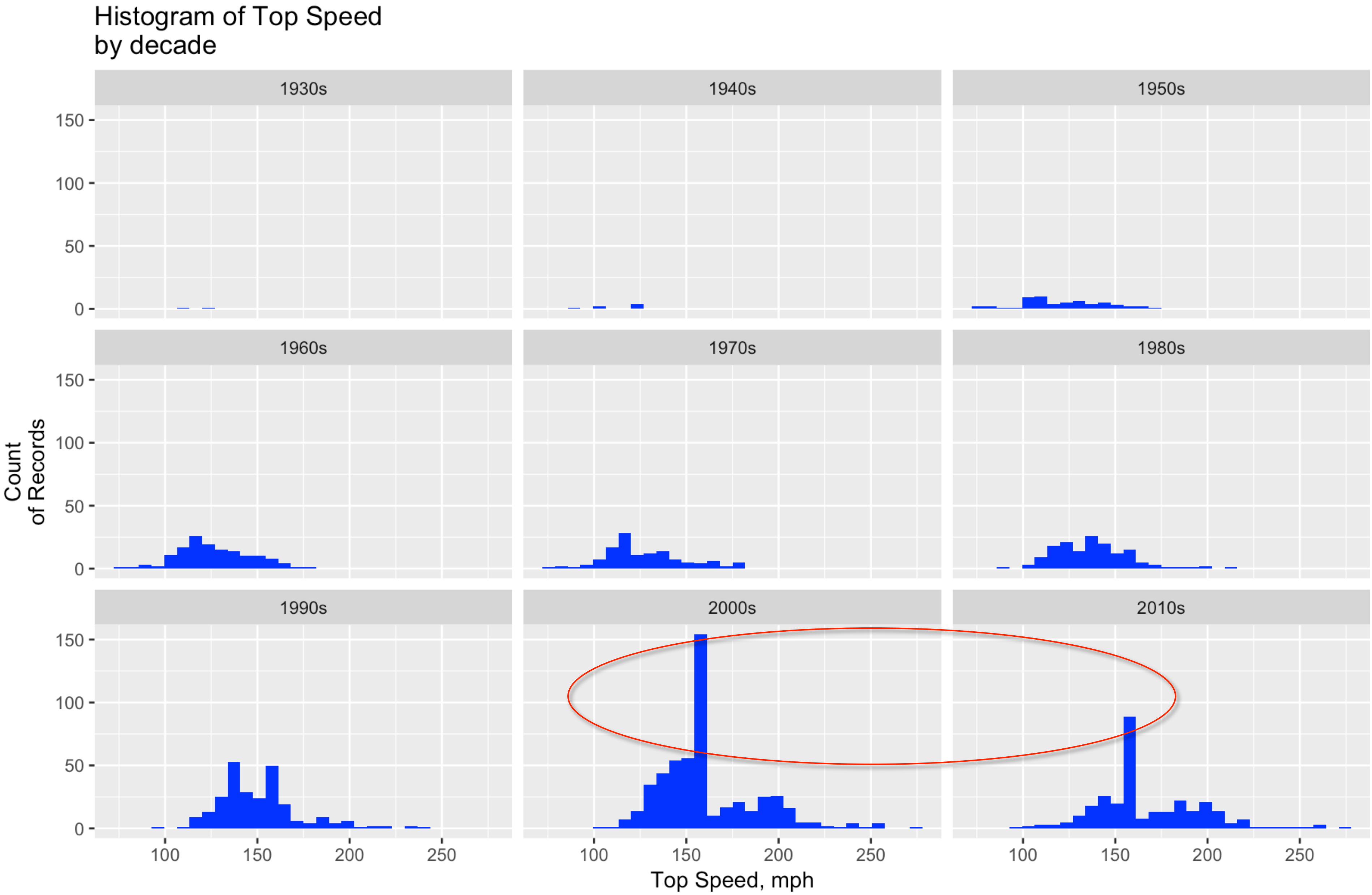
Speed between 149 and 159



When did the speed limit appear?

- Use faceting to look at different decades
- Use the variable top_speed_mph

```
ggplot(data=df.car_spec_data, aes(x=top_speed_mph)) +
        geom_histogram(fill=”blue") +
        ggtitle("Histogram of Top Speed\nby decade") +
        labs(x="Top Speed, mph", y="Count\nof Records") +
        facet_wrap(~decade)
```


Histogram of Top Speed
by decade
1930s
1940s
1950s
1960s
1970s
1980s
1990s
2000s
2010s
Count
of Records
0
50
100
150
100
150
200
250
Top Speed, mph

Do all companies have the same policy about speed limit control?

Search which car companies are limiting car speeds.

Use dplyr verbs chained together and piping %>%

1. Filter the data selecting cars with a top speed of 155 and made after 1990
2. group the data by car manufacturer. This information is in variable make_nm
3. count the number of cars.

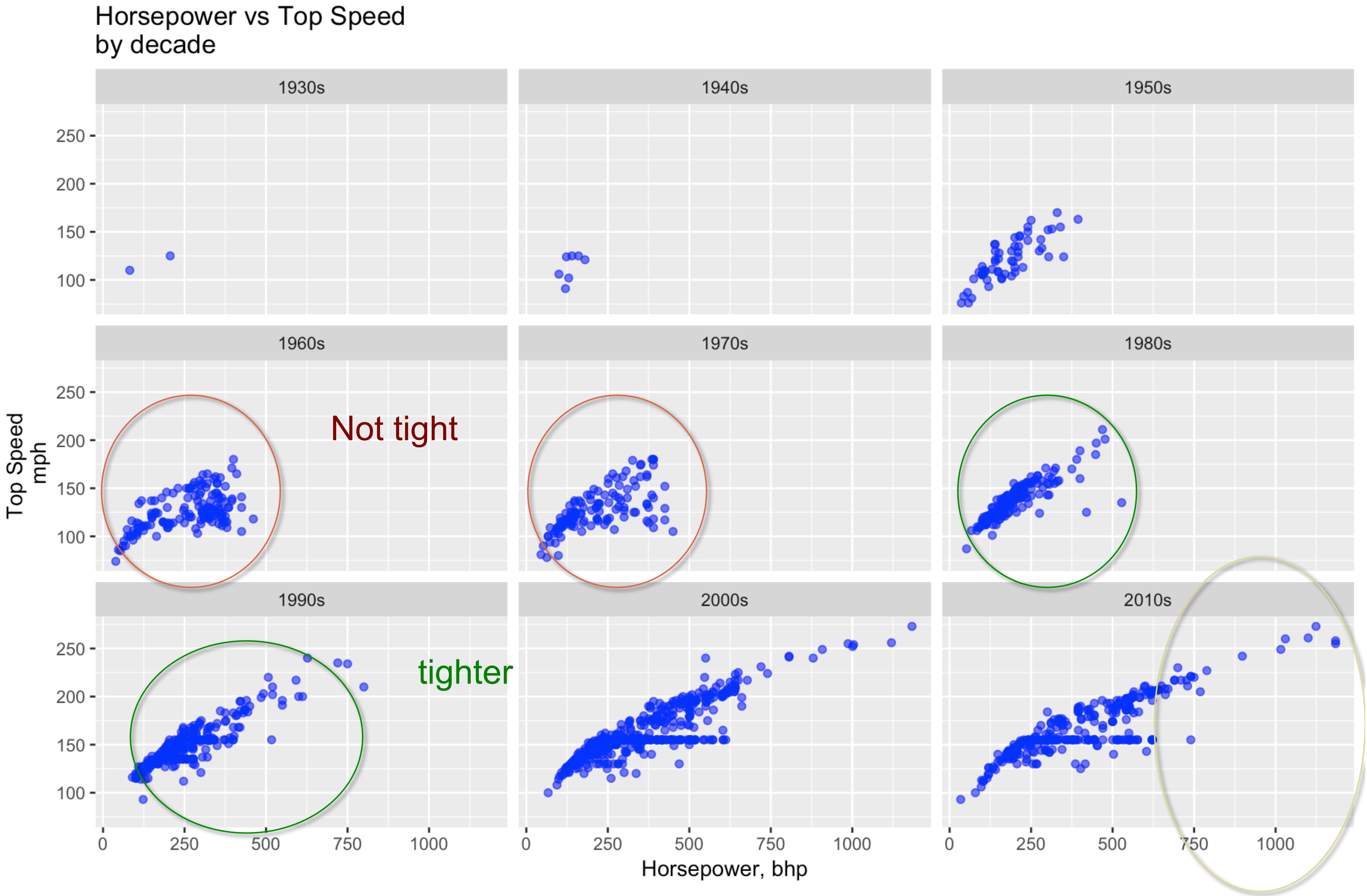
Do all companies have the same policy about speed limit control?

```
df.car_spec_data %>%
        filter(top_speed_mph == 155 & year>=1990) %>%
        group_by(make_nm) %>%
        summarize(count_speed_controlled = n()) %>%
        arrange(desc(count_speed_controlled))
```

```
make_nm        count_speed_controlled
   <fct>                           <int>
 1 BMW                                53
 2 Audi                               51
 3 Mercedes                           41
 4 Jaguar                             14
 5 Nissan                              9
 6 Subaru                              7
 7 Volkswagen(VW)                      7
 8 Volvo                               7
 9 Ford                                5
10 Mitsubishi                          5
# ... with 27 more rows
```

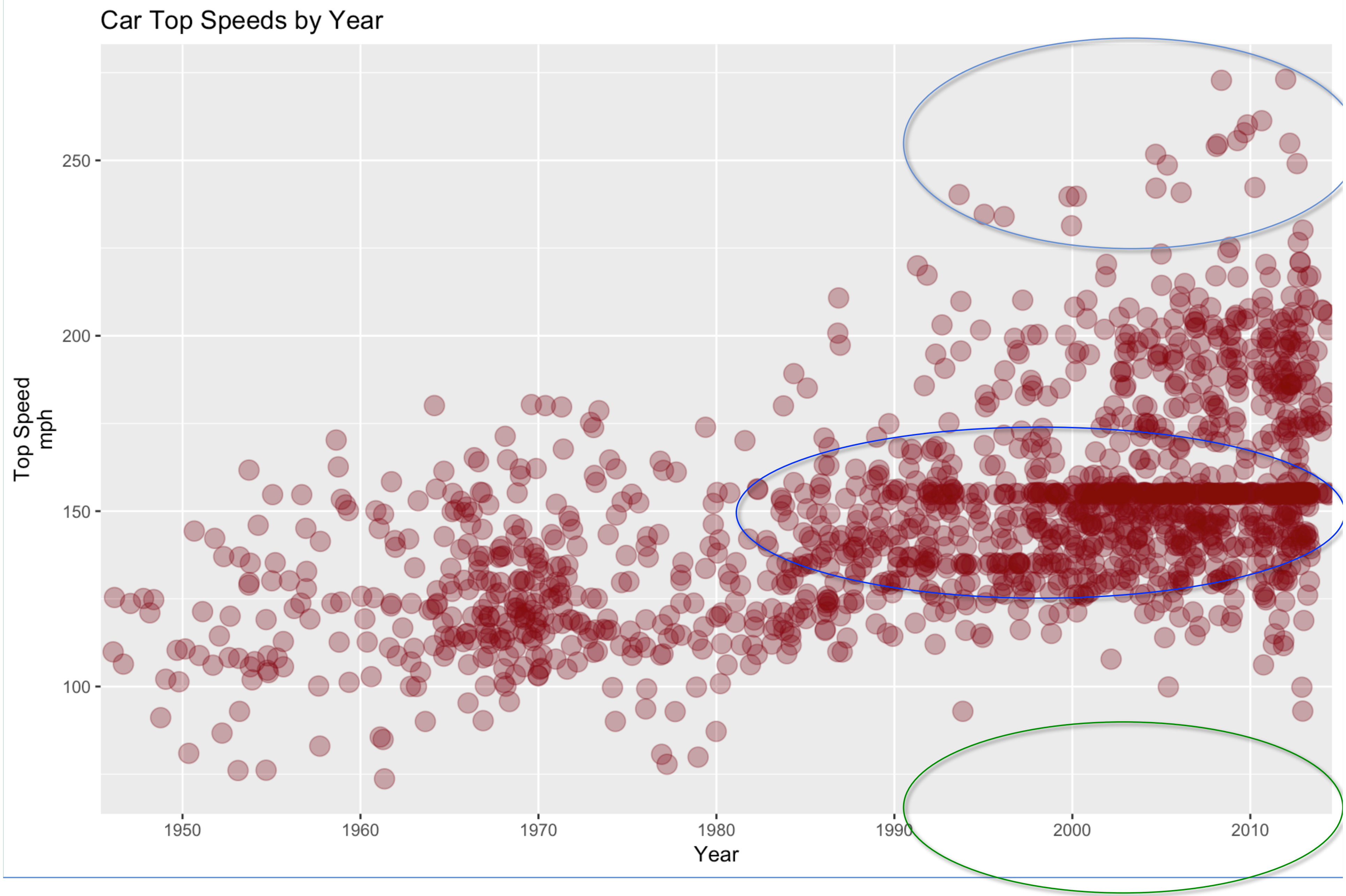
Faceting for searching for relationships

```
ggplot(data=df.car_spec_data, aes(x=horsepower_bhp,
y=top_speed_mph)) +
        geom_point(alpha=.6,color=“blue”) +
        facet_wrap(~decade) +
        ggtitle("Horsepower vs Top Speed\nby decade") +
        labs(x="Horsepower, bhp", y="Top Speed\n mph")
```


Horsepower vs Top Speed
by decade
1930s
1940s
1950s
1960s
1970s
1980s
1990s
2000s
2010s
Top Speed
mph
250
200
150
100
0
250
500
750
1000
Horsepower, bhp
Not tight
tighter

Increase of Speed with the years

```
ggplot(data=df.car_spec_data, aes(x=year,
y=df.car_spec_data$top_speed_mph)) +
        geom_point(alpha=.35, size=4.5, color="#880011", position =
position_jitter()) +
        scale_x_discrete(breaks =
c("1950","1960","1970","1980","1990","2000","2010")) +
        ggtitle("Car Top Speeds by Year") +
        labs(x="Year" ,y="Top Speed\nmph")
```

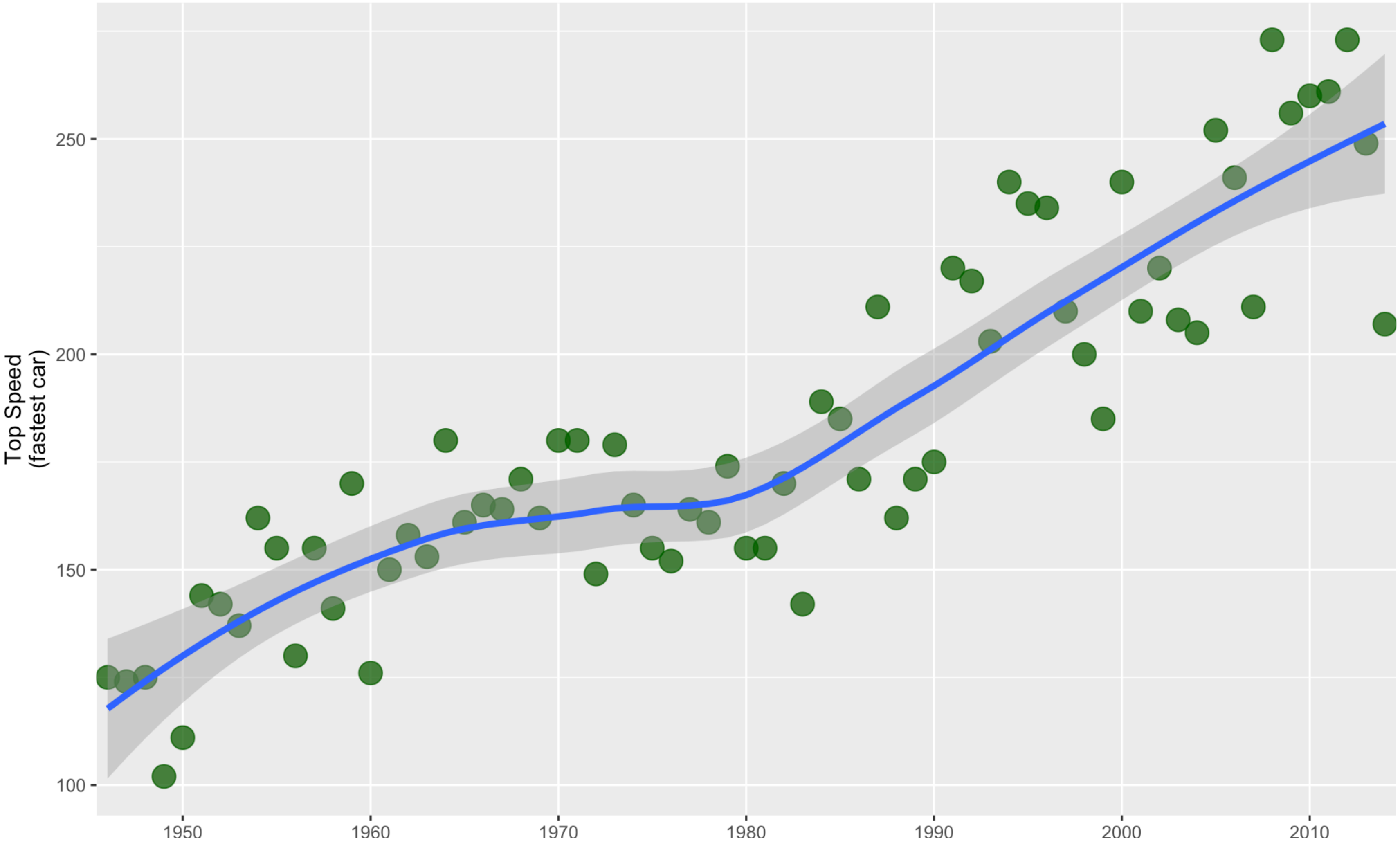

Car Top Speeds by Year
Top Speed
mph
250
200
150
100
1950
1960
1970
1980
1990
2000
2010
Year

Show this trend more clearly

- Show the fastest car of each type by year
- Tips:
  - group by year
  - Take into account the missing values
  - top_speed_mph is the variable containing the speed data
  - Make a geom_point() graph

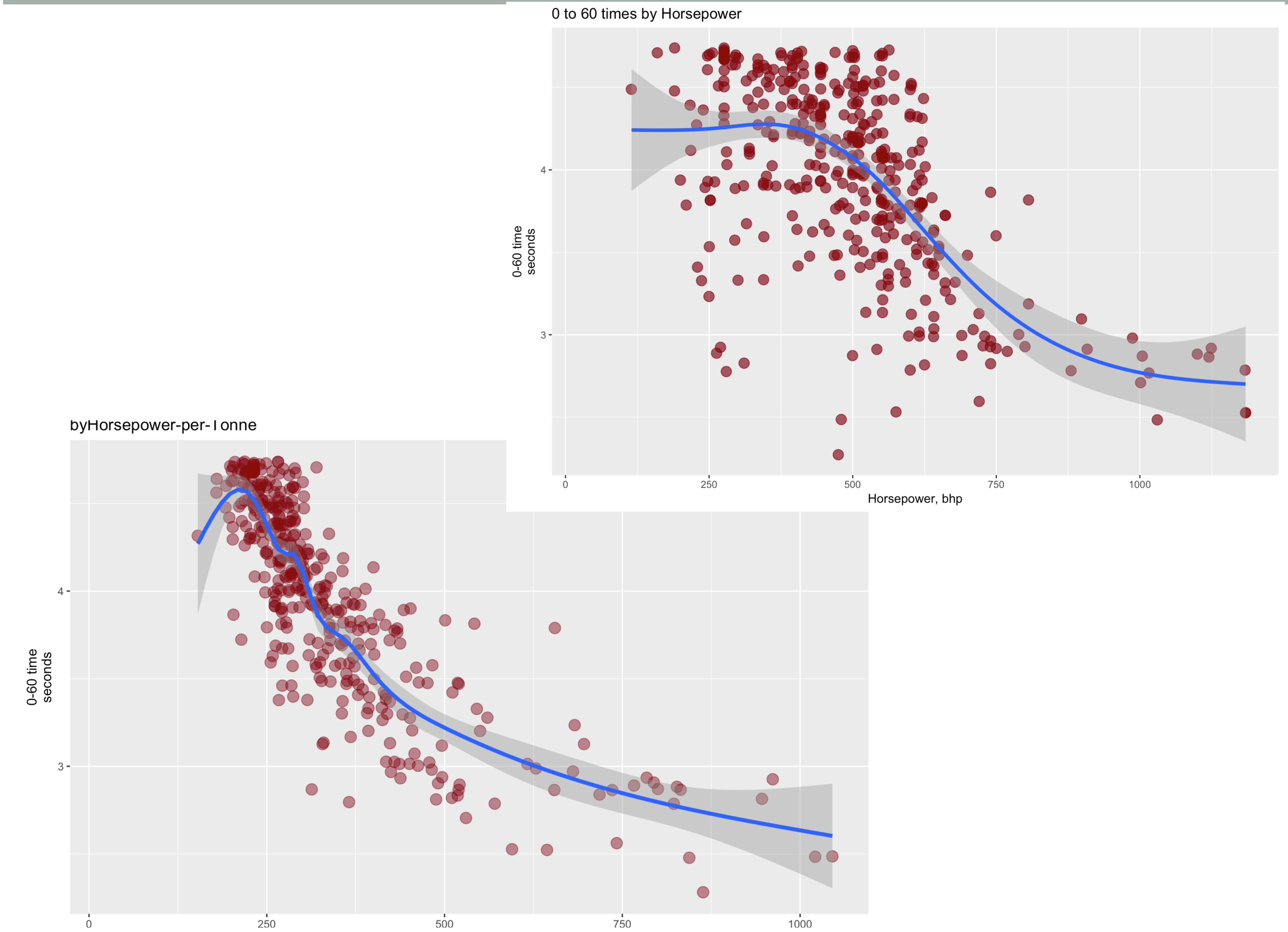
Maximun speed by Year

```
df.car_spec_data %>%
        group_by(year) %>%
        summarize(max_speed = max(top_speed_mph, na.rm=TRUE))%>%
        ggplot(aes(x=year,y=max_speed,group=1)) +
                geom_point(size=5, alpha=.8, color="#880011") +
                stat_smooth(method="auto",size=1.5) +
                scale_x_discrete(breaks =
        c("1950","1960","1970","1980","1990","2000","2010")) +
                ggtitle("Speed of Year's Fastest Car by Year") +
                labs(x="Year",y="Top Speed\n(fastest car)")
```


Top Speed
(fastest car)
250
200
150
100
1950
1960
1970
1980
1990
2000
2010

More hyphotesis

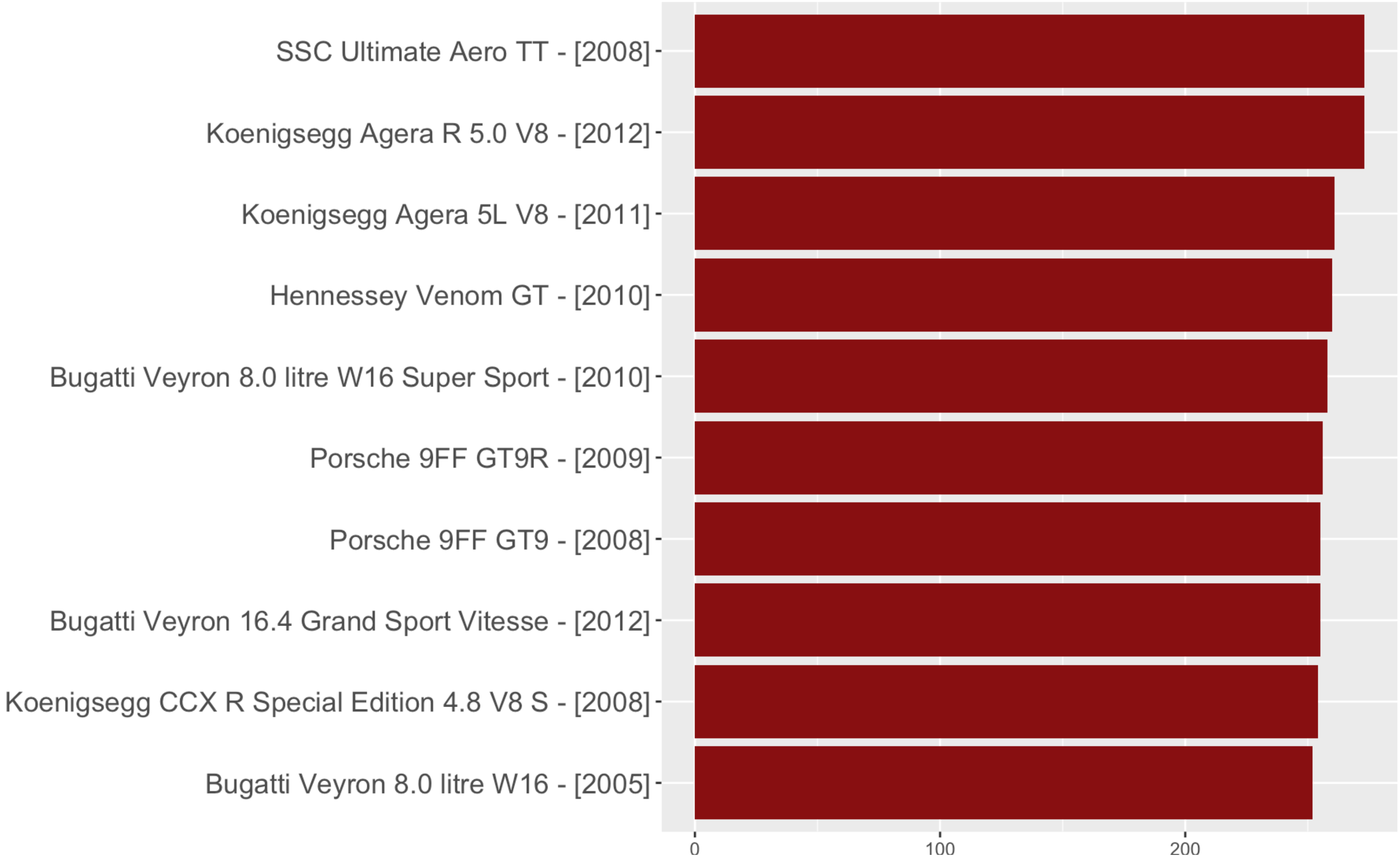
- Is there a relationship between the acceleration (0-to-60) and the power (horsepower_bhp)
- Is only dependent on the power or could the weight of the car be involved (tonne)


0 to 60 times by Horsepower
0-60 time
seconds
4
3
0
250
500
750
1000
Horsepower, bhp
byHorsepower-per-Tonne
0-60 time
seconds
4
3
0
250
500
750
1000

Calculate which are the fastest cars

- Make a subset of the autos and their speed
- Make a ranking of descending order and select the fastest 10
- Make a bar graph

Fastest cars


SSC Ultimate Aero TT - [2008]
Koenigsegg Agera R 5.0 V8 - [2012]
Koenigsegg Agera 5L V8 - [2011]
Hennessey Venom GT - [2010]
Bugatti Veyron 8.0 litre W16 Super Sport - [2010]
Porsche 9FF GT9R - [2009]
Porsche 9FF GT9 - [2008]
Bugatti Veyron 16.4 Grand Sport Vitesse - [2012]
Koenigsegg CCX R Special Edition 4.8 V8 S - [2008]
Bugatti Veyron 8.0 litre W16 - [2005]