Cars93 Dataset Analysis

Sarthak Dahal

Loading the necessary libraries: MASS for the Cars93 dataset and tidyverse for data manipulation and visualization.

```r
library(MASS)
str(Cars93)
```

```
## 'data.frame':    93 obs. of  27 variables:
##  $ Manufacturer      : Factor w/ 32 levels "Acura","Audi",..: 1 1 2 2 3 4 4 4 4 5 ...
##  $ Model             : Factor w/ 93 levels "100","190E","240",..: 49 56 9 1 6 24 54 74 73 35 ...
##  $ Type              : Factor w/ 6 levels "Compact","Large",..: 4 3 1 3 3 3 2 2 3 2 ...
##  $ Min.Price         : num  12.9 29.2 25.9 30.8 23.7 14.2 19.9 22.6 26.3 33 ...
##  $ Price             : num  15.9 33.9 29.1 37.7 30 15.7 20.8 23.7 26.3 34.7 ...
##  $ Max.Price         : num  18.8 38.7 32.3 44.6 36.2 17.3 21.7 24.9 26.3 36.3 ...
##  $ MPG.city          : int  25 18 20 19 22 22 19 16 19 16 ...
##  $ MPG.highway       : int  31 25 26 26 30 31 28 25 27 25 ...
##  $ AirBags           : Factor w/ 3 levels "Driver & Passenger",..: 3 1 2 1 2 2 2 2 2 2 ...
##  $ DriveTrain        : Factor w/ 3 levels "4WD","Front",..: 2 2 2 2 3 2 2 3 2 2 ...
##  $ Cylinders         : Factor w/ 6 levels "3","4","5","6",..: 2 4 4 4 2 2 4 4 4 5 ...
##  $ EngineSize        : num  1.8 3.2 2.8 2.8 3.5 2.2 3.8 5.7 3.8 4.9 ...
##  $ Horsepower        : int  140 200 172 172 208 110 170 180 170 200 ...
##  $ RPM               : int  6300 5500 5500 5500 5700 5200 4800 4000 4800 4100 ...
##  $ Rev.per.mile      : int  2890 2335 2280 2535 2545 2565 1570 1320 1690 1510 ...
##  $ Man.trans.avail   : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 1 1 1 1 ...
##  $ Fuel.tank.capacity: num  13.2 18 16.9 21.1 21.1 16.4 18 23 18.8 18 ...
##  $ Passengers        : int  5 5 5 6 4 6 6 6 5 6 ...
##  $ Length            : int  177 195 180 193 186 189 200 216 198 206 ...
##  $ Wheelbase         : int  102 115 102 106 109 105 111 116 108 114 ...
##  $ Width             : int  68 71 67 70 69 69 74 78 73 73 ...
##  $ Turn.circle       : int  37 38 37 37 39 41 42 45 41 43 ...
##  $ Rear.seat.room    : num  26.5 30 28 31 27 28 30.5 30.5 26.5 35 ...
##  $ Luggage.room      : int  11 15 14 17 13 16 17 21 14 18 ...
##  $ Weight            : int  2705 3560 3375 3405 3640 2880 3470 4105 3495 3620 ...
##  $ Origin            : Factor w/ 2 levels "USA","non-USA": 2 2 2 2 2 1 1 1 1 1 ...
##  $ Make              : Factor w/ 93 levels "Acura Integra",..: 1 2 4 3 5 6 7 9 8 10 ...
```

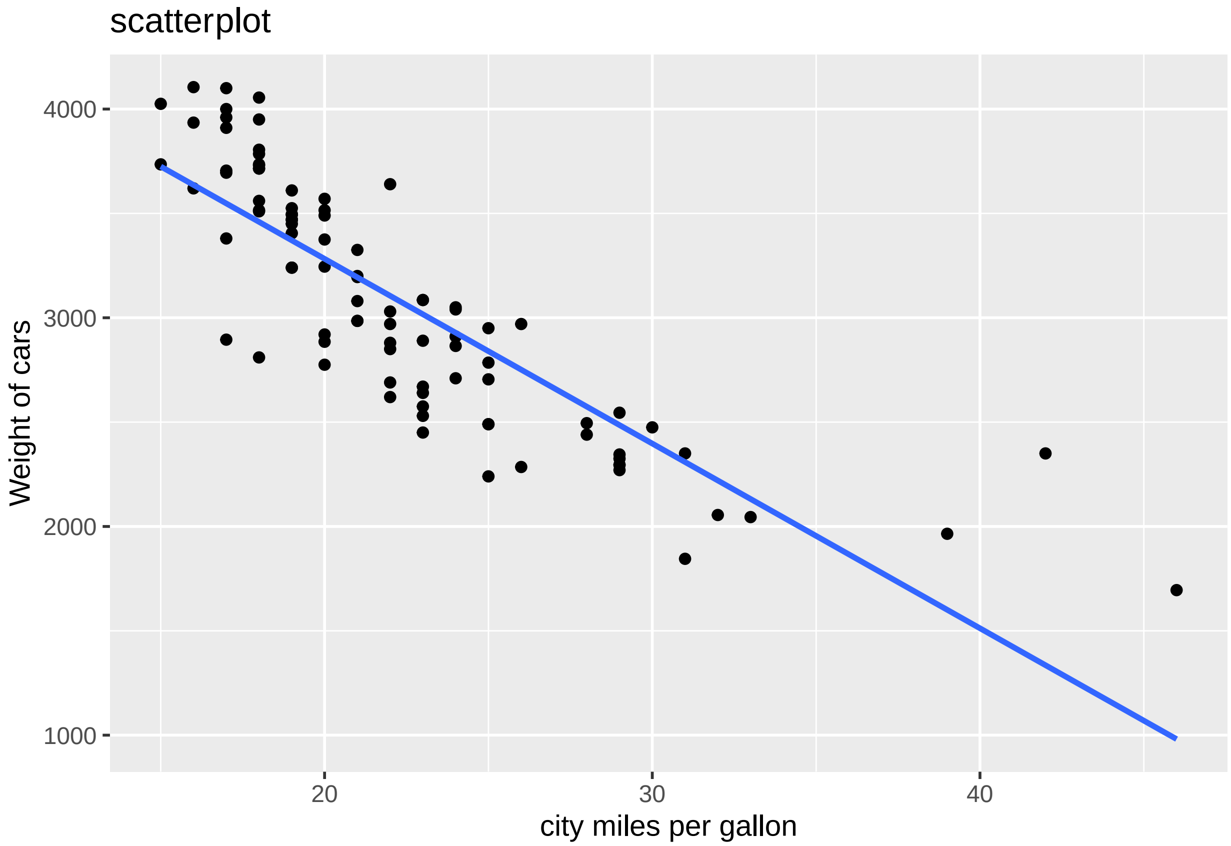
```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

Plot to visualize the relationship between MPG.city (miles per gallon in city) and Weight:

```
ggplot(Cars93, aes(x=MPG.city, y=Weight))+geom_point()+geom_smooth(method='lm',se=FALSE)+
  labs(x='city miles per gallon',y='Weight of cars', title='scatterplot')
```

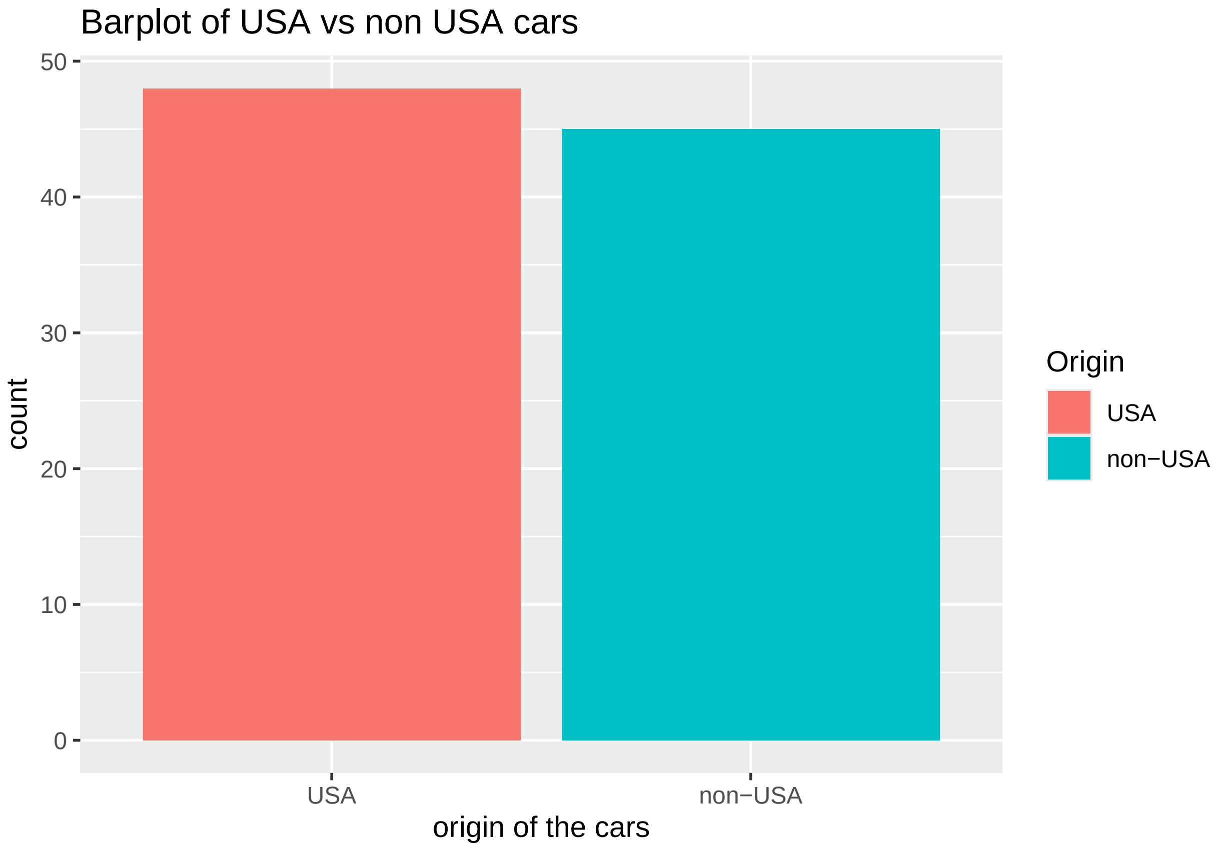
```
## `geom_smooth()` using formula = 'y ~ x'
```



The scatter plot shows how the weight of the cars varies with city miles per gallon. The linear regression line shows whether there is a positive or negative correlation.

Bar plot to show cars from the USA compared to non-USA origins in the dataset:

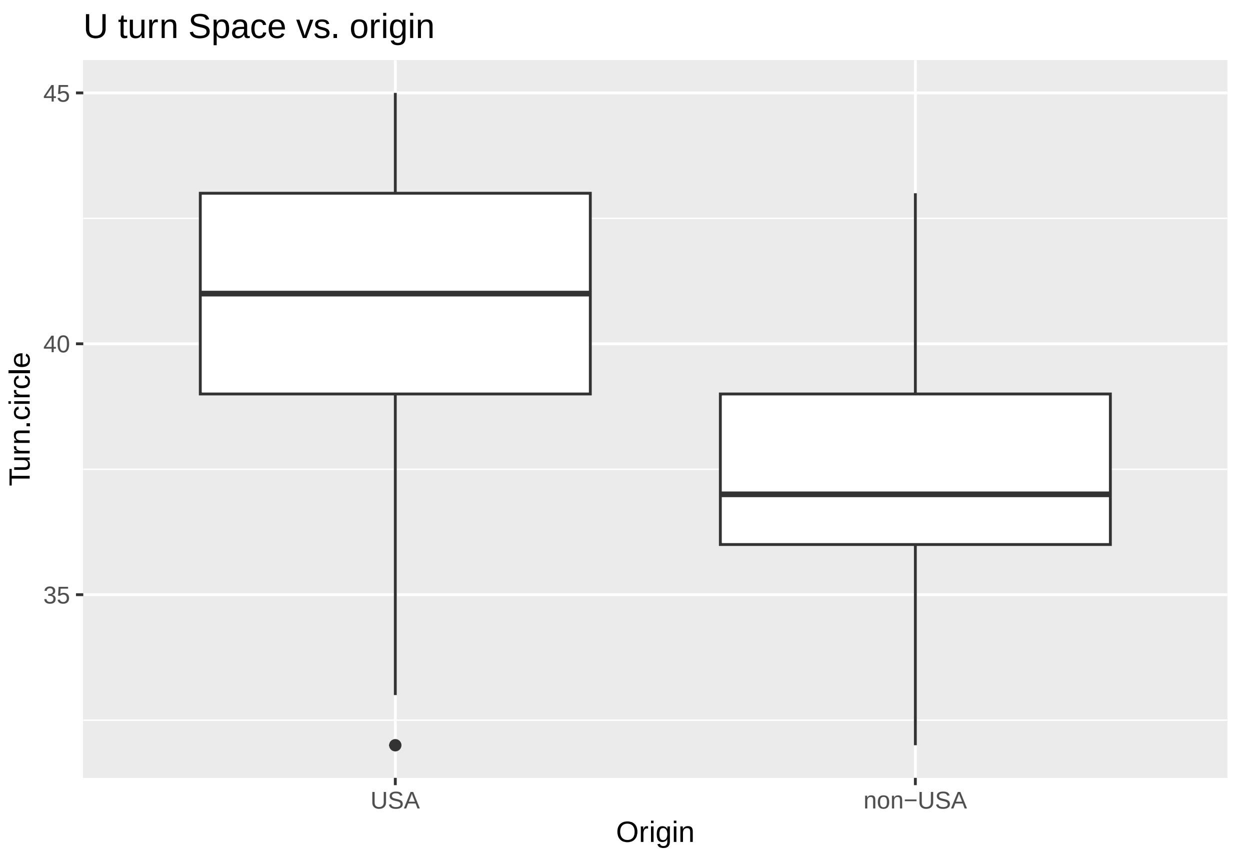
```
ggplot(Cars93,aes(x=Origin,fill=Origin))+geom_bar()+ggtitle('Barplot of USA vs non USA cars')+
  xlab('origin of the cars')
```



The plor reveals a dominance of American manufacturers in the dataset. This simple visual count suggests a possible bias toward American-made cars.

A box plot to compare the U-turn space between cars from different origins:

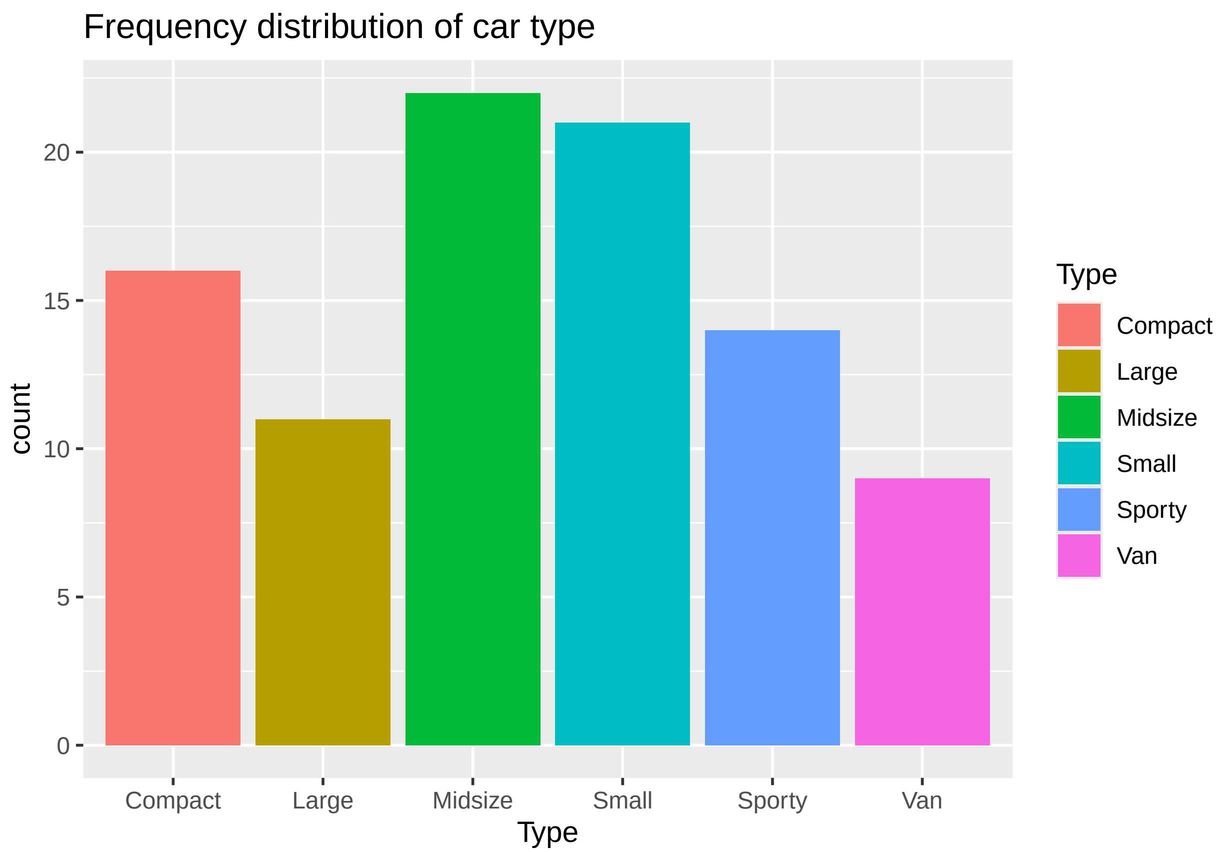
```r
ggplot(Cars93, aes(x=Origin, y=Turn.circle))+geom_boxplot()+ggtitle('U turn Space vs. origin')
```



This box plot shows that the median U-turn space is generally larger for USA cars compared to non-USA cars. This suggests that American cars, on average, require more space for U-turns. The presence of outliers in both groups suggests that there is variability within each group.

A bar plot to display the frequency distribution of different car types:

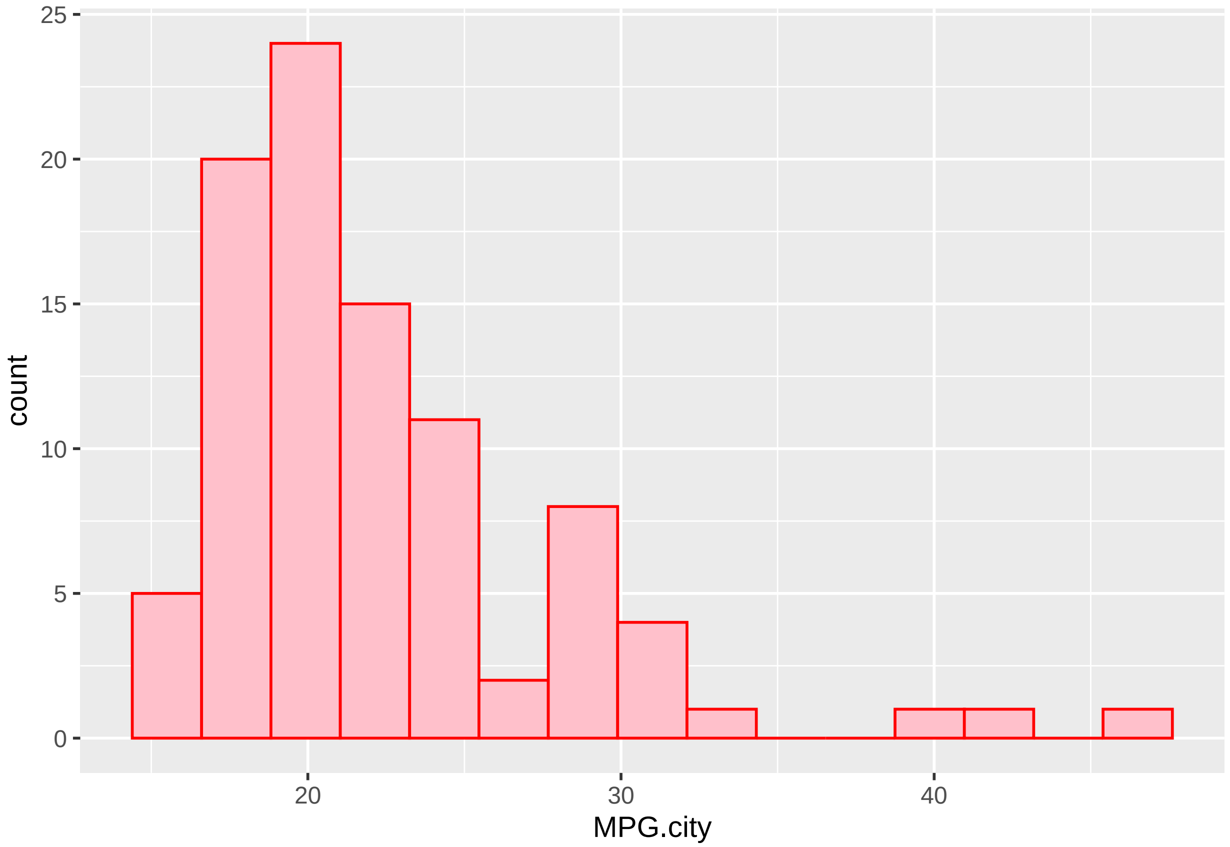
```r
ggplot(Cars93, aes(x=Type, fill=Type))+geom_bar()+ggtitle('Frequency distribution of car type')
```



This distribution shows that Mid-size and Compact cars are the most frequent car types in the dataset. Sporty and Large cars are less common.

A histogram to visualize the distribution of city miles per gallon (MPG):

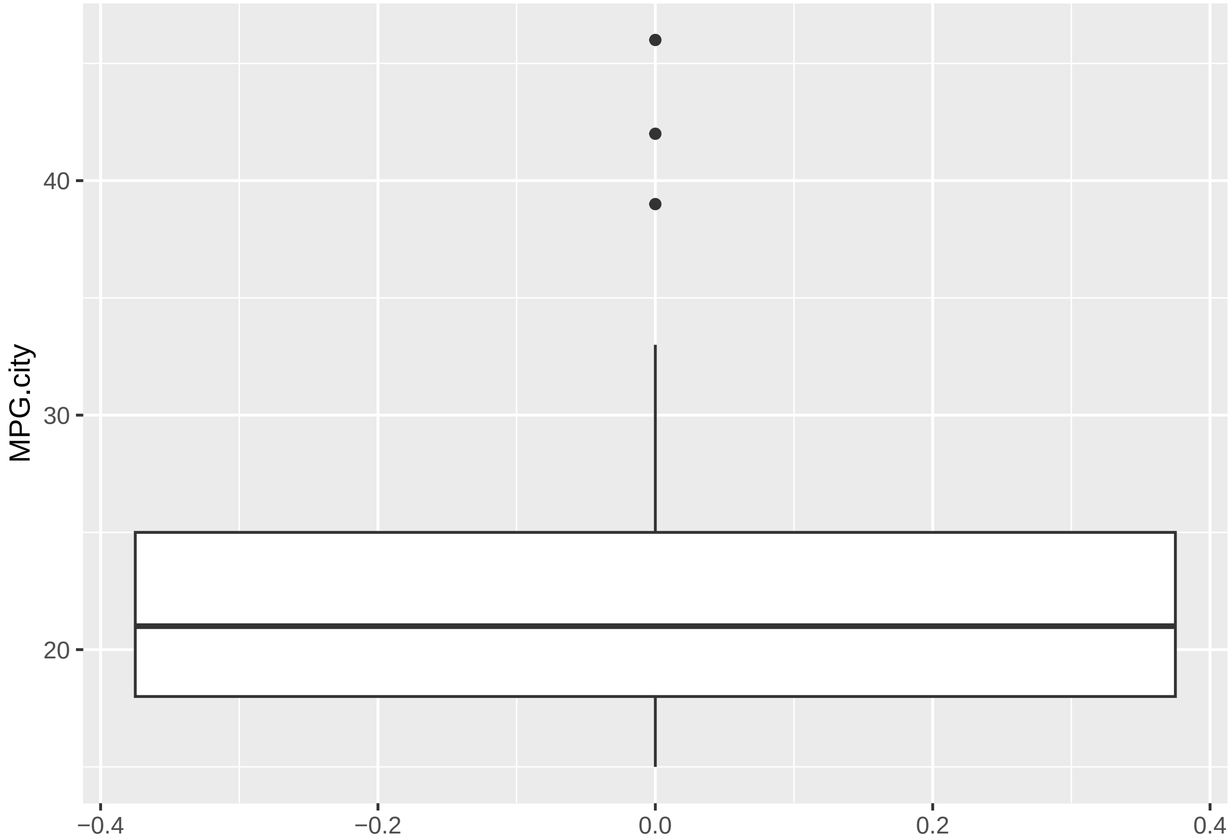
```
ggplot(Cars93, aes(MPG.city))+geom_histogram(bins=15, color = 'red', fill='pink')
```



This histogram shows that most cars get around 20-30 miles per gallon in the city, with a peak between 20 and 25 MPG. Very few cars fall at the extremes, indicating that highly fuel-efficient or inefficient cars are rare in this dataset.

A box plot to display the distribution of city miles per gallon, flipped horizontally:

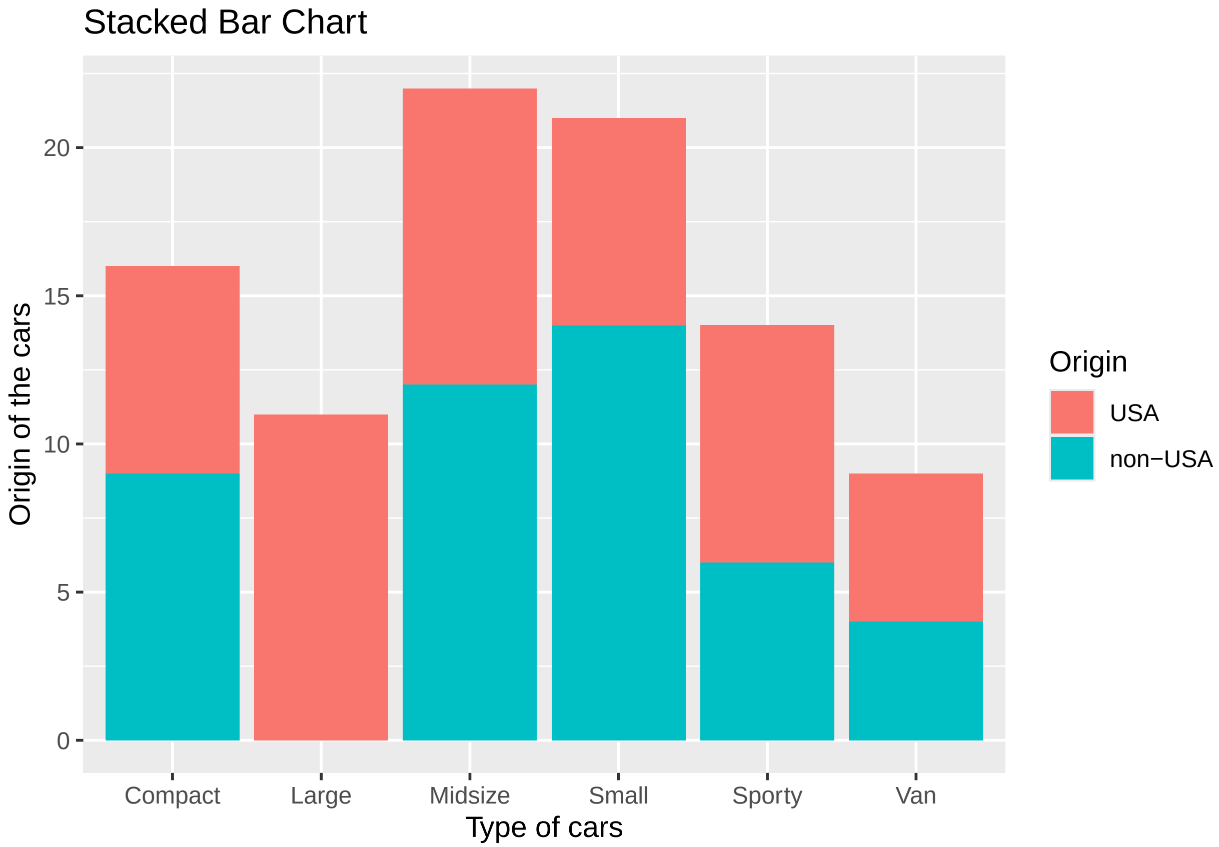
```r
ggplot(Cars93, aes(x=MPG.city))+geom_boxplot()+coord_flip()
```



The median city MPG is around 23-25 MPG, with a somewhat narrow interquartile range, indicating that most cars have similar fuel efficiency in city driving. However, there are a few outliers on the higher end.

A stacked bar chart to show the breakdown of car types by origin:

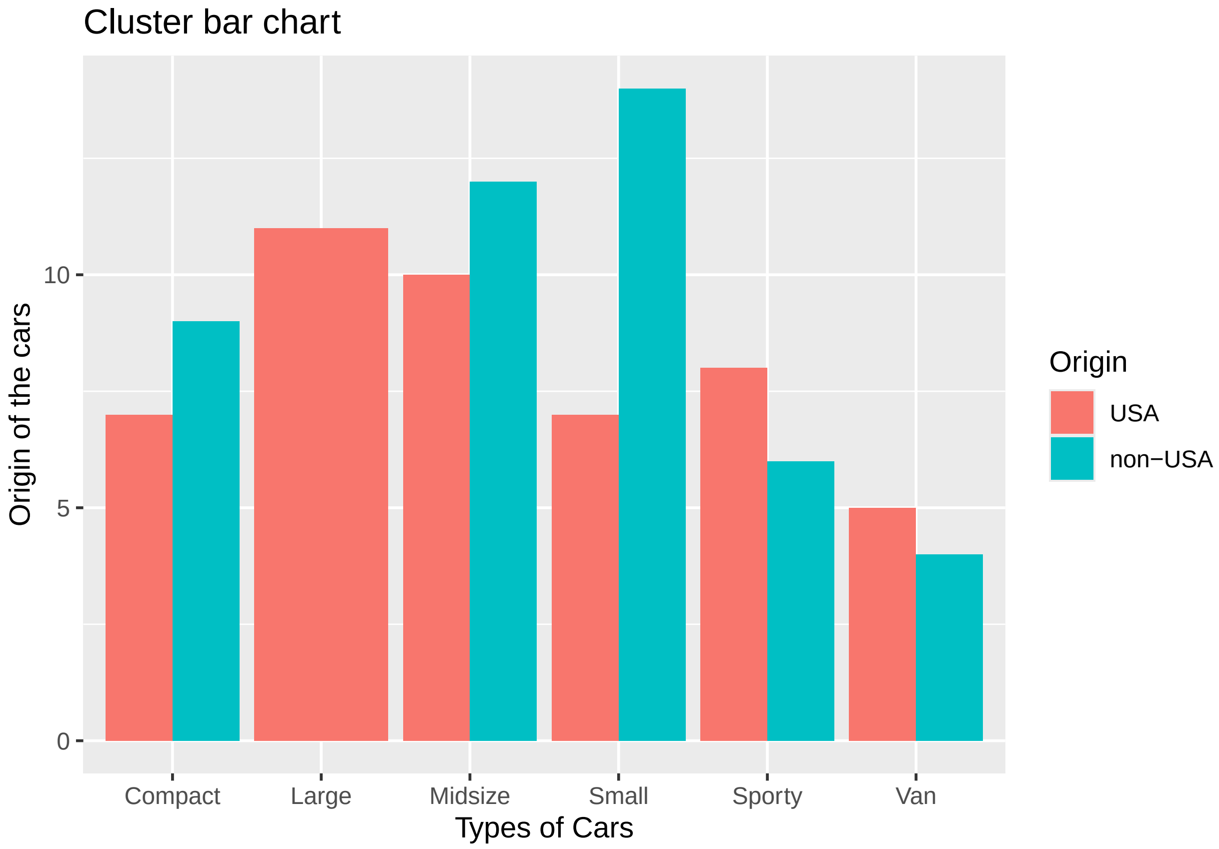
```r
ggplot(Cars93, aes(x=Type, fill=Origin))+geom_bar()+
  labs(x="Type of cars", y="Origin of the cars", title="Stacked Bar Chart")
```



The stacked bar chart shows that USA cars dominate most types, particularly Mid-size and Compact vehicles. Non-USA cars are more common in the Sporty and Small categories. This highlights production differences by origin across car types.

A cluster bar chart to show car types by origin for direct comparison:

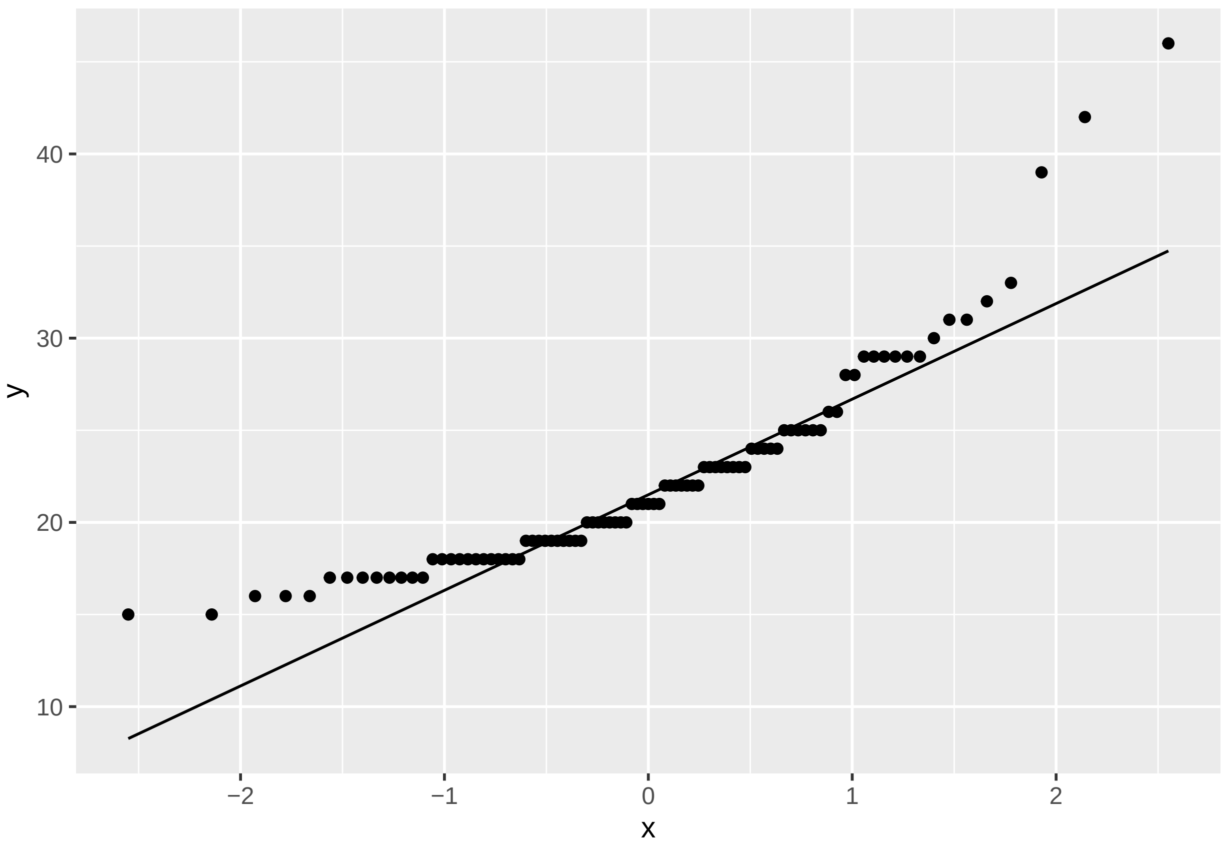
```r
ggplot(Cars93, aes(x=Type, fill=Origin))+geom_bar(position="dodge")+
  labs(x="Types of Cars", y = "Origin of the cars", title ="Cluster bar chart")
```



The comparison shows that USA cars dominate the Mid-size, Large, and Compact categories, while non-USA cars have a stronger presence in Sporty and Small types.

A QQ plot to visualize if the MPG.city variable has a normal distribution:

```r
ggplot(Cars93, aes(sample = MPG.city)) +
  stat_qq() +
  stat_qq_line()
```



The QQ plot shows some deviation from the normal line, particularly at the tails, indicating that the distribution of city MPG is not perfectly normal.

Shapiro-Wilk test to statistically test the normality of the MPG.city variable:

```r
shapiro.test(Cars93$MPG.city)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Cars93$MPG.city
## W = 0.85831, p-value = 5.763e-08
```

The p-value from the Shapiro-Wilk test is significant ($p < 0.05$), which confirms that the MPG.city variable does not follow a normal distribution. This finding is consistent with the deviations observed in the QQ plot.