IMS Chap 4 & 5: Exploring Data (Part 1)

Packages Needed To Recreate Code on Slides

```
library(tidyverse) # for ggplot functions/plotting
library(openintro) # for data set
```

Warning: It is not expected that you understand all the R code in this presentation right now. You will go over more R code in SDS 100 in the coming weeks. However, you are welcome to try to make these plots on your own.

Data Frame We Will Be Using

```r
# Preview the data
head(cars93)
```

```
# A tibble: 6 x 6
  type    price mpg_city drive_train passengers weight
  <fct>   <dbl>    <int> <fct>             <int>  <int>
1 small    15.9       25 front                 5   2705
2 midsize  33.9       18 front                 5   3560
3 midsize  37.7       19 front                 6   3405
4 midsize  30         22 rear                  4   3640
5 midsize  15.7       22 front                 6   2880
6 large    20.8       19 front                 6   3470
```

Data Frame We Will Be Using

These cars represent a random sample for 1993 models that were in both *Consumer Reports* and *PACE Buying Guide*.

- `type`: The vehicle type
- `price`: Vehicle price (USD).
- `mpg_city`: Vehicle mileage in city (miles per gallon).
- `drive_train`: Vehicle drive train with levels `4WD`, `front`, and `rear`.
- `passengers`: The vehicle passenger capacity.
- `weight`: Vehicle weight (lbs).

54 cases (rows) and 6 variables (columns)

What is the purpose of exploratory data analysis (EDA)?

- Exploratory data analysis (EDA) refers to the practice of reducing and summarizing data. EDA refers to both
  - numerical **summary statistics**
  - data visualizations or **graphs**.
- It is the statistics version of *tl;dr*
- We ultimately want to learn about the **distribution** of the variables in the data frame.
- A **distribution** is the map of the different possible values, and associated probabilities.

One Categorical Variable

One Categorical Variable - Summary Statistics

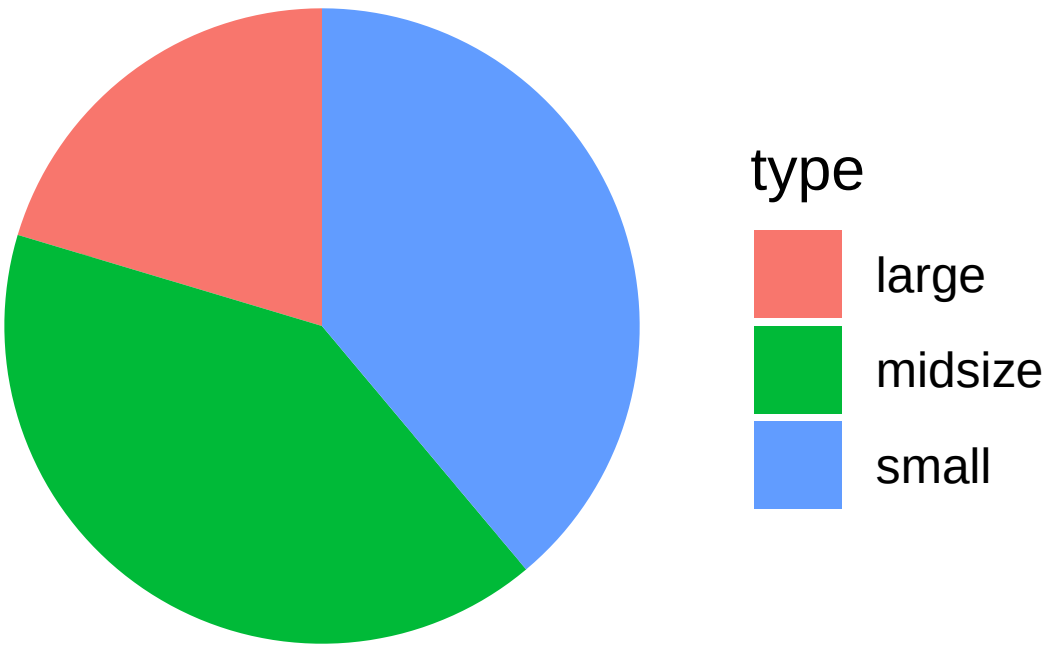
We typically summarize one categorical using a frequency table.

```
table(cars93$type)
```

```
large midsize   small
   11      22      21
```

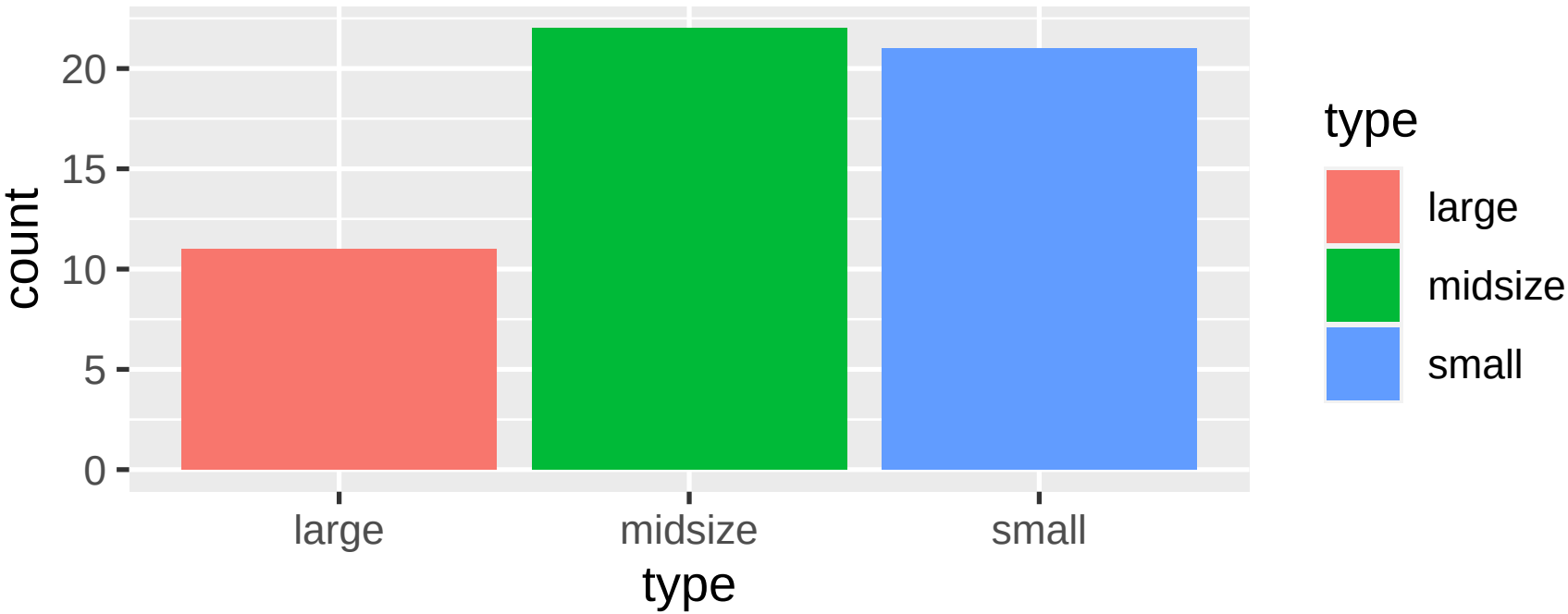
One Categorical Variable - Pie Chart

```
cars93 %>%
  count(type) %>%
  ggplot(aes(x="", y=n, fill=type)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0)+
  theme_void()
```



One Categorical Variable - Bar Chart

```
ggplot(cars93, aes(x = type, fill = type)) +
  geom_bar()
```



What are some differences between a pie chart and a bar chart?

One Numerical Variable

Summary Statistics - Measures of Center

A measure of center statistic tells us what a *typical* or *average* value is a for a numerical variable.

- **Mean**: Sum of all the values, divided by total number of values (n).

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Median**: Splits the data in half. 50% of the data fall below this value and 50% fall above it. (If n is even take the mean of the two center values).
- **Mode**: The most common value for the variable.

Summary Statistics - Measures of Spread

A measure of spread statistic roughly describes how far away a value is from the center.

- **Variance:**

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

- **Standard deviation:** $s = \sqrt{s^2}$
- **Range**: maximum - minimum

Summary Statistics - Measures of Spread

Why use variance, why not use *average absolute mean deviation*?

$$? = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

- Historically, people did use this!
- Absolute values are hard though...squared terms are nicer.
- Variance naturally occurs in important theorems.
- An error that is twice is large is often twice as bad, a square term reflects this.

Summary Statistics - Measures of Spread

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

Why do we have $n-1$ in the denominator of the variance formula instead of n?

- Better statistical properties!

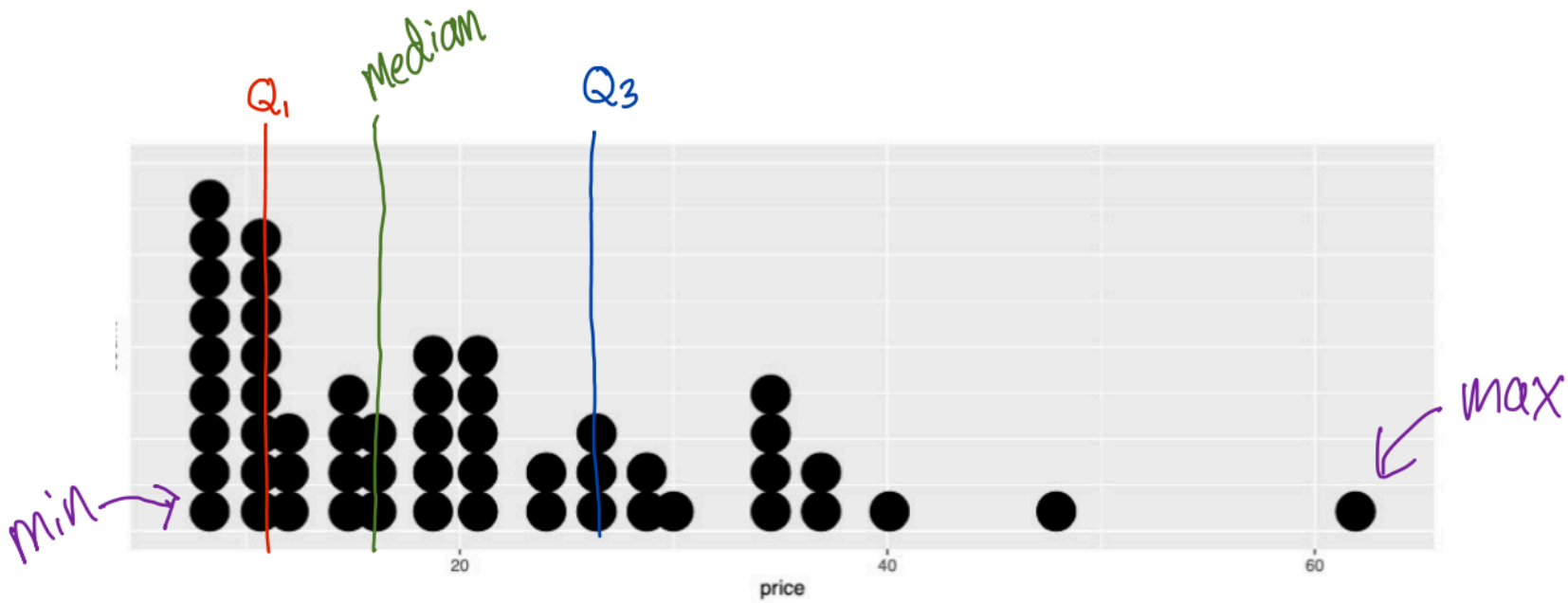
Summary Statistics - Measures of Spread

Quartiles are also a useful tool when considering measures of spread.

- 1st Quartile (Q_1): 25% of observations lie below this value.
- 2nd Quartile: the median
- 3rd Quartile (Q_3): 75% of observations lie below this value
- 4th Quartile: the maximum

Last measure of spread is **Interquartile Range (IQR)**: $Q_3 - Q_1$

Summary Statistics - Measures of Spread




```
quantile(cars93$price)
```

```
   0%    25%    50%    75%   100%
 7.40  10.95  17.25  26.25  61.90
```

Robust Summary Statistics

- There are many summary statistics for a single numerical variable.
- We typically use the mean and variance the most when describing data, however, median and IQR are very important too!
- The median and IQR are called *robust* statistics. Can we guess why they are called robust statistics?

Robust Summary Statistics

- There are many summary statistics for a single numerical variable.
- We typically use the mean and variance the most when describing data, however, median and IQR are very important too!
- The median and IQR are called *robust* statistics. Can we guess why they are called robust statistics?
  - They are not sensitive to **outliers**, or extreme values.

Summary Statistics - So what do we report?

For numerical variables we often report one of the following sets of summary statistics.

- Mean and variance
- 6 (or 5) Number Summary
  - Minimum, 1st Quartile, Median, Mean, 3rd Quartile, Maximum

We are not limited to this though!

Summary Statistics - Note on Notation

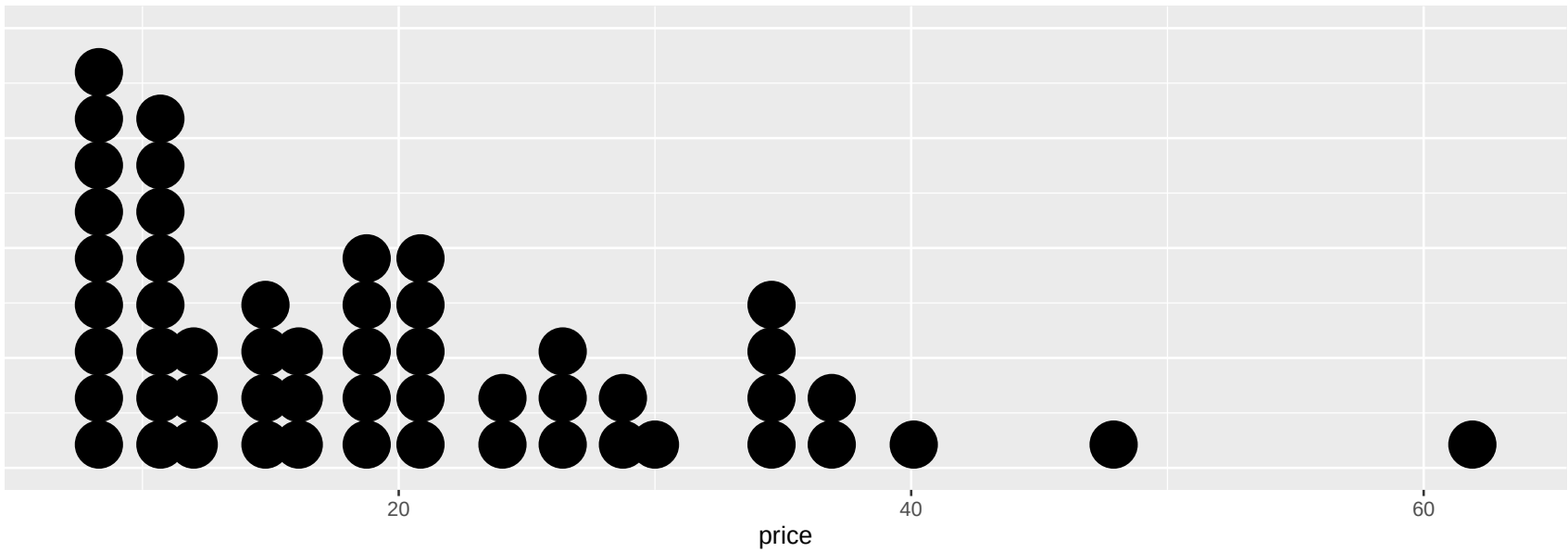
We have used $\overline{x}$ and s^2 for the sample mean and variance. These are *estimated* values. That is, they are calculated with data.

We may see later μ and σ^2 for the population mean and variance. These are considered the *true* values of the population. That is, they are the values of the mean and variance if we took a census (which almost never happens).

One Numerical Variable - Dot Plot

- Useful for small data sets.
- See every value.

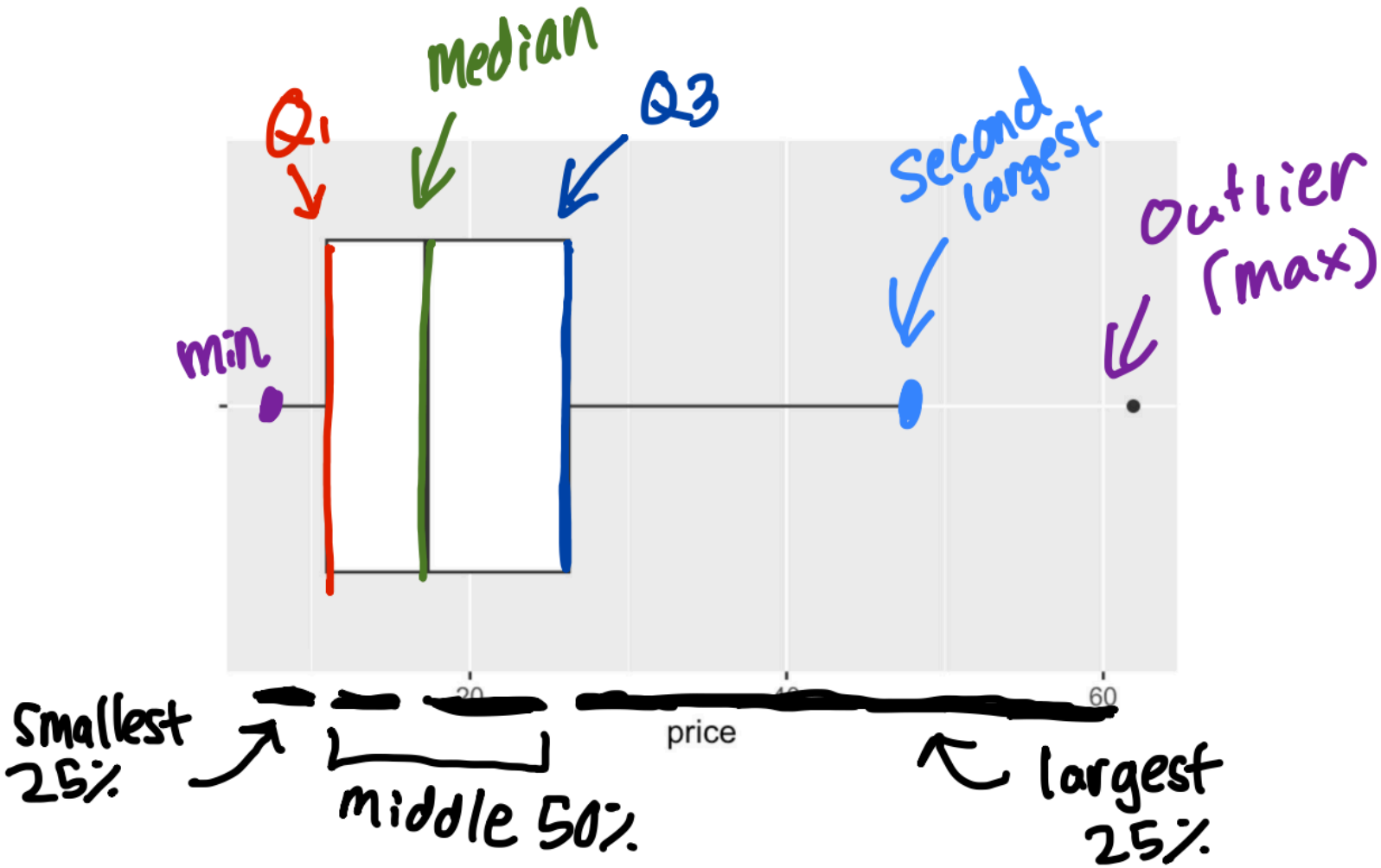
```
ggplot(cars93, aes(x= price))+
  geom_dotplot() + labs(y = "") +
  theme(axis.text.y=element_blank(),
        axis.ticks.y=element_blank())
```



One Numerical Variable - Boxplot

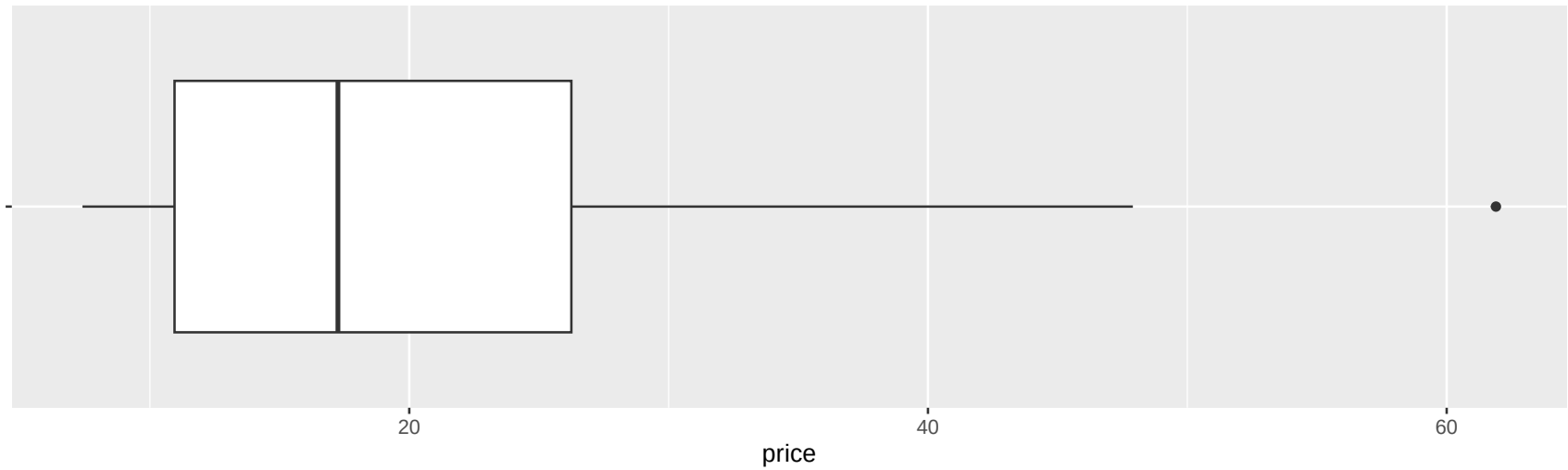
- The outer boundaries of the box are Q_1 and Q_3
- The middle line in the box is the median
- The whiskers extend from the box to the maximum and minimum values, excluding *outliers*.
- See helpful information at a glance.

One Numerical Variable - Boxplot



One Numerical Variable - Boxplot

```
ggplot(cars93, aes(x= price, y = ""))+
  geom_boxplot() + labs(y = "")
```

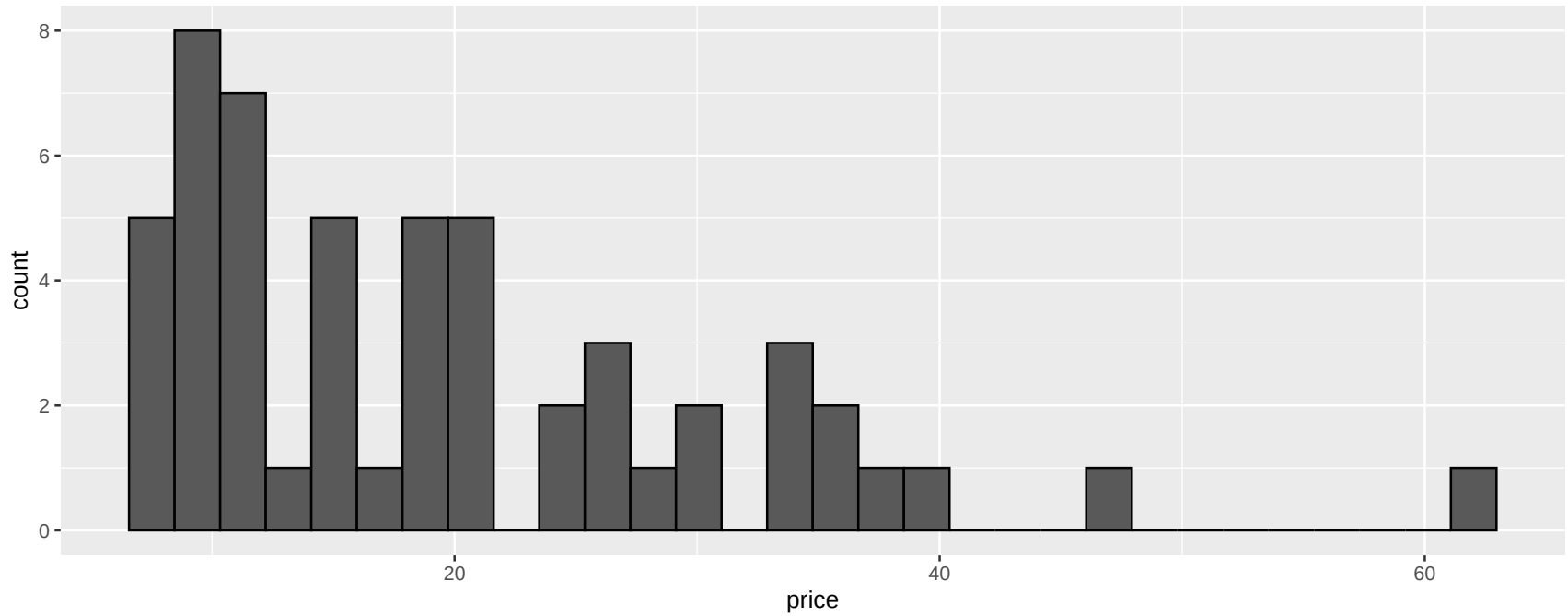


One Numerical Variable - Histogram

- Create *bins* to group close observations together. The bins are on the x axis.
- Observations that fall on the boundary of a bin are allocated to the lower bin.
- Count the number of observations within each bin. Frequencies are on the y axis.
- Useful for large data sets.
- A **mode** is represented by a prominent peak in the distribution.

One Numerical Variable - Histogram

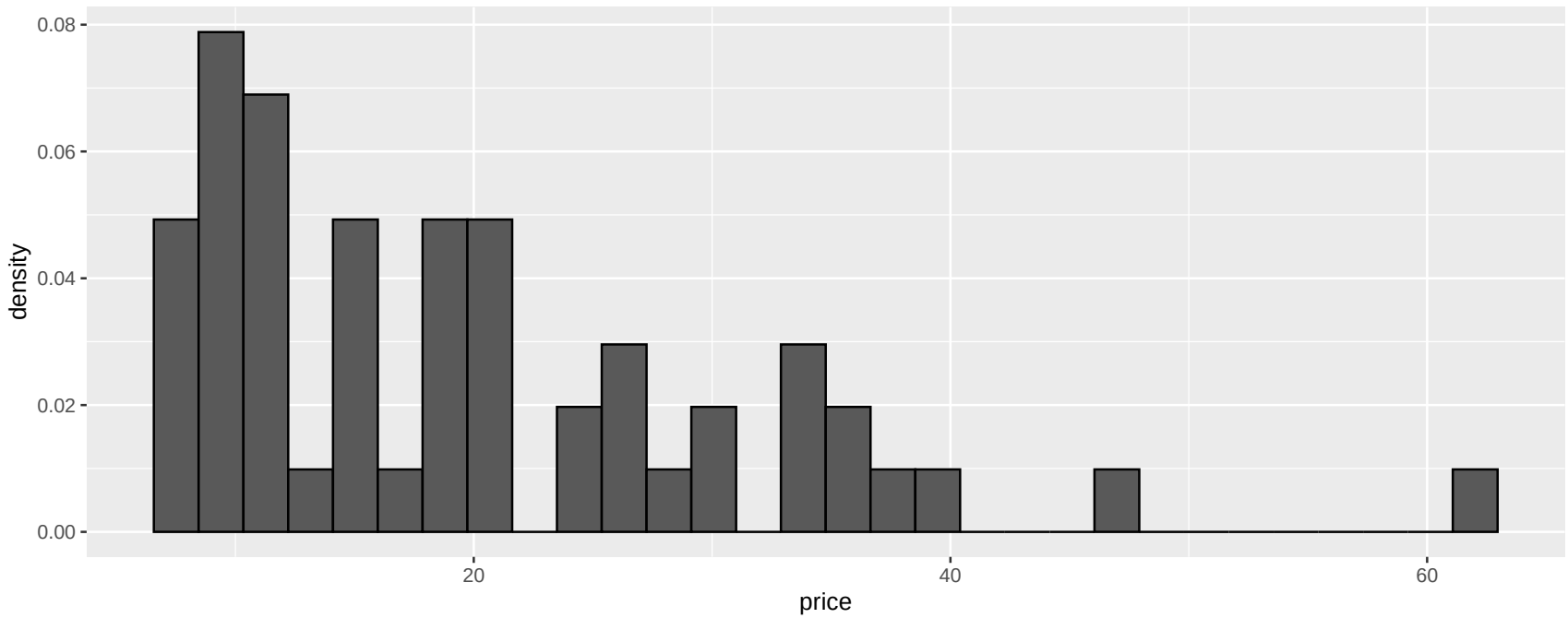
```
ggplot(cars93, aes(x = price))+
  geom_histogram(color = "black")
```



One Numerical Variable - Histogram

Sometimes, instead of counts, we plot the proportion of observations within each bin.

```
ggplot(cars93, aes(x = price, y =..density..))+
  geom_histogram(color = "black")
```

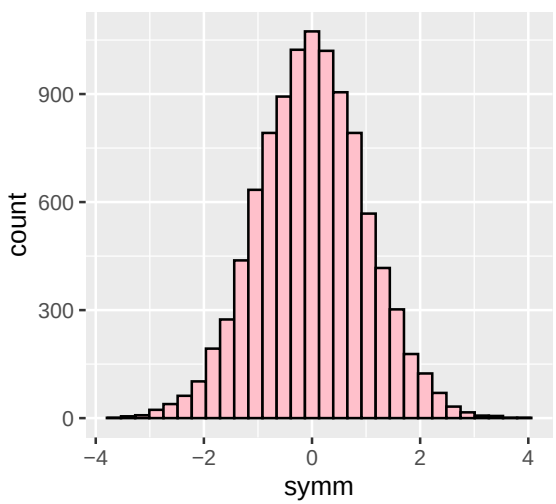


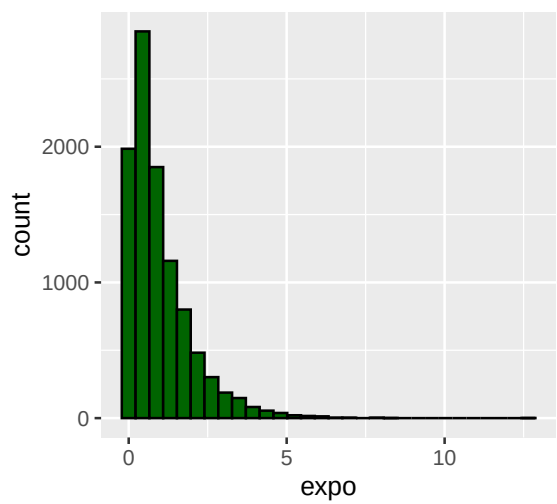
One Numerical Variable Plots

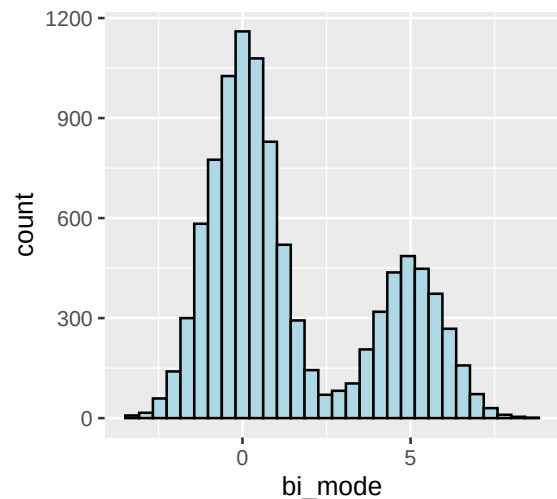
What should we look out for plots of numerical variables?

- **Skewness**: a measure of symmetry. We say a distribution is *symmetric* or *skewed*.
- **Modality:** how many peaks ("modes") the distribution has. We say a distribution is *unimodal*, *bimodal*, and *multimodal*.

There are no *exact* guidelines for determining skewness or modality (or outliers!). Rules of thumbs exist, but for sample data often we resort to context, expertise, and best guesses.



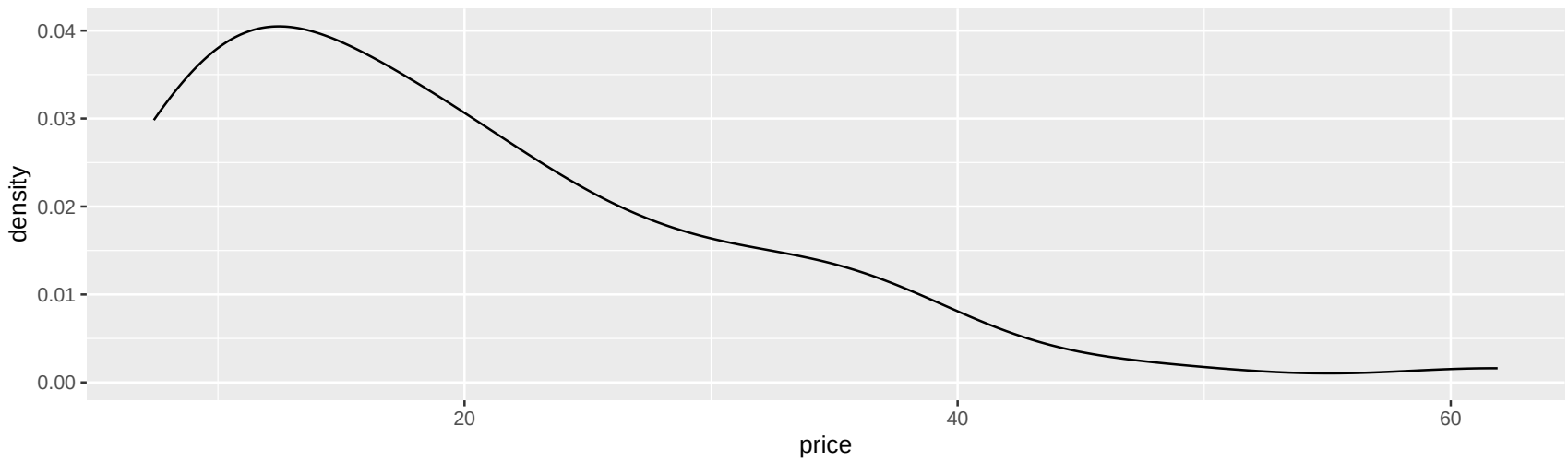




One Numerical Variable - Density Plot

- Density plots are similar to histograms, but they instead smooth out the bins.
- This can be useful when looking at multiple groups at once (more on that next time!)

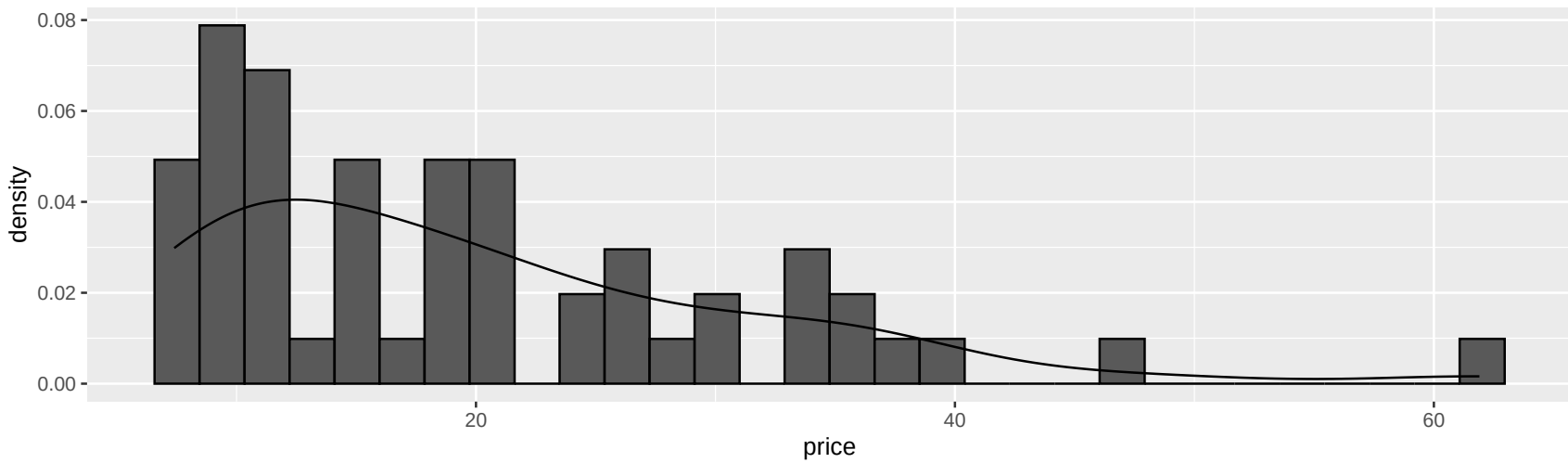
```
ggplot(cars93, aes(x = price))+
  geom_density()
```



One Numerical Variable - Density Plot

Density plots can sometimes be too smooth, or not smooth enough!

```
ggplot(cars93, aes(x = price))+
  geom_histogram(color = "black", aes(y =..density..))+
  geom_density()
```



A Few Notes About R

A Few Notes About R

If your numeric variables and categorical variables are stored as the right object type

- numeric variable $\rightarrow$ numeric vector in R
- categorical variable $\rightarrow$ factor or character vector in R

then the summary statistics can often be completed in one step with the `summary()` function.

Graphs typically take more effort.

Calculate Summary Statistics

```
# Try this on your machine!
# (Make sure you have loaded the openintro package)
summary(cars93)
```