Introduction to Data Science: Exploratory Data Analysis

Héctor Corrada Bravo

University of Maryland, College Park, USA

2019-08-14

CBCB Center for Bioinformatics & Computational Biology

University of Maryland

Exploratory Data Analysis

What to do with a dataset before modeling using Statistics or Machine Learning.

- Better understand the data at hand,
- help us make decisions about appropriate modeling methods,
- helpful data transformations that may be helpful to do.

Exploratory Data Analysis

There are many instances where statistical data modeling is not required to tell a clear and convincing story with data.

Many times an effective visualization can lead to convincing conclusions.

Exploratory Data Analysis

Goal Perform an initial exploration of attributes/variables across entities/observations.

We will concentrate on exploration of single or pairs of variables.

Later on in the course we will see *dimensionality reduction* methods that are useful in exploration of more than two variables at a time.

Exploratory Data Analysis

Computing summary statistics

- how to interpret them
- understand properties of attributes.

Data transformations

- change properties of variables to help in visualization or modeling.

First, how to use visualization for exploratory data analysis.

Exploratory Data Analysis

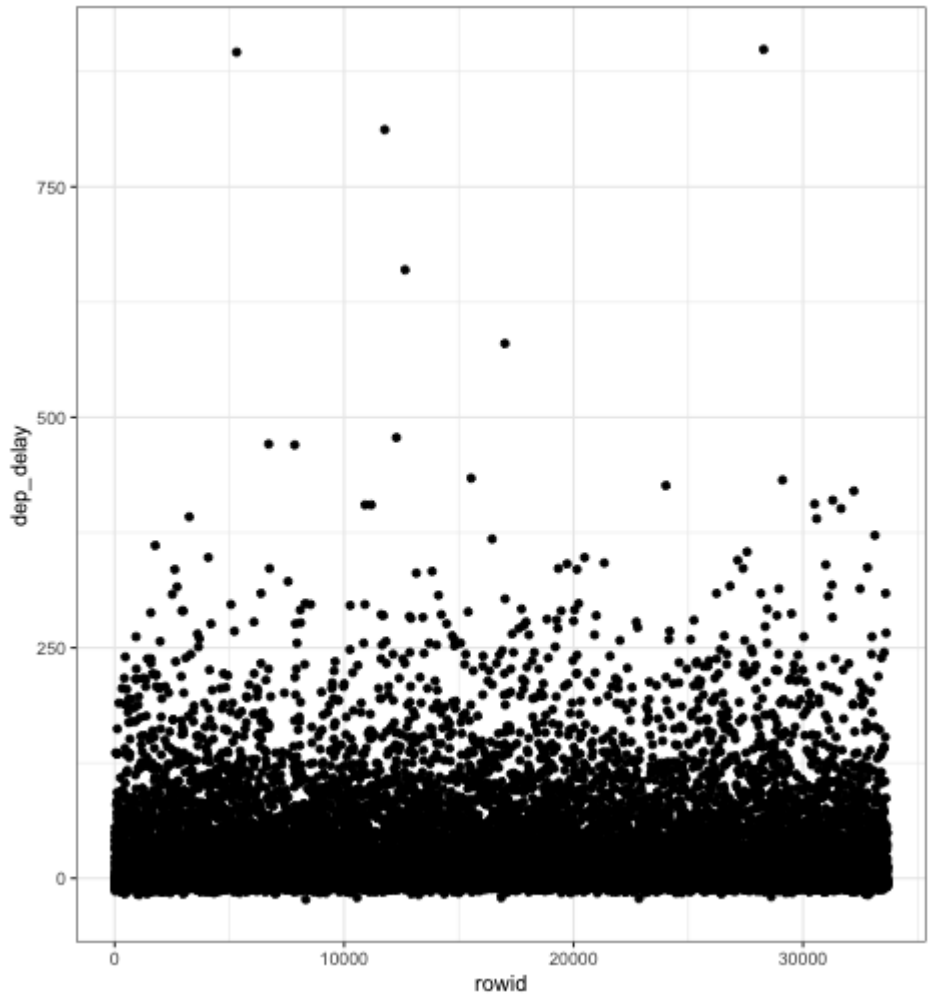
Ultimately, the purpose of EDA is to spot problems in data (as part of data wrangling) and understand variable properties like:

- central trends (mean)
- spread (variance)
- skew
- outliers

This will help us think of possible modeling strategies (e.g., probability distributions)

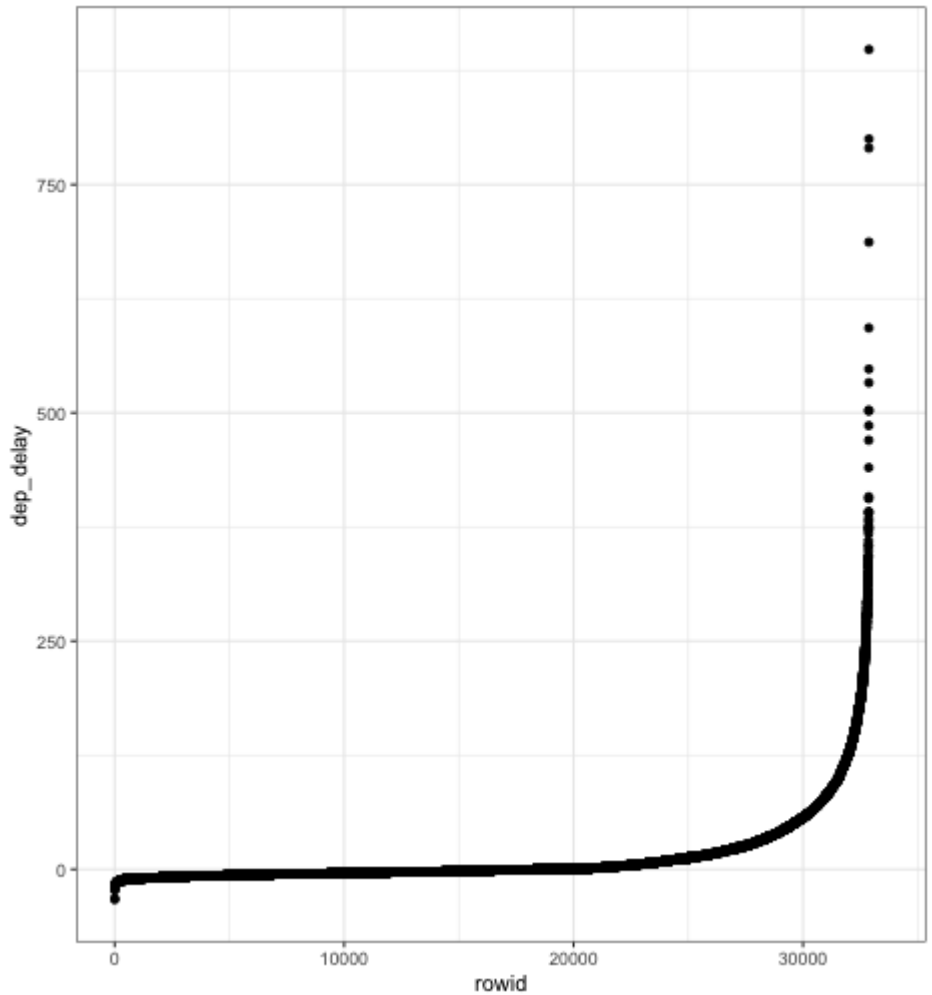
Visualization of single variables

```
flights %>%
  sample_frac(.1) %>%
  rowid_to_column() %>%
  ggplot(aes(x=rowid, y=dep_delay)) +
    geom_point()
```



Visualization of single variables

```
flights %>%
  sample_frac(.1) %>%
  arrange(dep_delay) %>%
  rowid_to_column() %>%
  ggplot(aes(x=rowid, y=dep_delay)) +
    geom_point()
```



Visualization of single variables

What can we make of that plot now? Start thinking of *central tendency*, *spread* and *skew* as you look at that plot.

Let's now create a graphical summary of that variable to incorporate observations made from this initial plot.

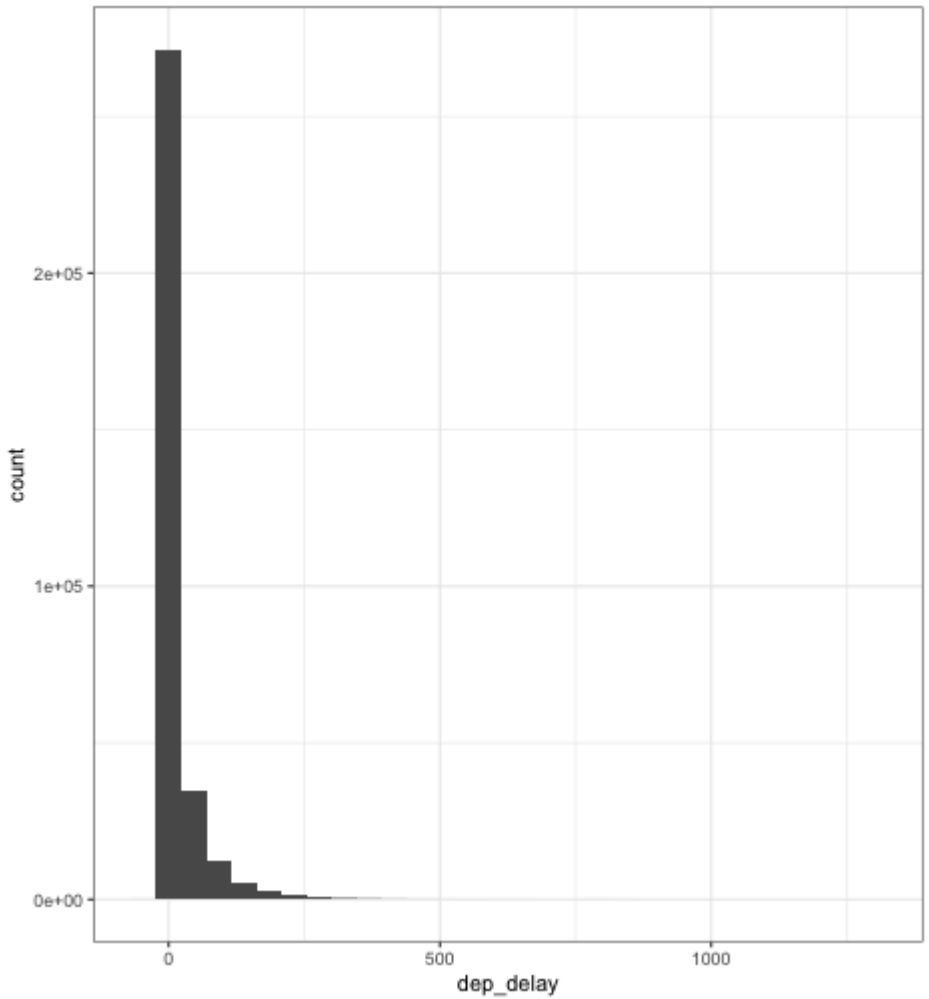
Let's start with a *histogram*: it divides the *range* of the `dep_delay` attribute into **equal-sized** bins, then plots the number of observations within each bin.

Visualization of single variables

```
flights %>%

  ggplot(aes(x=dep_delay)) +

    geom_histogram()
```



Visualization of single variables

Density plot

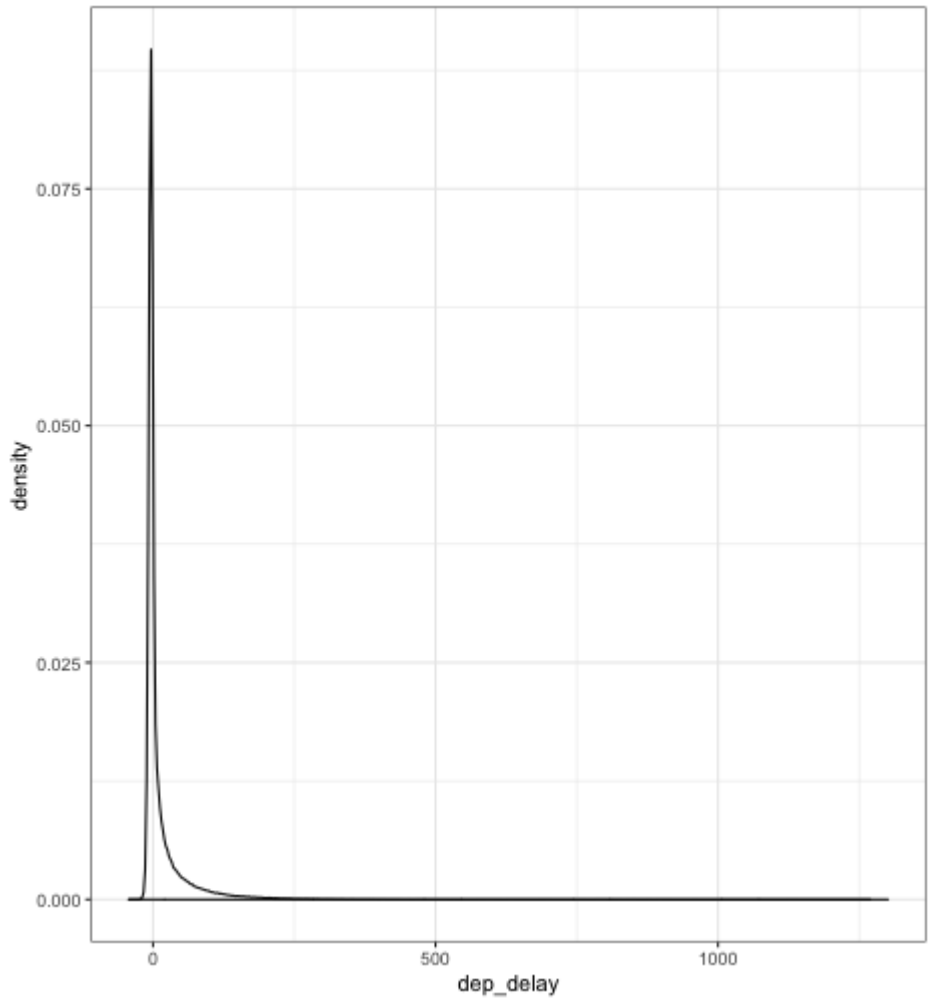
We can (conceptually) make the bins as small as possible and get a smooth curve that describes the *distribution* of values of the `dep_delay` variable.

Visualization of single variables

```
flights %>%

  ggplot(aes(x=dep_delay)) +

    geom_density()
```



Visualization of single variables

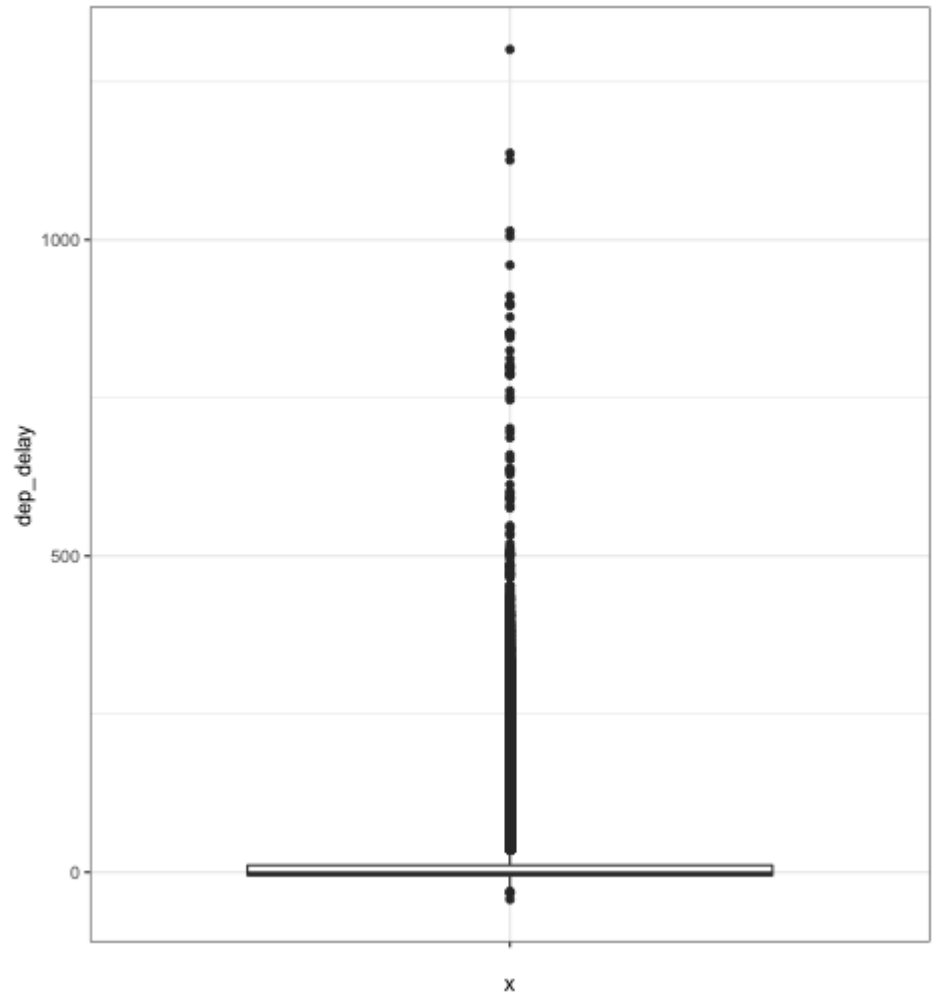
Boxplot Succint graphical summary of the distribution of a variable.

Visualization of single variables

```
flights %>%

   ggplot(aes(x='',y=dep_delay)) +

    geom_boxplot()
```

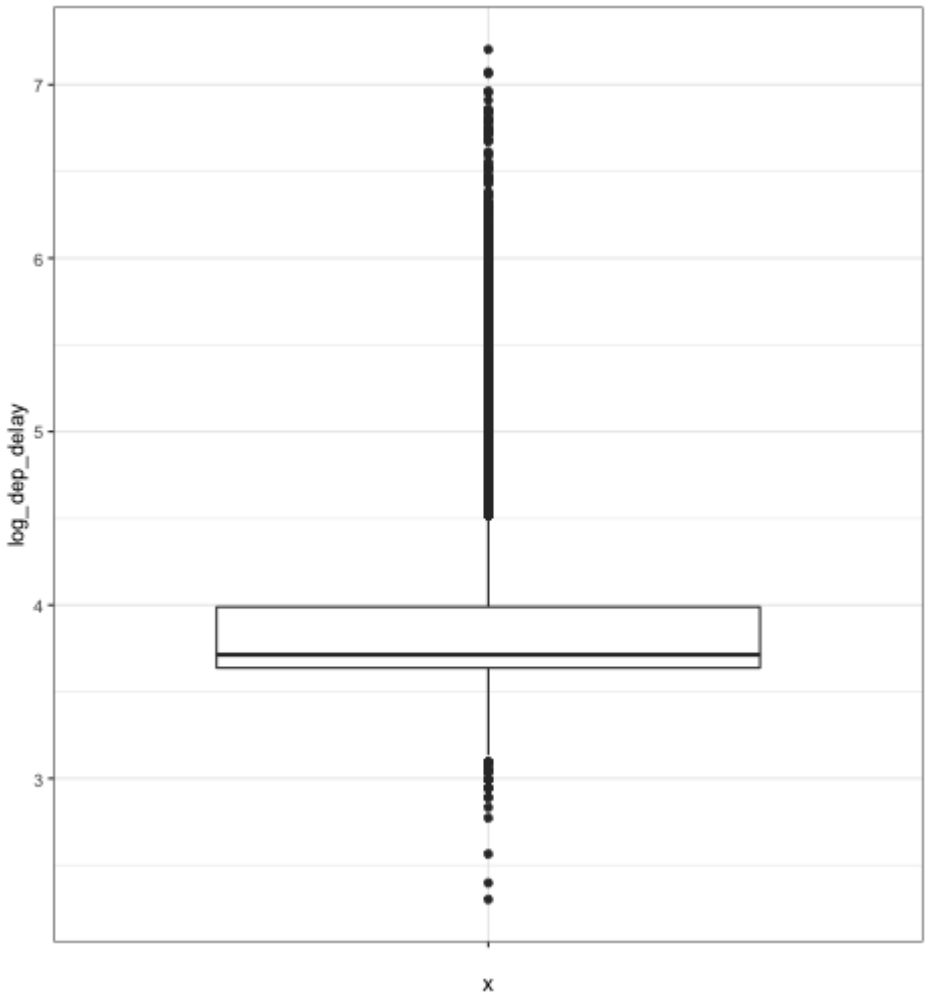


Visualization of single variables

That's not very clear to see, so let's do a *logarithmic* transformation of this data to see distribution better.

Visualization of single variables

```
flights %>%
  mutate(min_delay=min(dep_delay, na.rm=TRUE
  mutate(log_dep_delay = log(dep_delay - min
  ggplot(aes(x='', y=log_dep_delay)) +
    geom_boxplot()
```



Visualization of single variables

So what does this represent?

(a) central tendency (using the median) is represented by the black line within the box,

(b) spread (using inter-quartile range) is represented by the box and whiskers.

(c) outliers (data that is *unusually* outside the spread of the data)

Visualization of pairs of variables

How do each of the distributional properties we care about (central trend, spread and skew) of the values of an attribute change based on the value of a different attribute?

Suppose we want to see the relationship between `dep_delay`, a *numeric* variable, and `origin`, a *categorical* variable.

Visualization of pairs of variables

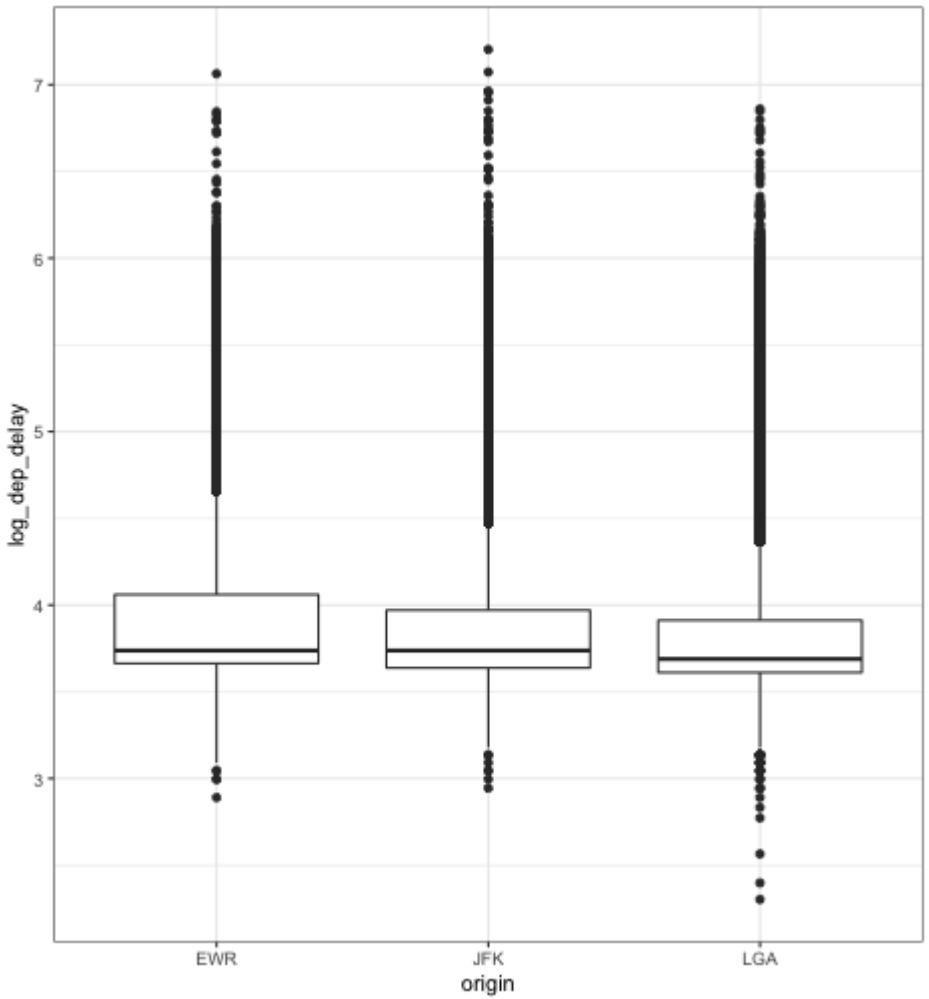
Previously, we saw used `group_by-summarize` operations to compute attribute summaries based on the value of another attribute.

We also called this *conditioning*. In visualization we can start thinking about conditioning as we saw before.

Here is how we can see a plot of the distribution of departure delays *conditioned* on origin airport.

Visualization of pairs of variables

```
flights %>%
  mutate(min_delay = min(dep_delay, na.rm=TR
  mutate(log_dep_delay = log(dep_delay - min
  ggplot(aes(x=origin, y=log_dep_delay)) +
    geom_boxplot()
```



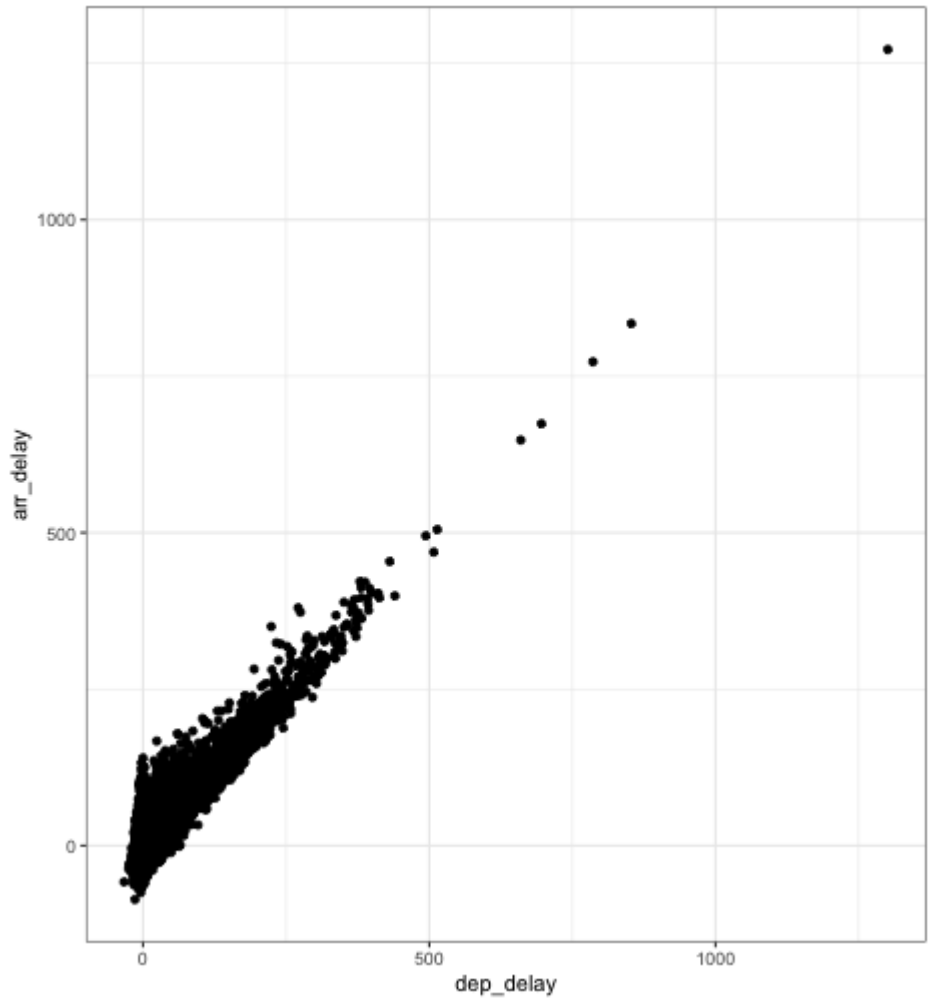
Visualization of pairs of variables

For pairs of continuous variables, the most useful visualization is the scatter plot.

This gives an idea of how one variable varies (in terms of central trend, variance and skew) conditioned on another variable.

Visualization of pairs of variables

```
flights %>%
  sample_frac(.1) %>%
  ggplot(aes(x=dep_delay, y=arr_delay)) +
    geom_point()
```



EDA with the grammar of graphics

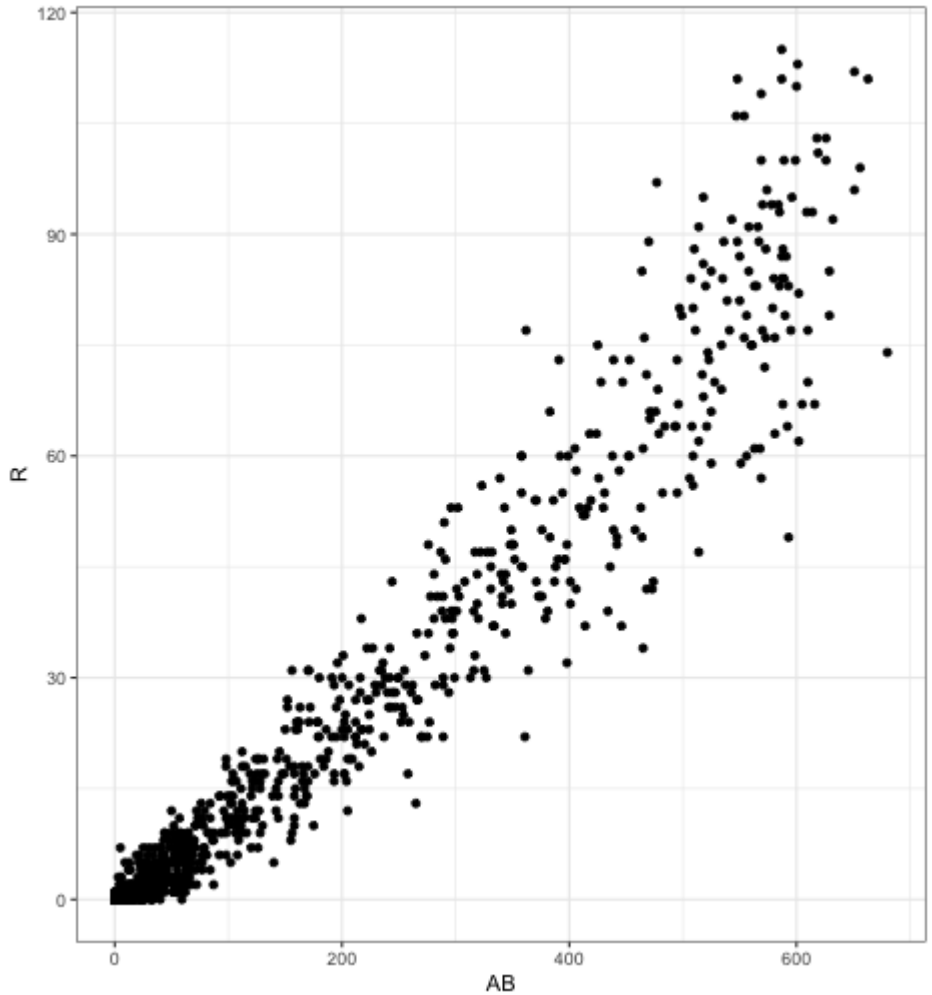
While we have seen a basic repertoire of graphics it's easier to proceed if we have a bit more formal way of thinking about graphics and plots.

The central premise is to characterize the building pieces behind plots:

1. The data that goes into a plot, works best when data is tidy
2. The mapping between data and *aesthetic* attributes
3. The *geometric* representation of these attributes

EDA with the grammar of graphics

```
batting %>%
  filter(yearID == "2010") %>%
  ggplot(aes(x=AB, y=R)) +
    geom_point()
```



EDA with the grammar of graphics

Data: Batting table filtering for year

Aesthetic attributes:

- x-axis mapped to variables `AB`
- y-axis mapped to variable `R`

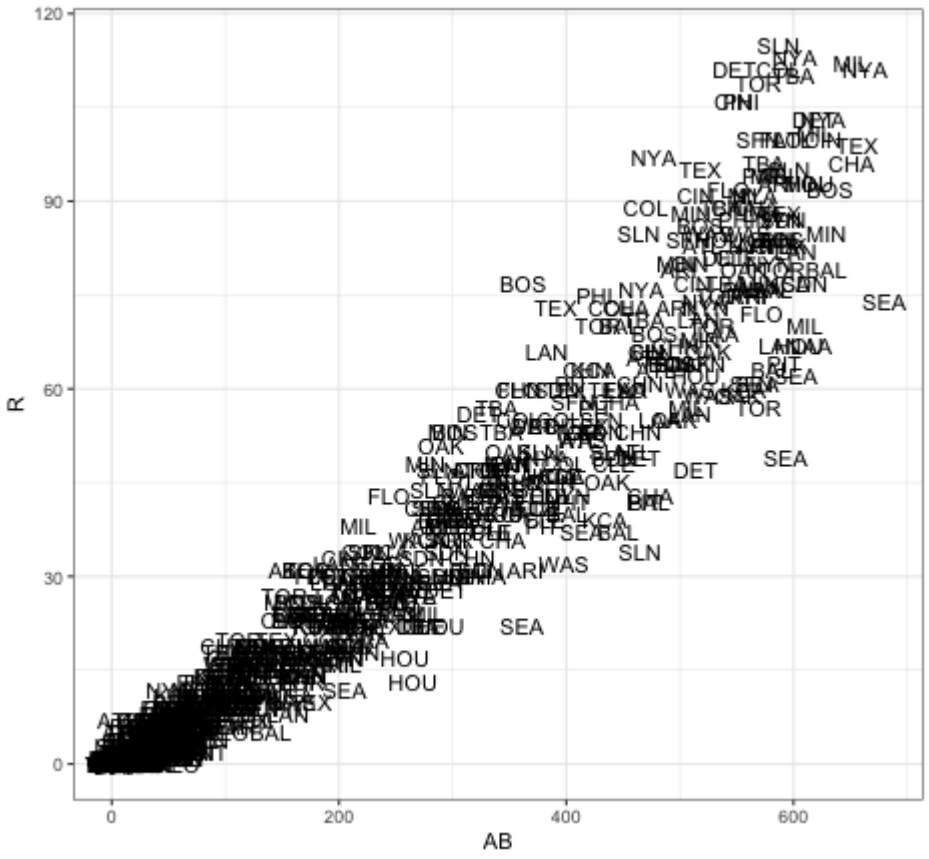
Geometric Representation: points!

Now, you can cleanly distinguish the constituent parts of the plot.

EDA with the grammar of graphics

E.g., change the geometric representation

```
batting %>%
  filter(yearID == "2010") %>%
  ggplot(aes(x=AB, y=R, label=teamID)) +
    geom_text()
```



EDA with the grammar of graphics

E.g., change the data.

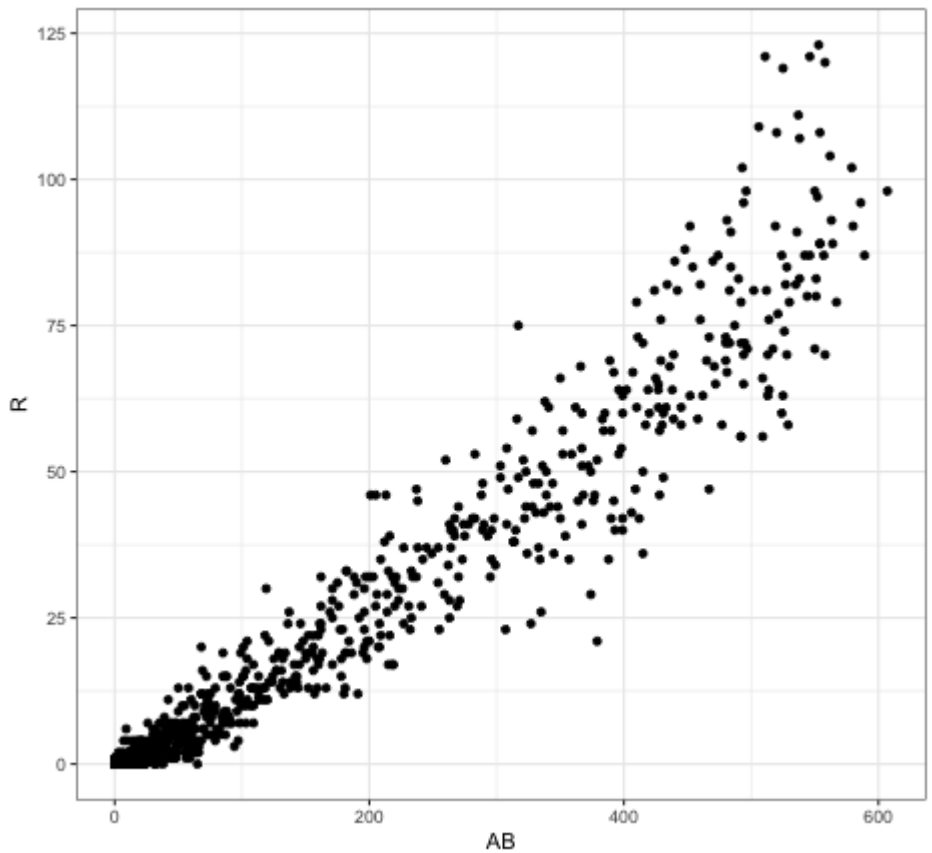
```
# scatter plot of at bats vs. runs for 1995

batting %>%

  filter(yearID == "1995") %>%

  ggplot(aes(x=AB, y=R)) +

    geom_point()
```



EDA with the grammar of graphics

E.g., change the aesthetic.

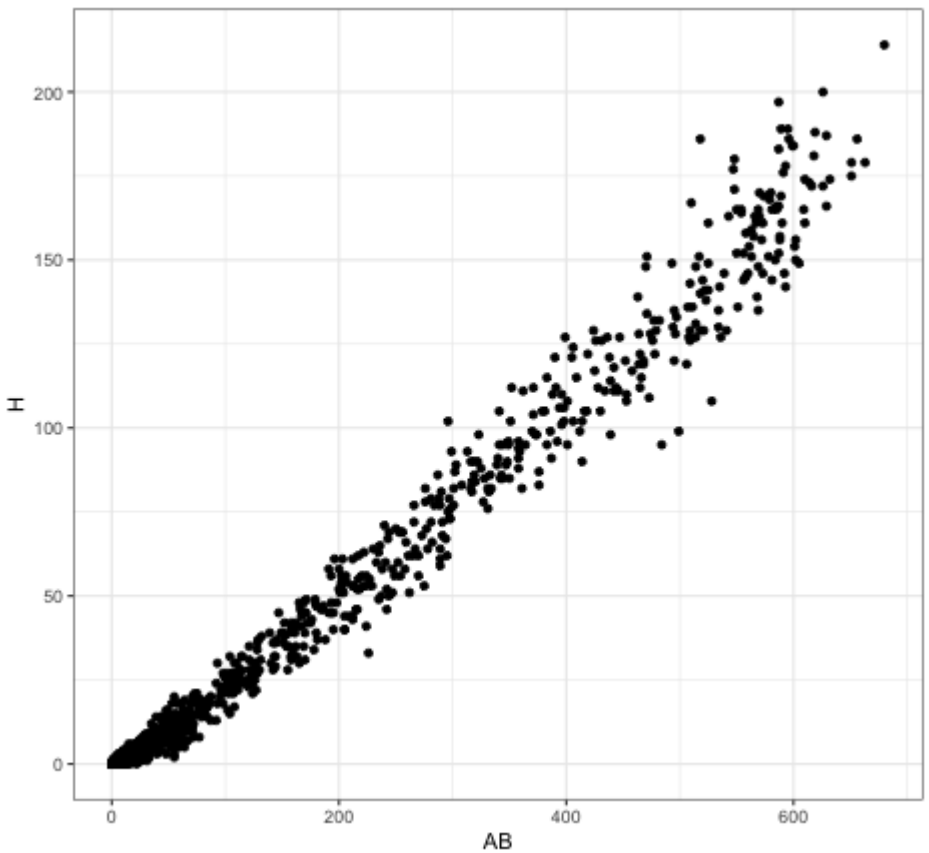
```
# scatter plot of at bats vs. hits for 2010

batting %>%

  filter(yearID == "2010") %>%

  ggplot(aes(x=AB, y=H)) +

    geom_point()
```



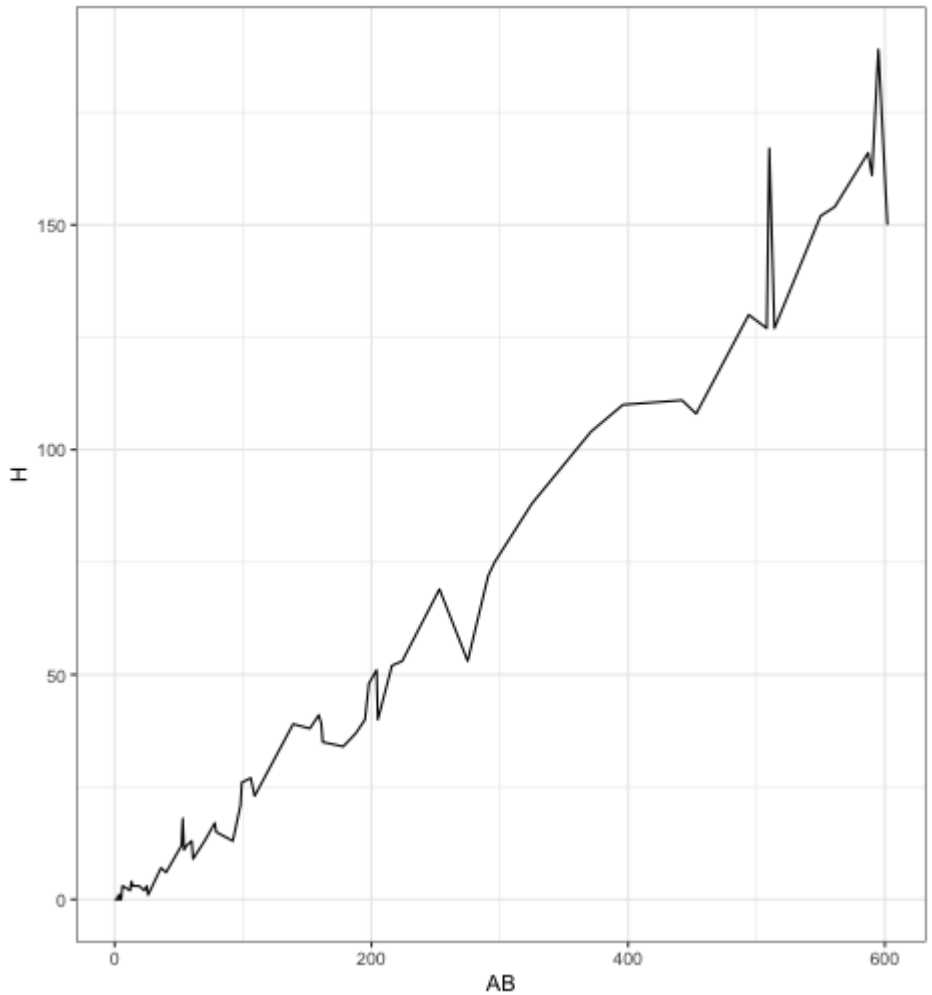
EDA with the grammar of graphics

Let's make a line plot

What do we change? (data, aesthetic or geometry?)

EDA with the grammar of graphics

```
batting %>%
  filter(yearID == "2010") %>%
  sample_n(100) %>%
  ggplot(aes(x=AB, y=H)) +
    geom_line()
```



EDA with the grammar of graphics

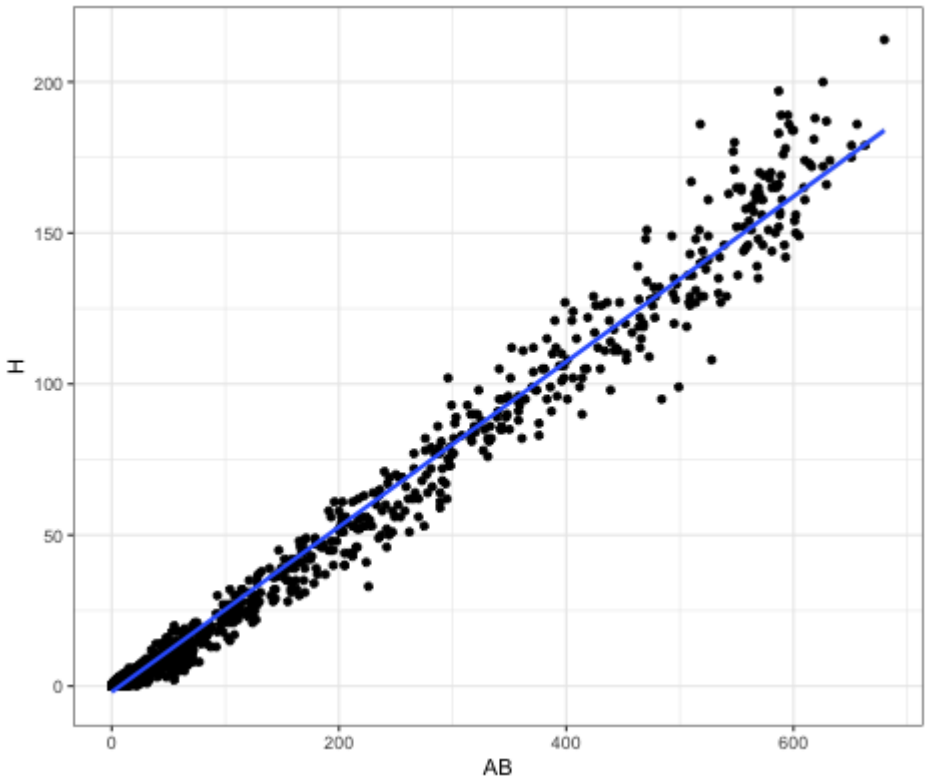
Let's add a regression line

What do we add? (data, aesthetic or geometry?)

EDA with the grammar of graphics

What can we see about central trend, variation and skew with this plot?

```
batting %>%
  filter(yearID == "2010") %>%
  ggplot(aes(x=AB, y=H)) +
    geom_point() +
    geom_smooth(method=lm)
```

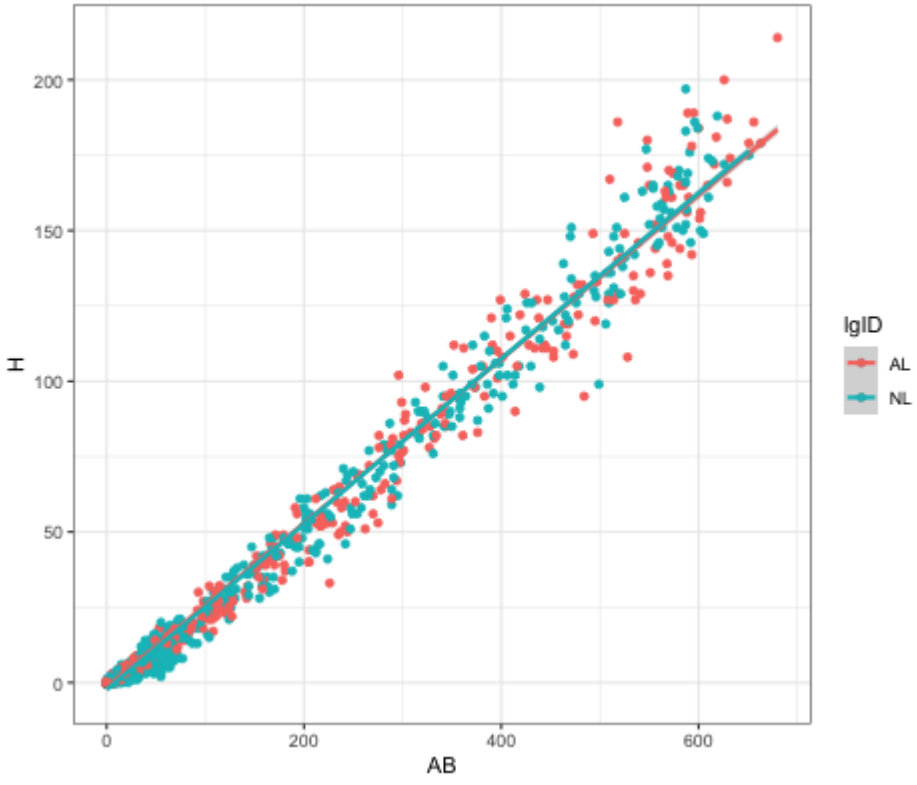


EDA with the grammar of graphics

Using other aesthetics we can incorporate information from other variables.

Color: color by categorical variable

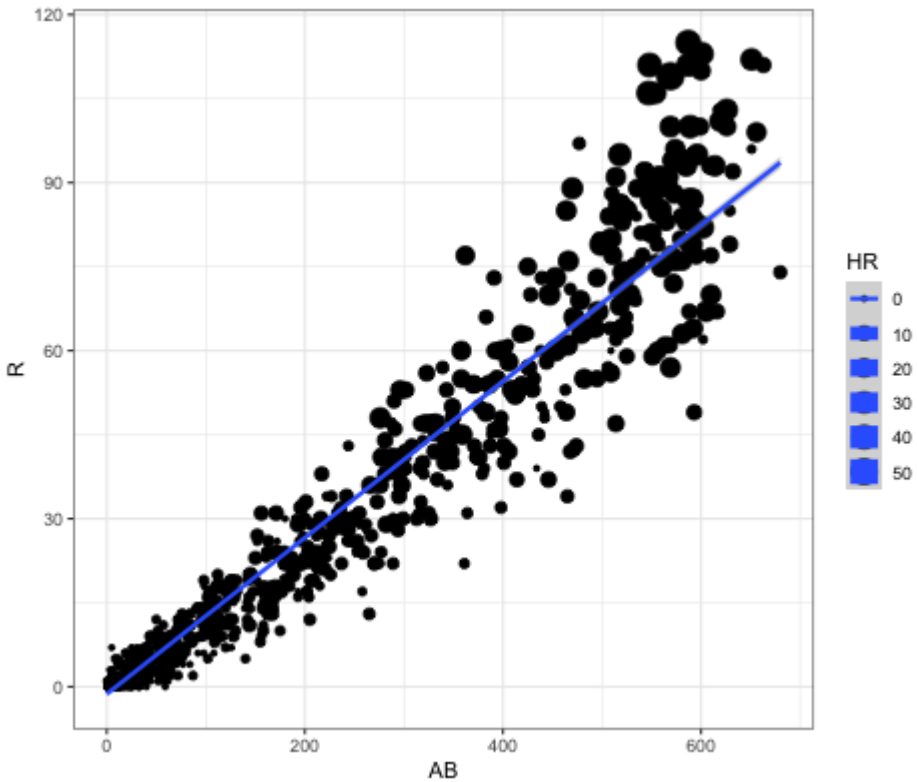
```
batting %>%
  filter(yearID == "2010") %>%
  ggplot(aes(x=AB, y=H, color=lgID)) +
    geom_point() +
    geom_smooth(method=lm)
```



EDA with the grammar of graphics

Size: size by (continuous) numeric variable

```
batting %>%
  filter(yearID == "2010") %>%
  ggplot(aes(x=AB, y=R, size=HR)) +
    geom_point() +
    geom_smooth(method=lm)
```



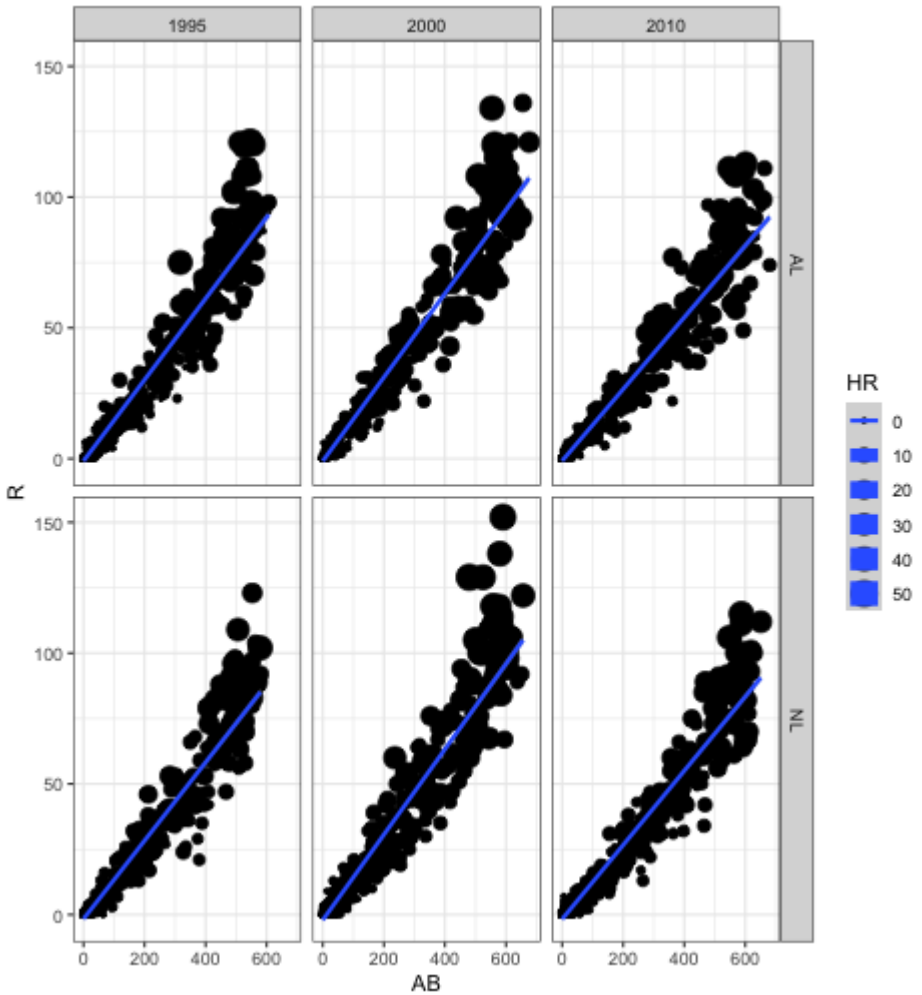
EDA with the grammar of graphics

Faceting

The last major component of exploratory analysis called `faceting` in visualization, corresponds to `conditioning` in statistical modeling, we've seen it as the motivation of `grouping` when wrangling data.

EDA with the grammar of graphics

```
batting %>%
  filter(yearID %in% c("1995", "2000", "2010
  ggplot(aes(x=AB, y=R, size=HR)) +
    facet_grid(lgID~yearID) +
    geom_point() +
    geom_smooth(method=lm)
```



Exploratory Data Analysis: Summary Statistics

Let's continue our discussion of Exploratory Data Analysis.

In the previous section we saw ways of visualizing attributes (variables) using plots to start understanding properties of how data is distributed.

In this section, we start discussing statistical summaries of data to quantify properties that we observed using visual summaries and representations.

Exploratory Data Analysis: Summary Statistics

Remember that one purpose of EDA is to spot problems in data (as part of data wrangling) and understand variable properties like:

- central trends (mean)
- spread (variance)
- skew
- suggest possible modeling strategies (e.g., probability distributions)

Exploratory Data Analysis: Summary Statistics

One last note on EDA.

John W. Tukey was an exceptional scientist/mathematician, who had profound impact on statistics and Computer Science.

A lot of what we cover in EDA is based on his groundbreaking work.

https://www.stat.berkeley.edu/~brill/Papers/life.pdf.

Exploratory Data Analysis: Summary Statistics

Range

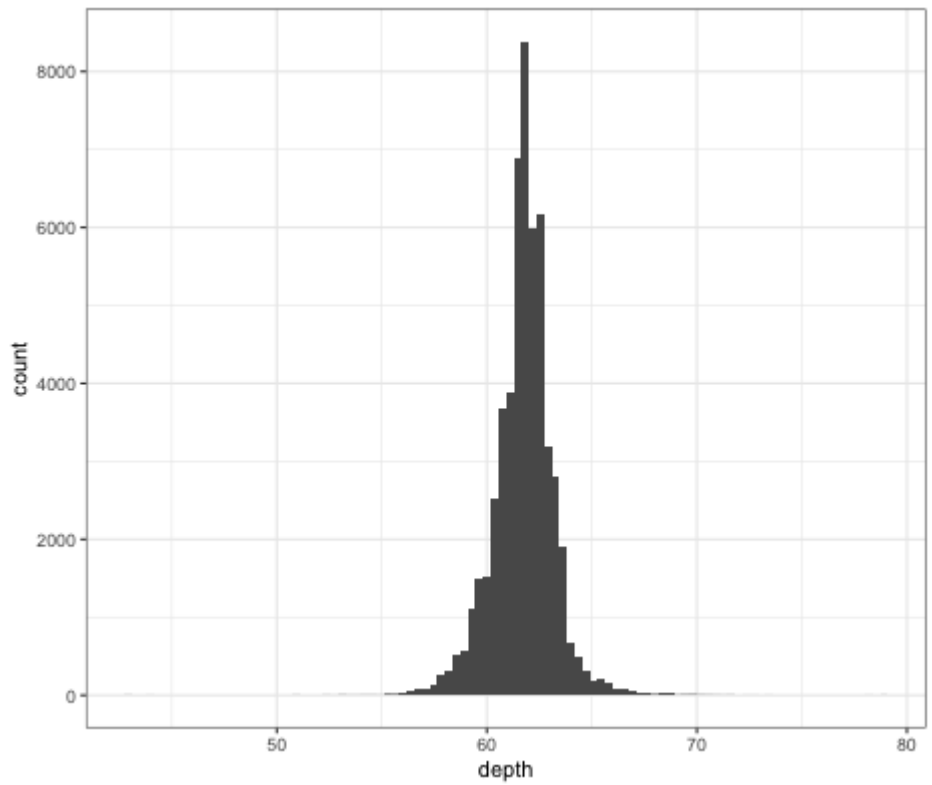
Part of our goal is to understand how variables are distributed in a given dataset.

Note, again, that we are not using *distributed* in a formal mathematical (or probabilistic) sense.

All statements we are making here are based on data at hand, so we could refer to this as the *empirical distribution* of data.

Exploratory Data Analysis: Summary Statistics

Let's use a dataset on diamond characteristics as an example.



Exploratory Data Analysis: Summary Statistics

Notation

We assume that we have data across n entitites (or observational units) for p attributes.

In this dataset $n = 53940$ and $p = 10$.

However, let's consider a single attribute, and denote the data for that attribute (or variable) as $x_1, x_2, \ldots, x_n$.

Exploratory Data Analysis: Summary Statistics

Since we want to understand how data is distributed across a *range*, we should first define the range.

```
diamonds %>%
  summarize(min_depth = min(depth), max_depth = max(depth))
```

```
## # A tibble: 1 x 2
##   min_depth max_depth
##       <dbl>     <dbl>
## 1        43        79
```

Exploratory Data Analysis: Summary Statistics

We use notation $x_{(1)}$ and $x_{(n)}$ to denote the minimum and maximum statistics.

In general, we use notation $x_{(q)}$ for the rank statistics, e.g., the qth largest value in the data.

Exploratory Data Analysis: Summary Statistics

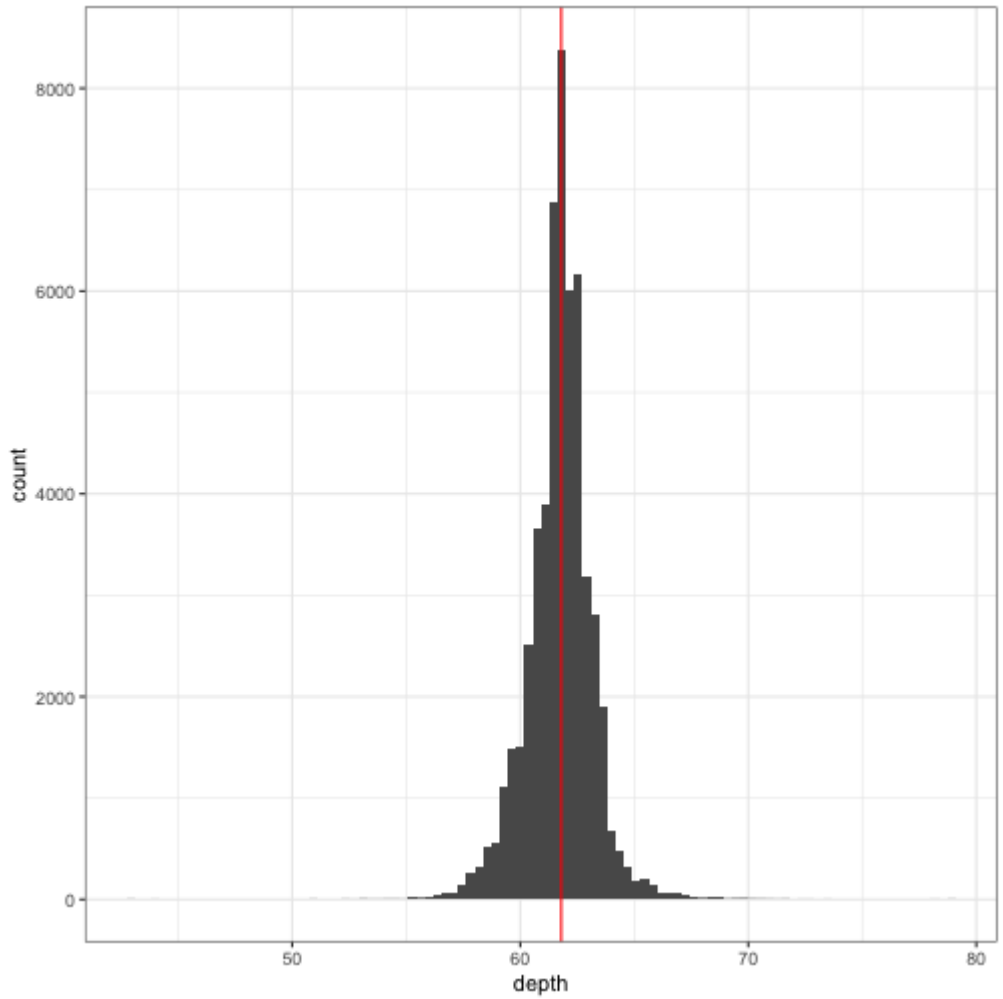
Central Tendency

Now that we know the range over which data is distributed, we can figure out a first summary of data is distributed across this range.

Let's start with the *center* of the data: the *median* is a statistic defined such that half of the data has a smaller value.

We can use notation $x_{(n/2)}$ (a rank statistic) to represent the median.

Exploratory Data Analysis: Summary Statistics



Exploratory Data Analysis: Summary Statistics

Derivation of the mean as central tendency statistic

Best known statistic for central tendency is the *mean*, or average of the data: $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$. It turns out that in this case, we can be a bit more formal about "center" means in this case.

Let's say that the *center* of a dataset is a point in the range of the data that is *close* to the data.

To say that something is *close* we need a measure of *distance*.

Exploratory Data Analysis: Summary Statistics

So for two points x_1 and x_2 what should we use for distance?

The distance between data point x_1 and x_2 is $(x_1 - x_2)^2$.

Exploratory Data Analysis: Summary Statistics

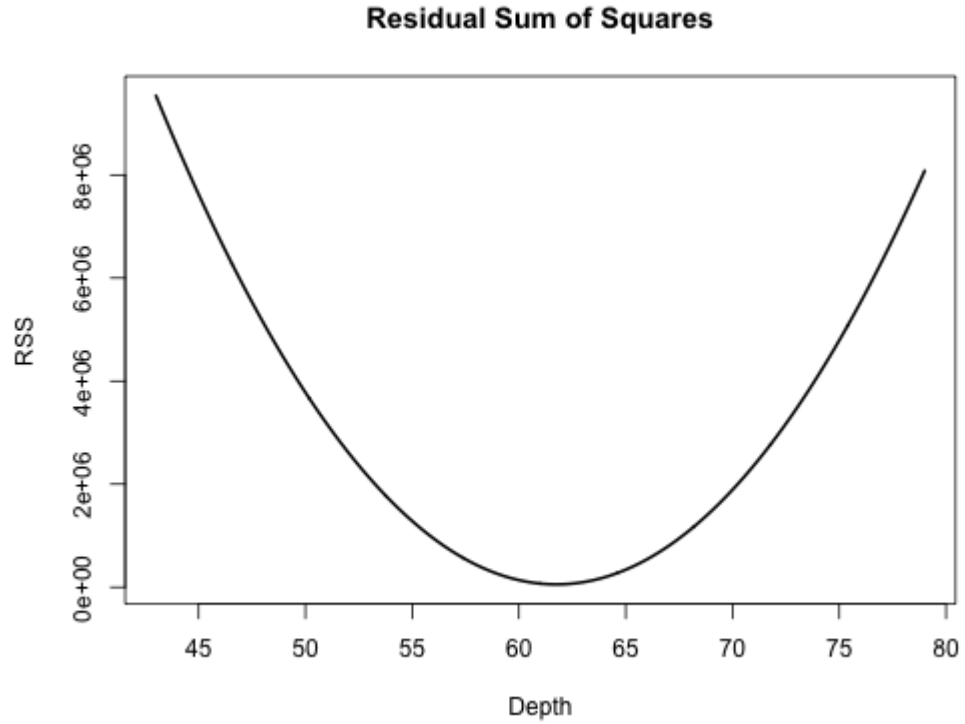
So, to define the *center*, let's build a criterion based on this distance by adding this distance across all points in our dataset:

$$RSS(\mu) = \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Here RSS means *residual sum of squares*, and we μ to stand for candidate values of *center*.

Exploratory Data Analysis: Summary Statistics

We can plot RSS for different values of μ:



Exploratory Data Analysis: Summary Statistics

Now, what should our "center" estimate be?

We want a value that is *close* to the data based on RSS!

So we need to find the value in the range that minimizes RSS.

Exploratory Data Analysis: Summary Statistics

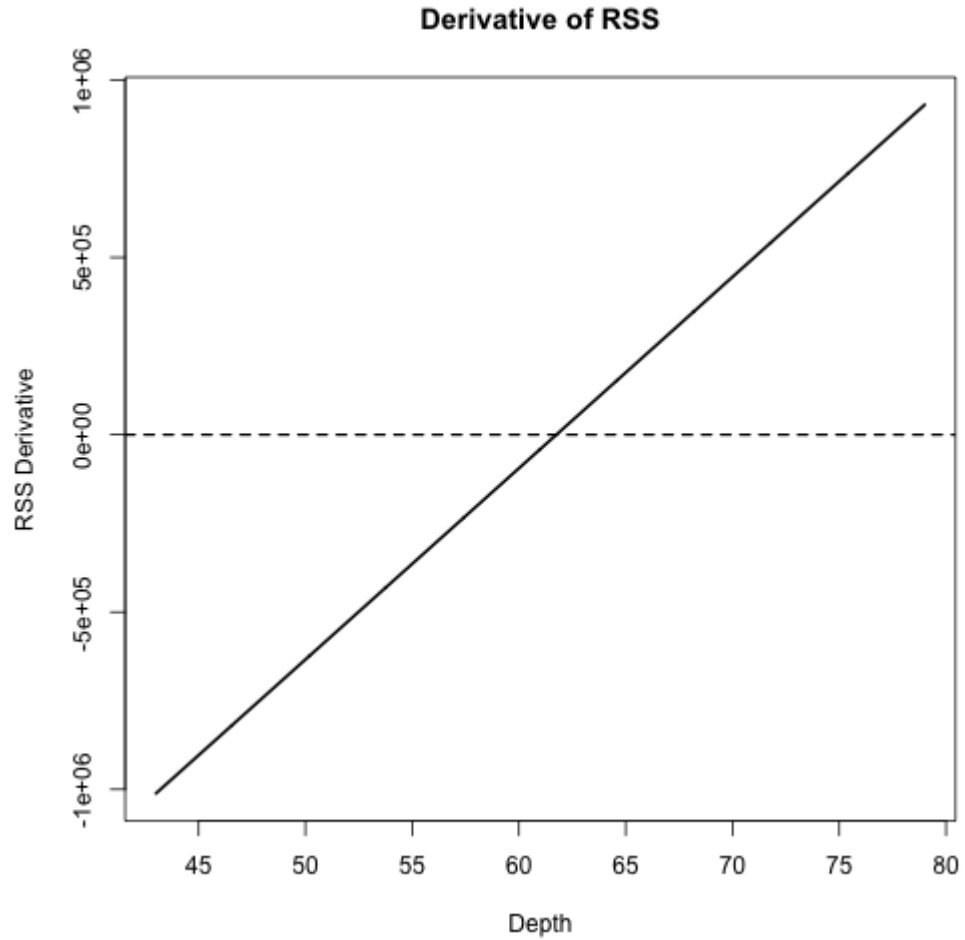
From calculus, we know that a necessary condition for the minimizer $\hat{\mu}$ of RSS is that the derivative of RSS is zero at that point.

So, the strategy to minimize RSS is to compute its derivative, and find the value of μ where it equals zero.

Exploratory Data Analysis: Summary Statistics

$$\begin{aligned}
\frac{\partial}{\partial \mu} \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 &= \frac{1}{2} \sum_{i=1}^{n} \frac{\partial}{\partial \mu} (x_i - \mu)^2 \text{ (sum rule)} \\
&= \sum_{i=1}^{n} \mu - \sum_{i=1}^{n} x_i \\
&= n\mu - \sum_{i=1}^{n} x_i
\end{aligned}$$

Exploratory Data Analysis: Summary Statistics



Exploratory Data Analysis: Summary Statistics

Next, we set that equal to zero and find the value of μ that solves that equation:

$$\begin{aligned} \frac{\partial}{\partial \mu} &= 0 \Rightarrow \\ n\mu &= \sum_{i=1}^{n} x_i \Rightarrow \\ \mu &= \frac{1}{n} \sum_{i=1}^{n} x_i \end{aligned}$$

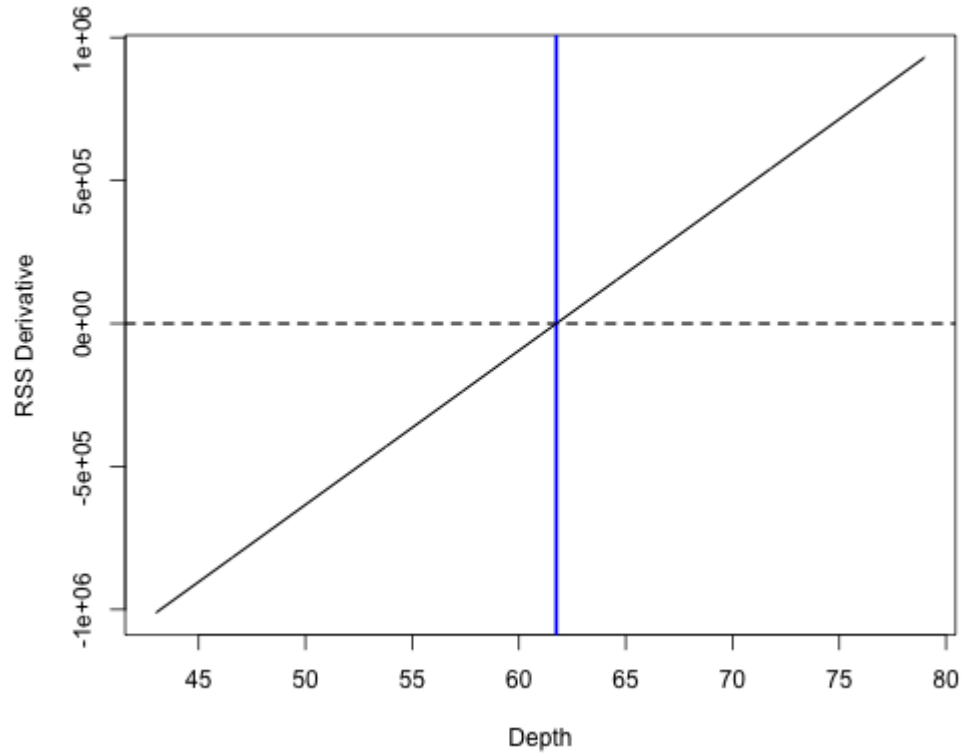
Exploratory Data Analysis: Summary Statistics

The fact you should remember:

The mean is the value that minimizes RSS for a vector of attribute values

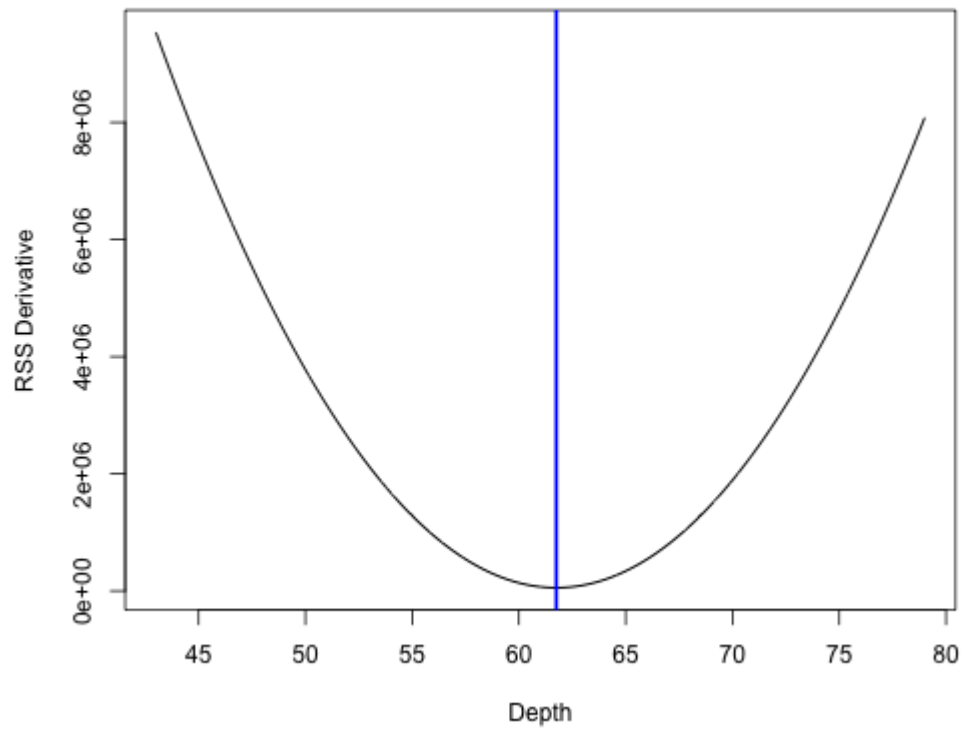
Exploratory Data Analysis: Summary Statistics

It equals the value where the derivative of RSS is 0:



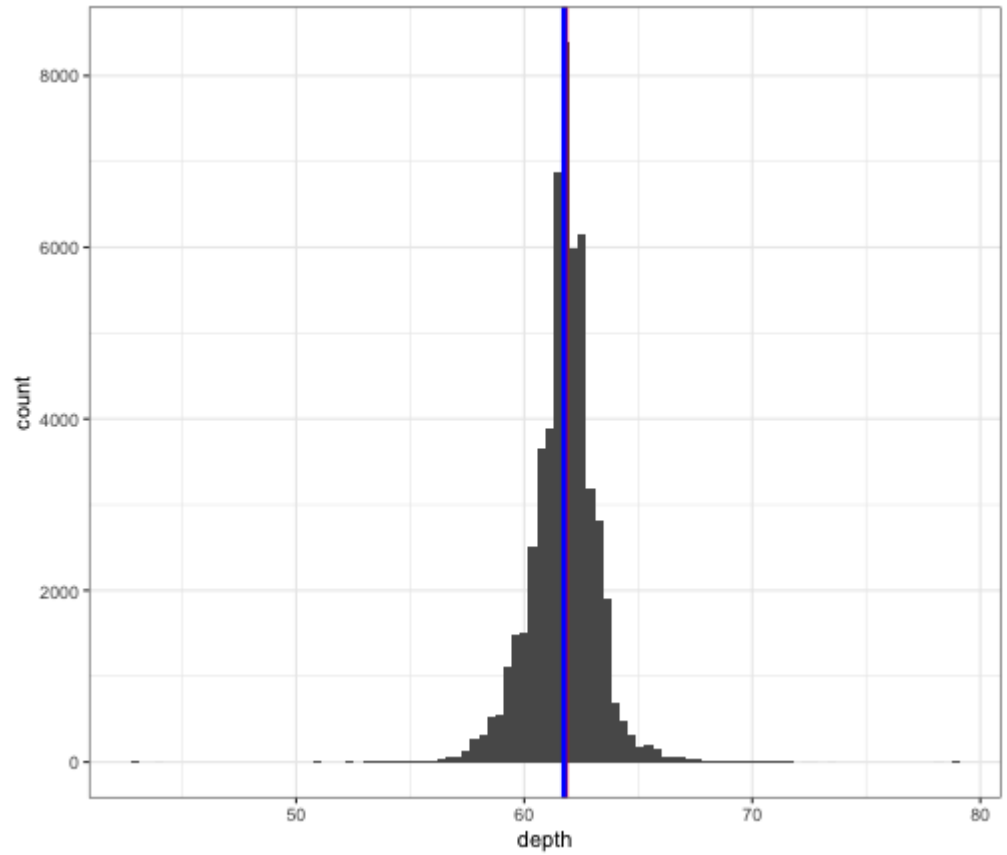
Exploratory Data Analysis: Summary Statistics

It is the value that minimizes RSS:



Exploratory Data Analysis: Summary Statistics

And it serves as an estimate of central tendency of the dataset:



Exploratory Data Analysis: Summary Statistics

Note that in this dataset the mean and median are not exactly equal, but are very close:

```
diamonds %>%

  summarize(mean_depth = mean(depth), median_depth = median(depth))
```

```
## # A tibble: 1 x 2

##   mean_depth median_depth

##        <dbl>        <dbl>

## 1       61.7         61.8
```

Exploratory Data Analysis: Summary Statistics

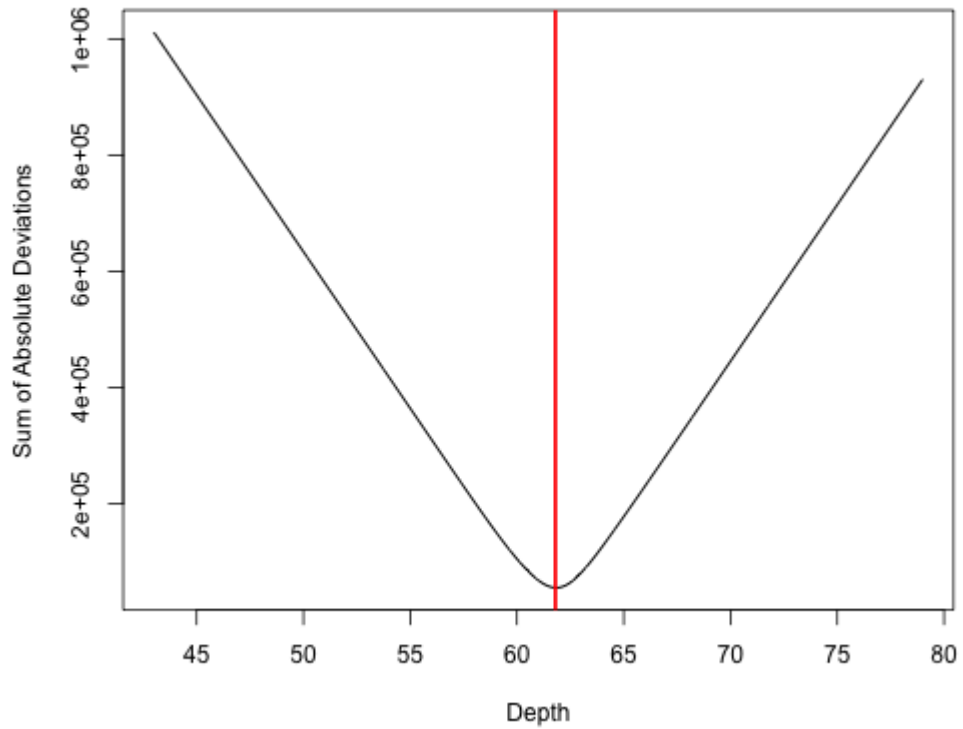
There is a similar argument to define the median as a measure of *center*.

In this case, instead of using RSS we use a different criterion: the sum of absolute deviations

$$SAD(m) = \sum_{i=1}^{n} |x_i - m|.$$

The median is the minimizer of this criterion.

Exploratory Data Analysis: Summary Statistics



Exploratory Data Analysis: Summary Statistics

Spread

Now that we have a measure of center, we can now discuss how data is *spread* around that center.

Exploratory Data Analysis: Summary Statistics

Variance

For the mean, we have a convenient way of describing this: the average distance (using squared difference) from the mean. We call this the *variance* of the data:

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

Exploratory Data Analysis: Summary Statistics

You will also see it with a slightly different constant in the front for technical reasons that we may discuss later on:

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

Exploratory Data Analysis: Summary Statistics

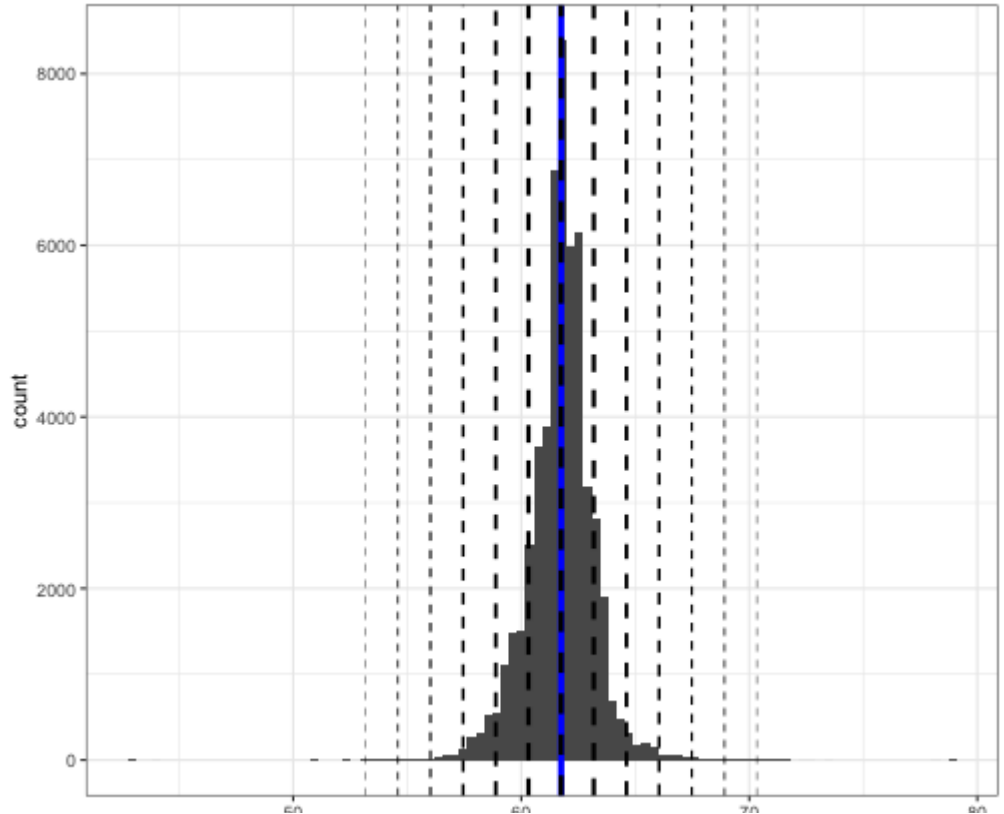
Variance is a commonly used statistic for spread but it has the disadvantage that its units are not easy to conceptualize (e.g., squared diamond depth).

A spread statistic that is in the same units as the data is the *standard deviation*, which is just the squared root of variance:

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

Exploratory Data Analysis: Summary Statistics

We can also use *standard deviations* as an interpretable unit of how far a given data point is from the mean:



Exploratory Data Analysis: Summary Statistics

As a rough guide, we can use "standard deviations away from the mean" as a measure of spread as follows:

SDs	proportion	Interpretation
1	0.68	68% of the data is within $\pm$ 1 sds
2	0.95	95% of the data is within $\pm$ 2 sds
3	0.9973	99.73% of the data is within $\pm$ 3 sds
4	0.999937	99.9937% of the data is within $\pm$ 4 sds
5	0.9999994	99.999943% of the data is within $\pm$ 5 sds
6	1	99.9999998% of the data is within $\pm$ 6 sds

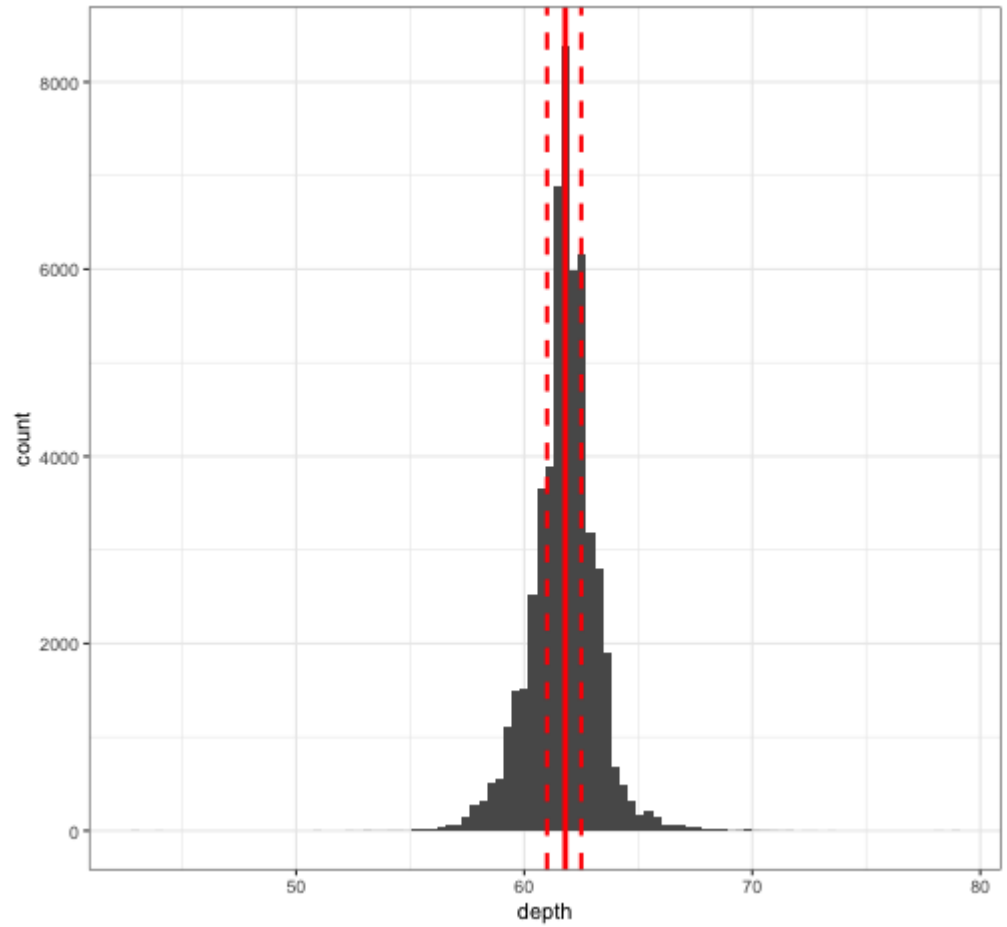
Exploratory Data Analysis: Summary Statistics

Spread estimates using rank statistics

Just like we saw how the median is a rank statistic used to describe central tendency, we can also use rank statistics to describe spread.

For this we use two more rank statistics: the first and third *quartiles*, $x_{(n/4)}$ and $x_{(3n/4)}$ respectively.

Exploratory Data Analysis: Summary Statistics



Exploratory Data Analysis: Summary Statistics

Note, the five order statistics we have seen so far: minimum, maximum, median and first and third quartiles are so frequently used that this is exactly what `R` uses by default as a `summary` of a numeric vector of data (along with the mean):

```
summary(diamonds$depth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   43.00   61.00   61.80   61.75   62.50   79.00
```

Exploratory Data Analysis: Summary Statistics

This five-number summary are also all of the statistics used to construct a boxplot to summarize data distribution.

In particular, the *inter-quartile range*, which is defined as the difference between the third and first quartile: $\mathrm{IQR}(x) = x_{(3n/4)} - x_{(1n/4)}$ gives a measure of spread.

Exploratory Data Analysis: Summary Statistics

The interpretation here is that half the data is within the IQR around the median.

```
diamonds %>%
  summarize(sd_depth = sd(depth), iqr_depth = IQR(depth))
```

```
## # A tibble: 1 x 2
##   sd_depth iqr_depth
##      <dbl>     <dbl>
## 1     1.43       1.5
```

Exploratory Data Analysis: Summary Statistics

Outliers

We can use estimates of spread to identify outlier values in a dataset. Given an estimate of spread based on the techniques we've just seen, we can identify values that are *unusually* far away from the center of the distribution.

Exploratory Data Analysis: Summary Statistics

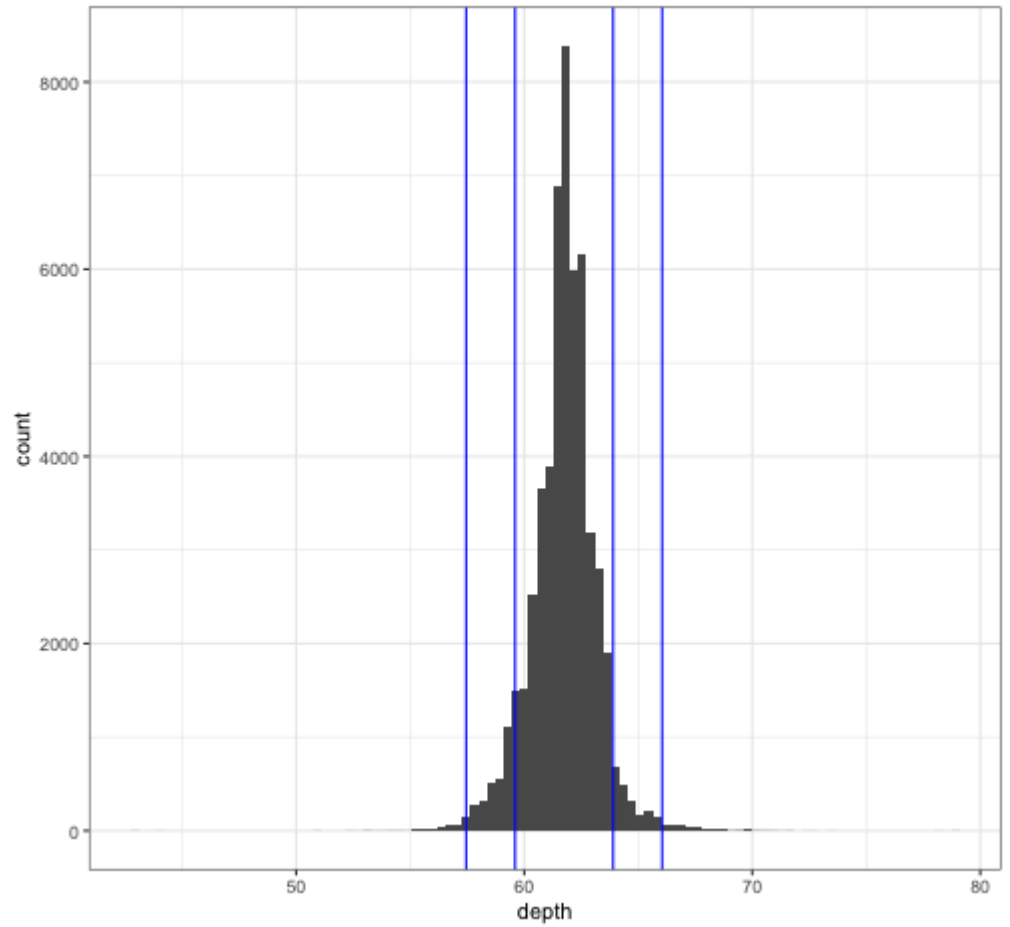
One often cited rule of thumb is based on using standard deviation estimates. We can identify outliers as the set

$$\mathbf{outliers}_{\mathrm{sd}}(x) = \{x_j \,|\, |x_j| > \overline{x} + k \times \mathrm{sd}(x)\}$$

where $\overline{x}$ is the sample mean of the data and $\mathrm{sd}(x)$ it's standard deviation.

Multiplier k determines if we are identifying (in Tukey's nomenclature) *outliers* or points that are *far out*.

Exploratory Data Analysis: Summary Statistics



Exploratory Data Analysis: Summary Statistics

While this method works relatively well in practice, it presents a fundamental problem.

Severe outliers can significantly affect spread estimates based on standard deviation.

Specifically, spread estimates will be inflated in the presence of severe outliers.

Exploratory Data Analysis: Summary Statistics

To circumvent this problem, we use rank-based estimates of spread to identify outliers as:

$$ \mathrm{outliers*{IQR}}(x) = {x_j \, / \, x_j < x{(1/4)} - k \times \mathrm{IQR}(x) \; \mathrm{ or } \; xj > x{(3/4)} + k \times \mathrm{IQR}(x)} $$

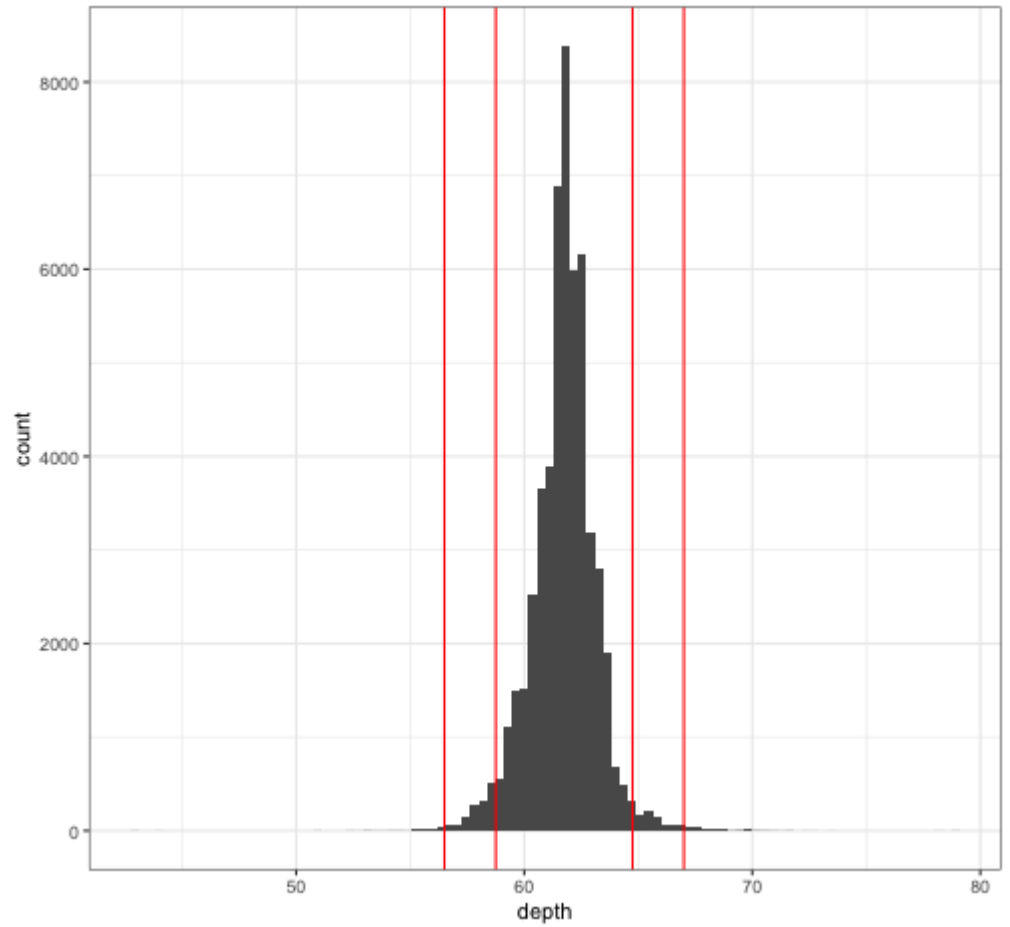
This is usually referred to as the *Tukey outlier rule*, with multiplier k serving the same role as before.

Exploratory Data Analysis: Summary Statistics

We use the IQR here because it is less susceptible to be inflated by severe outliers in the dataset.

It also works better for skewed data than the method based on standard deviation.

Exploratory Data Analysis: Summary Statistics



Exploratory Data Analysis: Summary Statistics

Skew

The five-number summary can be used to understand if data is skewed.

Consider the differences between the first and third quartiles to the median.

Exploratory Data Analysis: Summary Statistics

```
diamonds %>%
  summarize(med_depth = median(depth),
            q1_depth = quantile(depth, 1/4),
            q3_depth = quantile(depth, 3/4)) %>%
  mutate(d1_depth = med_depth - q1_depth,
         d2_depth = q3_depth - med_depth) %>%
  select(d1_depth, d2_depth)
```

```
## # A tibble: 1 x 2
##   d1_depth d2_depth
##      <dbl>    <dbl>
## 1    0.800      0.7
```

Exploratory Data Analysis: Summary Statistics

If one of these differences is larger than the other, then that indicates that this dataset might be skewed.

The range of data on one side of the median is longer (or shorter) than the range of data on the other side of the median.

Exploratory Data Analysis: Summary Statistics

Covariance and correlation

The scatter plot is a visual way of observing relationships between pairs of variables.

Like descriptions of distributions of single variables, we would like to construct statistics that summarize the relationship between two variables quantitatively.

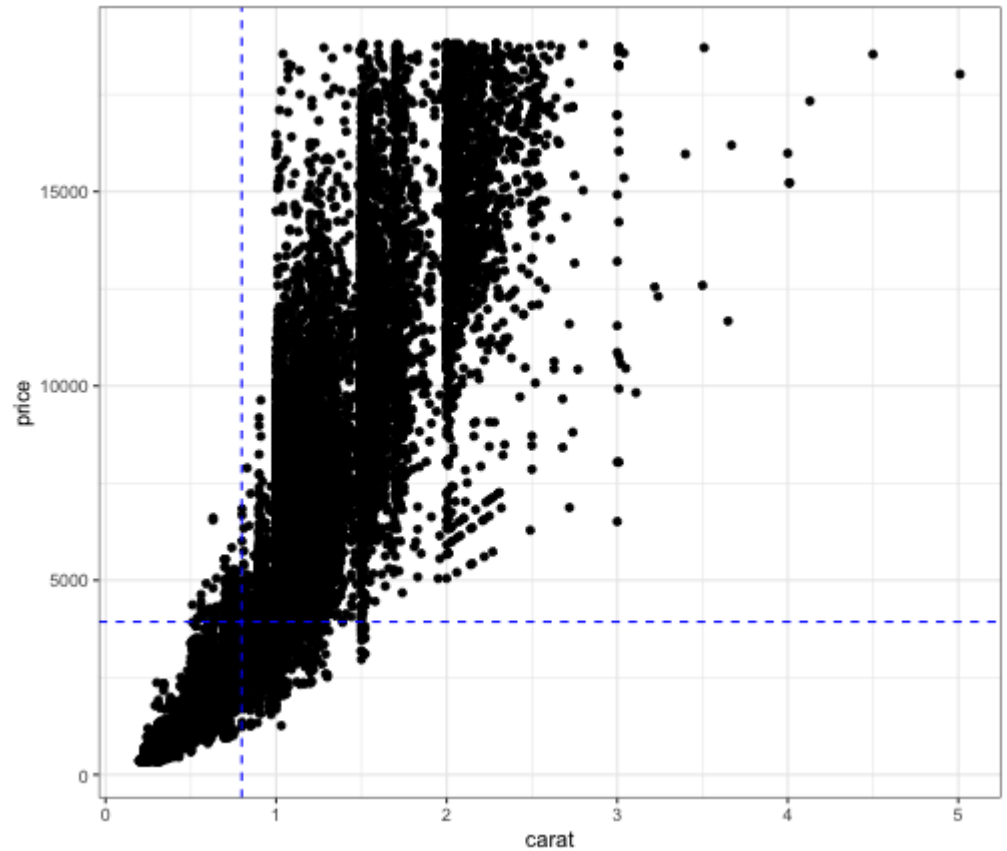
To do this we will extend our notion of *spread* (or variation of data around the mean) to the notion of *co-variation*: do pairs of variables vary around the mean in the same way.

Exploratory Data Analysis: Summary Statistics

Consider now data for two variables over the same n entities: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

For example, for each diamond, we have `carat` and `price` as two variables.

Exploratory Data Analysis: Summary Statistics



Exploratory Data Analysis: Summary Statistics

We want to capture the relationship: does x_i vary in the same direction and scale away from its mean as y_i?

This leads to *covariance*

$$cov(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Exploratory Data Analysis: Summary Statistics

Just like variance, we have an issue with units and interpretation for covariance, so we introduce *correlation* (formally, Pearson's correlation coefficient) to summarize this relationship in a *unit-less* way:

$$cor(x, y) = \frac{cov(x, y)}{sd(x)sd(y)}$$

Exploratory Data Analysis: Summary Statistics

As before, we can also use rank statistics to define a measure of how two variables are associated.

One of these, *Spearman correlation* is commonly used.

It is defined as the Pearson correlation coefficient of the ranks (rather than actual values) of pairs of variables.

Exploratory Data Analysis: Summary Statistics

Summary

EDA: visual and computational methods to describe the distribution of data attributes over a range of values

Grammar of graphics as effective tool for visual EDA

Statistical summaries that directly establish properties of data distribution