Boxplot: Side-By-Side Boxplots

Learning Objective: Compare and contrast distributions (of quantitative data) from two or more groups, and produce a brief summary, interpreting your findings in context.

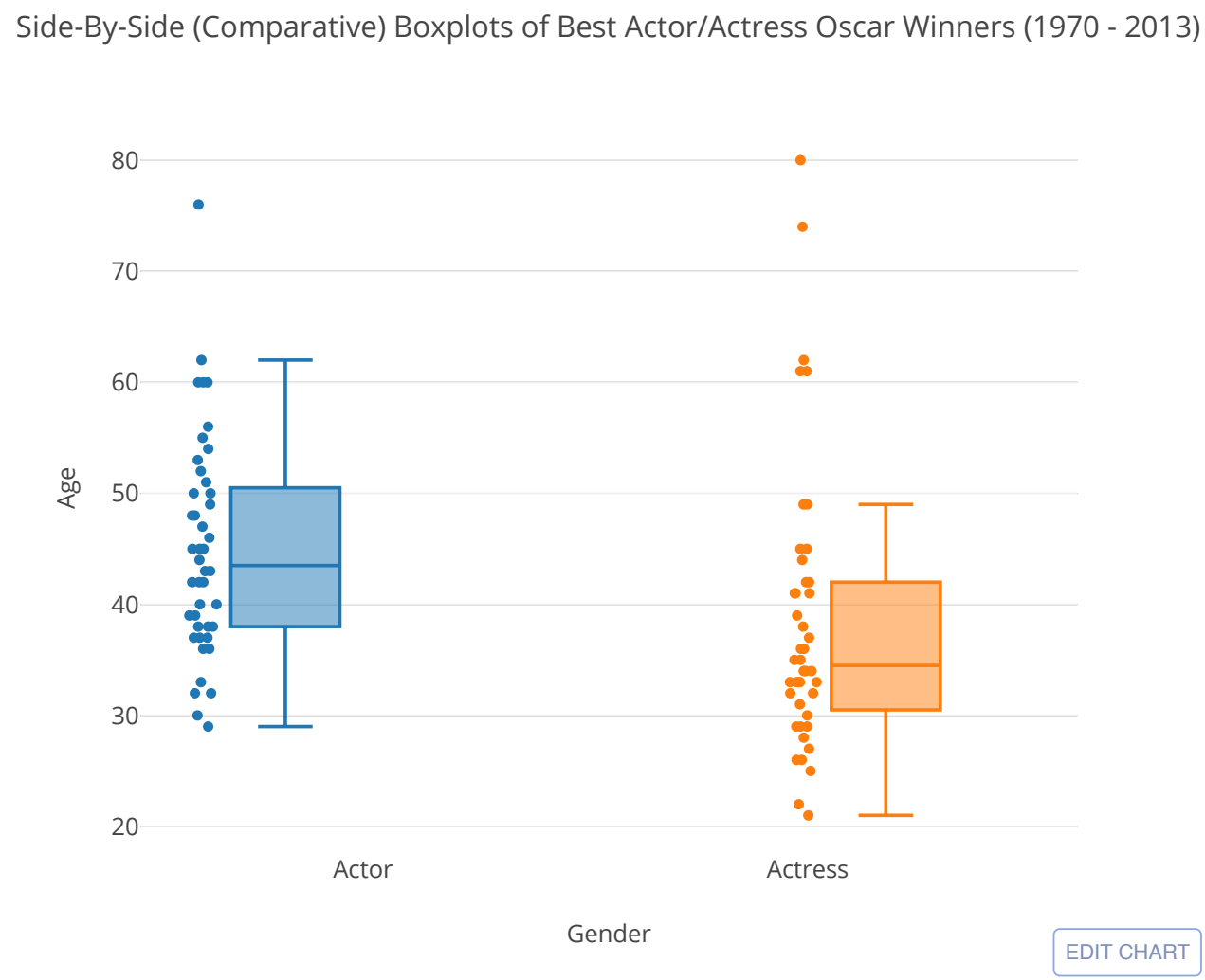
Side-By-Side (Comparative) Boxplots

As we learned in the beginning of this module, the distribution of a quantitative variable is best represented graphically by a histogram. Boxplots are most useful when presented side-by-side for comparing and contrasting distributions from two or more groups.

Example: Best Actor/Actress Oscar Winners

So far we have examined the age distributions of Oscar winners for males and females separately.

It will be interesting to *compare* the age distributions of actors and actresses who won best acting Oscars. To do that we will look at side-by-side boxplots of the age distributions by gender.



Recall also that we found the five-number summary and means for both distributions. Here are the results for the Best Actor and Best Actress datasets:

- Actors: min = 31, Q1 = 38, M = 43.5, Q3 = 50.5, Max = 76
- Actresses: min = 21, Q1 = 30.5, M = 34.5, Q3 = 42, Max = 80

Based on the graph and numerical measures, we can make the following comparison between the two distributions:

Center: The graph reveals that the age distribution of the males is higher than the females' age distribution. This is supported by the numerical measures. The median age for females (34.5) is lower than for the males (43.5). Actually, it should be noted that even the third quartile of the females' distribution (42) is lower than the median age for males. We therefore conclude that in general, actresses win the Best Actress Oscar at a younger age than actors do.

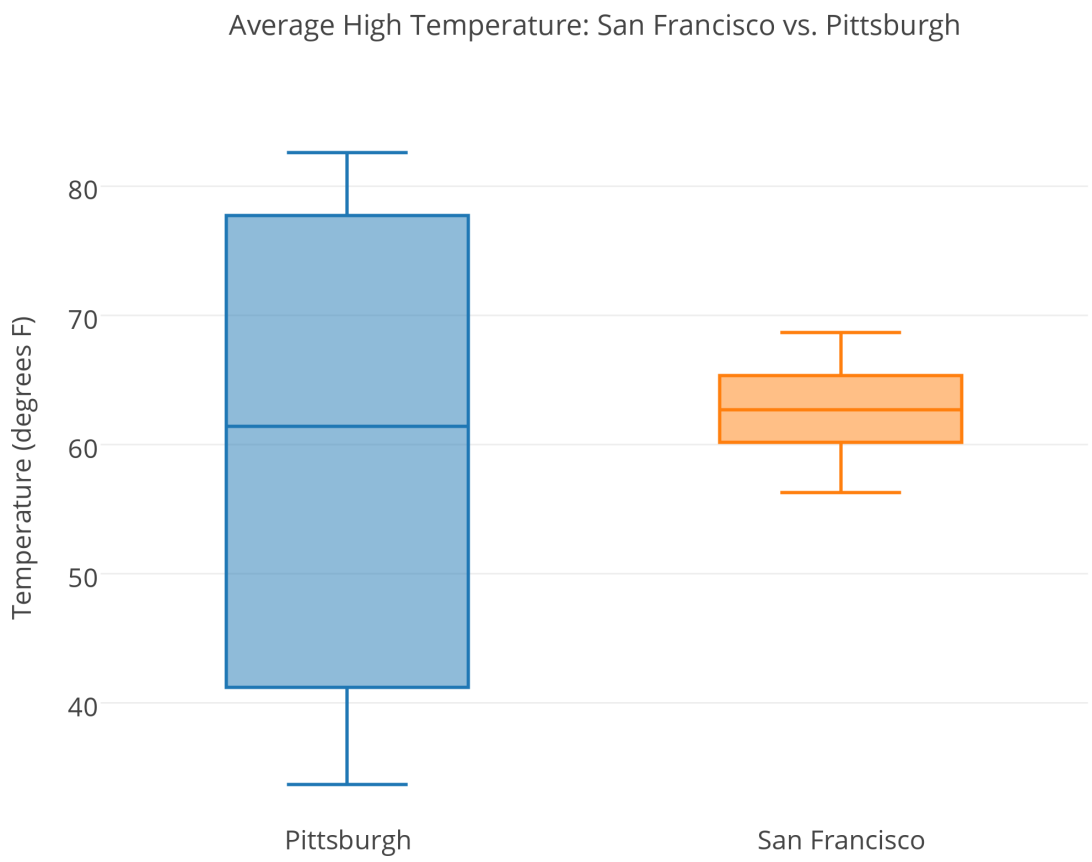
Spread: Judging by the range of the data, there is much more variability in the females' distribution (range = 59) than there is in the males' distribution (range = 47). On the other hand, if we look at the IQR, which measures the variability only among the middle 50% of the distribution, we see slightly

more spread in the ages of males (IQR = 12.5) than females (IQR = 11.5). We conclude that among all the winners, the actors' ages are more alike than the actresses' ages. However, the middle 50% of the age distribution of actresses is more homogeneous than the actors' age distribution.

Outliers: We see that we have outliers in both distributions. There is only one high outlier in the actors' distribution (76, Henry Fonda, *On Golden Pond*), compared with five high outliers in the actresses' distribution.

Example: Temperature of Pittsburgh vs. San Francisco

In order to compare the average high temperatures of Pittsburgh to those in San Francisco we will look at the following side-by-side boxplots, and supplement the graph with the descriptive statistics of each of the two distributions.



Statistic	San Francisco	Pittsburgh
min	56.3	33.7
Q1	60.2	41.2

Median	62.7	61.4
Q3	65.35	77.75
Max	68.7	82.6

When looking at the graph, the similarities and differences between the two distributions are striking. Both distributions have roughly the same center (medians are 61.4 for Pitt, and 62.7 for San Francisco). However, the temperatures in Pittsburgh have a much larger variability than the temperatures in San Francisco (Range: 49 vs. 12. IQR: 36.5 vs. 5).

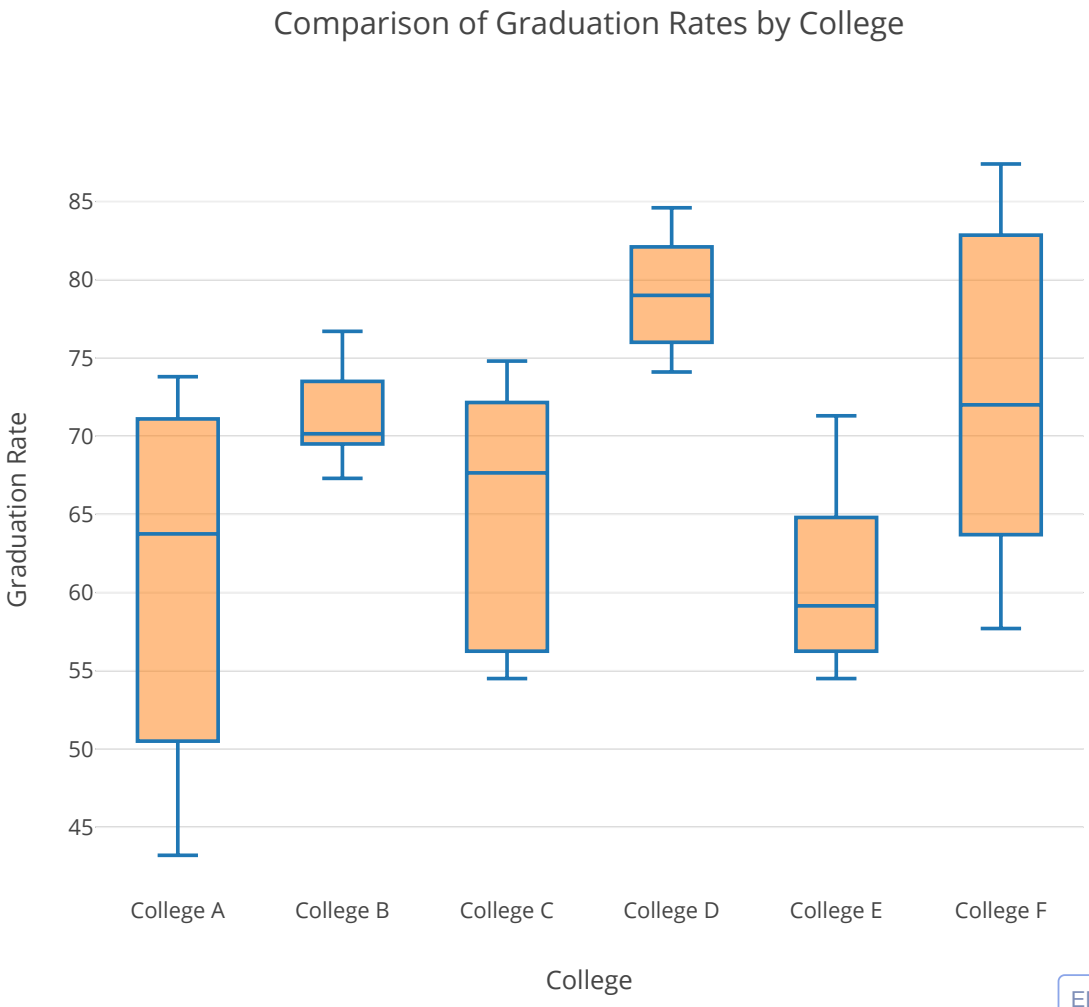
The practical interpretation of the results we got is that the weather in San Francisco is much more consistent than the weather in Pittsburgh, which varies a lot during the year. Also, because the temperatures in San Francisco vary so little during the year, knowing that the median temperature is around 63 is actually very informative. On the other hand, knowing that the median temperature in Pittsburgh is around 61 is practically useless, since temperatures vary so much during the year, and can get much warmer or much colder.

Note that this example provides more intuition about variability by interpreting small variability as consistency, and large variability as lack of consistency. Also, through this example we learned that the center of the distribution is more meaningful as a typical value for the distribution when there is little variability (or, as statisticians say, little "noise") around it. When there is large variability, the center loses its practical meaning as a typical value.

Scenario: Gradution Rate

The percentage of each entering Freshman class that graduated on time was recorded for each of six colleges at a major university over a period of several years. (Source: This data is distributed with the software package, Data Desk. (1993). Ithaca, NY: Data Description, Inc., and appears in http://lib.stat.cmu.edu/DASL/)

In order to compare the graduation rates among the different colleges, we created side-by-side boxplots (graduation rate by college), and supplemented the graph with numerical measures.



Summary Data	College A	College B	College C	College D	College E	College F
Min	43.20	67.30	54.50	74.10	54.50	57.50
Q1	50.95	69.55	56.58	76.65	56.88	65.05
Median	63.75	70.15	67.65	79.00	59.15	72.00
Q3	70.50	73.05	71.58	81.10	63.70	81.28
Max	73.80	76.70	74.80	84.60	71.30	87.40

Learn By Doing

1/1 point (graded)

Based on the boxplots and data, which of the six colleges has the best on-time graduation rate?

- College A
- College B
- College C
- College D ✔
- College E
- College F

Answer
Correct:
This college has the largest median graduation rates (79%), but it also has the smallest variation in graduation rates over the years (range = 10.5%, IQR = 4.45%). This means that even in years when college D has a relatively small graduation rate, it is not MUCH smaller than the median (min = 74.1%, Median = 79%), and is still higher than most graduation rates at the other colleges. hese data suggest that College D has the best on-time graduation rate.

Submit

Let's Summarize

- The five-number summary of a distribution consists of the median (M), the two quartiles (Q1, Q3) and the extremes (min, Max).
- The five-number summary provides a complete numerical description of a distribution. The median describes the center, and the extremes (which give the range) and the quartiles (which give the IQR) describe the spread.
- The boxplot graphically represents the distribution of a quantitative variable by visually displaying the five number summary and any observation that was classified as a suspected outlier using the 1.5 (IQR) criterion.
- Boxplots are most useful when presented side-by-side to compare and contrast distributions from two or more groups.