Name: Key

Problem Set 1-4: Boxplots, 5 number summary, and Outliers

1. The 5-number summary is usually listed as min, Q_1, median, Q_3, max. Each one of the 5 has at least one synonym. List one for each.

min - smallest data entry, 0th percentile
Q_1 - 25th percentile
median - middle, 50th percentile
Q_3 - 75th percentile
max - largest data entry, 100th percentile

2. Open the "Mean Hospital Stay" data set in Fathom.

a. "Analyze the Data"

skew right
mean: 7.994

min: 5.8
Q_1: 6.7
med: 7.1
Q_3: 8.2
max: 18.7

IQR: 8.2 - 6.7 = 1.5
1.5IQR: 2.25
upper cut: 8.2 + 2.25 = 10.45
lower cut: 6.7 - 2.25 = 4.45

upper outliers:
MA, NJ, SD, ND, PA, MT, MT

no lower outliers

b. Write a complete sentence describing what Q_1 means within the context of this problem.

In 1995, 25% of the states had a mean hospital stay of less than 6.7 days.

c. What percentage of the states have a mean stay between Q_1 and Q_3?

75% - 25%

50%

d. Which measure of center, mean or median, seems more appropriate with this data set and why?

Median, because the upper outliers affect the mean.

e. Can you make any geographic generalizations about the states that have comparatively long vs. short mean stays in hospitals?

Colder/northern states tend to have longer hospital stays

ex] height of the members of my immediate family (4 people)

3. Give an example of a data set where a box plot is not appropriate.

anything with a really small sample size
ex] batting averages for 1 baseball team (9 players) (starters only)

4. What shape of a box and whisker plot or dot plot will result in the median being appreciably different from the mean, even when there are no outliers?

Anything with skew → the more skew the bigger the difference.

5. : Open the Fathom data set called "Airplanes"

a. Create a box and whisker plot of the attribute "costph": cost of operating in dollars per hour. Without doing any calculations, predict which will be bigger, the mean or the median, just by looking at the graph. Write a sentence which explains your reasoning.

It looks like the mean will be greater than the median because most of the data is sitting b/w Q_2 and Q_3 indicating skew to the right.

b. "Analyze the data" for the attribute "costph": cost of operating in dollars per hour. Check to see if your conjecture in (a) was correct.

shape is skew right
5# summary
min: 1409
Q1: 2002
med: 2675
Q_3 = 5237
max = 6859
mean = 3518.4

IQR = 3235
1.5 · IQR = 4852.5

Upper Cut = 5237 + 4852.5
= 10,089.5

Lower Cut = 2002 - 4852.5
= -2850.5

There are no outliers.

6. Open the Fathom data set called "2010 Midsize car fuel economy".

a. Create a box and whisker plot of fuel economy. Based on the box and whisker plot alone, are there any outliers? If so, which car(s)?

yes there is an outlier
Toyota Prius 4 cyl

b. Test for outliers using the 1.5 x IQR and verify algebraically your results in part (a).

IQR = 31 - 24 = 7
1.5 · IQR = 7 · 1.5 = 10.5
Upper Cut = 31 + 10.5 = 41.5 → since 48 > 4.5, The Toyota Prius is an outlier
Lower Cut = 24 - 10.5 = 13.5 → since there are no data entries < 13.5 there are no outliers on the lower end