Boxplots and Outliers

Earl F Glynn

Kansas City R Users Group

Beginner's Workshop
4 Oct 2014

Gist: https://gist.github.com/EarlGlynn/a13b651289eff61a2201

Outline

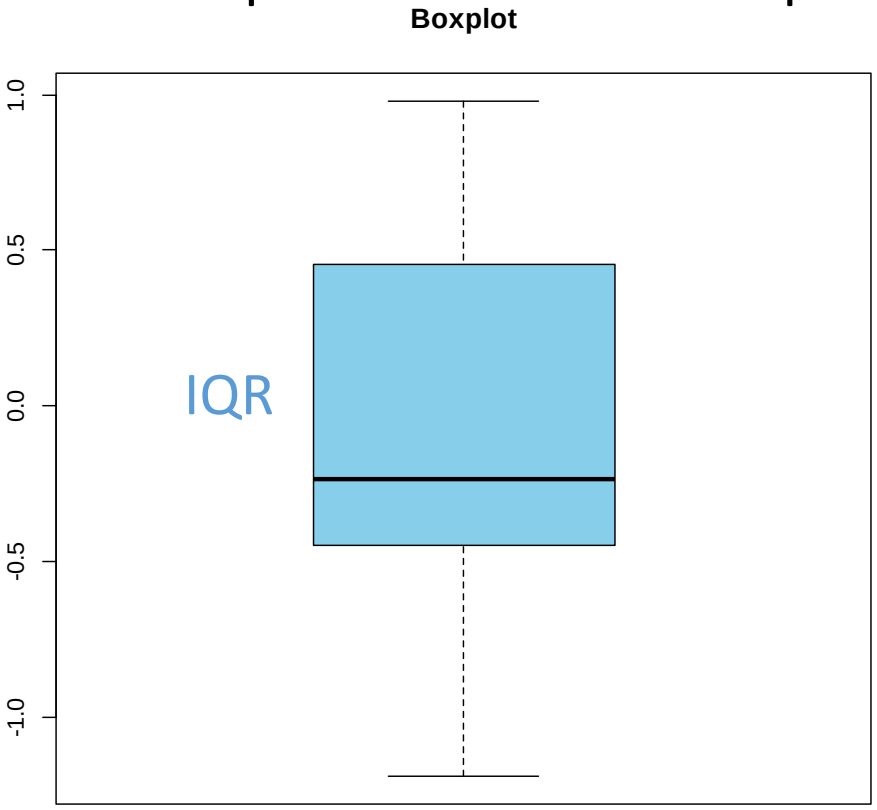
- Fivenum quartile summary
- Interquartile range (IQR)
- Boxplot: visual display of fivenum summary
- Outliers
- “Notched” boxplots
- Examples of using boxplots to identify outliers

“fivenum” Quartile Summary

```
> set.seed(19)
>
> x <- rnorm(11)
> x
 [1] -1.1894537  0.3885812 -0.3443333 -0.5478961  0.9806622 -0.2366460
 [7]  0.8097397 -0.7447795 -0.2597870 -0.1830838  0.5186300
> sort(x)
 [1] -1.1894537 -0.7447795 -0.5478961 -0.3443333 -0.2597870 -0.2366460
 [7] -0.1830838  0.3885812  0.5186300  0.8097397  0.9806622
>
> min(x)
[1] -1.189454
> mean(sort(x)[3:4])
[1] -0.4461147
> median(x)
[1] -0.236646
> mean(sort(x)[8:9])
[1] 0.4536056
> max(x)
[1] 0.9806622
>
> quantile(x, probs=seq(0,1,0.25))
        0%        25%        50%        75%       100%
-1.1894537 -0.4461147 -0.2366460  0.4536056  0.9806622
> fivenum(x)
[1] -1.1894537 -0.4461147 -0.2366460  0.4536056  0.9806622
```

IQR = Interquartile Range = Middle 50%

Boxplot: Visual Display of "fivenum" summary

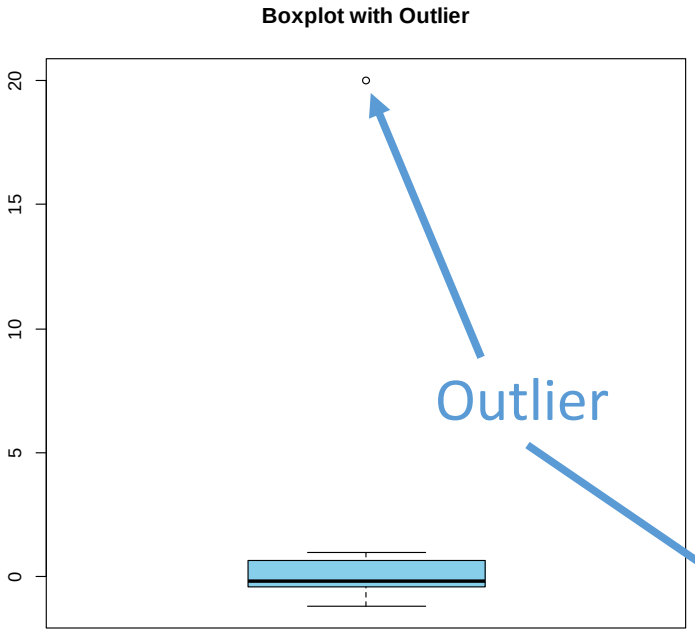



```
> sort(x)
 [1] -1.1894537 -0.7447795 -0.5478961 -0.3443333 -0.2597870 -0.2366460
 [7] -0.1830838  0.3885812  0.5186300  0.8097397  0.9806622
>
> fivenum(x)
[1] -1.1894537 -0.4461147 -0.2366460  0.4536056  0.9806622
>
> boxplot(x, col="skyblue", main="Boxplot")
> boxplot.stats(x)
$stats
[1] -1.1894537 -0.4461147 -0.2366460  0.4536056  0.9806622

$n
[1] 11

$conf
[1] -0.6652619  0.1919699

$out
numeric(0)
```

Tukey, John W., *Exploratory Data Analysis*, Addison-Wesley Publishing Company, 1977, Section 2B, "Hinges and 5-number summaries."

Boxplot with Outlier

What happens if one value is “bad”? Let’s replace x[3] with value 20.




```
> x[3] <- 20  # Introduce outlier
> sort(x)
 [1] -1.1894537 -0.7447795 -0.5478961 -0.2597870 -0.2366460 -0.1830838
 [7]  0.3885812  0.5186300  0.8097397  0.9806622 20.0000000
> fivenum(x)  # Shows new max
[1] -1.1894537 -0.4038416 -0.1830838  0.6641848 20.0000000
> boxplot(x, col="skyblue", main="Boxplot with Outlier")
> boxplot.stats(x)
$stats
[1] -1.1894537 -0.4038416 -0.1830838  0.6641848  0.9806622

$n
[1] 11

$conf
[1] -0.6918787  0.3257111

$out
[1] 20
```

Median is more “robust” measure of central tendency than mean

```
> # Central tendency
> set.seed(19)
> x <- rnorm(11)
> x
 [1] -1.1894537  0.3885812 -0.3443333 -0.5478961  0.9806622 -0.2366460
 [7]  0.8097397 -0.7447795 -0.2597870 -0.1830838  0.5186300
> mean(x)
[1] -0.07348786
> median(x)
[1] -0.236646
>
> x[3] <- 20  # Introduce outlier
> mean(x)
[1] 1.775997
> median(x)
[1] -0.1830838
```

IQR is more “robust” measure of dispersion than standard deviation

```
> # Dispersion
> set.seed(19)
> x <- rnorm(11)
> x
 [1] -1.1894537  0.3885812 -0.3443333 -0.5478961  0.9806622 -0.2366460
 [7]  0.8097397 -0.7447795 -0.2597870 -0.1830838  0.5186300
> sd(x)
[1] 0.6725479
> diff(boxplot.stats(x)$stats[c(2,4)])  # IQR
[1] 0.8997204
>
> x[3] <- 20  # Introduce outlier
> sd(x)
[1] 6.080857
> diff(boxplot.stats(x)$stats[c(2,4)])  # IQR
[1] 1.068026
```

Boxplot (median, IQR) vs. Normal Distribution (mean, standard deviation)

- Boxplot makes no assumptions about probability distribution.
- IQR contains 50% of data.
- If normal data, ± 1 standard deviation contains ~68% of data.
- Median, IQR more “robust” than mean, standard deviation

How are Outliers Defined?

- Look at the code: boxplot.stats

```
> boxplot.stats
function (x, coef = 1.5, do.conf = TRUE, do.out = TRUE)
{
    if (coef < 0)
        stop("'coef' must not be negative")
    nna <- !is.na(x)
    n <- sum(nna)
    stats <- stats::fivenum(x, na.rm = TRUE)
    iqr <- diff(stats[c(2, 4)])
    if (coef == 0)
        do.out <- FALSE
    else {
        out <- if (!is.na(iqr)) {
            x < (stats[2L] - coef * iqr) | x > (stats[4L] + coef *
                iqr)
        }
        else !is.finite(x)
        if (any(out[nna], na.rm = TRUE))
            stats[c(1, 5)] <- range(x[!out], na.rm = TRUE)
    }
    conf <- if (do.conf)
        stats[3L] + c(-1.58, 1.58) * iqr/sqrt(n)
```

How are Outliers Defined?

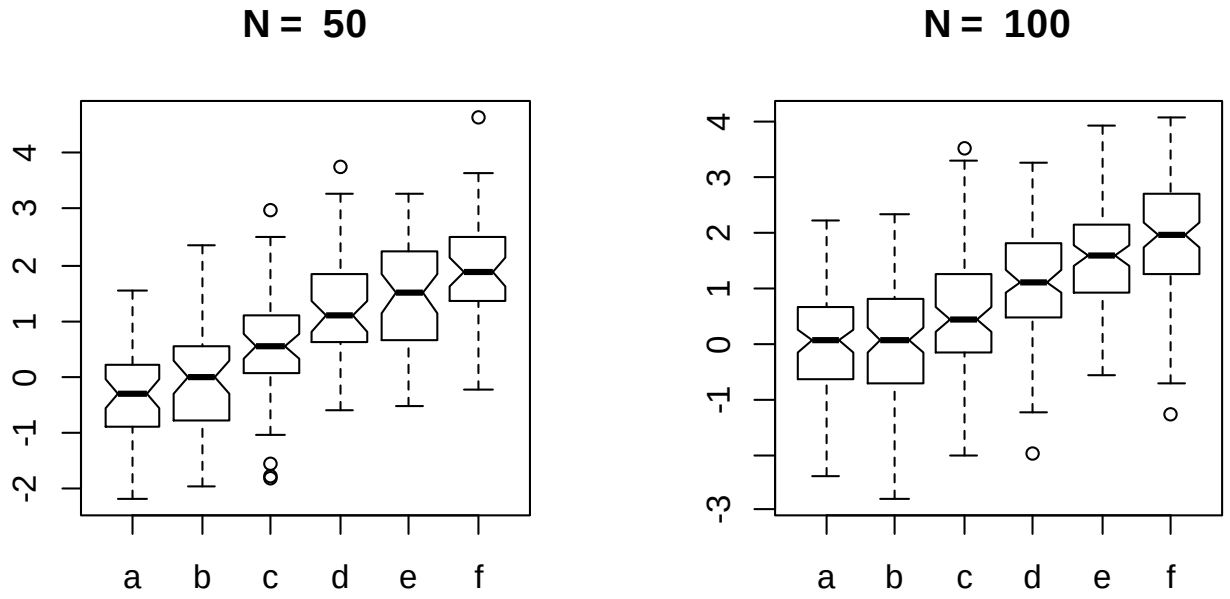
- Simplified version

```
> coef <- 1.5  # default value
>
> stats <- stats::fivenum(x, na.rm=TRUE)
> stats
[1] -1.1894537 -0.4038416 -0.1830838  0.6641848 20.0000000
>
> iqr <- diff(stats[c(2,4)])  # interquartile range
> iqr
[1] 1.068026
>
> out <- x < (stats[2L] - coef*iqr) | x > (stats[4L] + coef*iqr)
> which(out)
[1] 3
>
> stats[c(1,5)] <- range(x[!out], na.rm=TRUE)  # update without outliers
> stats
[1] -1.1894537 -0.4038416 -0.1830838  0.6641848  0.9806622
```

Notched Boxplots

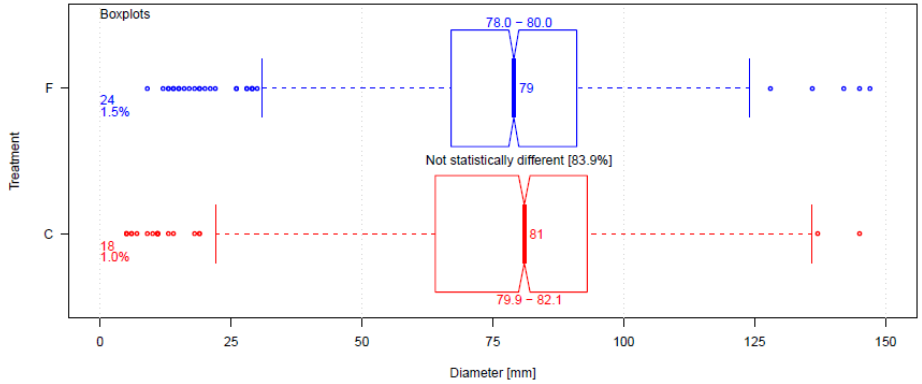
```
> # Statistics:  An introduction using R by Michael J. Crawley, p. 297.
>
> set.seed(11)
>
> plotboxes <- function(N)
+ {
+
+   a <- rnorm(N, mean=0.00, sd=1)
+   b <- rnorm(N, mean=0.00, sd=1)
+   c <- rnorm(N, mean=0.50, sd=1)
+   d <- rnorm(N, mean=1.00, sd=1)
+   e <- rnorm(N, mean=1.50, sd=1)
+   f <- rnorm(N, mean=2.00, sd=1)
+
+   boxplot(data.frame(a,b,c,d,e,f),
+     notch=TRUE, main=paste("N = ", N))
+
+ }
>
> par(mfrow=c(1,2))
> plotboxes(50)
> plotboxes(100)
```

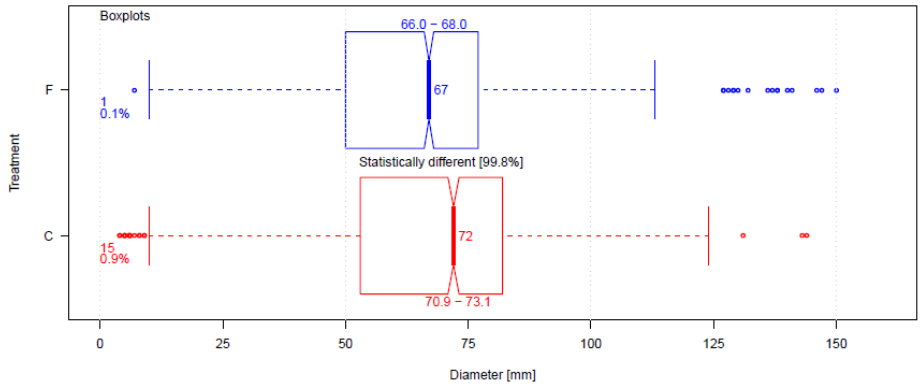
Notched Boxplots



The notch=TRUE option allows significance testing of the difference in medians. Where the notches do not overlap the medians are significantly different at an $\alpha = 0.05$ significance level. When the notches overlap, there is no significant difference between the medians.

“Notched” Boxplots





From InnoCentive.com submission

Examples of Using Boxplots to Identify Outliers

- Identify “problem” images in Kaggle competition facial images
- Congressional disbursements
- “PULSE” diagrams to study political money
- Shawnee County, KS public salaries

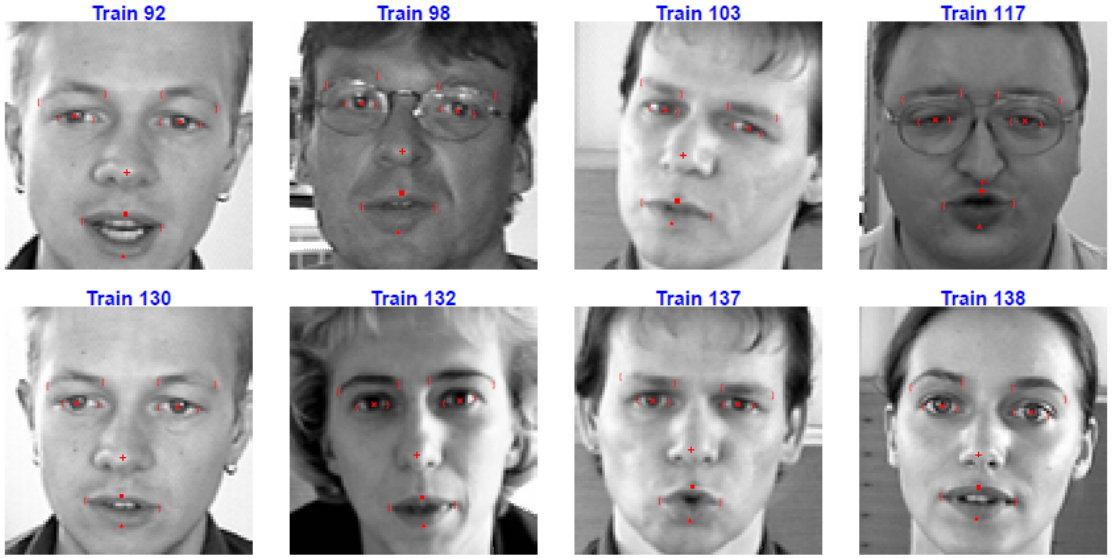
Kaggle Competition: Facial Keypoints Detection

http://www.kaggle.com/

Knowledge • 49 teams

Facial Keypoints Detection

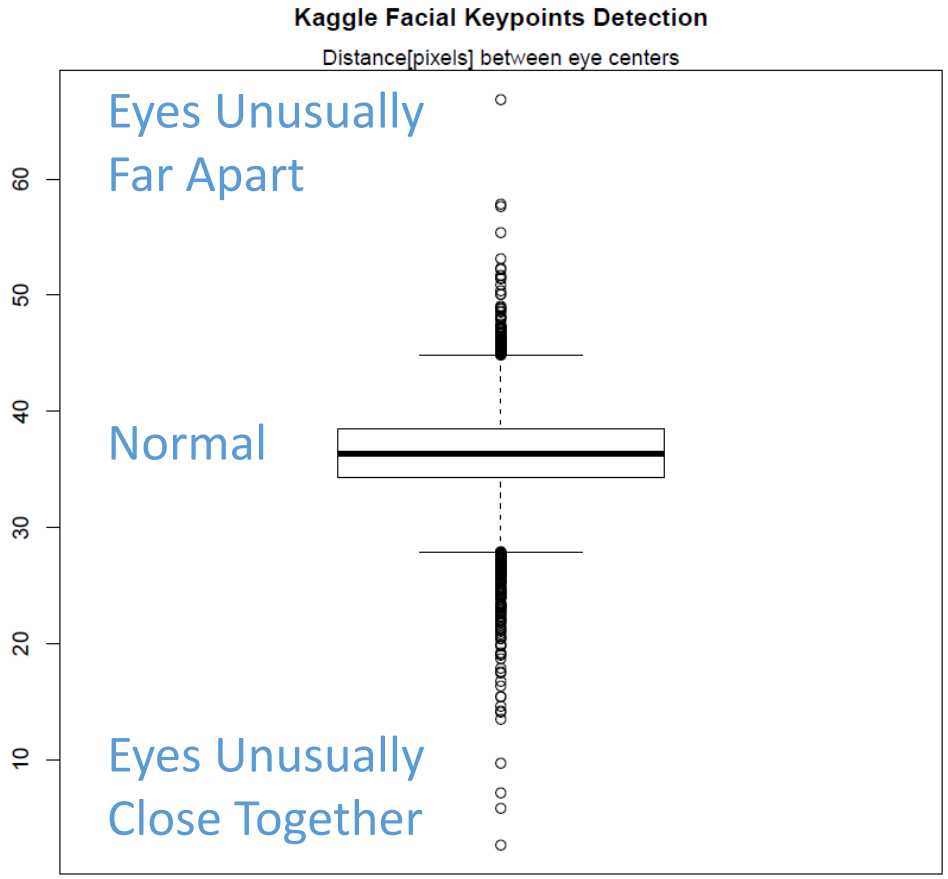
Tue 7 May 2013 — Wed 31 Dec 2014 (2 months to go)



Images:
7049 Train
1783 Test

96 x 96 pixels

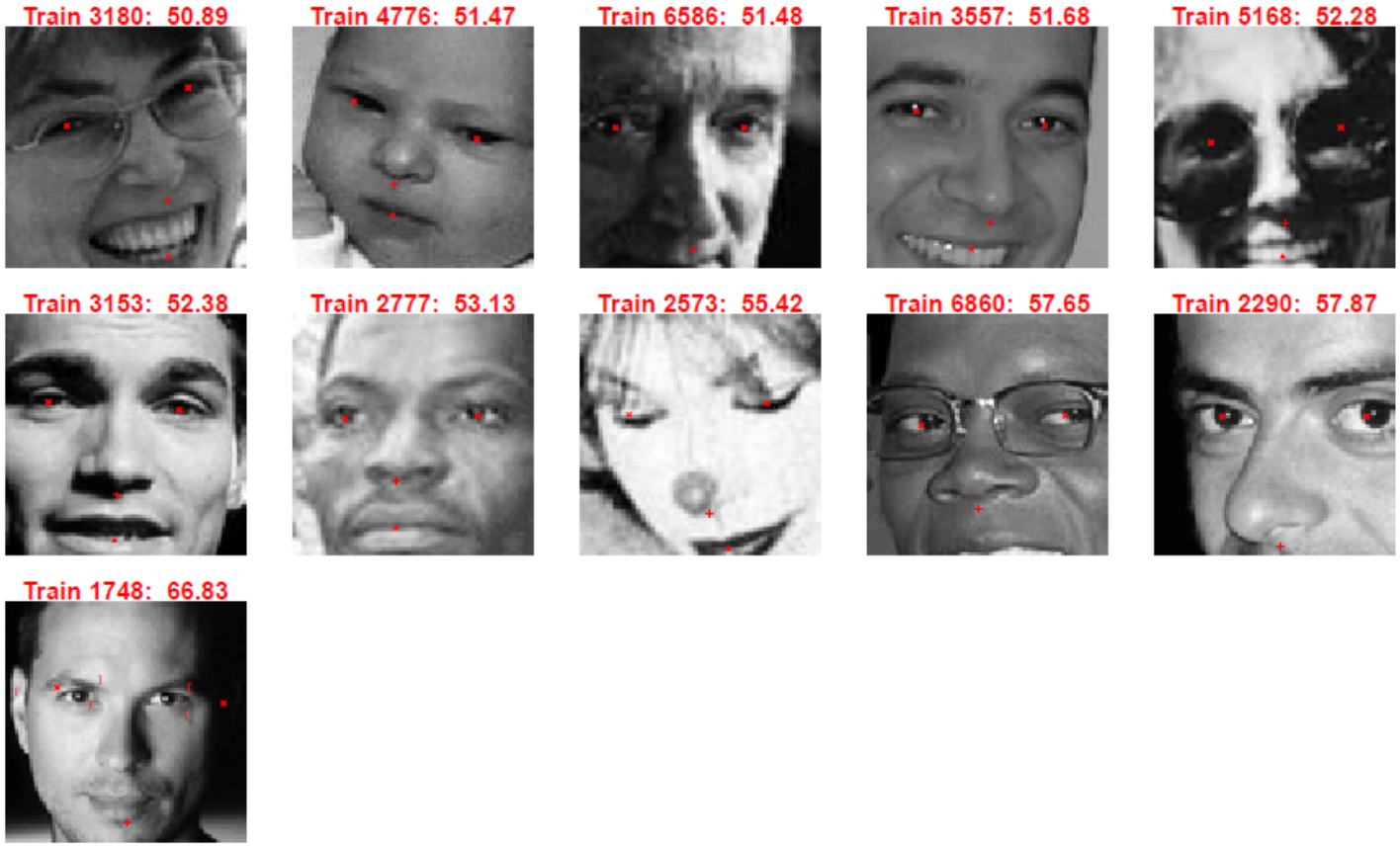
How can “problem” images be identified?



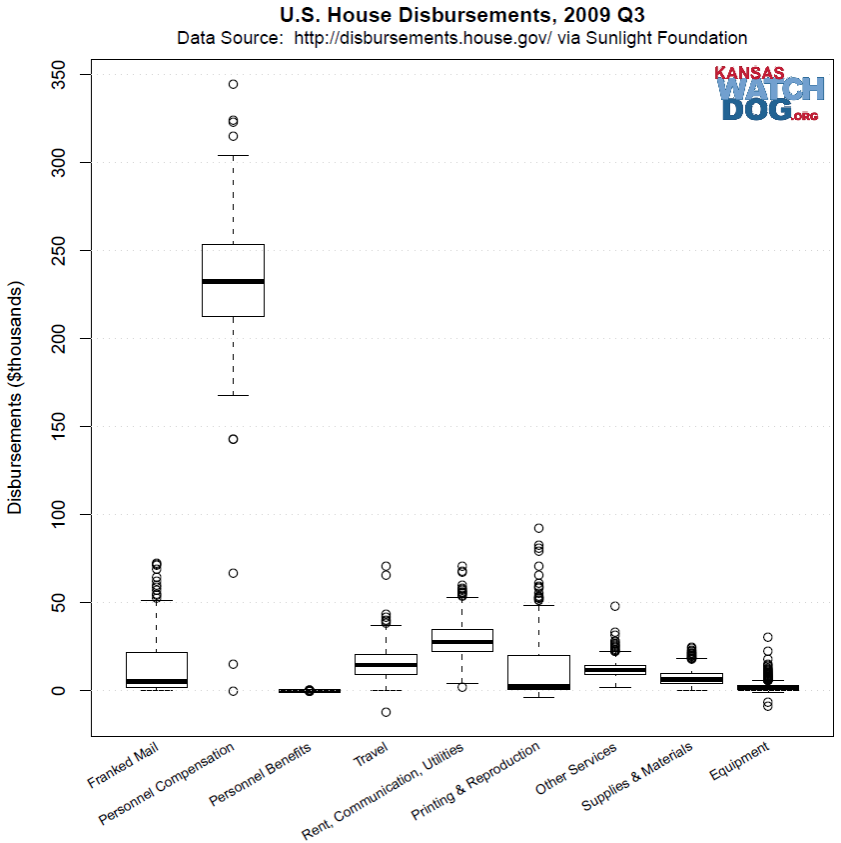
Normal

Eyes unusually close together

Eyes unusually far apart

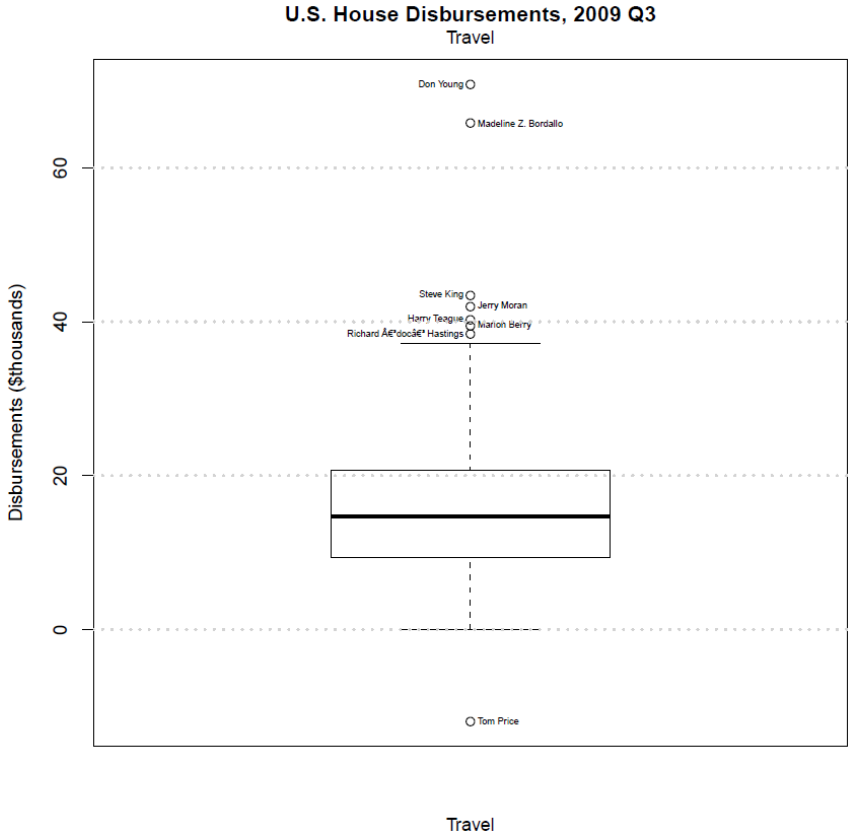


Congressional Disbursements



What expenses are “normal”?

Congressional Disbursements



What expenses are “normal”?

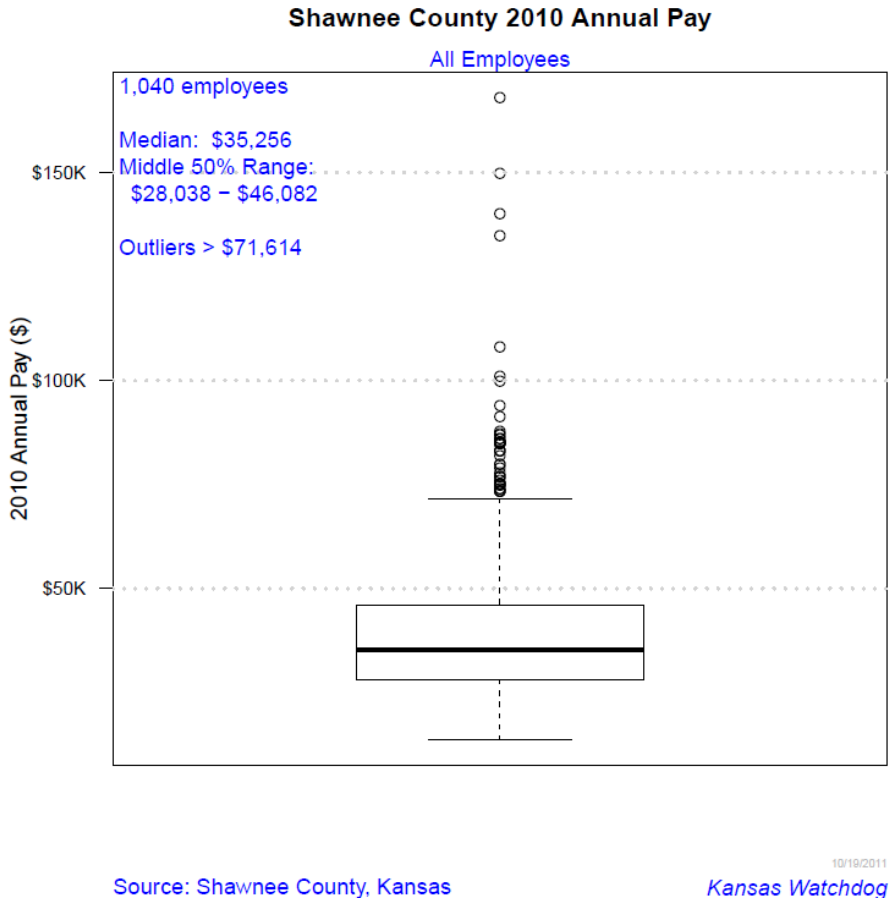
“PULSE” Diagrams to Study Political Money

No longer online at FollowTheMoney.org

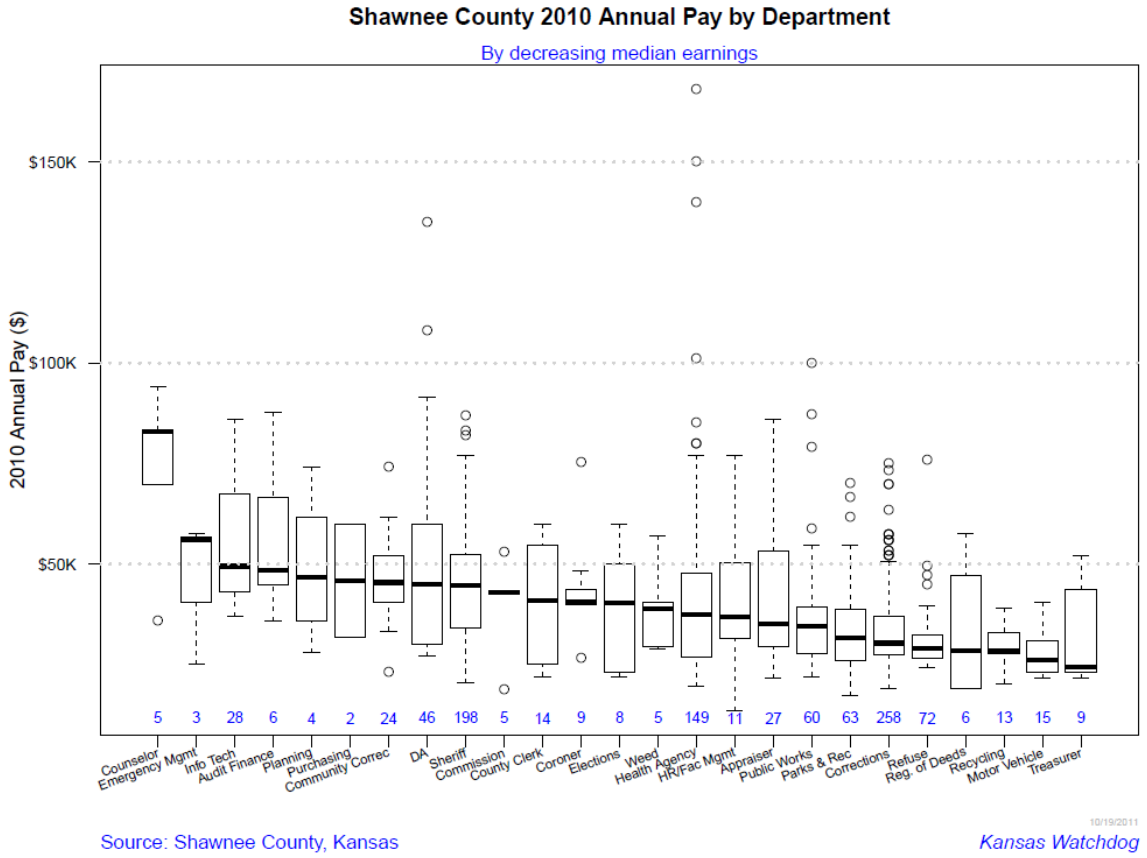
PULSE = Political Contribution Logarithmic Scatterplot Profile

Kansas Senate Winners and Losers in 2008

Shawnee County, Kansas Public Salaries



Shawnee County, Kansas Public Salaries



Take Away

- Great tools for exploratory data analysis:
  - fivenum summary (or use quantile function)
  - boxplot summary
- Median and IQR robust statistics for any distribution
- Outliers: bad data or possibly something interesting