Lect1-Intro

```
#View(iris)
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

A certain reaction was run several times using each of two catalysts, A and B. The catalysts were supposed to control the yield of an undesirable side product. Results, in units of ounces yield, for 4 runs of catalyst A and 6 runs of catalyst B are as follows:

CATALYST A:

4.4 3.4 2.6 3.8

CATALYST B:

3.4 1.1 2.9 5.5 6.4 5.0

ANSWER THE FOLLOWING QUESTIONS:

1. What type of data is this?
2. Does there appear to be a (meaningful) difference between the two catalysts? Why?
3. What would help you answer question 2 more confidently?
4. What is the range of values for each catalyst?
5. What is the median value for each catalyst?
6. What is the mean value for each catalyst?
7. What is the standard deviation for each catalyst?

Entering the data

```
Yield_A<-c(4.4, 3.4, 2.6, 3.8)
Yield_A
```

```
## [1] 4.4 3.4 2.6 3.8
```

```
Yield_B<-c(3.4, 1.1, 2.9, 5.5, 6.4, 5.0)
Yield_B
```

```
## [1] 3.4 1.1 2.9 5.5 6.4 5.0
```

```
#You do not need to combine your data into a dataframe, but this can be a useful form for seeing the re
all_Yields<-c(Yield_A, Yield_B)
all_Yields
```

```
##  [1] 4.4 3.4 2.6 3.8 3.4 1.1 2.9 5.5 6.4 5.0
```

```
Catalyst<-c(rep("A", times=4), rep("B", times=6))
Catalyst
```

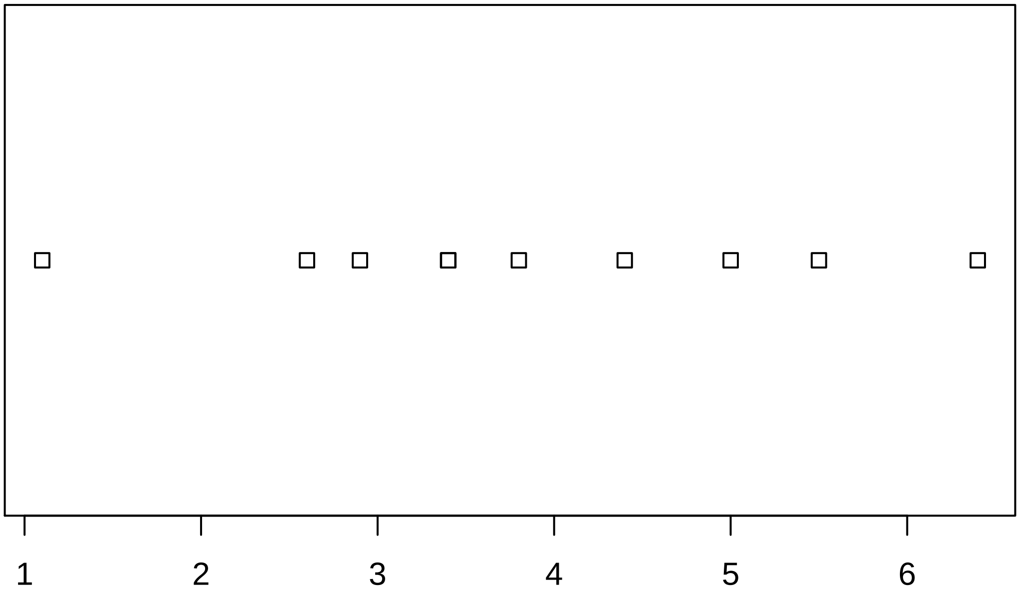
```
##  [1] "A" "A" "A" "A" "B" "B" "B" "B" "B" "B"
```

```
Yield_data<-data.frame(Catalyst, all_Yields)
#View(Yield_data)
str(Yield_data)
```

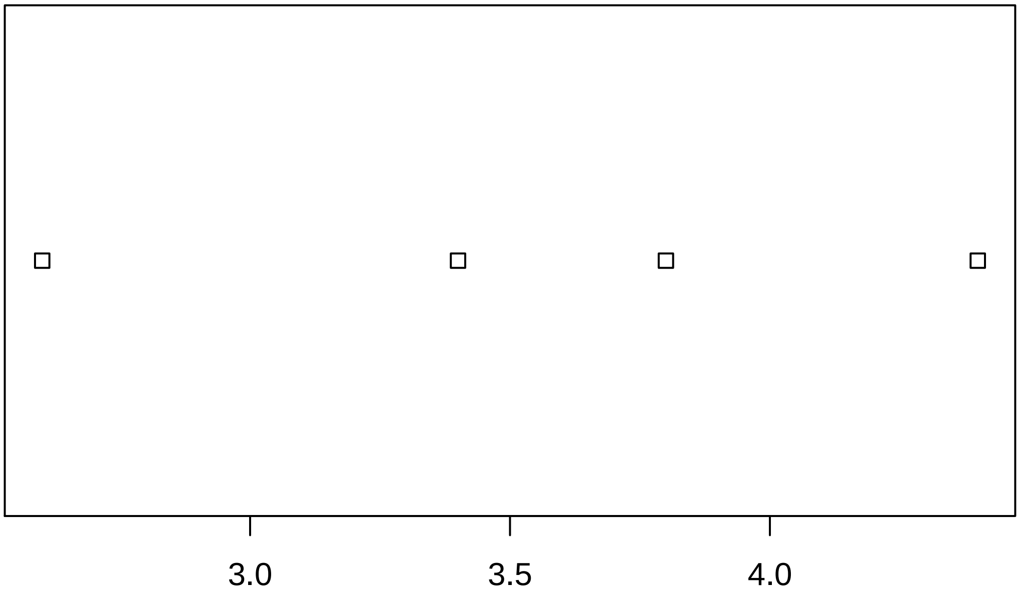
```
## 'data.frame':    10 obs. of  2 variables:
##  $ Catalyst  : Factor w/ 2 levels "A","B": 1 1 1 1 2 2 2 2 2 2
##  $ all_Yields: num  4.4 3.4 2.6 3.8 3.4 1.1 2.9 5.5 6.4 5
```

Graphing the data using dotplots (since small numeric data sets)

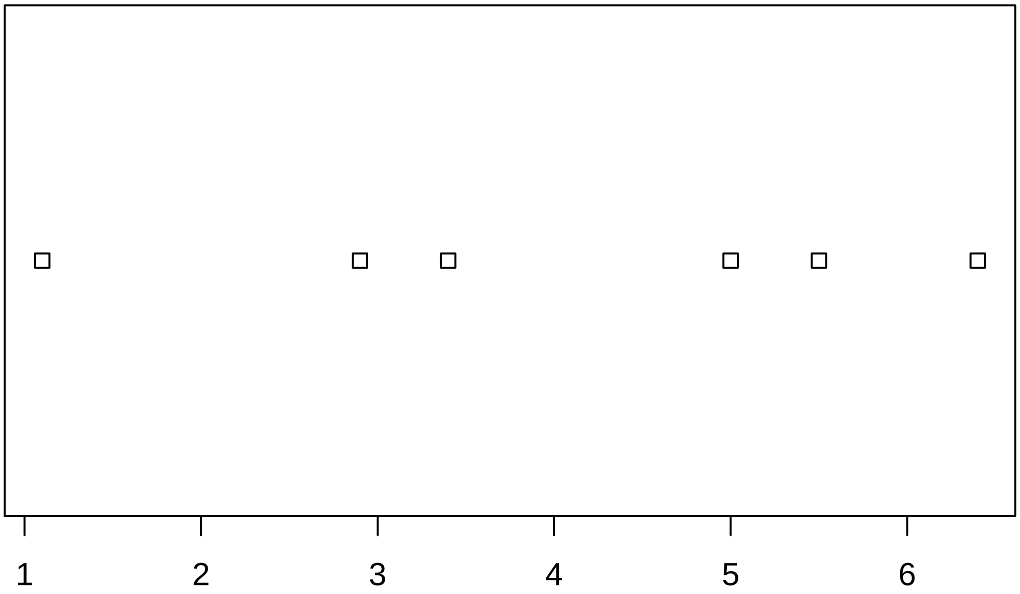
```
?stripchart   #Gives me info on the stripchart function

stripchart(all_Yields)  #dot plot of all yields
```



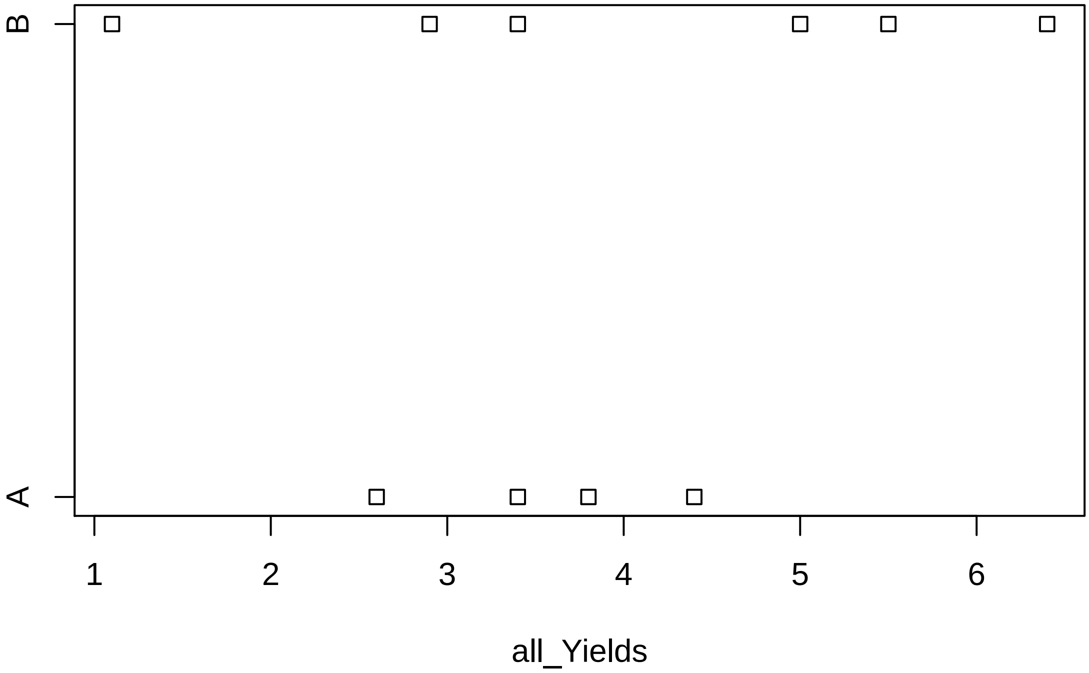
```
stripchart(Yield_A)  #dot plot of Catalyst A yields
```



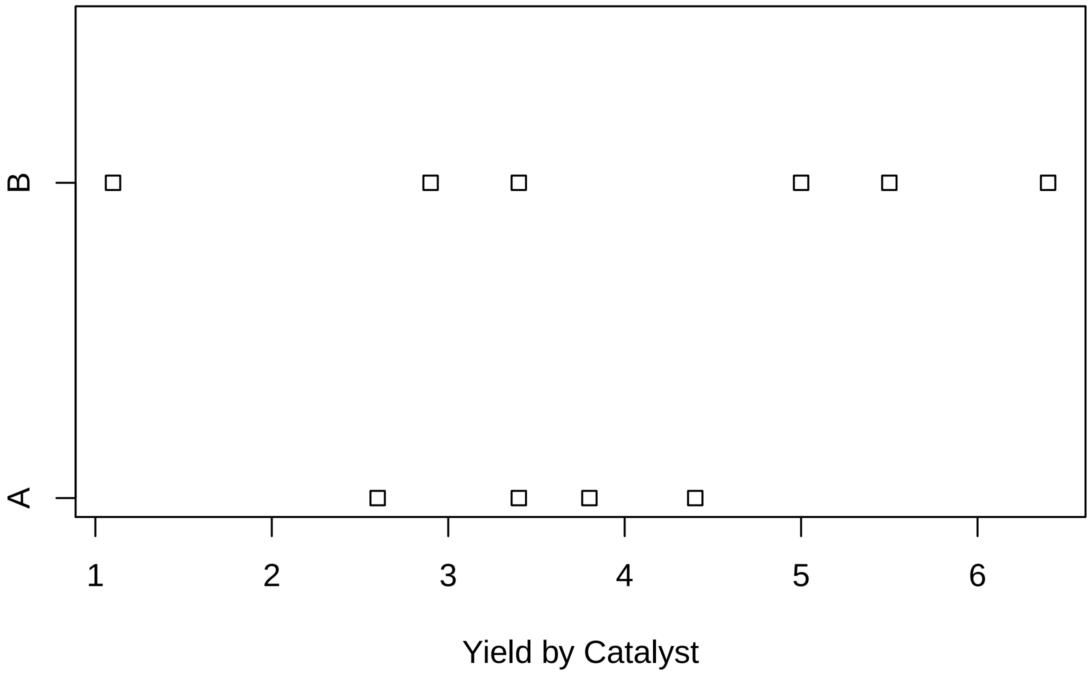
```r
stripchart(Yield_B)  #dot plot of Catalyst B yields
```



```r
stripchart(all_Yields~Catalyst) #dot plot of all yields broken out by catalyst
```



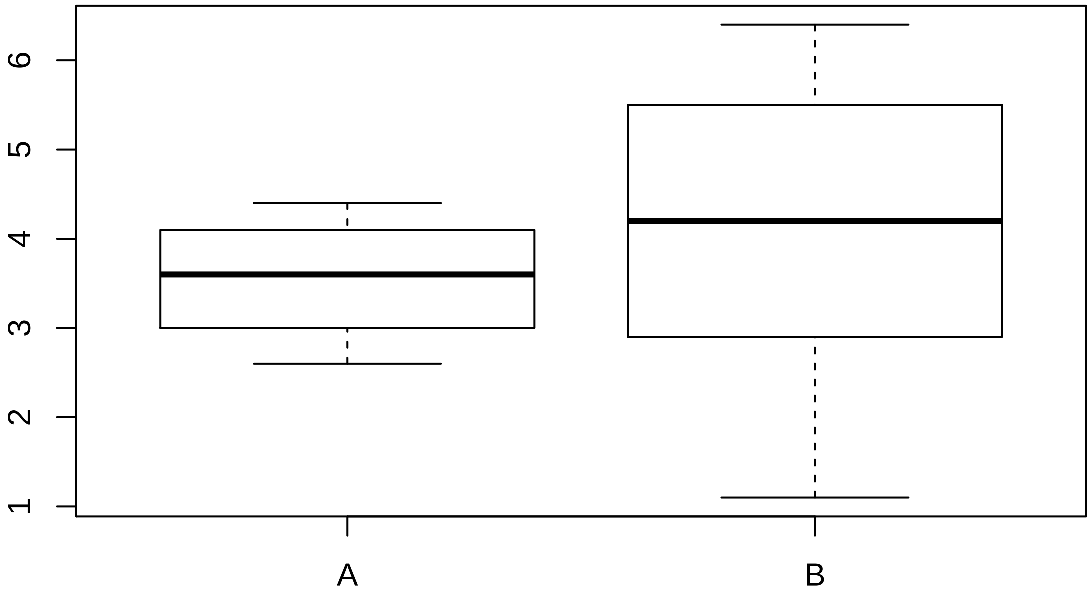
```r
stripchart(all_Yields~Catalyst, data=Yield_data, xlab="Yield by Catalyst", method="stack")
```



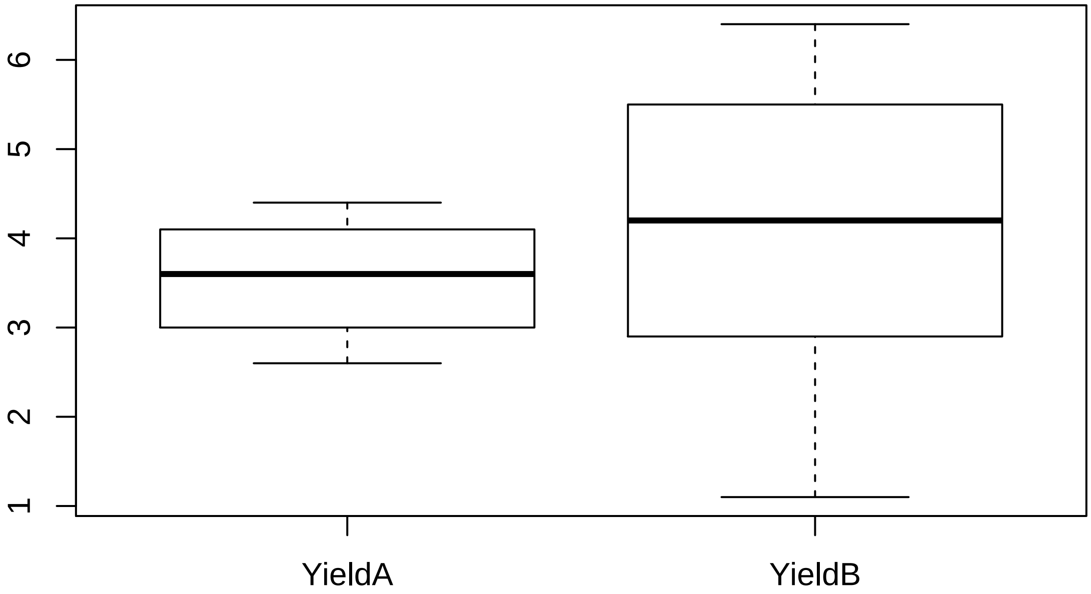
```
#Update label on x axis to be more descriptive than variable name
```

What do we lose with boxplots?

```
boxplot(all_Yields~Catalyst)
```



```
#Or

boxplot(Yield_A, Yield_B, names=c("YieldA", "YieldB"))
```



Estimate some Descriptive Statistics (range, mean, median, sd) for each of the Catalyst yields before calculating them below.

Calculating some Descriptive Statistics Using R

```
range(Yield_A)
```

```
## [1] 2.6 4.4
```

```
range_A<-range(Yield_A)
range_A
```

```
## [1] 2.6 4.4
```

```
range_A[2]-range_A[1]
```

```
## [1] 1.8
```

```
median(Yield_A)
```

```
## [1] 3.6
```

```
mean(Yield_A)
```

```
## [1] 3.55
```

```
sd(Yield_A)  #This function calculates the standard deviation of a sample - more on this later
```

```
## [1] 0.7549834
```

```
#Calculate SD by hand?

range_B<-range(Yield_B)
range_B[2]-range_B[1]
```

```
## [1] 5.3
```

```
median(Yield_B)
```

```
## [1] 4.2
```

```r
mean(Yield_B)
```

```
## [1] 4.05
```

```r
sd(Yield_B)  #sample SD
```

```
## [1] 1.948076
```

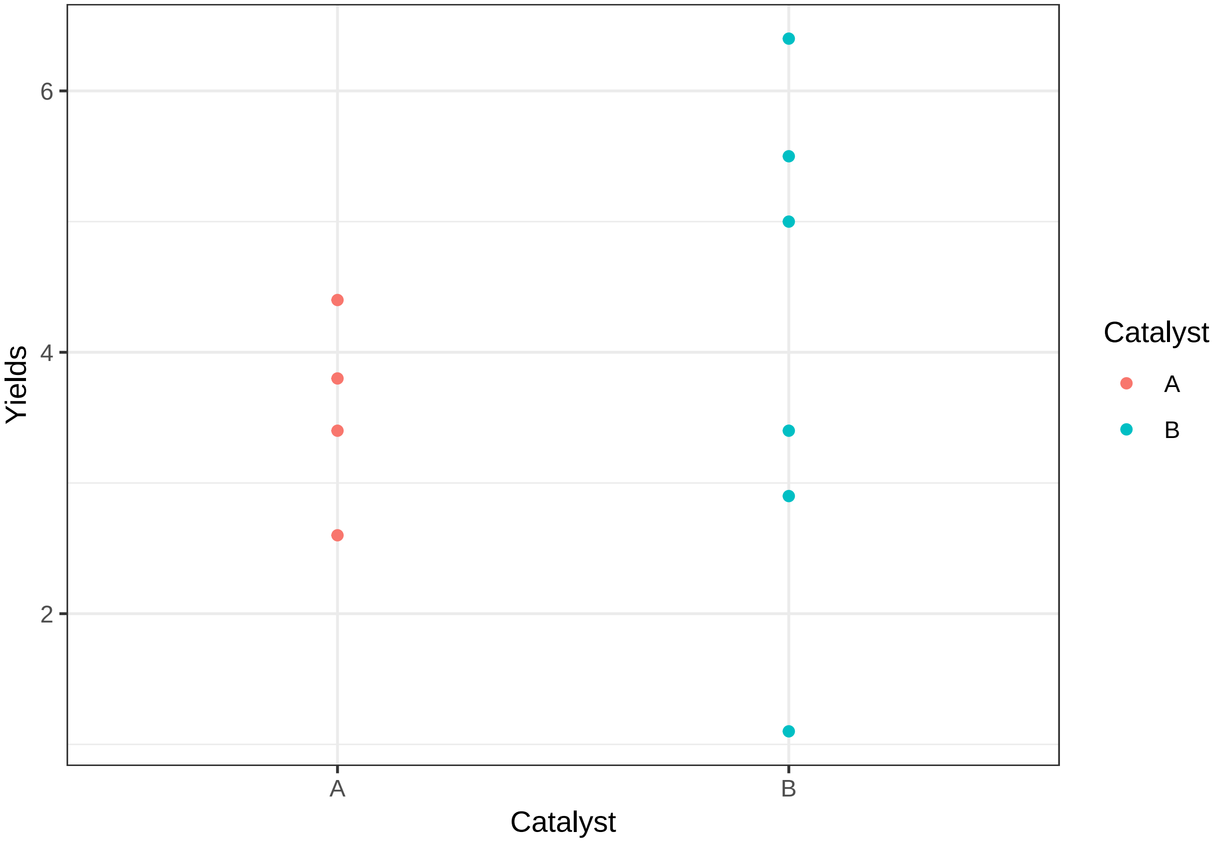
Suped-up graphing

```r
#install.packages("ggplot2")
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```r
ggplot(data=Yield_data, aes(x=Catalyst, y=all_Yields, color=Catalyst))+
  geom_point()+
  theme_bw()+
  ylab("Yields")
```


Yields
A
B
Catalyst
Catalyst
● A
● B