Ve572 Lecture 4

Manuel and Jing

UM-SJTU Joint Institute

May 29, 2018

Q: Do you think my midterms were/are going to be too difficult?

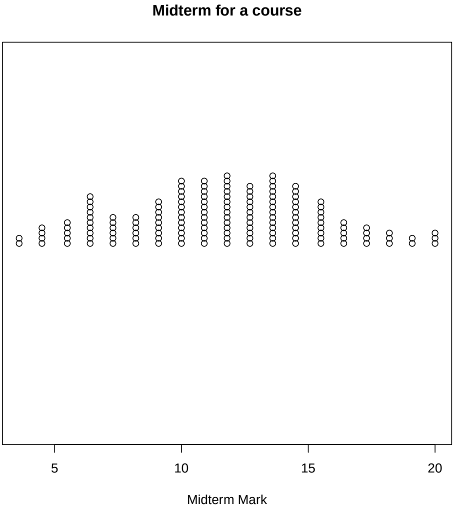
```
> course.df = read.table("~/Desktop/course.txt",
+                        header = TRUE)
> nrow(course.df); head(course.df, 5)
```

```
[1] 146
  Exam Gender Attend Assign Midterm
1   42   Male    Yes   17.2     9.1
2   58 Female    Yes   17.2    13.6
3   81 Female    Yes   17.2    14.5
4   86 Female    Yes   19.6    19.1
5   35   Male     No    8.0     8.2
```

```
> sapply(course.df, class)
```

```
     Exam    Gender    Attend    Assign   Midterm
"integer"  "factor"  "factor" "numeric" "numeric"
```

- Dot plot or strip chart



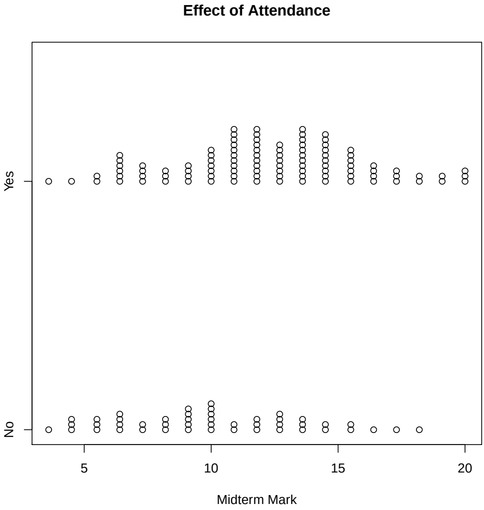
```
> # pdf()       ## Create a pdf file of the plot
> stripchart(course.df$Midterm,
+             method = "stack",
+             pch = 1,
+             main = "Midterm for a course",
+             xlab = "Midterm Mark")
> # dev.off()   ## Close the pdf file
```

Q: Do you attend my class regularly?

```
> stripchart(Midterm~Attend,
+             data = course.df,
+             method = "stack",
+             pch = 1,
+             main = "Effect of Attendance",
+             xlab = "Midterm Mark")
```

Q: Do you think it matters?

- Multiple strip charts of a numeric variable by a factor variable.



- A strip chart is not always appropriate!

```
> midcounts.df =
+   as.data.frame(table(course.df$Midterm))
> colnames(midcounts.df)[1] = c("Midterm")
> str(midcounts.df, vec.len = 2)
```

```
'data.frame':   19 obs. of  2 variables:
 $ Midterm: num  3.6 4.5 5.5 6.4 7.3 ...
 $ Counts : int  2 4 5 10 6 ...
```

```
> summary(midcounts.df$Counts)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   4.000   6.000   7.684  12.000  14.000
```

Q: What shall we do if there are too many distinct marks or too many students?

- For example, strip charts are not appropriate for the following tmp.

```
> tmp = rnorm(1000)     # A random sample of Normal
> length(unique(tmp))
```

```
[1] 1000
```

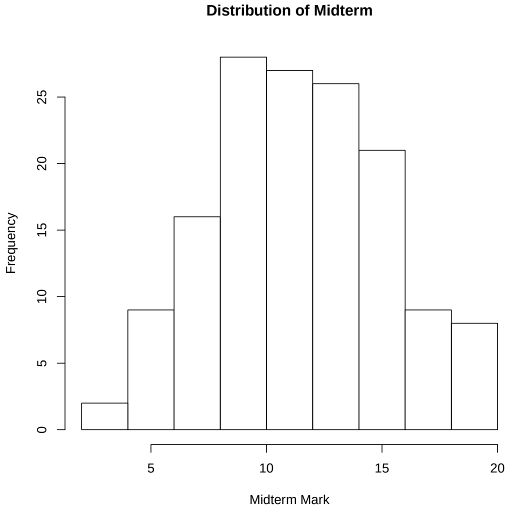
```
> # A random sample of Chi-Squared
> tmp = round(rchisq(n = 1e4, df = 30))
> length(unique(tmp))
```

```
[1] 72
```

```
> freq.df = as.data.frame(table(tmp))
> summary(freq.df$Freq)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    9.75   83.50  138.90  246.00  419.00
```

- Histogram


Distribution of Midterm
Frequency
0
5
10
15
20
25
5
10
15
20
Midterm Mark

```
> hist(course.df$Midterm,
+       main = "Distribution of Midterm",
+       xlab = "Midterm Mark")
```

- Often histograms are normalised, i.e.

  Area underneath is one

  and are companied by a kernel density estimation (KDE)

  $$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K(x - x_i; h)$$

  where x_i denotes the ith data point, and h is known as bandwidth.
- It estimates the probability density function of the distribution from which the observed sample comes, it is an inference made about the population.

- In this context, the kernel function K can be the density function of any symmetric continuous distribution, e.g. the gaussian kernel

$$K(x - x_i, h) = \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{(x - x_i)^2}{2h^2}\right)$$

- Intuitively, KDE is an average of density functions, one for each data point.

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K(x - x_i; h)$$

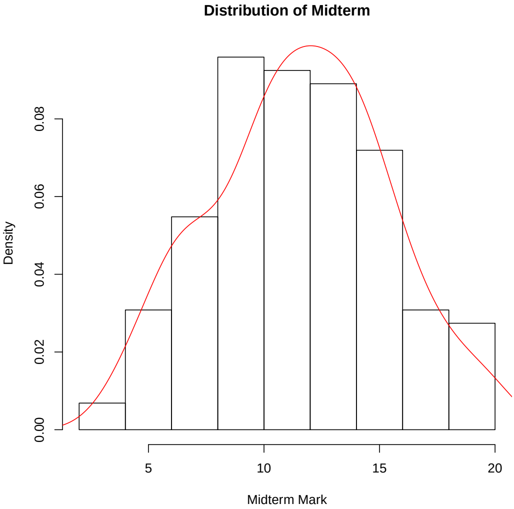
- There are various methods to choice h based on the observed data.

```
> bw.nrd0(course.df$Midterm) # Default in R
```

```
[1] 1.255401
```

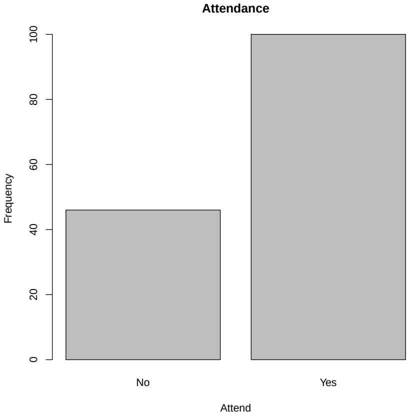
```
> hist(course.df$Midterm, probability = TRUE,
+       main = "Distribution of Midterm",
+       xlab = "Midterm Mark")
> lines(density(course.df$Midterm), col = "red")
```

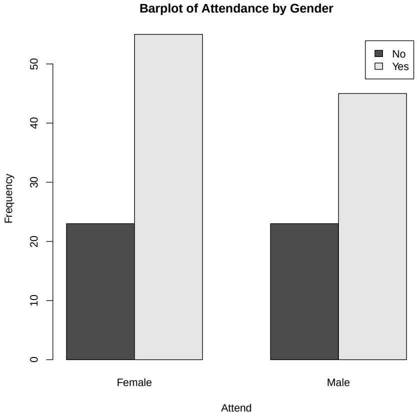
- Normalised histogram with a kernel density estimation



Q: Can we use a histogram to visualise a discrete/factor variable?

- Barplots





```
> gradetable = table(course.df$Attend)
> gradetable
```

```
 No Yes
 46 100
```

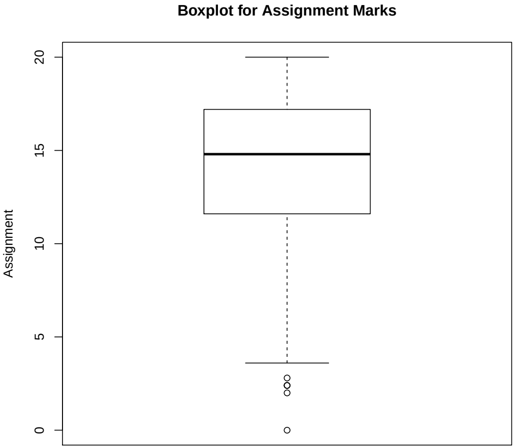
```
> barplot(gradetable, main = "Attendance",
+         ylab = "Frequency", xlab = "Attend")
```

```
> (gradetable = with(course.df,
+                     table(Attend, Gender)))
```

```
      Gender
Attend Female Male
   No      23   23
   Yes     55   45
```

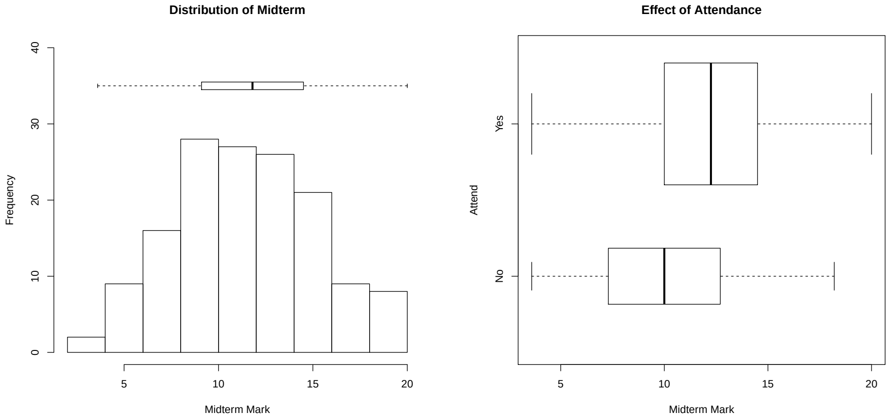
```
> barplot(gradetable, legend = TRUE, beside = TRUE,
+         main = "Barplot of Attendance by Gender",
+         ylab = "Frequency", xlab = "Attend")
```

- Box plot



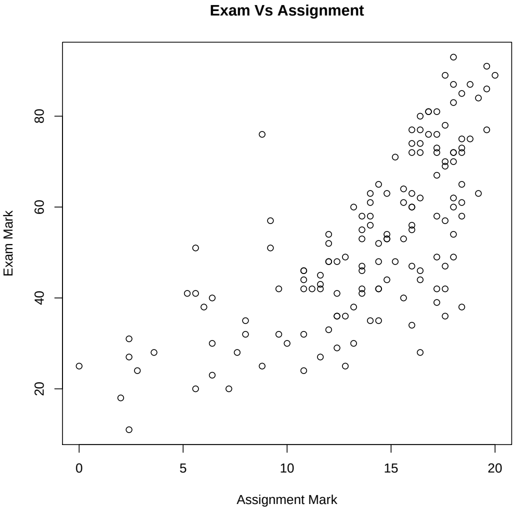
```
> boxplot(course.df$Assign,
+         main = "Boxplot for Assignment Marks",
+         ylab = "Assignment")

> hist(course.df$Midterm,
+      main = "Distribution of Midterm",
+      xlab = "Midterm Mark", ylim = c(0,40))
>
> boxplot(course.df$Midterm, boxwex = 2,
+         horizontal = TRUE, at = 35,
+         add = TRUE, axes = FALSE)

> prop.vec = c(gradetable[[1]], gradetable[[2]])
> boxplot(Midterm~Attend, data = course.df,
+         width = prop.vec,
+         main = "Effect of Attendance",
+         xlab = "Midterm Mark", ylab = "Attend",
+         horizontal = TRUE)
```

- More box plots



Q: Do you think assignments and exams are correlated?

- Scatter plot



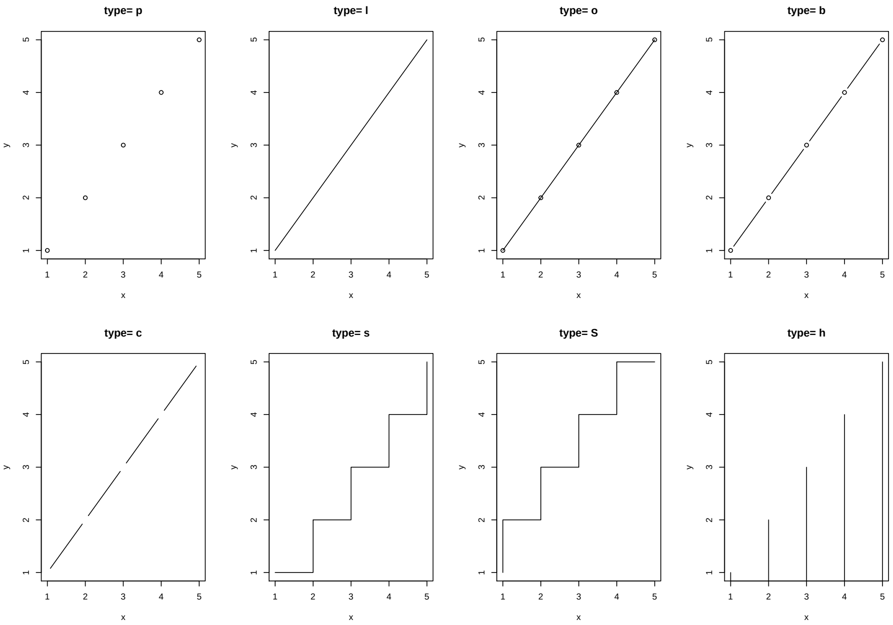
```
> plot(course.df$Assign, course.df$Exam,
+       main = "Exam Vs Assignment",
+       xlab = "Assignment Mark",
+       ylab = "Exam Mark")

> y.vec = course.df$Attend == "Yes"
> class(y.vec);
```

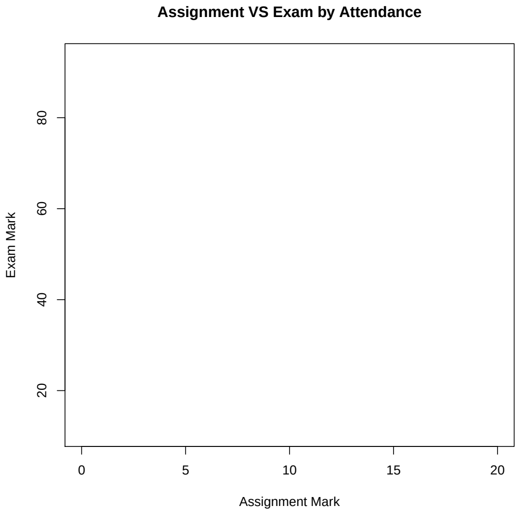
```
[1] "logical"
```

```
> plot(course.df$Assign, course.df$Exam,
+       main = "Assignment VS Exam by Attendance",
+       xlab = "Assignment Mark",
+       ylab = "Exam Mark",
+       type = "n")
```

- There are 9 values `type` can take.

```
> # create some data
> x = 1:5; y = x
>
> # plotting symbol, color, and asp
> par(pch=1, col=1, pty = "m")
>
> # all plots on one page
> par(mfrow=c(2,4))
>
> opts = c("p","l","o","b","c","s","S","h")
>
> for(i in 1:length(opts)){
+
+   heading = paste("type=", opts[i])
+
+   plot(x, y, type = opts[i], main = heading)
+
+ }
```


type= p
type= l
type= o
type= b
type= c
type= s
type= S
type= h
x
y
1
2
3
4
5



- `type ="n"` tells R to produce the "correct" frame without plotting the data.

- Continue with our original plot

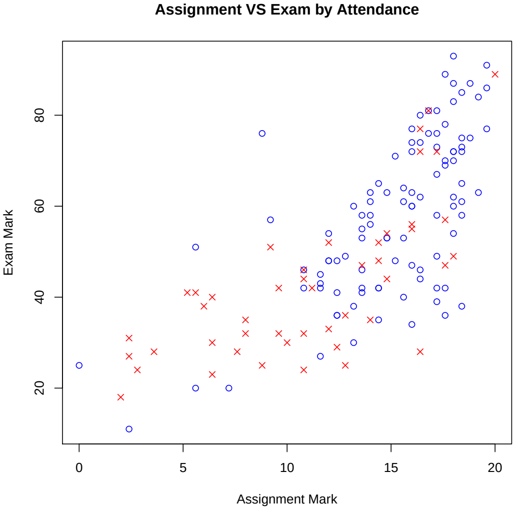
```
> plot(course.df$Assign, course.df$Exam,
+       main = "Assignment VS Exam by Attendance",
+       xlab = "Assignment Mark",
+       ylab = "Exam Mark",
+       type = "n")
```

- We can add things to the empty plot.

```
> points(course.df$Assign[y.vec],
+         course.df$Exam[y.vec],
+         col = 4, pch = 1)
```

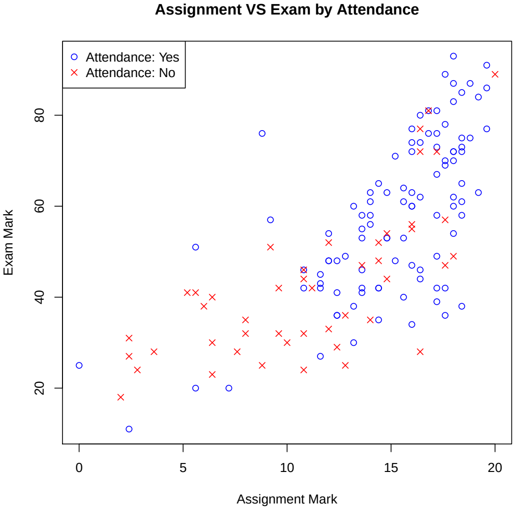
- Different plotting symbol and different colour

```
> points(course.df$Assign[!y.vec],
+         course.df$Exam[!y.vec],
+         col = 2, pch = 4)
```



- We need a label to tell what is what.

```
> legend("topleft", legend =
+           c("Attendance: Yes", "Attendance: No"),
+        col = c(4,2), pch = c(1,4))
```



Q: Do cars with big engines use more fuel than cars with small engines?

- R has a dataset, mpg, about fuel economy, amongst the variables, we have the followings
  - displ

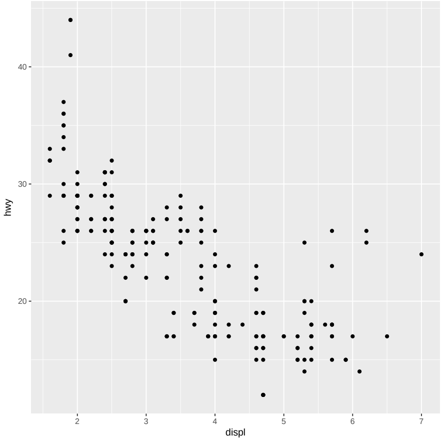
    Engine size

    (liters)
  - hwy

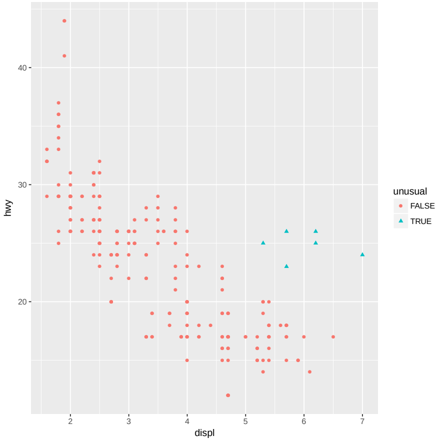
    Fuel usage on the highway

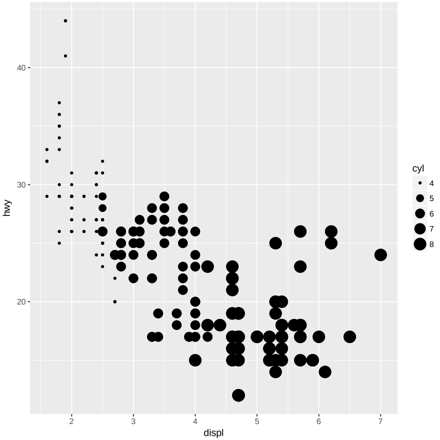
    (miles per gallon)

Q: What can we see from the plot?



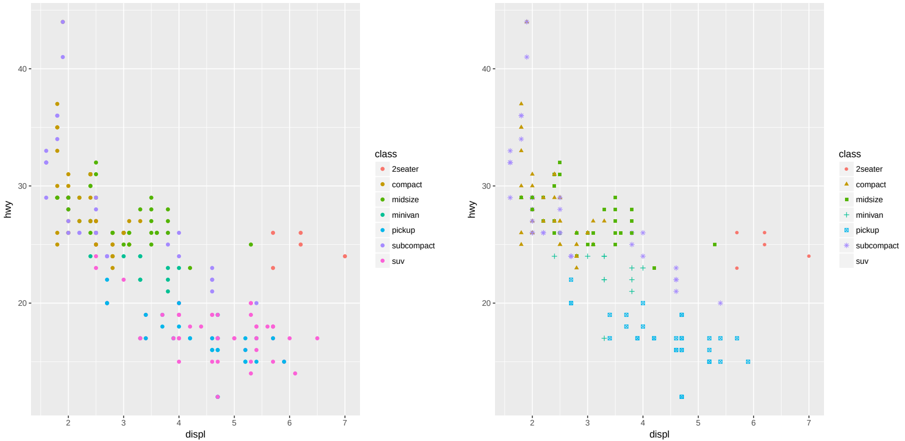
```
> library(ggplot2); help(mpg)
> ggplot(data = mpg) + geom_point(
+   mapping = aes(x = displ, y = hwy))
```





```
> unusual = mpg$displ >= 5.0 & mpg$hwy > 20
> ggplot(data = mpg) + geom_point(
+   mapping = aes(x = displ, y = hwy,
+                          color = unusual, shape = unusual))
> ggplot(data = mpg) + geom_point(
+   mapping = aes(x = displ, y = hwy, size = cyl))
```

- Scatter plots of Fuel usage VS Engine size by class

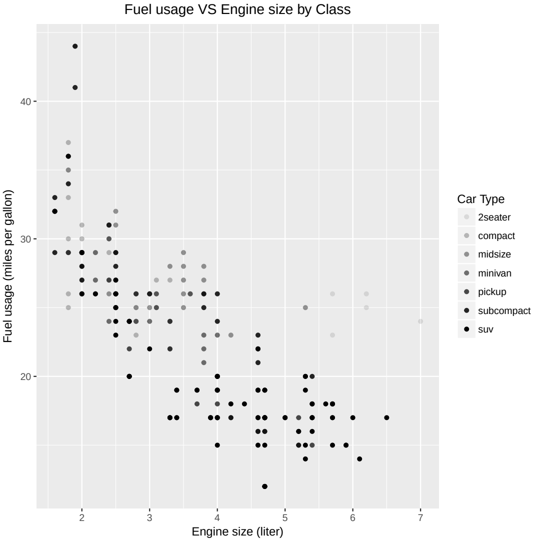


Q: Do you notice any problem?

```
> ggplot(data = mpg) + geom_point(
+   mapping = aes(x = displ, y = hwy,
+   color = class))
> ggplot(data = mpg) + geom_point(
+   mapping = aes(x = displ, y = hwy,
+   color = class, shape = class))
```

```
## Warning messages:
## 1: The shape palette can deal with a maximum of
## 6 discrete values because more than 6 becomes
## difficult to discriminate; you have 7. Consider
## specifying shapes manually if you must have them
## 2: Removed 62 rows containing missing values.
```

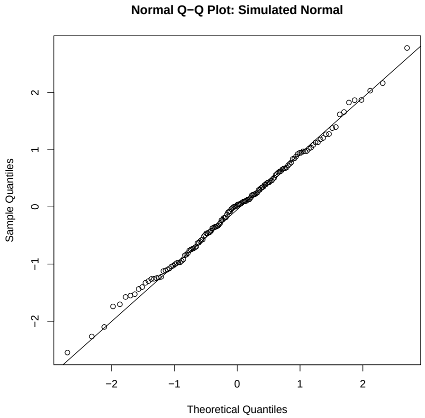
```
> ggplot(data = mpg) + geom_point(
+   aes(x = displ, y = hwy, alpha = class)) +
+   ggtitle("Fuel usage VS Engine size by Class") +
+   theme(plot.title = element_text(hjust = 0.5)) +
+   labs(y = "Fuel usage (miles per gallon)",
+   x = "Engine size (liter)", alpha = "Car Type")
```

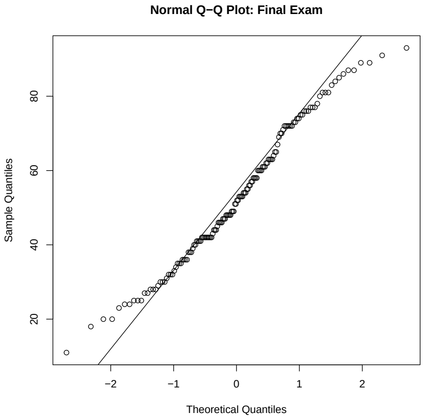


- For more information on ggplots, see HERE

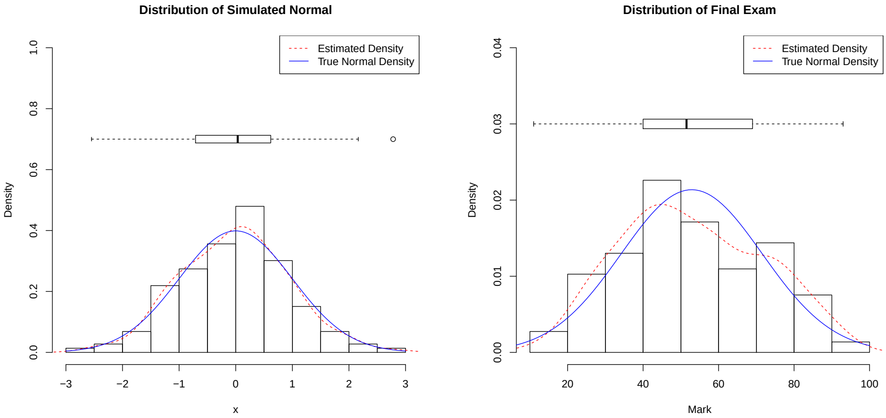
- Normal QQ plot

```
> qqnorm({nor = rnorm(146)}, main = ""); qqline(nor)
> title("Normal Q-Q Plot: Simulated Normal")
> qqnorm({exam = course.df$Exam},main = "")
> qqline(exam); title("Normal Q-Q Plot: Final Exam")
```





- Noticed the difference, and what QQ plots reveal.

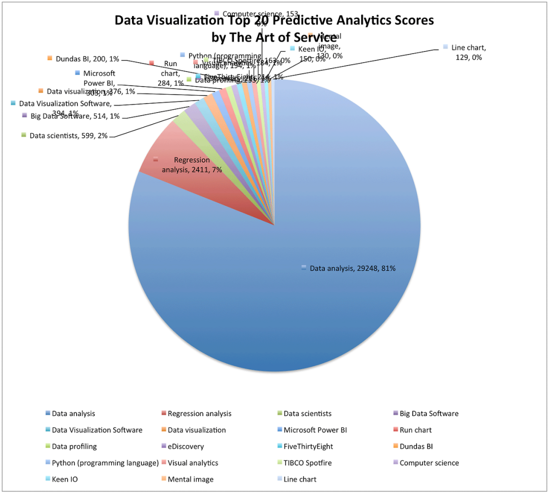


```
> hist(nor, probability = TRUE,
+      main = "Distribution of Simulated Normal",
+      xlab = "x", ylim = c(0,1))
>
> boxplot(nor, boxwex = 0.05, horizontal = TRUE,
+         at = 0.7, add = TRUE, axes = FALSE)
>
> lines(density(nor), col = 2, lty = 2)
> tmp = seq(-3, 3, length = 100)
> lines(tmp, dnorm(tmp), col = 4)
>
> legend("topright",
+        legend = c("Estimated Density",
+                   "True Normal Density"),
+        col = c(2,4), lty = c(2,1))
```

```
> hist(exam, probability = TRUE,
+      main = "Distribution of Final Exam",
+      xlab = "Mark", ylim = c(0,0.04))
>
> boxplot(exam, boxwex = 0.0025,
+         horizontal = TRUE, at = 0.03,
+         add = TRUE, axes = FALSE)
>
> lines(density(exam), col = 2, lty = 2)
> lines({tmp = 0:100},
+       dnorm(tmp, mean = mean(exam),
+             sd = sd(exam)),
+       col = 4)
>
> legend("topright",
+        legend = c("Estimated Density",
+                   "True Normal Density"),
+        col = c(2,4), lty = c(2,1))
```

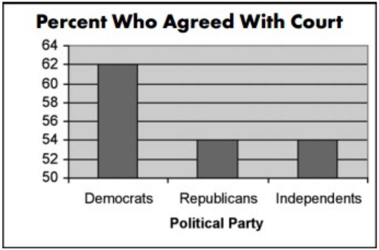
Bad Idea!

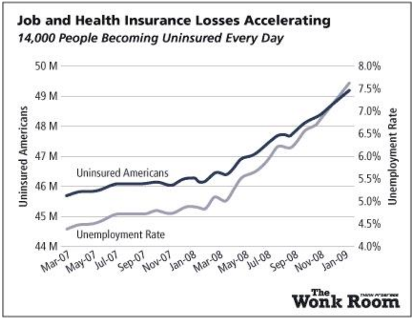
- Not informative



Bad Idea!

Q: What is the problem here?

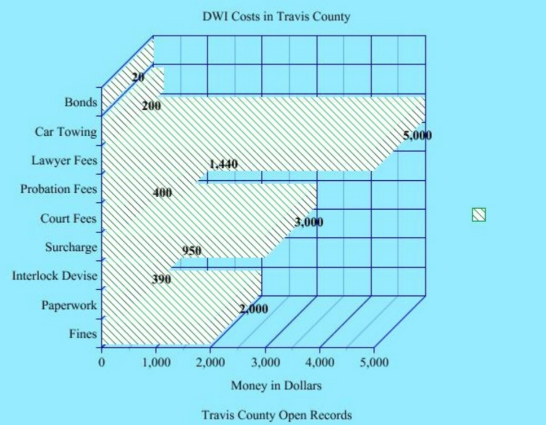




- Misleading scale and truncation.

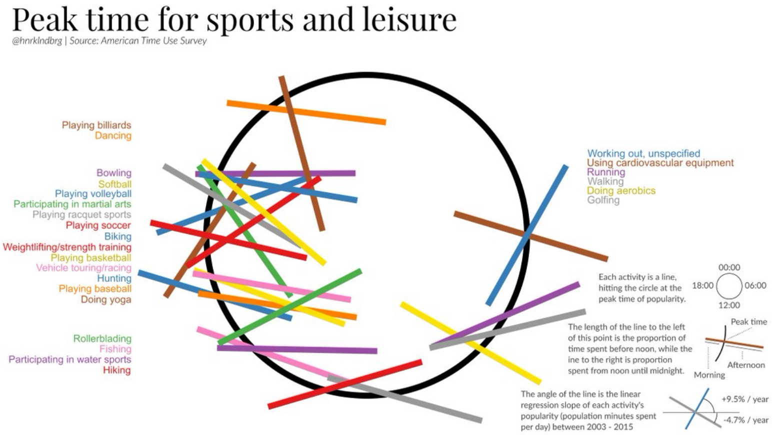
Bad Idea!

- 3D insanity



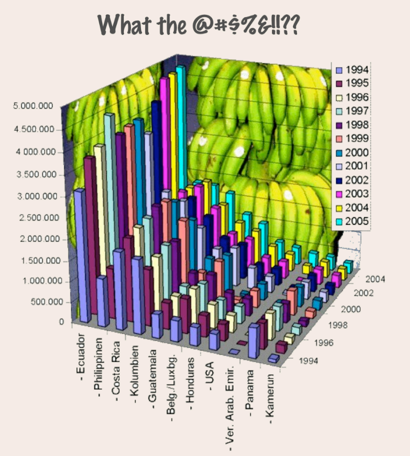
Bad Idea!

- Stay away from 4 dimensional plots!



Bad Idea!

- Who let the monkeys out?


What the @#$%&!!??
- Ecuador
- Philippinen
- Costa Rica
- Kolumbien
- Guatemala
- Belg./Luxbg.
- Honduras
- USA
- Ver. Arab. Emir.
- Panama
- Kamerun