Advanced Analytics with R

Exploring Data

What we have learned

Data overview with *summary*

```
custdata<-read.table('custdata.tsv',header=T,sep = '\t')

summary(custdata)
```

Data Problem – Missing Data

- A very common phenomenon
- Don’t guess
- Find out the cause from the data source

Data with missing information

- custdata[!complete.cases(custdata),]

What to do with missing data?

Missing Categorical Data – Simple Solution

- custdata[is.na(custdata$is.employed),]

```
custdata$employment.status.fix<-ifelse(is.na(custdata$is.employed),"Unknown",
                                       ifelse(custdata$is.employed==T,"employed","unemployed"))
```

Missing Numerical Data – Missing Randomly

- Assign mean value to the missing data
- Assign value to missing data based on relationship between this variable and other related variable(s)
  - Income vs age
  - Income vs Profession
  - …

Missing Numerical Data – Missing Systematically

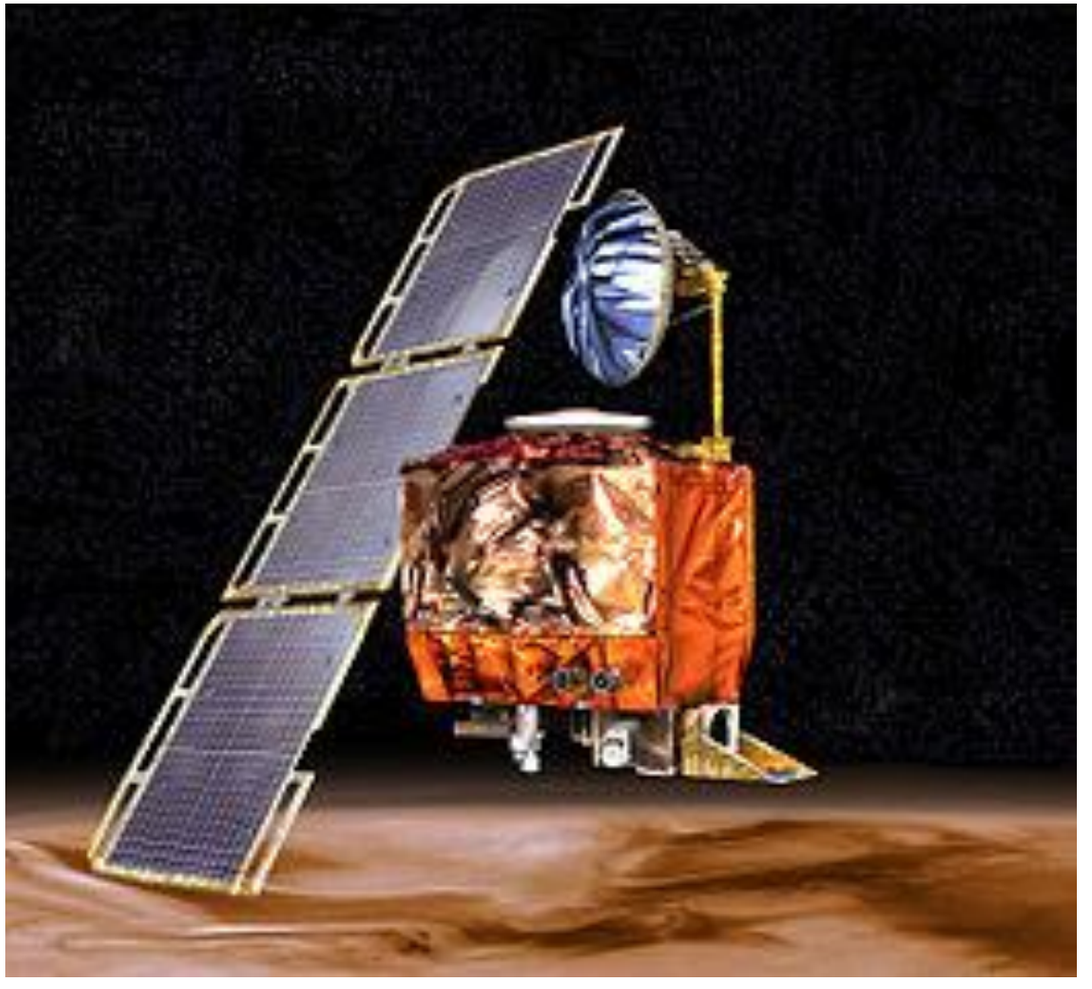
- Convert numerical data into categorical data

```
breaks <- c(0, 10000, 50000,100000,250000, 1000000)
Income.groups <- custdata$Income, breaks=breaks, include.lowest = T)
Income.groups <- as.character(Income.groups)
Income.groups <- ifelse(is.na(Income.groups), “No Income”, Income.groups)
```

- Assign zeros to the missing data

```
custdata$missingIncome<-is.na(custdata$income)
custdata$income.fix<-ifelse(is.na(custdata$income),0,custdata$income)
```

- Data entry problem
- Logical error
- Outdated
- Different standard



Data Range

```
summary(custdata$income)
```

Stephen Few

Numbers have an important story to tell. They rely on you to give them a clear and convincing voice.

‘Rogue train’ to blame for signal interference, disruptions on Circle Line

The train had faulty signalling hardware that affected the communications of other trains travelling in its vicinity, said LTA and SMRT on Friday (Nov 11).

By Lianne Chia **Posted** 11 Nov 2016 11:36 **Updated** 11 Nov 2016 22:39

VIDEOS PHOTOS

Circle Line disruption at Bishan Station. (Photo: Calvin Hui)

SINGAPORE: It was a rogue train with faulty signalling hardware that was the cause of the wireless signal interference on the Circle Line, revealed the Land Transport Authority (LTA) and train operator SMRT on Friday (Nov 11).

The intermittent hardware failure between Nov 2 and Nov 6 caused about 100 occurrences of loss of signalling communications on trains travelling in the proximity of the train, identified as Passenger Vehicle 46 (PV46).

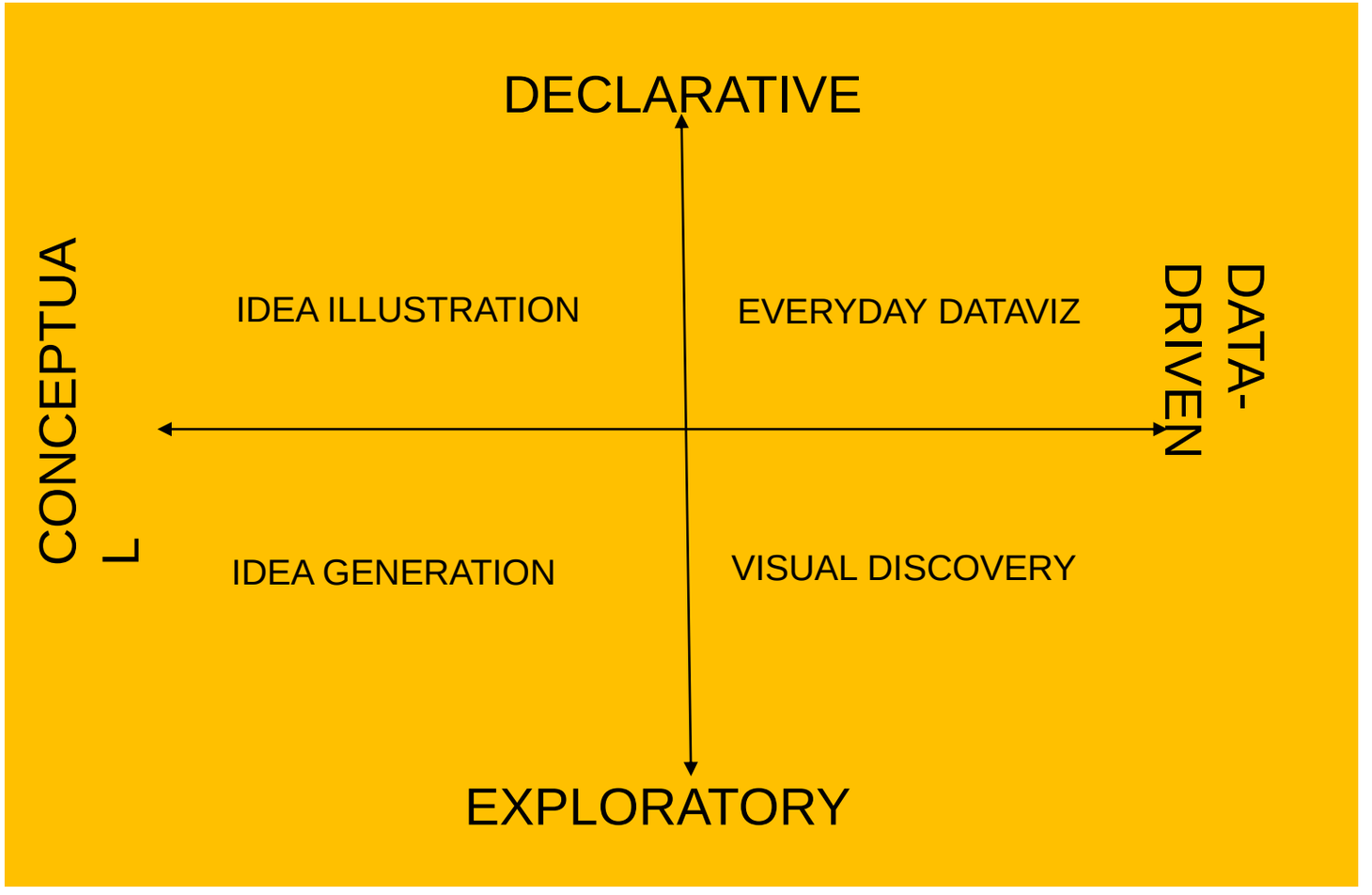
- https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a#.epr8cuq0d

Not long ago, the ability to create smart data visualizations, or dataviz, was a nice-to-have skill. For the most part, it benefited design- and data-minded managers who made a deliberate decision to invest in acquiring it. That’s changed. Now visual communication is a must-have skill for all managers, because more and more often, it’s the only way to make sense of the work they do.

Scott Berinato

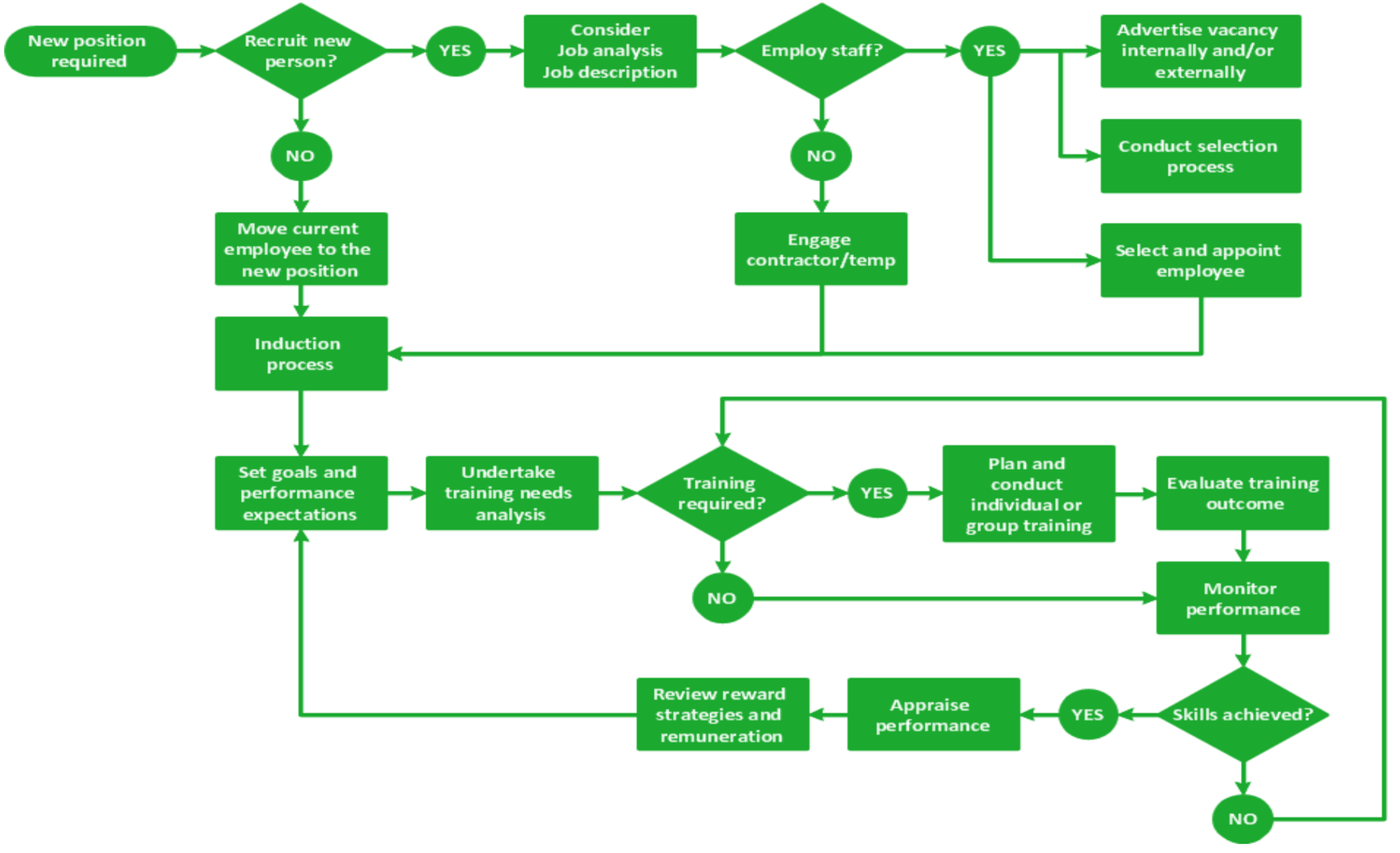
Senior editor at Harvard Business Review

Four types of data visualization



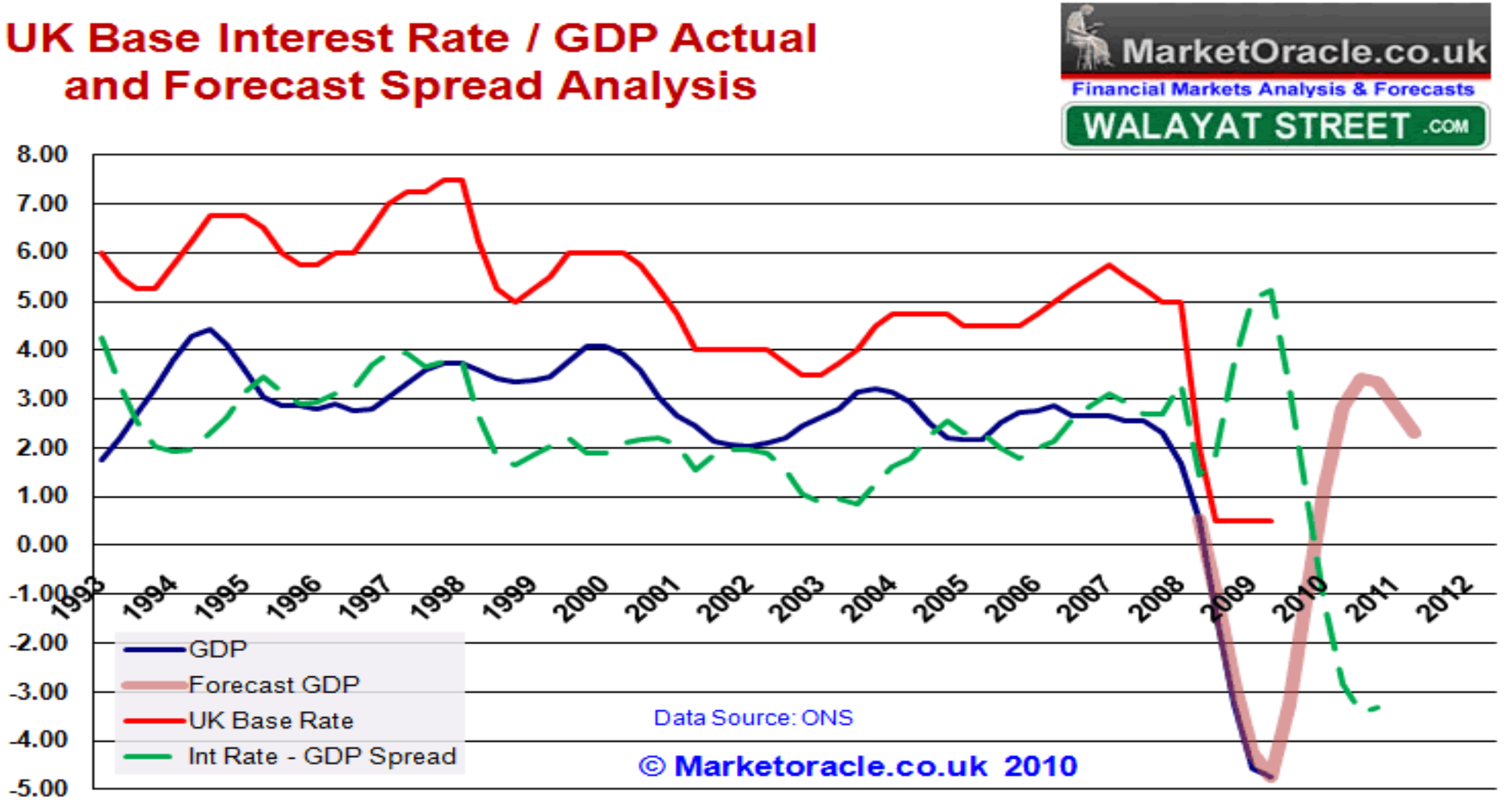
Source: “Visualizations that really work”, Harvard Business Review, June 2016

Idea illustration



Idea Generation

Daily Dataviz



Visual Discovery

- https://www.newyorkfed.org/data-and-statistics/data-visualization/index.html

8 Core principles of data visualization – Stephen Few

Simplify	• Good data visualization captures the essence of data
Compare	• We need to be able to compare our data visualization side by side
Attend	• The tool needs to make it easy for us to attend to the data that's really important.
Explore	• It should allow us explore data and discover things
View Diversely	• Different views of the same data provide different insights
Ask Why	• More than knowing "What's happening", we need to know "Why it's happening".
Be Skeptical	• Don't simply accept the first answer we get
Respond	• Be able to share our data that leads to global enlightenment

- https://www.youtube.com/watch?v=Nrm5ubKuGmw

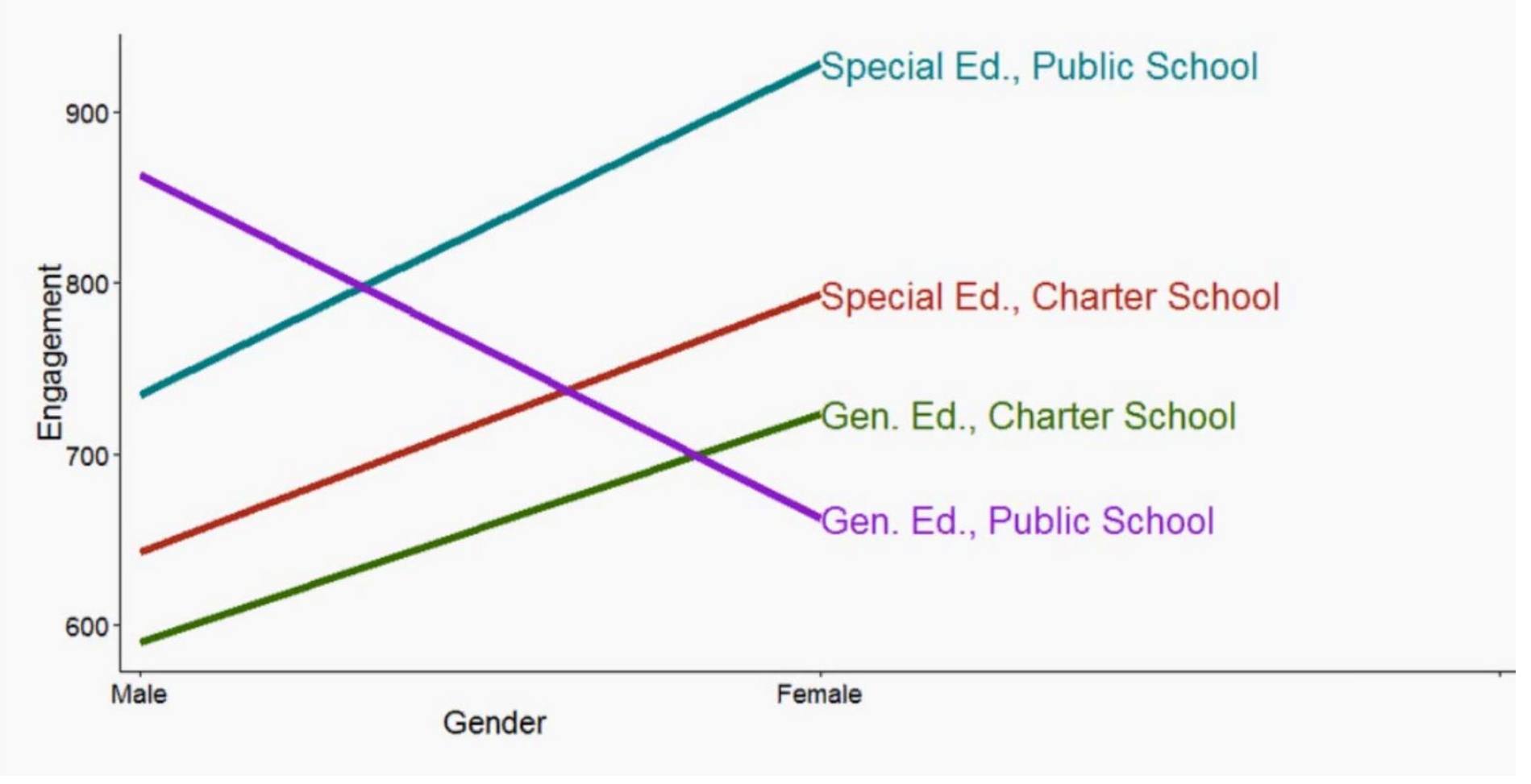

Fire
life

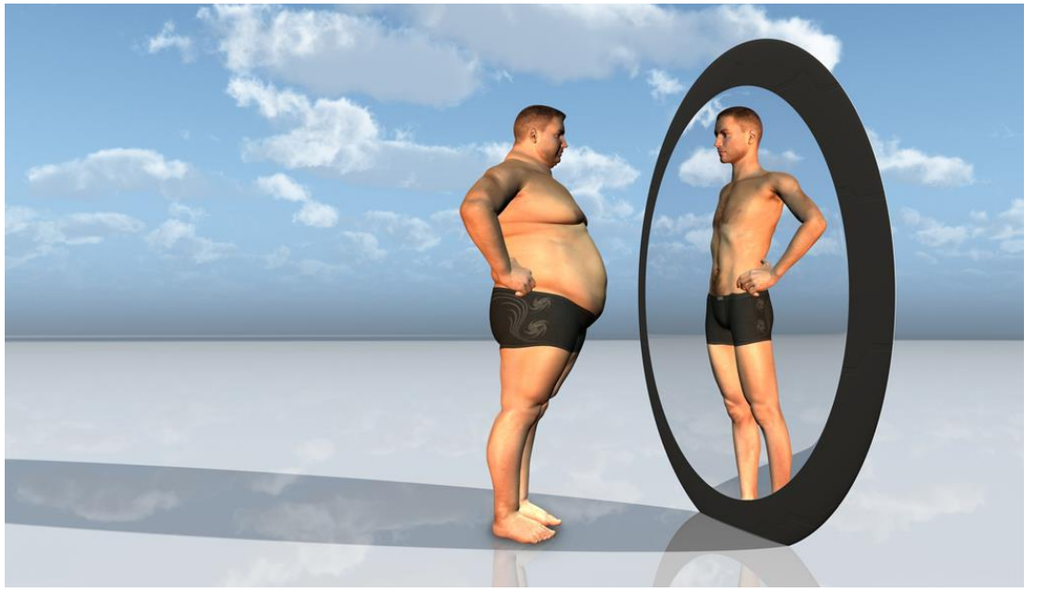
Those born in odd days close your eyes

Average Student Engagement Level

	Special Ed.		Gen. Ed.	
School Type	Male	Female	Male	Female
Charter	643	793	590	724
Public	735	928	863	662

Those born in even days close your eyes





Visualizing One Variable

Distribution for a single variable

- What is the peak value of the distribution?
- How many peaks are there in the distribution
- How normal is the data?
- How concentrated or dispersed is the data?

Basic charts for distribution

- ggplot(custdata) + geom_histogram(aes(x=age))

- ggplot(custdata) + geom_density(aes(x=age))

- ggplot(custdata) + geom_boxplot(aes(x="age",y=age))

Make it a little nicer

- ggplot(custdata) + geom_histogram(aes(x=age), fill='blue')
- ggplot(custdata) + geom_histogram(aes(x=age), colour='blue')
- ggplot(custdata) + geom_boxplot(aes(x="age",y=age), outlier.colour='red')
- ggplot(custdata) + geom_boxplot(aes(x="age",y=age), width=0.5)

Distribution of categorical variable

- ggplot(custdata) + geom_bar(aes(x=marital.stat), fill='blue', width=0.5)

Another example

```
ggplot(custdata) + geom_bar(aes(x=state.of.res), fill="lightblue") + coord_flip()
```

Better view

```
#Obtain the statistics
stats<-as.data.frame(table(custdata$state.of.res))

#Change the default column names to  make them more meaningful
colnames(stats) <- c("state.of.res","count")

#Sort the data by count
statsf <- stats[order(-stats$count),]

#plot the bar chart
ggplot(statsf) + geom_bar(aes(x=state.of.res, y=count), stat="identity",
fill="lightblue")+ coord_flip()+ theme(axis.text.y=element_text(size = rel(0.8)))
```

Relationship between Two Variables

Relationship between variables

- Is there relationship between two variables?
- Is the relationship strong?
- Is the relationship linear or not linear?

Line graph

```
y<-runif(100)

x<-qnorm(y)

ggplot(data.frame(x=x,y=y),aes(x=x,y=y) )+geom_line()
```

Scatter plot

```
ggplot(custdata,aes(x=age,y=income)) + geom_point() + ylim(-5000,200000)
```

```
custdata2 <- subset(custdata, (custdata$age>0 & custdata$age<100 & custdata$income>0))
ggplot(custdata2,aes(x=age,y=income)) + geom_point() + ylim(0,200000)
```

```
ggplot(custdata2,aes(x=age,y=income)) + geom_point() + ylim(0,200000) + stat_smooth(method="lm")
```

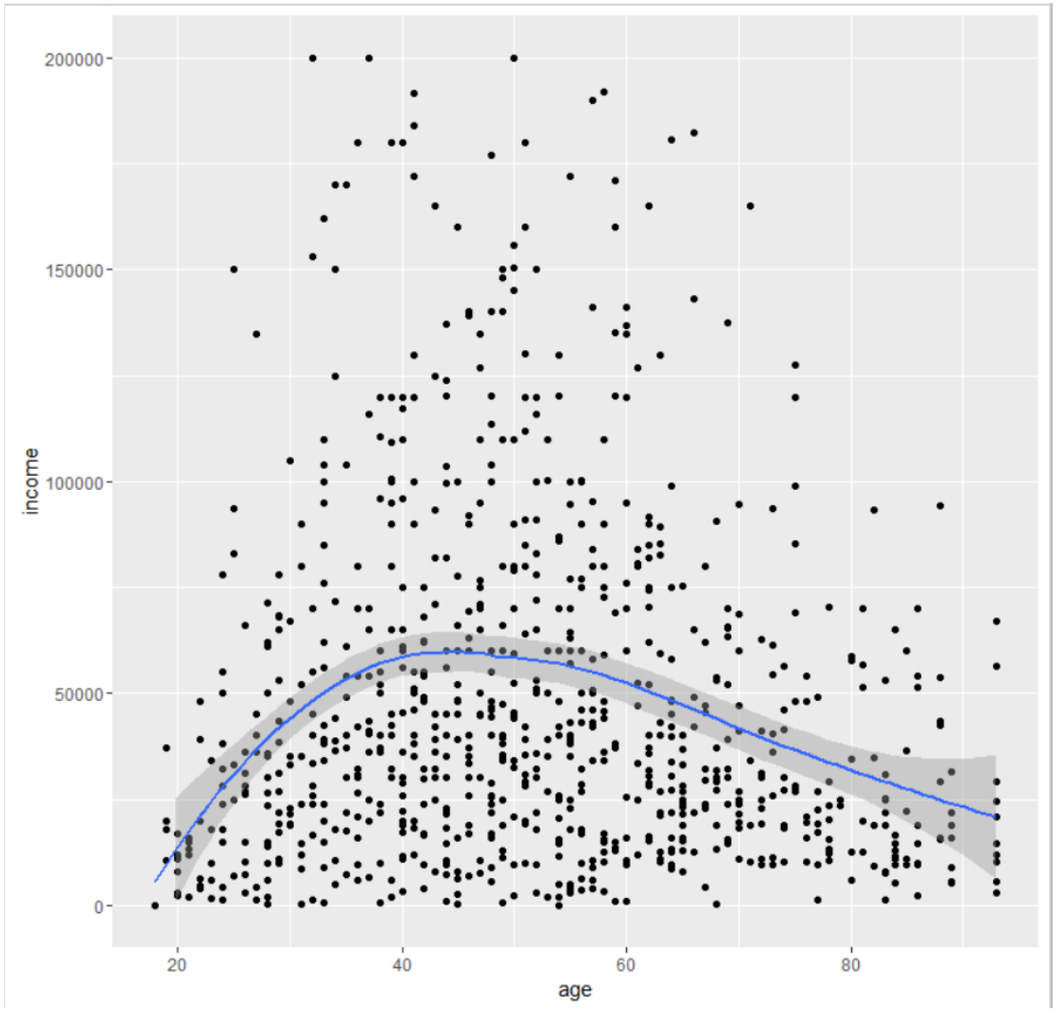
```
ggplot(custdata2,aes(x=age,y=income)) + geom_point() + ylim(0,200000) + geom_smooth()
```

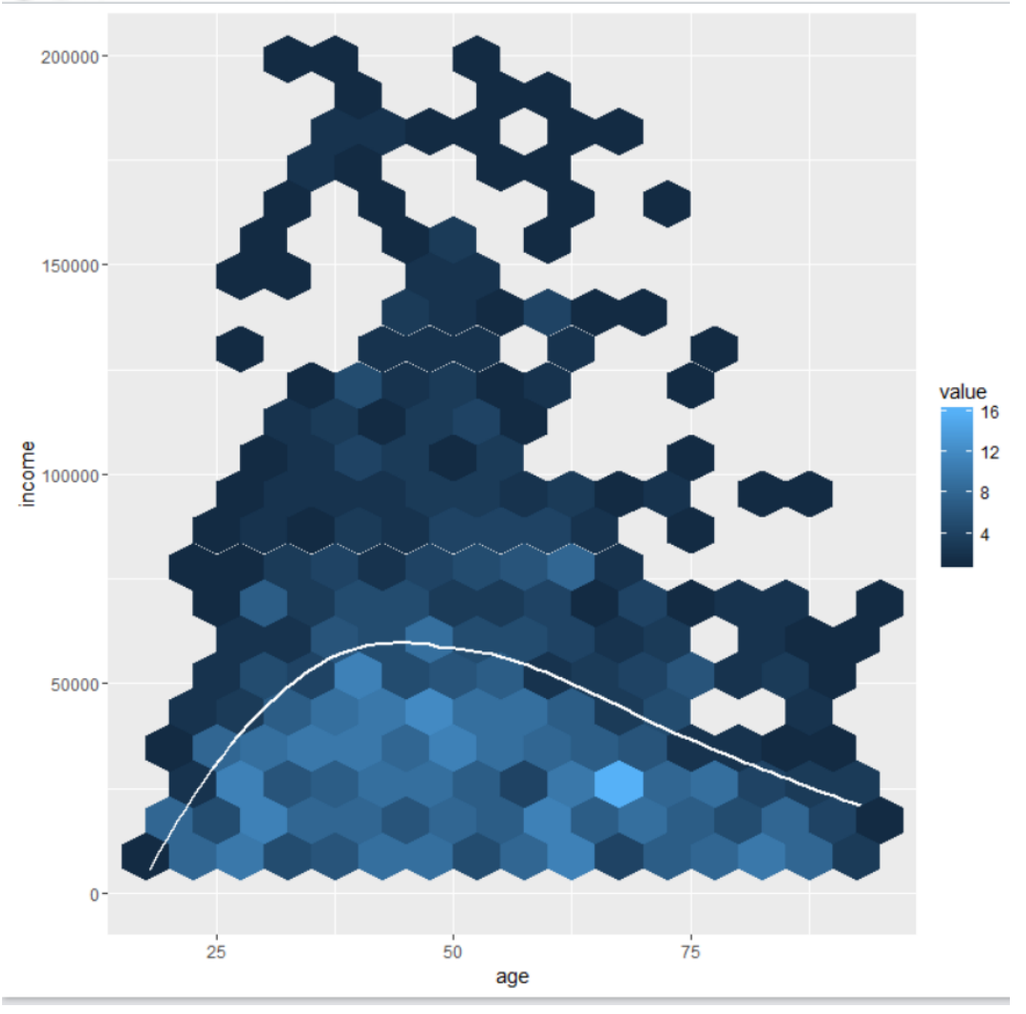
```
ggplot(custdata2, aes(x=age, y=as.numeric(health.ins))) +
geom_point(position=position_jitter(w=0.05,h=0.05)) + geom_smooth()
```

Enhancement of scatter plot

```
library(hexbin)

ggplot(custdata2, aes(x=age,y=income)) + geom_hex(binwidth=c(5,10000)) +
geom_smooth(color="white",se=F) + ylim(0,200000)
```

Comparison


200000
150000
100000
50000
0
income
20
40
60
80
age


200000
150000
100000
50000
0
income
25
50
75
age
value
16
12
8
4

Relationship between Categorical Variables

```
ggplot(custdata) + geom_bar(aes(x=marital.stat,fill=health.ins))
```

```
ggplot(custdata) + geom_bar(aes(x=marital.stat,fill=health.ins),position = "dodge")
```

```
ggplot(custdata) + geom_bar(aes(x=marital.stat,fill=health.ins),position = "fill")
```

```
ggplot(custdata, aes(x=marital.stat)) + geom_bar(aes(fill=health.ins), position="fill") +
geom_point(aes(y=-0.05), size=0.75,alpha=0.3,position=position_jitter(h=0.01))
```

Exercise

- Adding rug to bar chart is not a very straightforward visual comparison. Please find a way to use a combination of stacked bar chart and line chart to have a better simultaneous view of both the population in each category and the ratio of insured to uninsured.

Data Transformation

Why data transformation?

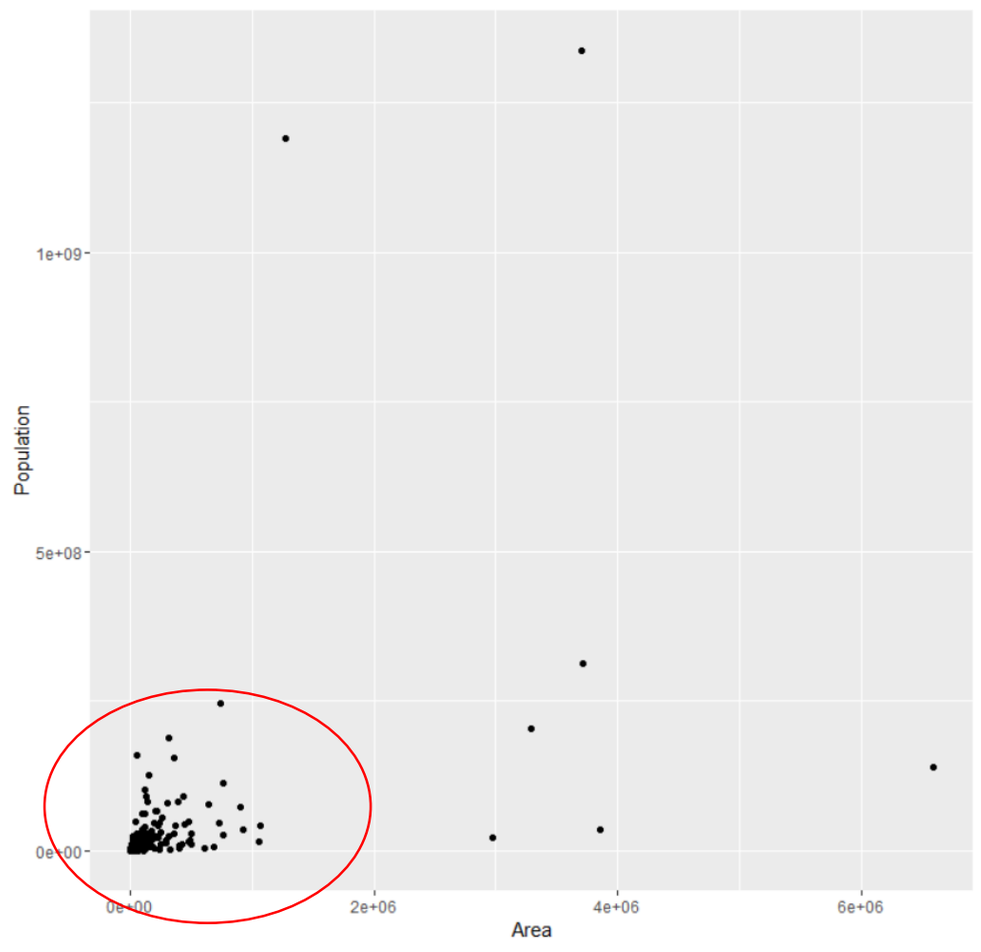
- Modelling needs
- Better visualization
- Better interpretation

World Population vs Area

```
library(XML)
library(RCurl)

theurl <- "http://www.infoplease.com/ipa/A0004379.html"
urldata <- getURL(theurl)
data <- readHTMLTable(urldata,  stringsAsFactors = FALSE)

df<-as.data.frame(data[1])
df<-df[-c(1,199),]
colnames(df) <- c("Country","Population","Area")
```

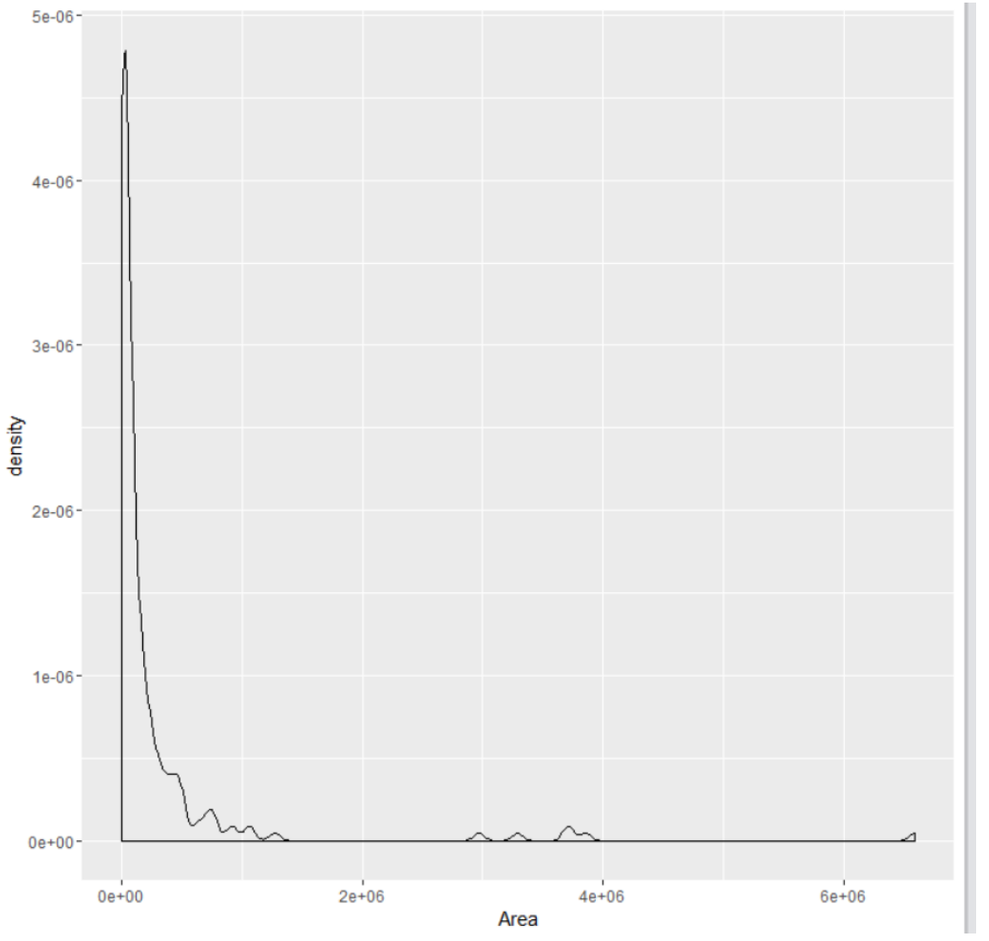
Visualizing data

```
df$Population <- gsub(",","",df$Population)
df$Population <- as.numeric(df$Population)
df$Area <- gsub(",","",df$Area)
df$Area <- as.numeric(df$Area)

ggplot(df,aes(x=Area,y=Population)) + geom_point()
```



Lognormally distributed Data

```
ggplot(df) + geom_density(aes(x=Area))
```



Wide applications of log normal

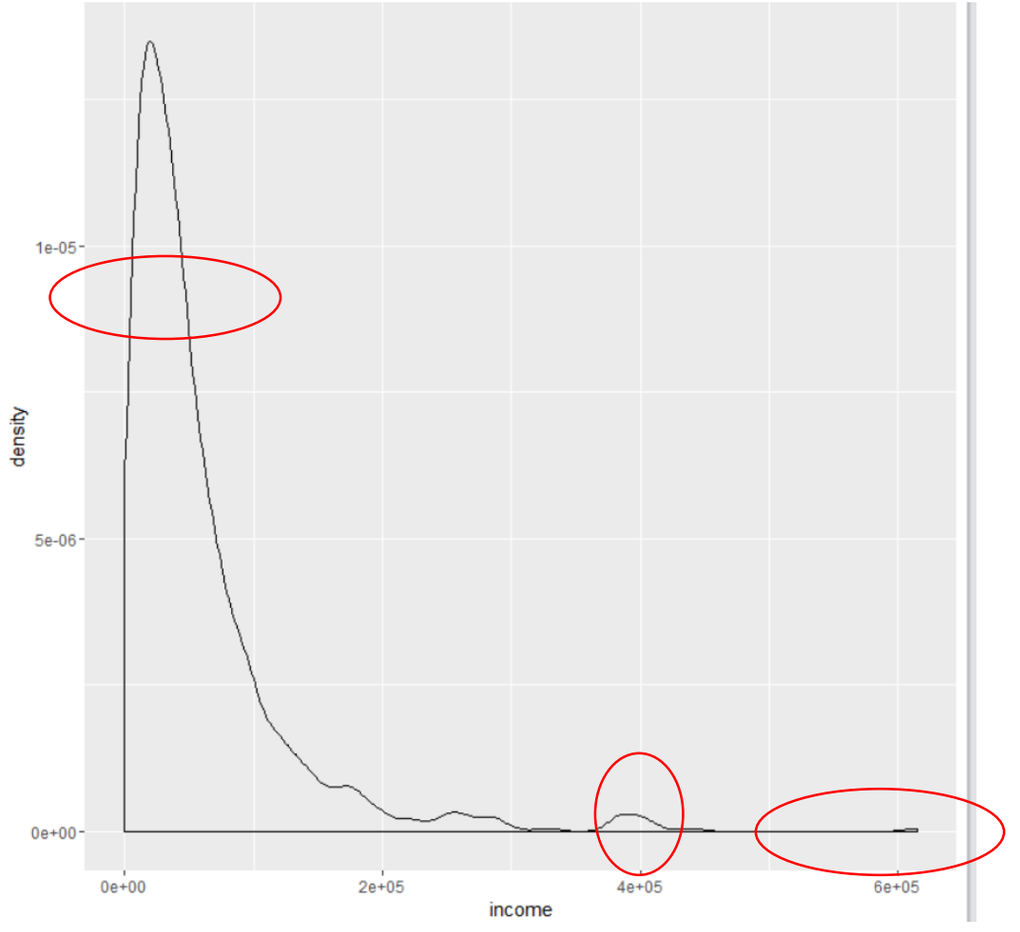
- The length of comments posted in Internet discussion forums follows a log-normal distribution.
- Measures of size of living tissue (length, skin area, weight)
- The income of 97%–99% of the population is distributed log-normally.
- City sizes
- Changes in the *logarithm* of exchange rates, price indices, and stock market indices are assumed normal

What do you see now?

```
signedlog10 = function(x) { + ifelse(abs(x) <= 1, 0, sign(x)*log10(abs(x))) + }

df$Log.Area <- signedlog10(df$Area)

df$Log.Population <- signedlog10(df$Population)

ggplot(df,aes(x=Log.Area,y=Log.Population)) + geom_point() + geom_smooth(method="lm")
```

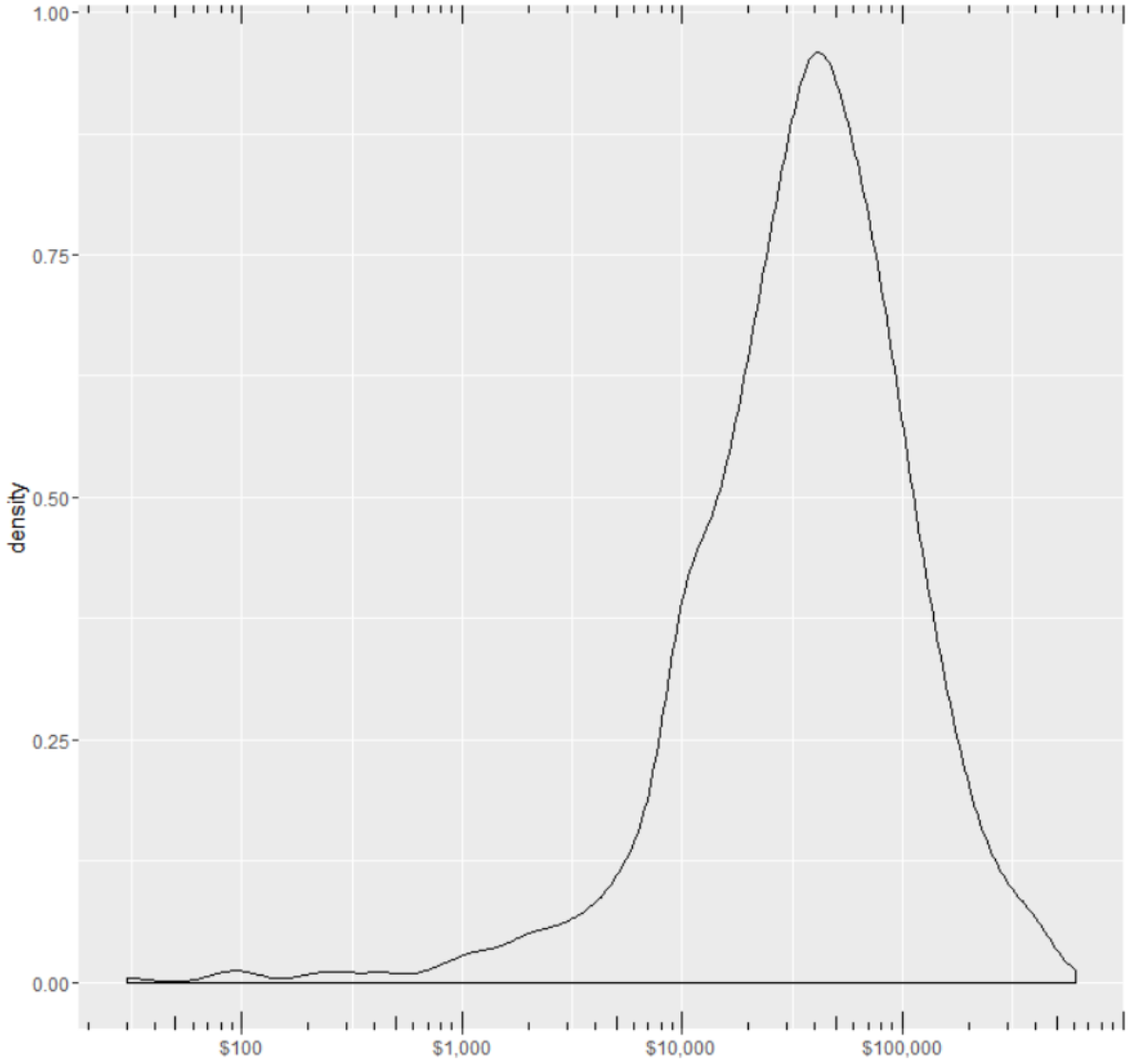
Visualizing Income

```
custdata<-custdata[custdata$income>0,]

ggplot(custdata) + geom_density(aes(x=income))
```



Visualizing Income

```
ggplot(custdata) + geom_density(aes(x=income)) + scale_x_log10(breaks=c(100,1000,10000,100000),labels=dollar) + annotation_logticks(sides="bt")
```

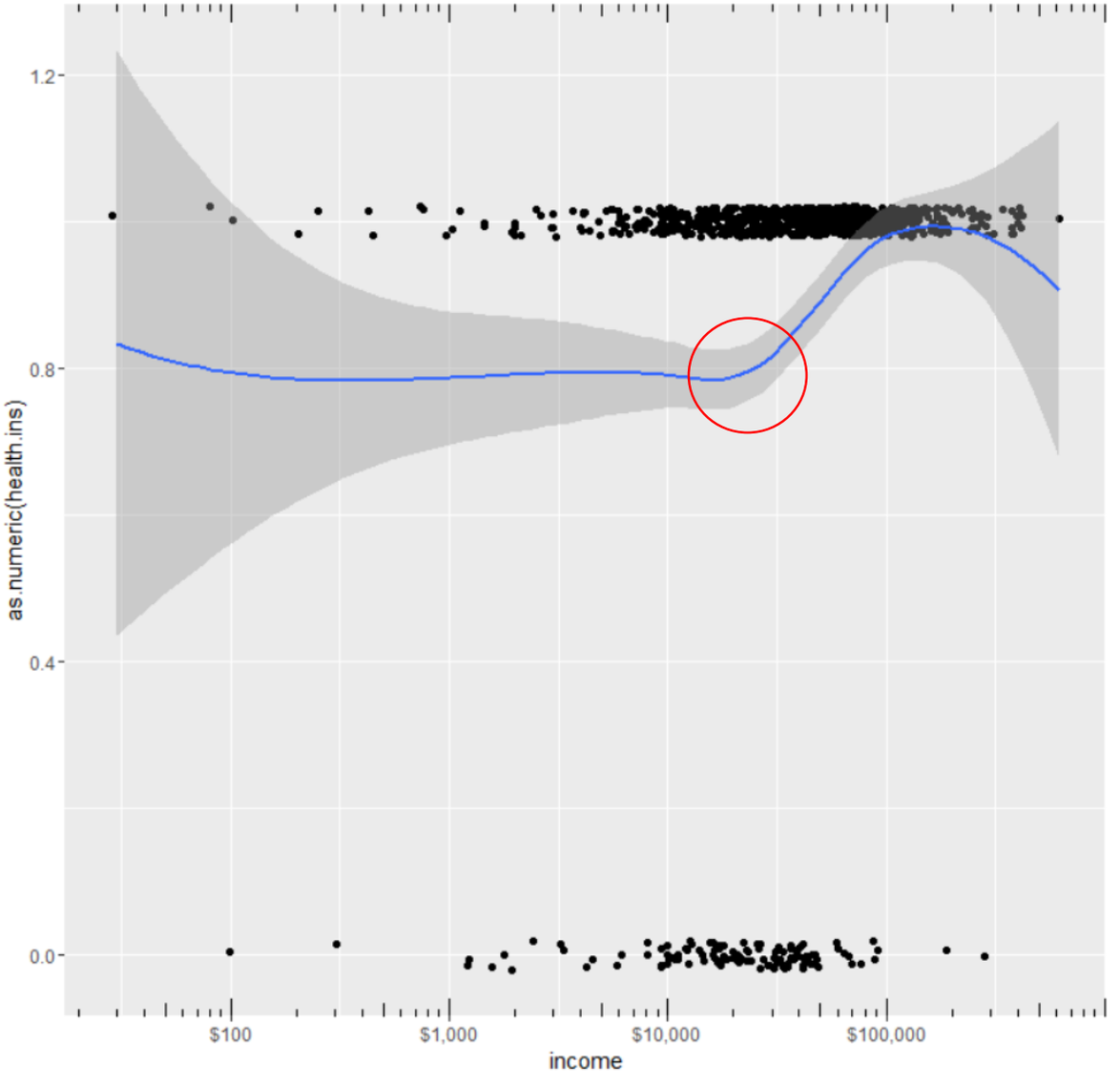


Scaling

- What is the difference between age group of “0-2” vs “20-22”?

Converting Continuous Variable into Categorical

```
ggplot(custdata2, aes(x=income, y=as.numeric(health.ins))) + scale_x_log10(breaks=c(100,1000,10000,100000), labels=dollar)
+ geom_point(position=position_jitter(w=0.05,h=0.05)) + geom_smooth() + annotation_logticks(sides="bt")
```



Converting Continuous Variable into Categorical

```
custdata$income.group <- ifelse(custdata$income < 20000, "LOW", "HIGH")
custdata$income.group <- as.factor(custdata$income.group)
```

Normalize Data – Why?

One day, I became millionaire!!!

Normalize Data – How?

- Download median income by state data from https://en.wikipedia.org/wiki/List_of_U.S._states_by_income

```
custdata<-merge(custdata, medianincome, by.x="state.of.res", by.y="State")
custdata$income.normalised <- with(custdata, income/Median.Income)
```