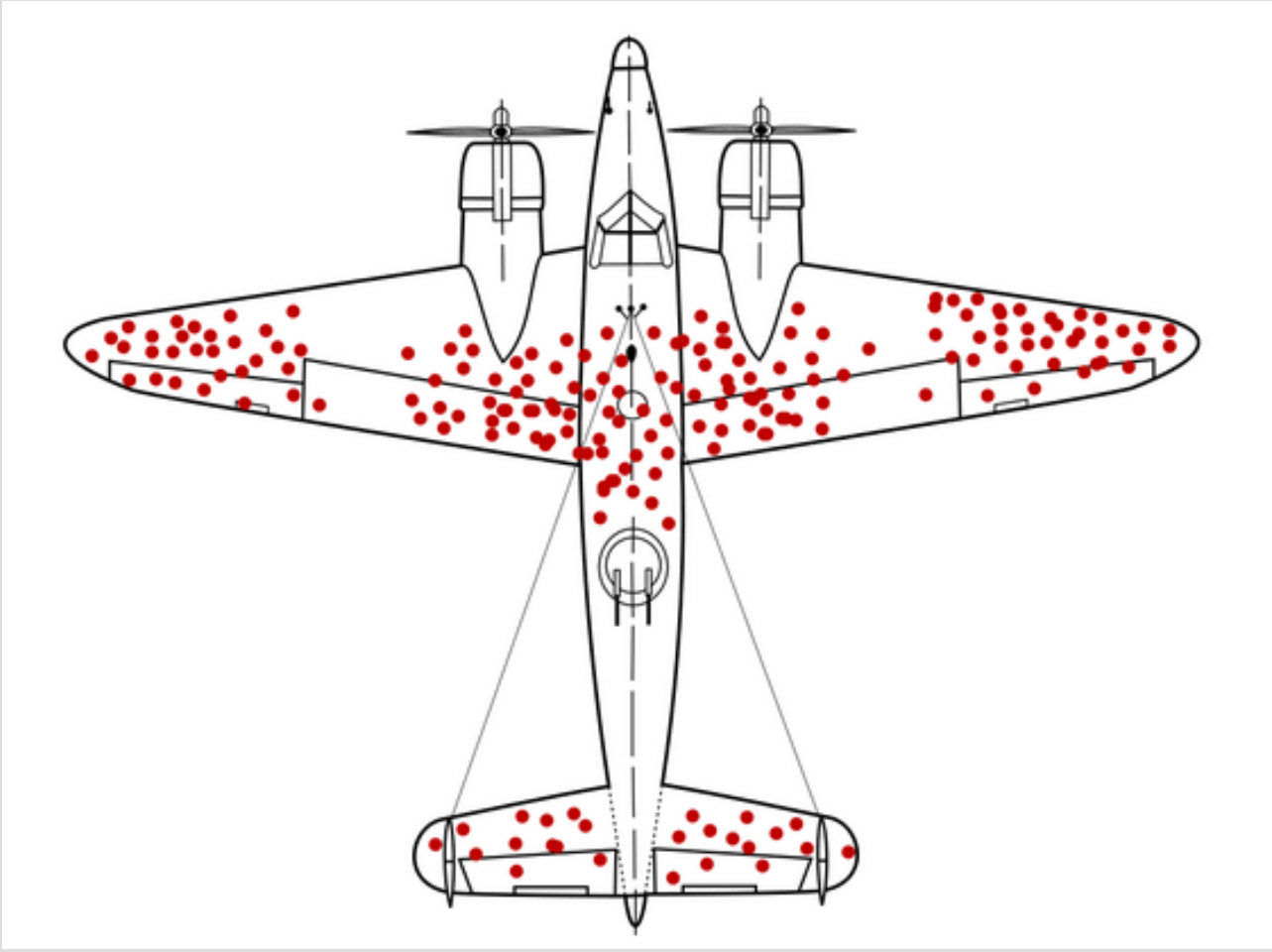
Data Fallacies, Tricks and Data Workshop



In WWII, some planes never came back, and some came back with bullet holes. Here is the distribution of bullet holes. How would you reinforce the plane to increase the survival rate?

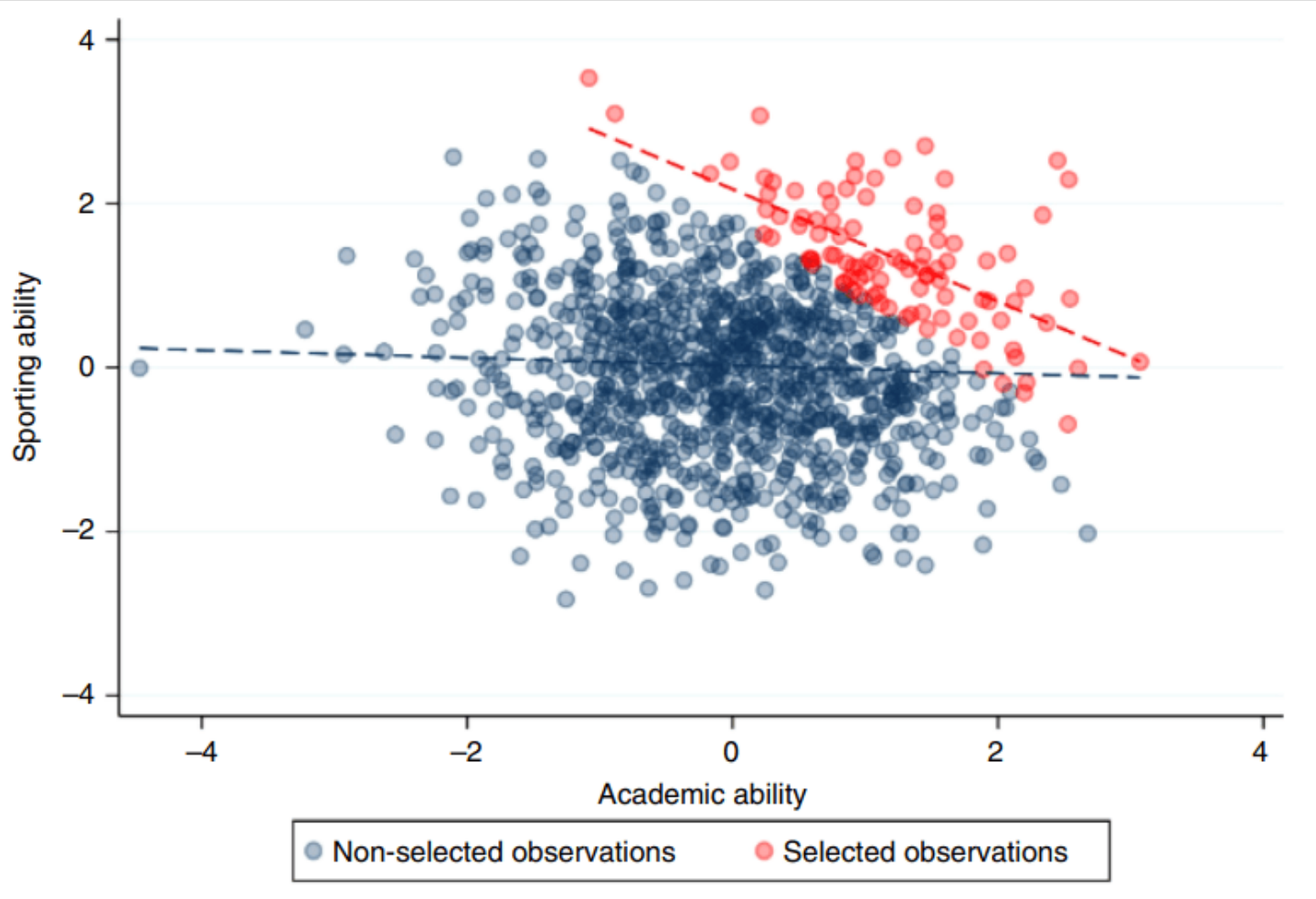
You should reinforce places where there are no holes, because if the plane will never come back when it has holes in these places! This is called the survivalship bias.

We only see part of the dataset. The other part is missing!

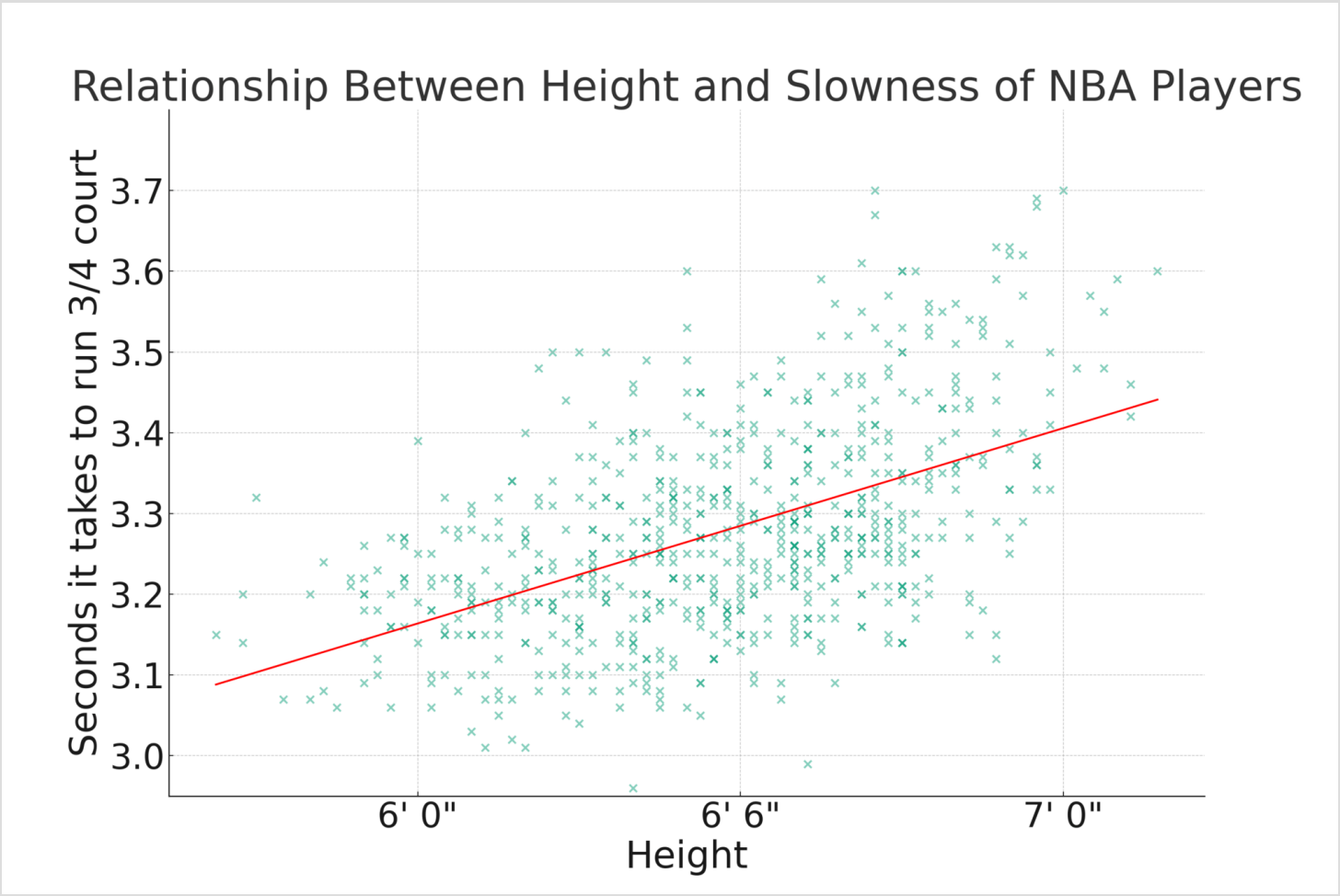
When you survey existing consumers, you don’t know why consumers are not buying your products. You don’t get honest opinion on your product development.

This is the survivalship bias: We only see successful ones in the data but do not see failed ones in the data.

Can you come up with other examples of survivalship bias?



Sporting ability is negatively correlated with academic ability.



Taller NBA players run more slowly.

David recently transferred from the Marketing programme to the Management programme. Consequently, the average IQ of students in both the Marketing and Management programmes has increased.

Why?

Simpson’s Paradox

You are making a comparison between two hospitals:

- Hospital A: Among each 1,000 patients, 900 survived.
- Hospital B: Among each 1,000 patients, 800 survived.

Which hospital will you choose, and why?

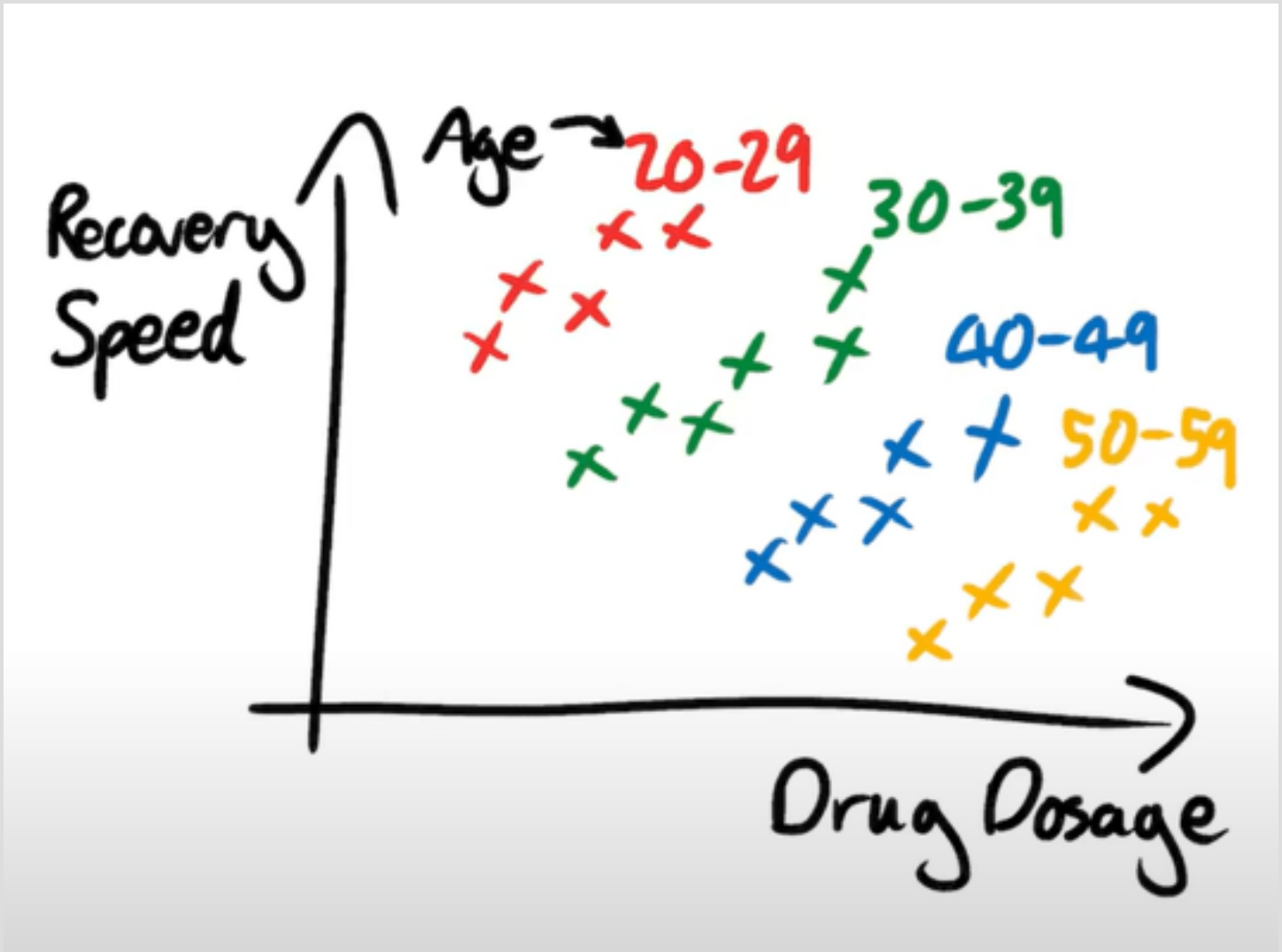
Let us take a closer look at the data...

- Hospital A has 100 severe patients, among them 30 survived. It has 900 mild patients, among them 870 survived.
- Hospital B has 400 severe patients, among them 210 survived. It has 600 mild patients, among them 590 survived.

Location	Total Price (Billion)	Units	Average Price (MM)
Suburban	$10	3,500	2.9
Downtown	$100	21,000	4.8
Total	$110	24,500	4.5

Location	Total Price (Billion)	Units	Average Price (MM)
Suburban	$50	15,000	3.3
Downtown	$100	20,000	5.0
Total	$150	35,000	4.3

Do patients recovery more slowly when taking more drugs?



Again, Simpson's Paradox

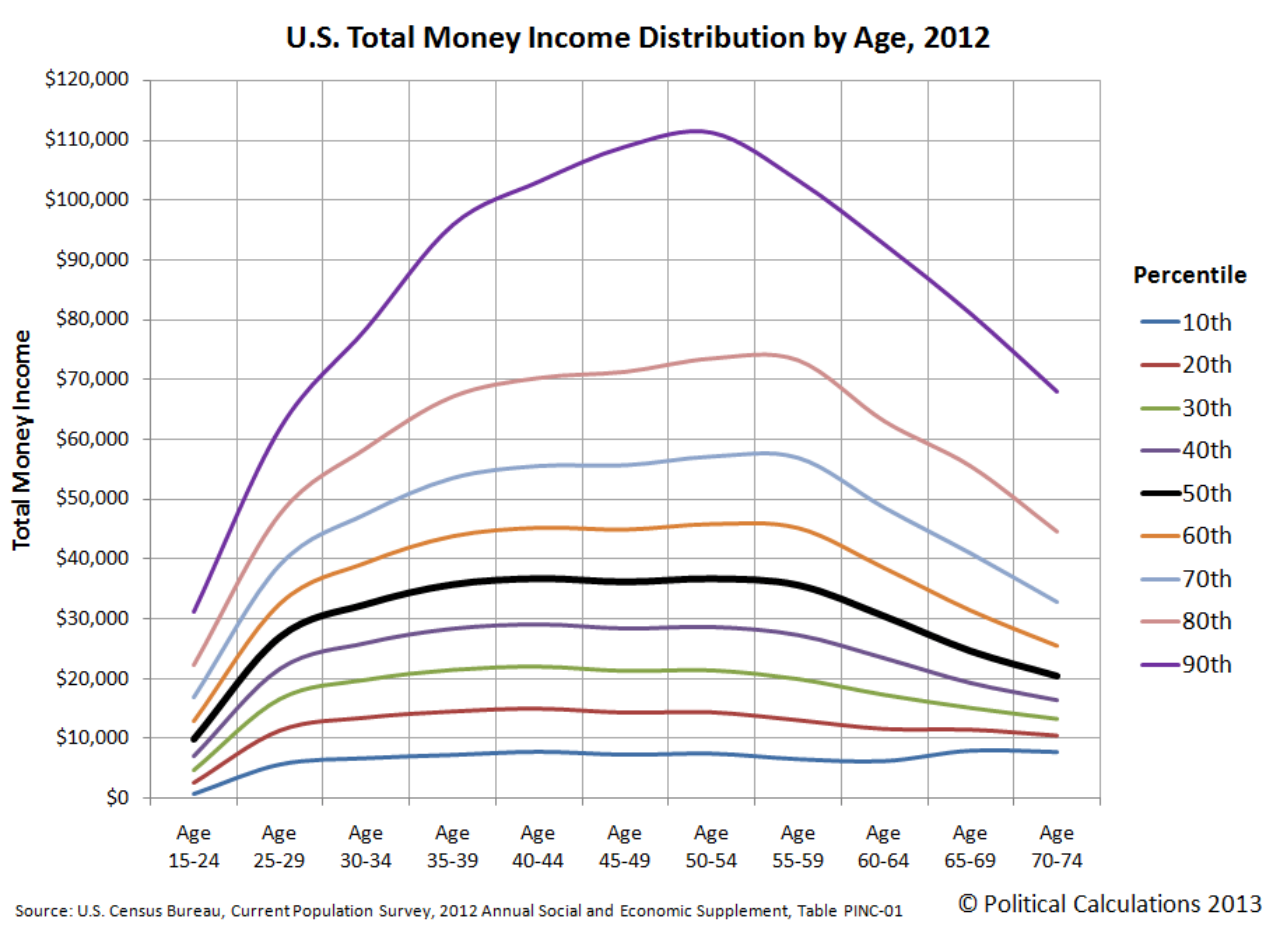
https://www.youtube.com/embed/ebEkn-BiW5k?enablejsapi=1

Demonstration 示例

```r
library(ggplot2)
library(dplyr)
mydata <-
read.csv("https://ximarketing.github.io/class/simpson.csv",
                    fileEncoding = "UTF-8-BOM")
head(mydata)
mydata %>% ggplot(aes(x= Price, y= Demand))+geom_point()+
geom_smooth(method='lm')
mydata %>% ggplot(aes(x= Price, y= Demand, group= Country, col=
Country))+geom_point()+ geom_smooth(method='lm', col='black')
result = lm(Demand ~ Price, data = mydata)
summary(result)
result = lm(Demand ~ Price + factor(Country), data = mydata)
summary(result)
```

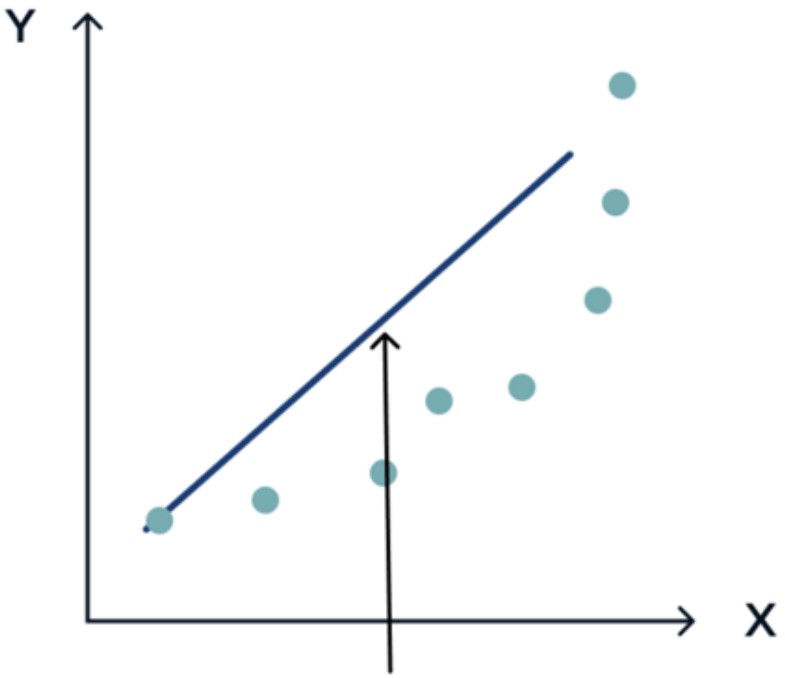
Data Tricks

How does income change with age?



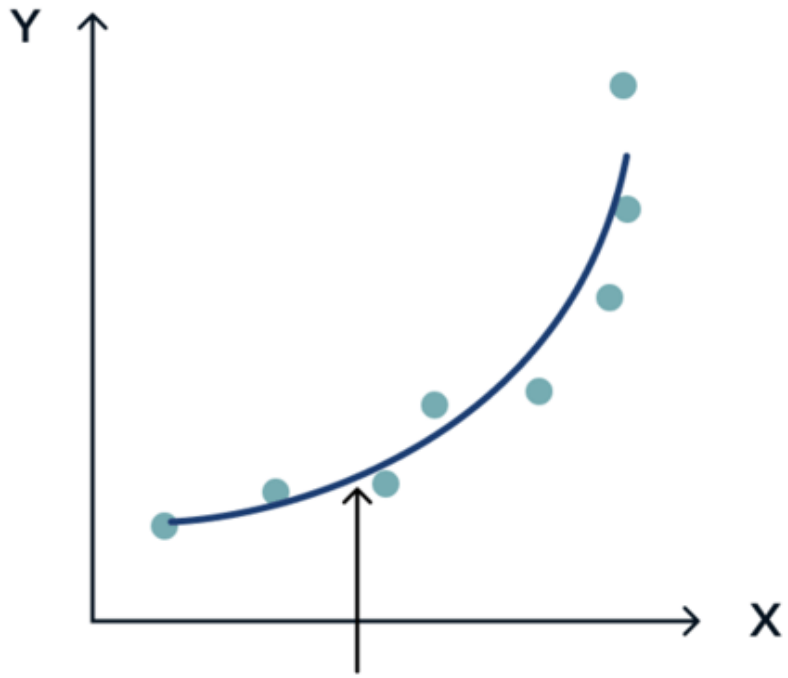
If you run a linear regression, you find that income either increases or decreases with age. But this does not capture the nonlinear relationship between the two variables. What should you do in this scenario?

Simple linear model



$$y=b_0+b_1x$$

Polynomial model



$$y=b_0+b_1x+b_2x_1^2$$

Quadratic Regression

Suppose that we want to see how Y changes nonlinearly with X, we can run the following quadratic regression (as opposed to linear regression):

$$Y = a + b_1 X + b_2 X^2$$

You can further extend the model to include cubic terms etc.

Crowdfunding: An Example

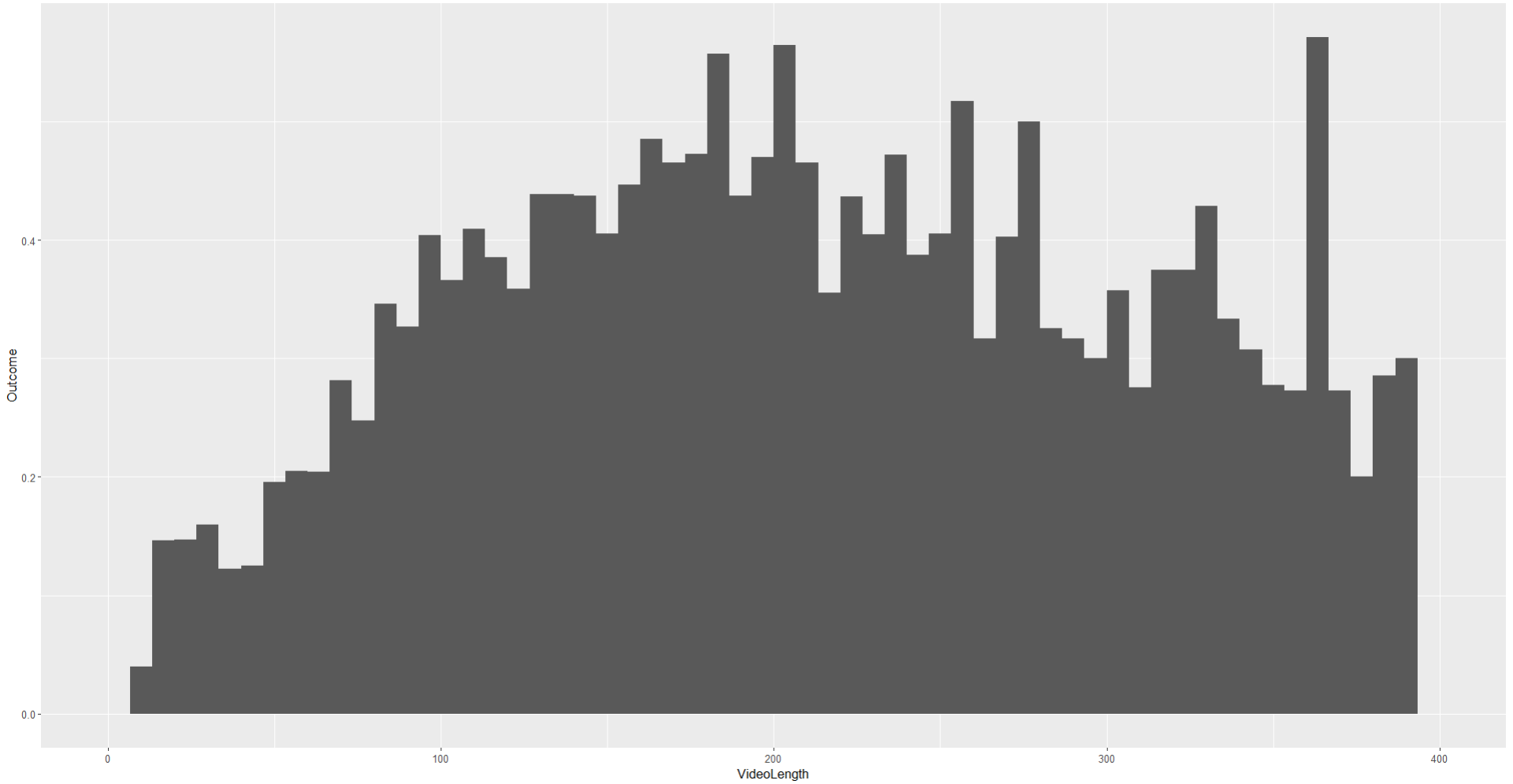
We want to investigate the relationship between video length and the chance of success. Let us prepare the data:

```
1 mydata <-
  read.csv("https://ximarketing.github.io/class/Kickstarter-
  Project.csv", fileEncoding = "UTF-8-BOM")
2 subdata = subset(mydata, IsVideoAvailable == 1)
```

Crowdfunding: An Example

We want to investigate the relationship between video length and the chance of success. Let us prepare the data:

```
1 library(ggplot2)
2 ggplot(subdata,  mapping = aes(VideoLength, Outcome)) +
3 stat_summary_bin(fun.y="mean", geom="bar", bins=60)+xlim(0,
  400)
```


Outcome
0.4
0.2
0.0
0
100
200
300
400
VideoLength

Crowdfunding: An Example

The relationship between video length and project success appears to be nonlinear. Shorter videos can enhance the success rate as their length increases; however, excessively long videos do not provide additional benefits to the project.

Crowdfunding: An Example

Let us try the following logistic regression:

$$\Pr[\text{Success}] = \frac{1}{1 + \exp(-(a + b_1 \times \text{Length} + b_2 \times \text{Length}^2))}$$

Consider the following code:

```
1 logit <- glm(Outcome ~ VideoLength + I(VideoLength^2), data =
  subdata, family = "binomial")
2 summary(logit)
```

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.192e+00  8.958e-02 -13.307  < 2e-16 ***
VideoLength        6.541e-03  8.102e-04   8.074 6.81e-16 ***
I(VideoLength^2) -1.056e-05  1.584e-06  -6.666 2.63e-11 ***
```

Question: What is the optimal video length?

Optimal video length

Recall your high-school mathematics: A quadratic function $f = b_2 x^2 + b_1 x + a, b_2 < 0$ is maximized when

$$x = -\frac{b_1}{2b_2}$$

In our example, $b_2 = -1.056 \times 10^{-5}$ and $b_1 = 6.541 \times 10^{-3}$. With a little bit calculation you can find out that the optimal video length is about 300 seconds (i.e., 5 minutes).

Question

Suppose that you want to predict students' performance in exam. Two factors come into play: IQ and Hours of Study.

- A student with a higher IQ is more clever, and gets higher grades on average.
- A student who studies longer hours understands the content better, and gets higher grades on average.

Question

Let's run the following linear regression:

$$\text{Grades}_i = a + b_1 \times \text{IQ}_i + b_2 \times \text{Hours}_i$$

Is anything missing from the regression?

Question

Consider two types of students: High IQ students and low IQ students. High IQ students are clever, and they study more efficiently. That is, when a high IQ student studies for one hour, they learn more than a low IQ student who also studies for one hour.

How to incorporate this into our regression model?

Question

We consider the interaction between IQ and Hours of Study:

$$\text{Grades}_i = a + b_1 \times \text{IQ}_i + b_2 \times \text{Hours}_i + b_3 \times \text{IQ}_i \times \text{Hours}_i$$

Suppose that you find out $b_3 > 0$, what does it imply?

Interaction Effects

$$\text{Grades}_i = 10 + 0.2 \times \text{IQ}_i + 4 \times \text{Hours}_i + 0.01 \times \text{IQ}_i \times \text{Hours}_i$$

- Alice has an IQ 120. If she studies 8 hours, she will get 75.6. If she studies 9 hours, she will get 80.8. For Alice, one extra hour of study improves her grades by 5.2.
- Bob has an IQ 80. If he studies 8 hours, he will get 64.4. If he studies 9 hours, he will get 69.2. For Bob, one extra hour of study improves his grades by 4.8.
- Alice studies more efficiently than Bob!

Interaction Effects

Suppose that your dependent variable is a programmer's salary. You have two independent variables: the programmer's knowledge of Python and his/her knowledge of R. You find that

$$\text{Salary}_i = 1 + 3 \times \text{Python}_i + 2 \times \text{R}_i - 0.5 \times \text{Python}_i \times \text{R}_i$$

How would you interpret this result?

Interaction Effects

$$\text{Salary}_i = 1 + 3 \times \text{Python}_i + 2 \times \text{R}_i - 0.5 \times \text{Python}_i \times \text{R}_i$$

If you know more about Python, you can make a higher salary.

If you know more about R, you can maker a higher salary.

However, if you already know Python well, then knowing more about R does not help much, and vice versa.

This result suggests that Python and R are substitutes: After learning about one language, learning about the other does not help you much.

Interaction Effects

Suppose that your dependent variable is a person's health score. You have two independent variables: the amount of swimming and running.

$$\text{Health}_i = 4 + 5 \times \text{Running}_i + 3 \times \text{Swimming}_i + 2 \times \text{Running}_i \times \text{Swimming}_i$$

How would you interpret this regression result?

Interaction Effects

Suppose that your dependent variable is a person's health score. You have two independent variables: the amount of running exercise and whether or not the person is overweight.

$$\text{Health}_i = 4 + 5 \times \text{Running}_i - 2 \times \text{Overweight}_i + 3 \times \text{Running}_i \times \text{Overweight}_i$$

How would you interpret this regression result?

A Crowdfunding Example

We want to investigate the relationship between funding outcome, the creators' experience and the crowdfunding video.

```
result = glm(Outcome ~ Created * IsVideoAvailable,
             family = "binomial", data = mydata)
summary(result)
```

A Crowdfunding Example

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.6976     0.1074 -25.115  < 2e-16 ***
Created                      0.3339     0.0577   5.786  7.2e-09 ***
IsVideoAvailable             2.0406     0.1121  18.207  < 2e-16 ***
Created:IsVideoAvailable    -0.1334     0.0606  -2.201   0.0277 *
```

What do you learn from the results?

Group Data Project:
HK Property Valuation

Most of you are renting a flat in HK.

Do you know the selling price of your flat?

It’s all available here!

How much has Hong Kong housing price increased since 1997?

Please make your guess!

中原城市領先指數 CCL

本週公佈 | 較上週 | 較上月

145.01 | ↑ 1.09% | ↑ 1.77%

每週五公佈 — 最新2025/12/19公佈，反映2025/12/08至2025/12/14(預計簽署正式買賣合約時段)的二手私人住宅樓價。一般在簽署臨時買賣合約後14日內簽署正式買賣合約

1997年7月第1週指數為100點

查詢過往數據

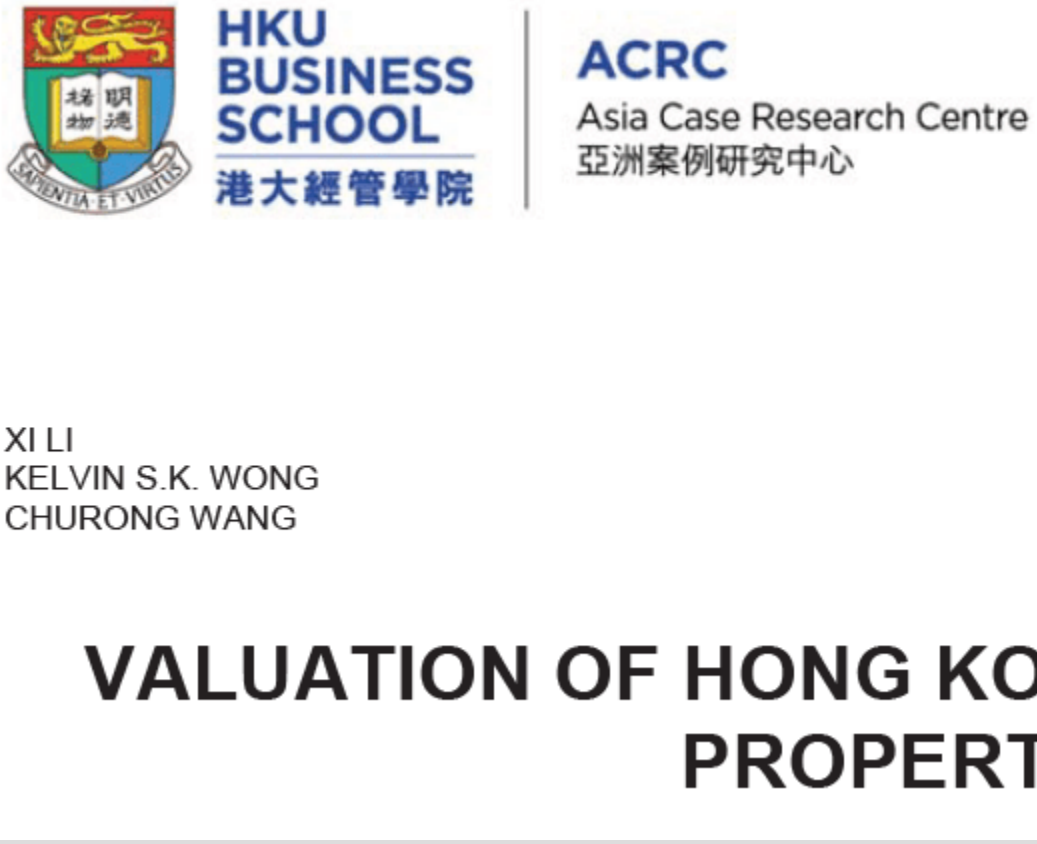
About 45%

Guess

In Hong Kong, what do people care about most when buying or renting a flat?

Valuation of Hong Kong Residential Property

In this project, we want to understand the HK real estate market. We have collaborated with Centaline (中原地產), one of the largest property agencies in Hong Kong, to get the property transaction data in Hong Kong.



Kelvin Wong is Professor of Real Estate at the University of Hong Kong. Churong Wang is currently my PhD student.

We will be using a data platfrom for your data project

- Please sign up an account at http://acrc.hku.hk/ using your HKU email address.
- Please add your coursepack using the link https://www.acrc.hku.hk/enrol/1000012200

Valuation of Hong Kong Residential Property

Loading the data:

```
1 mydata = read.csv('/dataset/Centaline/Centaline_train.csv',header=TRUE)
```

Valuation of Hong Kong Residential Property

Transaction_price: The transaction price of the property (in Hong Kong dollars). You may want to take the log transformation of this variable to analyze its percentage change.

Why do we take the log transformation?

```
mydata$LogPrice = log(mydata$Transaction_price)
```

Valuation of Hong Kong Residential Property

```
mydata$LogPrice = log(mydata$Transaction_price)
hist(mydata$Transaction_price, breaks = 100, xlim = c(0,
1e+08))
hist(mydata$LogPrice, breaks = 100)
```

Valuation of Hong Kong Residential Property

Transaction_year: The year in which the transaction takes place (e.g., 2020).

Transaction_month: The month in which the transaction takes place (e.g., 10 for October). When using this variable, you may want to take it as a fixed effect.

Valuation of Hong Kong Residential Property

Location and Estate: The location and estate for each property. Please do not use them in your data analysis.

HMA: It stands for "Housing Market Area", a term used to describe the area at which the property is located (e.g., Pok Fu Lam).

Developer: The developer of the property (e.g., Hang_Lung_Group for 恆隆集團). If the developer is a small developer not included in the dataset, then the value is "Other".

Valuation of Hong Kong Residential Property

Gross_size: 建築面積 in Chinese. It is measured in square foot. If data is unavailable for a property, then its Gross_size = -1.

Saleable_size: 使用面積 in Chinese. It is measured in square foot. If data is unavailable for a property, then its Saleable_size = -1.

No_of_rooms: The number of rooms in the property. 0 means studio; -1 means data is not available.

Floor: The floor of the property (10 for 10th floor).

Valuation of Hong Kong Residential Property

Region: The region of the property; it takes values Hong Kong, Kowloon and New Territories.

Primary_school:小學學區 in Chinese. Primary school Net divides Hong Kong's primary schools into 36 zones

Secondary_school: 中學學區 in Chinese. Secondary schools use a zoning system based on the 18 districts in Hong Kong.

Age_of_property: The age of the property in years; -1 means the property is not built yet (-1 對應樓花).

Valuation of Hong Kong Residential Property

Uncompleted: Whether the construction is completed. 0 means completed and 1 means under construction.

MTR_station: The name of the nearest MTR station. -1 means property is distant from all MTR stations.

Close_to_MTR: Whether the property is close an MTR station. 1 means close to and 0 means far from MTR stations.

Shopping_Mall, Swimming_Pool, Sport_facility, Club, Garden, Sauna_Shower, Playground, Cinema, Bar_karaoke, Study_Room, Ballroom: These are all binary variables. 1 means the amenity is available while 0 means there are no such amenities.

Valuation of Hong Kong Residential Property

District: One of Hong Kong's 18 districts.

Median_income: The median income in the HMA.

Median_age: The median age of residents in the HMA.

Population: The total population of the HMA.

Unit: Number of property units in the HMA.

Sample Codes (run on DAP)

```r
library(stargazer)
mydata = read.csv('/dataset/Centaline/Centaline_train.csv',header=TRUE)
head(mydata)
mydata$LogPrice = log(mydata$Transaction_price)
result <- lm(LogPrice ~ Age_of_property * Close_to_MTR, data = mydata)
summary(result)
```

What should we do in this project?

Each group should only ask one (big) research question in your project and answer it with data. Choose the right data analysis methods and come up with a good answer to your questions, with implications for sellers, buyers, developers, property agencies and the government.

What should we do in this project?

You need to include at least one interaction term or a square term in your analysis.

A full-mark example:

How does floor level affect housing pricing?

Floor Numbers and Housing Price

There is a significant drop in prices when floor is 13 or ends with 4. But not for 18.

Special Numbers and Housing Price

The higher the floor, the higher the unit price.

However, the marginal effect of floor on unit price is decreasing with the floor level.

Submission

To save your time, you only need to submit a few pages of slides (no more than 10 pages for main text + no more than 6 slides for appendix) to Moodle covering your research question(s), data analysis (e.g., regression equations), findings, and implications.

No reports/presentations.

Deadline: Jan 9, 2026

12:30 for Class A, 17:00 for Class B, and 21:30 for Class C

Next Week

We are going to work on LLMs.

Make sure you can access some advanced LLMs (e.g., ChatGPT). I am using Perplexity, which allows me to access multiple LLMs including GPT 5.2 and Gemini. You can enjoy a 12-month Education Free Trial.

The HKU AI platform may not be powerful enough.