




More on Regression Analysis

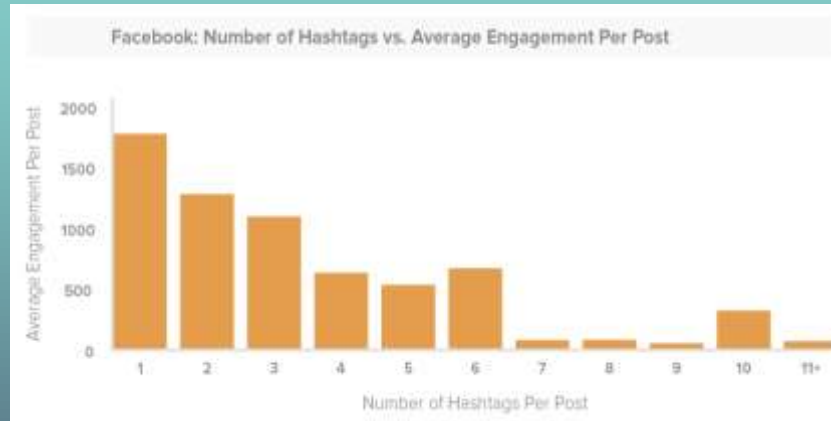


Question

Suppose that we want to estimate the relationship between the number of Facebook hashtags of a post (X) and the level user engagement (e.g., number of shares or likes, Y). What type of analysis should you do?



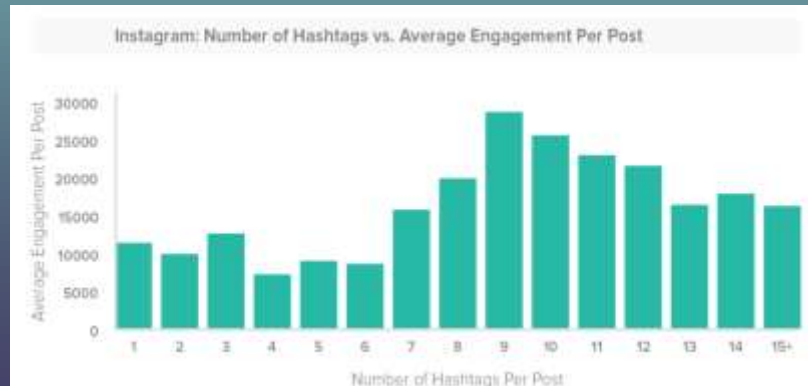
Question



Suppose this is your data. You may regress engagement on the number of hashtags, and find that the coefficient is negative: **Having more hashtags reduces user engagement.**

Question

You cannot run a simple linear regression here! Why?
The relationship between X and Y is nonlinear!
What should you do then?



Another Example

Income Changes Over the Course of an Individual's Life

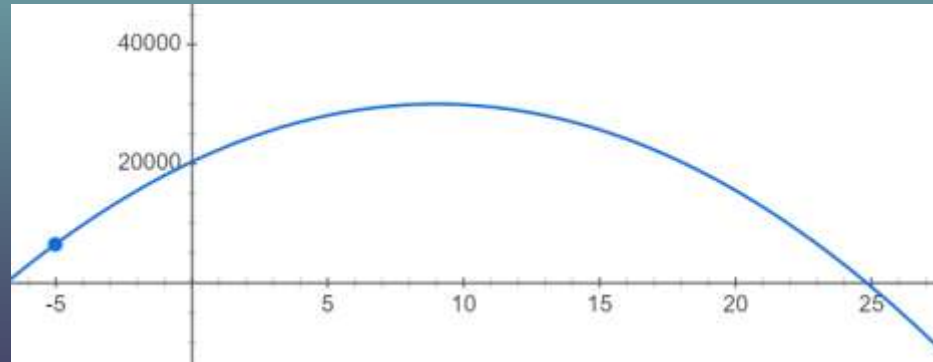
Average Adjusted Gross Income by Age



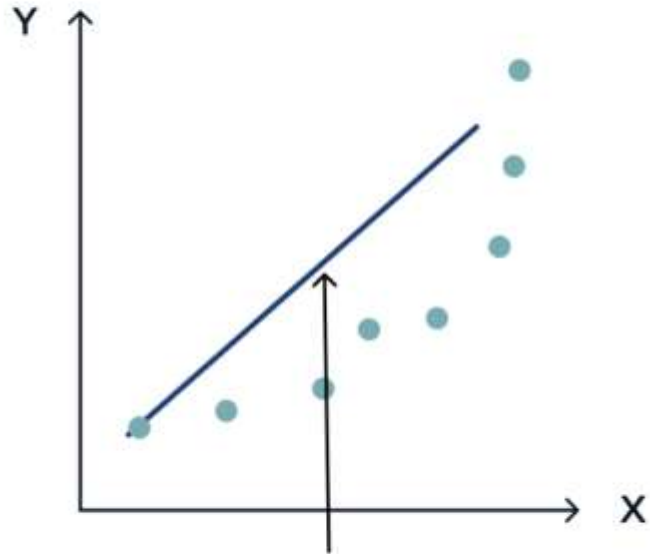
Source: Internal Revenue Service, "Table 1.5 All Returns: Sources of Income, Adjustments, and Tax Items, by Age, Tax Year 2016 (Filing Year 2017)"

Question

This is a nonlinear relationship between X and Y ! This can be captured by a quadratic form. Consider the following example.

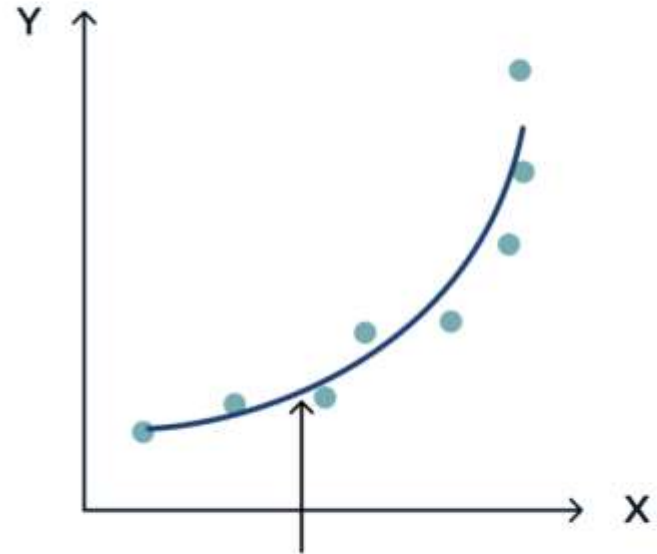


Simple linear model



$$y = b_0 + b_1x$$

Polynomial model



$$y = b_0 + b_1x + b_2x^2$$

Quadratic Regression in R

Suppose that we believe that the relationship between X and Y is quadratic (as opposed to linear). Then, we would like to regress Y on both X and X^2 . In this case, our regression question will be

$$Y = a + b_1X + b_2X^2$$

You can further extend the model to run cubic/polynomial regression...

$$Y = a + b_1X + b_2X^2 + b_3X^3 + \dots$$

POLYNOMIAL REGRESSION

$$P = 2x + 21x^2$$

$$\theta \quad |a \times b|$$



Crowdfunding: An example

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

```
mydata <-  
read.csv("https://ximarketing.github.io/class/Kickstart  
er-Project.csv", fileEncoding = "UTF-8-BOM")
```

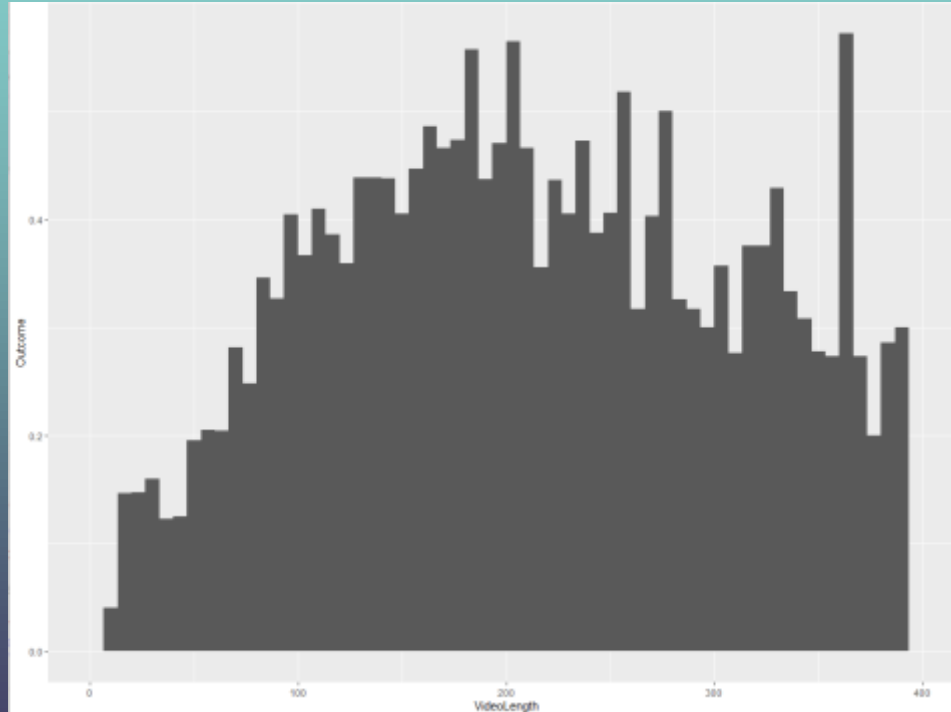
```
subdata = subset(mydata, IsVideoAvailable == 1)
```

Crowdfunding: An example

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

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


Crowdfunding: An example





Crowdfunding: An example

It seems that the relationship between the video length and project success is nonlinear: When video length is short, increasing video length improves the success rate. However, having a very lengthy video does not benefit the project either.



Crowdfunding: An example

Let us try the following logistic regression:

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

Crowdfunding: An example

Without quadratic term:

```
logit <- glm(Outcome ~ VideoLength, data = subdata,  
family = "binomial")  
summary(logit)
```

```
Coefficients:  
              Estimate Std. Error z value Pr(>|z|)  
(Intercept) -0.5689252  0.0481630 -11.812  <2e-16 ***  
VideoLength  0.0004219  0.0002285   1.846   0.0649 .  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The longer the video is, the more successful the project will be.

Crowdfunding: An example

With quadratic term:

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

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Crowdfunding: An example

We can further calculate the optimal length of the video: For a quadratic function $f = b_2x^2 + b_1x + a$ ($b_2 < 0$), the function is maximized when

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

If you forgot about it, please review your high school math. Link is [here](#).

In our regression, $b_2 = -1.056 \times 10^{-5}$ and $b_1 = 6.541 \times 10^{-3}$. Then, we can calculate that the optimal length of the video is around 300 seconds (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:

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

Are we missing anything?





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 studies for one hour.

How to incorporate this into our regression model?






Question

We consider the interaction between IQ and Hours of Study:

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

Suppose that we find out $b_3 > 0$, what does this imply?



Question

Suppose that

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

Consider two persons:

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 is more efficient than Bob!

Examples of Interaction Effects

Suppose that your dependent variable is a programmer's salary.

Suppose that you have two independent variables: the programmer's knowledge of Python and his/her knowledge of R.

We find that

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

How would you interpret this regression result?

Examples of Interaction Effects

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

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

If you know more about R, you can make 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 thing, learning about the other does not help you much.

Examples of Interaction Effects

Suppose that your dependent variable is a person's health score.

Suppose that you have two independent variables: the amount of swimming and running.

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

How would you interpret this regression result?

Examples of Interaction Effects

Suppose that your dependent variable is a person's health score.

Suppose that you have two independent variables: the amount of running exercise and whether or not the person is overweight.

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

How would you interpret this regression result?

$$\text{sex} = \begin{cases} 1, & \text{if} \\ 0, & \text{M} \end{cases}$$

Interact Dummy Variable

$$\text{wage}_i = \alpha + \beta_1 \text{educ}_i + \beta_2 \text{sex}_i + \beta_3 \text{sex}_i \text{educ}_i$$

$$\begin{aligned} \text{F: } \overline{\text{wage}}_F &= \alpha + \beta_1 \text{educ} + \beta_2 + \beta_3 \text{educ} \\ &= (\alpha + \beta_2) + (\beta_1 + \beta_3) \text{educ} \end{aligned}$$

$$\text{M: } \overline{\text{wage}}_M = \alpha + \beta_1 \text{educ}$$

Crowdfunding: An example

We want to investigate the relationship between the total funding, the creators' experience and the number of products offered. Let us prepare the data:

```
mydata <-  
read.csv("https://ximarketing.github.io/class/Kicks  
tarter-Project.csv", fileEncoding = "UTF-8-BOM")
```

```
mydata$LogFunding = log(mydata$FundingRaised + 1)
```

Crowdfunding: An example

We want to investigate the relationship between the total funding, the creators' experience and the number of products offered. Let us run a regression with an interaction term:

```
result = lm(LogFunding ~ Created *  
NumberOfProducts, data = mydata)  
summary(result)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.039097	0.071653	42.414	< 2e-16	***
Created	0.240761	0.042393	5.679	1.41e-08	***
NumberOfProducts	0.443064	0.008182	54.148	< 2e-16	***
Created:NumberOfProducts	-0.012090	0.005019	-2.409	0.016	*



Crowdfunding: An example

$$\text{LogFunding} = 3.04 + 0.24\text{Created} + 0.44\text{Number of Products} - 0.012\text{Created} \times \text{Number of Products}$$

What does this result tell us?






Crowdfunding: Another example

Let us explore something interesting.

We already know that in a crowdfunding project, putting your face in front of the camera makes the project more successful.

However, does it make a difference whether this is a female face or a male face? What's your intuition?



Crowdfunding: Another example

```
subdata = subset(mydata, IsVideoAvailable == 1)
result = lm(LogFunding ~ factor(Gender) * Human,
data = subdata)
summary(result)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.1768	0.3778	13.704	< 2e-16	***
factor(Gender)M	0.4710	0.4003	1.177	0.23939	
factor(Gender)U	1.9207	0.4057	4.734	2.26e-06	***
Human	2.3467	0.4137	5.672	1.48e-08	***
factor(Gender)M:Human	-1.1873	0.4411	-2.692	0.00713	**
factor(Gender)U:Human	-0.5688	0.4453	-1.277	0.20153	

Crowdfunding: Another example

This result tells us that, featuring a human in your video is beneficial. **Nonetheless, featuring a male is less helpful compared to featuring a female.**

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.1768	0.3778	13.704	< 2e-16	***
factor(Gender)M	0.4710	0.4003	1.177	0.23939	
factor(Gender)U	1.9207	0.4057	4.734	2.26e-06	***
Human	2.3467	0.4137	5.672	1.48e-08	***
factor(Gender)M:Human	-1.1873	0.4411	-2.692	0.00713	**
factor(Gender)U:Human	-0.5688	0.4453	-1.277	0.20153	



Exercise

Play with the Kickstarter dataset yourself and see if you can find any interesting interaction effects. Share it with us!






Individual Data Project



Task

You can either collect your data (e.g., searching for a dataset online, hiring a data scraper at Taobao, or collecting data yourself), or use the dataset provided below.

Work on the dataset to ask your own research questions and answer the research questions yourself.





Background

Here is background information of the dataset that you can use directly:

We all know that online reviews are important, and our purchase decisions are likely to be influenced by online reviews.



TripAdvisor

Founded in 2000, TripAdvisor is one of the leading online review platforms. It mainly focuses on hotel and restaurant reviews.



A Typical TripAdvisor Review

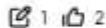


Reviewed 1 week ago  via mobile

Artistry Moments

Artistry Moments was impressively delicious what a memorable meal!

ohsiri



We have been to Mandarin Grill and Bar for a few times now. The ambiance the decor the food the service the staff are top notch.

Manolo and Chef Simon always look after all of us they are so detailed and professional we all are very impressed by their food and service very time we visit. The staff are very attentive so thoughtful, exceptionally good at what they do

Saturday and Sunday brunches are always one of our favorites place to go. Always good to be back

Show less

Date of visit: August 2021



[Ask ohsiri about Mandarin Grill + Bar at Mandarin Oriental, Hong Kong](#)

 1 Thank ohsiri





Data We Have

Reviews of Top 30 restaurants in the following cities:

New York, Las Vegas, Los Angeles, Chicago, Toronto, Vancouver, London, Sydney. 240 restaurants in total.

These are all English-speaking cities, and we only focus on reviews written in English that can be analyzed automatically.

Almost 150K reviews.






Data about the reviewer

Local: Whether or not the reviewer is a local resident (1 = local; 0 = nonlocal).

We incorporate this variables because local people's preference may be different from that of visitors.






Data about the reviewer

CountResaturant: How many restaurants the reviewer has been to.

This variable captures the experience of the reviewers. An experienced foodies may differ from a normal person (e.g., a foodie may be more critical).






Data about the reviewer

CountReview: How many reviews the reviewer has written.

This variable also captures the experience of the reviewers.
The more reviews written, the more experienced the reviewer.






Data about the reviewer

CountVotes: How many helpful votes the reviewer has received.

This variable also captures the experience of the reviewers.
The more votes, the more popular the reviewer.





Data about the review

Rating: The rating assigned by the reviewer.

TripAdvisor uses a 5-point ratings, with 5 being the best and 1 being the worst.






Data about the review

Helpful: The number of helpful votes the review received.

When a review receives more helpful votes, the review is more popular.





Data about the review

Mobile: Whether or not the review is typed from a mobile device (1 = mobile, 0 = nonmobile).

Mobile devices are small, and reviewers' behavior may be different when using mobile devices.





Data about the review

Photo: Number of photos in the review.

Date: The date the review was posted on TripAdvisor.

Data has format YYYY-MM-DD.





Data about the review

Title Length: The length of the review title.

Length: The length of the review body.

Both are measured in number of characters.





Data about the review

We also use sentiment analysis to capture the sentiment of the review:

Sentiment: the polarization of the review (-1; negative to 1; positive)

Subjectivity: the subjectivity of the review (0 = objective, 1 = subjective)






Data about the review

We also use sentiment analysis to capture the sentiment of the review:

Happy / Angry / Sad / Surprise: the emotion that is captured from consumer review. Each of them is a value between 0 to 1, and when the value is larger, it means the emotion is stronger.



Data about the review

We also analyze the content of photos of the review, if the review has at least one photo:

Menu: Whether or not there is a photo of the restaurant menu.

Building: Whether or not there is a photo of the restaurant building.

Meat: Whether or not there is a photo of meat.

Vegetable: Whether or not there is a photo of vegetable.

Person: Whether or not there is a photo of a person (e.g., a selfie)



Sample Questions

What makes a helpful review?

How does the reviewer's experience affect the characteristics of a review?

Is it true that "one picture is worth 1000 words?"





On Your Data Analysis

Try to incorporate at least one interaction effect in your data analysis.

Use quadratic terms and fixed effects whenever necessary.


Again, it would be desirable if you can come up with something surprising.





What should we do?


Each individual should only ask **one big or two small research questions** in your project. Quality beats quantity. Choose the right data analysis methods and come up with a good answer to your questions, with implications for the platform, business owners or consumers.



You need to submit:

To save your time, you only need to submit a few pages of slides (**no more than 12 slides if you use the TripAdvisor data; no more than 15 slides if you use your own dataset**) to Moodle covering your dataset (if it is not TripAdvisor data), research question(s), data analysis (e.g., regression equations), findings, and implications. No reports/presentations are needed!

Deadline:	Class A:	Jan 22,	12:30
	Class B:	Jan 22,	17:00
	Class C:	Jan 24,	17:00



Group Data Project II: HK Property Valuation





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.



Valuation of Hong Kong Residential Property

Accessing the data through the cloud platform:

<https://dap.acrc.hku.hk/hku-dap-client/#/Signin>

Loading the data:

```
df =  
read.csv('/dataset/Centaline_data/Centaline_data.csv',header=TRUE)
```

Property Data

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

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.**

Property Data

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”.

Property Data

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).

Property Data

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).



Property Data

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.





Property Data

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.





Property Data

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.





Property Data

Lon and Lat: The longitude and latitude of the property.

Distance_to_Central: The distance from the property to central, measured in meters.


Second_Hand: 1 means the property is second-hand while 0 means it is a new property.





What should we do?

Each group should only ask **one big or two small research questions** in your project. Quality beats quantity. 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.



You need to submit:

To save your time, you only need to submit a few pages of slides (**no more than 18 slides**) to Moodle covering your research question(s), data analysis (e.g., regression equations), findings, and implications. No reports/presentations are needed!

Deadline:	Class A:	Jan 22,	12:30
	Class B:	Jan 22,	17:00
	Class C:	Jan 24,	17:00