Investigate Employment Income and Place of Living by Computational Social Science Method

New- team:

Cheng Chung Yu

and

Lee Chun To

Works are divided equally to each author

<u>Introduction</u>

Gini coefficient is the measurement of the income distribution across a population which was found by the Italian statistician in 1912. It is often used as a standard of economic disparity, estimating income distribution or, less commonly, wealth distribution among a population. The coefficient ranges from 0 (or 0%) to 1 (or 100%), with 0 expressing perfect equality and 1 symbolizing absolute inequality. (Westfall,2020) According to research, Hong Kong Gini coefficient of single monthly household income rose from 0.533 in 2006 to 0.539 in 2016. Meanwhile, the Gini coefficient of post-tax post-social transfer monthly household earnings was 0.473 in 2016. It is more severe than other developed economies. For instance, Singapore was 0.356, the United States was 0.391, and the United Kingdom was 0.351. The median monthly household income of the first 10% is 44 times higher than that of the lowest in 2016. In 2006, this number was only 34 times. The number of low households has stretched up to five hundred and thirty thousand, while three hundred thousand residents are low-income families. Further, over 1.3 million citizens live in poverty. (OXFam Hong Kong,2018) Hong Kong also is the city with an enormous homeownership burden in the world. According to a survey report by Demographia, Hong Kong has been the most stressful city to afford estate prices for ten straight years.

The median household income of property prices has improved slightly to 20.8

times from 20.9 times the previous survey. The survey revealed that the median

property price in Hong Kong was 7.04 million, and the median household

income was 338,000. If the median household interest rate calculated, it means

that Hong Kong people have not to eat or drink for 20.8 years. (Shawna, 2020).

Since Hong Kong has extreme income inequality and a severe housing problem,

therefore, this paper is going to examine how employment income increase, as

well as the relationships between employment income and living condition.

Literature Review

Differential Income to Education

Education determines one's position in the labour market and affects the

individual's employment income situation in society. Education can be

considered as a critical role to drive economic growth (Krueger and Lindahl,

2001). Sylwester (2002) suggests that investing more resources on public

education can lower the income inequality for the countries in the following

years. It merely means that it influences society continually. Since education

brings impact on various sectors such as public health, reduction of crime rate,

the environment, etc., the relationship of attained education to the income of

individuals is needed to be found. According to Hallerod and Larsson's definition (2008); poor people represent people who insignificantly access economic resources so that they can just enjoy deficient levels of consumption goods and services. Turčínková and Stávková (2012) pointed out the effect of the attained level of education on the income situation of individuals and household. They used conventional EU to measure relative poverty, and four educational categories including primary or no education learned a trade through apprenticeship or lower secondary education or without a leaving certificate, full secondary or vocational or post-secondary education, and higher (tertiary) education. The results found the most vulnerable group comprises households with primary education or no education. It also concluded people who attained tertiary education level have better income situation than other categories of families with a lower level of attained education. Therefore, based on the research, the relationship between income inequality and the achieved level of education has been found.

Effect of Housing Deprivation to Income Level

There is some proof indicating that housing deprivation is strongly and directly related to income. However, the backgrounds of some families suggest that housing conditions are more complicated. Some families, significantly older

households, present a weak relationship between income and various deprivations. The research also proposes that a deprivation measure of the condition provides a further positive understanding of poverty since it controls for the experience of personal unfavourable housing conditions by higher-income groups. Indicators such as dampness, condensation, mildew, and housing community dissatisfaction were found to have healthier relationships with low income. This finding is uniform with data describing housing, health and well-being. (Littlewood & Munro, 1997)

Effect of Indoor Environment to Income Level

Another Study mainly examines the indoor environment and low-income housing. There is some proof indicating that housing deprivation is strongly and directly related to income. However, the backgrounds of some families suggest that housing conditions are more complicated. Some families, significantly older households, present a weak relationship between income and various deprivations. The research also proposes that a deprivation measure of the condition provides a further positive understanding of poverty since it controls for the experience of personal harmful housing conditions by higher-income

groups. This finding is uniform with data describing housing, health and well-being.

Further, several comparatively wealthy households live in a residence which lacks amenities or is unfit by the Below Tolerable Standard measure without being deprived in other respects. The results confirm that indoor environmental dangers are widespread in low-income houses and that these dangers clustered in developments. The majority of families showed three or more exposures in their places. These results are consistent with other studies that explain the relationship between housing situations and particular environmental vulnerabilities. Clustering by house site was evident for bugs, flaming byproducts, mould, and ventilation. Because both the flaming byproducts and ventilation indexes directly reflected physical characteristics of the units and their maintenance (residence of gas stoves and working bathroom and kitchen fans, respectively), the site result is not unexpected in these circumstances. For bugs and mould, these results are consistent with previous studies, which have shown that construction properties directly affect mould growth and pest infestation. The results also support the view that bypassing or mitigating these problems in multifamily residence may need work at the building level, besides, to perform at the resident level. Standard methods for

the application of combined pest management methods in multifamily buildings also require action at the resident and management levels. (Adamkiewicz et al., 2014)

Effect of the place of living affecting the occupation income

Educational and occupational preferences mean a great workers' career. Also, where workers choose to live in effect significantly in many occupations. As the analysis has shown, there is a wide variety of salary results across areas in the United States. Average pay is extensively higher in some places than in others, though the position of the highest salary differs depending on the occupation. Higher-income is sometimes partially balanced by a more expensive cost of living and taxes, depending on location. Still, the higher cost of living and taxes are adjusted in some places by nonwage amenities that may attract employees. (Lockwood et al., 2017)

Hypothesis

Hypothesis 1: Employment income would be higher with a higher educational level

Hypothesis 2: Professional occupation has positively associated with employment income

Hypothesis 3: Citizens with different backgrounds such as income, occupation and room belong to different social levels

Data

We collected data from the 2011 Population Census from the Hong Kong Census and Statistics Department, which contained up-to-date information of the Hong Kong population. Face-to-face interviews collected data, the postal return of completed short-form questionnaire and both long-form and short-form questionnaire through the Internet. The census data aimed at providing information for planning and policy formulation to the government, as well as business and research purposes. Due to the privacy issue, for the educational purpose, this study obtained 5% of the total census data. The sample dataset contained 112 variables, and the sample population was about 36878

respondents. Apart from necessary demographic data like age, we adopted

extra valid data including type of quarters, type of accommodation, no. of rooms,

monthly domestic household rent, current district of residence, educational

attainment (highest level completed), occupation, monthly income from primary

employment, and place of work to analyze the research.

<u>Research methods</u>

Different approaches are used to test our hypotheses and to make several

predictions. For supervised learning, we used regression analysis to examine

the relationship between attained education and employment income, and then

try to predict the income trending. The relationship between occupation and

employment income also studied by regression analysis to find what types of

the occupation have better income earning. Apart from regression analysis, for

the classification model, a k-nearest neighbour method is adopted. A k-nearest-

neighbour is a data classification algorithm that attempts to determine what

group a data point is by looking at the data points around it. We try to use the

K-NN method to classify which income groups respondents belong. For

unsupervised learning, we used k-mean clustering to identify and group the

respondents as different income groups. The technique makes inferences from

datasets using only input vectors by dividing them into other clusters without

referring to known, or labelled outcomes. A cluster refers to a collection of data

points aggregated together because of certain similarities. The use of this

method can help generalize several incomes groups and recognize the centroid

of each cluster.

Findings

Supervised Learning:

<u>Table 1</u>

The below each scatterplot includes fit lines, marginal box plots, conditioning

on a factor, and interactive point identification of income and education.

20: Secondary School Level
40: Craft Level
60: Higher Diploma and Sub-degree
80: Postgraduate
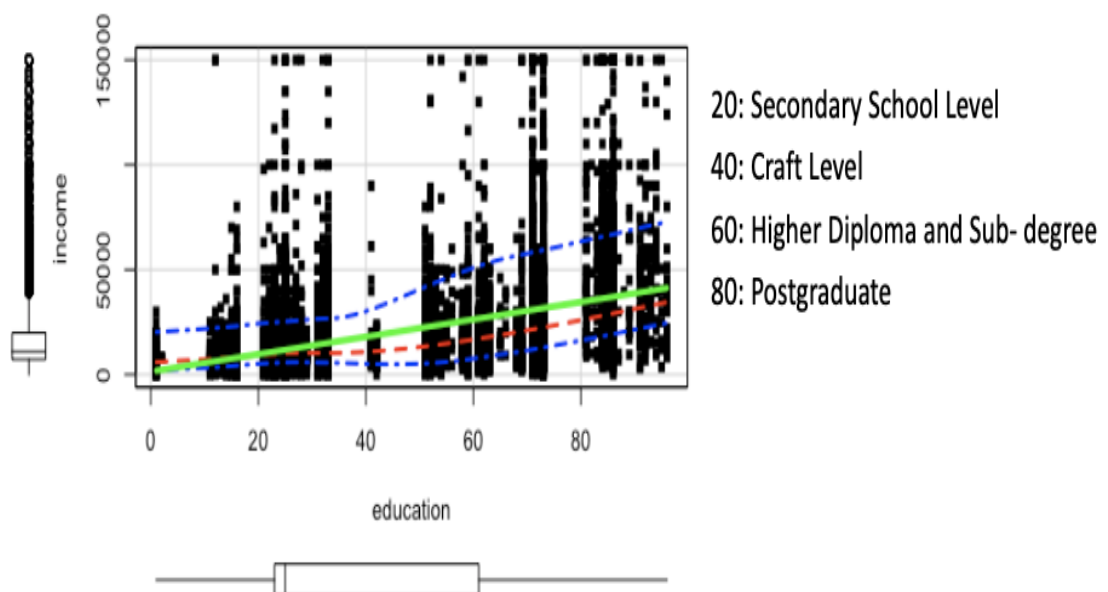
The below showed the linear model and its error.

```
MODEL INFO:
Observations: 18063
Dependent Variable: edu$`HKC2011$mearn`
Type: OLS linear regression

MODEL FIT:
F(1,18061) = 5644.907, p = 0.000
R² = 0.238
Adj. R² = 0.238

Standard errors: OLS
------------------------------------------------------------------------
                             Est.      2.5%      97.5%   t val.       p
------------------------- ---------- --------- ---------- -------- -------
(Intercept)               1443.241   959.578   1926.903    5.849   0.000
edu$`HKC2011$educnh         415.187   404.355    426.018   75.133   0.000
------------------------------------------------------------------------
```
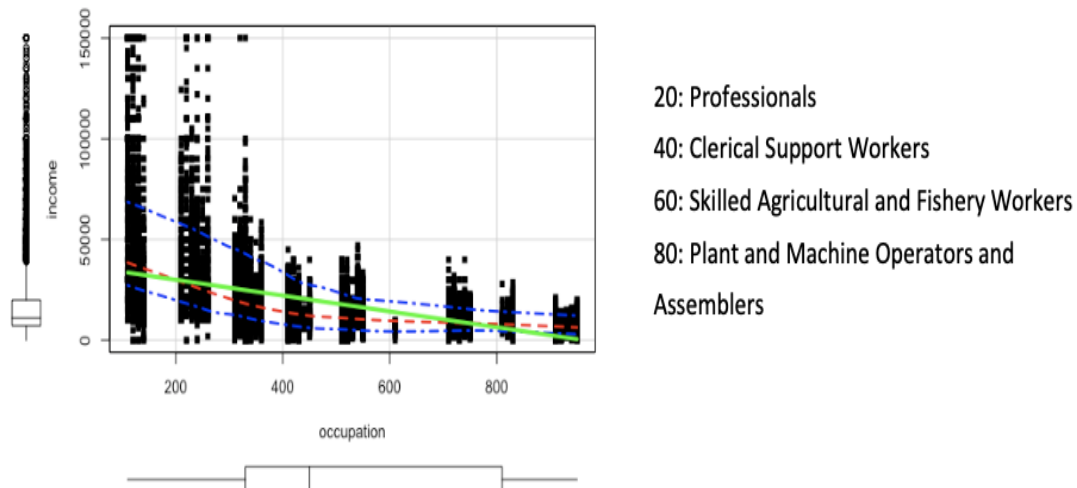
Regression analysis is adopted to find the relationship between monthly employment income and educational level. The table shows that monthly employment income is positively associated with the academic rank (B=415.187, t= 75.133, p= 0.000). The green line in the graph represents the fitted regression line. Although there are some cases that people with lower education can have high income received, generally, people with higher education such as a degree and postgraduate can have better income situations than other categories, holding other factors constant.

Table 2

The below scatterplot includes fit lines, marginal box plots, conditioning on a

factor, and interactive point identification of income and occupation.

20: Professionals

40: Clerical Support Workers

60: Skilled Agricultural and Fishery Workers

80: Plant and Machine Operators and

Assemblers

MODEL INFO:
Observations: 18063
Dependent Variable: income_o$`HKC2011$mearn`
Type: OLS linear regression

MODEL FIT:
$F(1,18061) = 7035.988$, *p* = 0.000
$R^2 = 0.280$
Adj. $R^2 = 0.280$

Standard errors: OLS

                                    Est.   2.5%   97.5%   t val.   p
--------------------------- ----------- ----------- ----------- --------- -------
(Intercept)   37910.719   37372.404   38449.034   138.039   0.000
income_o$`HKC2011$occup   -39.324   -40.243   -38.405   -83.881   0.000

Regression analysis is adopted to examine the relationship between monthly

employment income and occupation. The result shows that the monthly

employment income is significantly associated with occupation (B= -39.324, t=

-83.881, p= 0.000). The green line is the fitted regression line which used to predict the income situation among occupations. The kind of work divided into three main aspects, including managers and administrators, professionals, and workers and assemblers, respectively. The lower the code of occupations, the higher the monthly income respondents can generally receive. It means respondents who are working as management aspects can relatively receive a higher monthly income. Still, respondents who are low-skilled workers or assemblers can just receive a lower monthly income.

Table 3

```
   Cell Contents
|-----------------------|
|                     N |
|         N / Row Total |
|         N / Col Total |
|       N / Table Total |
|-----------------------|


Total Observations in Table:  5929


                   | Rmerge$`Observed `
Rmerge$`Predicted ` |         1 |          2 |          3 | Row Total |
-------------------|-----------|-----------|-----------|-----------|
                1 |       164 |         0 |         0 |       164 |
                   |     1.000 |     0.000 |     0.000 |     0.028 |
                   |     1.000 |     0.000 |     0.000 |           |
                   |     0.028 |     0.000 |     0.000 |           |
-------------------|-----------|-----------|-----------|-----------|
                2 |         0 |      4785 |         0 |      4785 |
                   |     0.000 |     1.000 |     0.000 |     0.807 |
                   |     0.000 |     1.000 |     0.000 |           |
                   |     0.000 |     0.807 |     0.000 |           |
-------------------|-----------|-----------|-----------|-----------|
                3 |         0 |         0 |       980 |       980 |
                   |     0.000 |     0.000 |     1.000 |     0.165 |
                   |     0.000 |     0.000 |     1.000 |           |
                   |     0.000 |     0.000 |     0.165 |           |
-------------------|-----------|-----------|-----------|-----------|
     Column Total |       164 |      4785 |       980 |      5929 |
                   |     0.028 |     0.807 |     0.165 |           |
-------------------|-----------|-----------|-----------|-----------|
```

```
!'S
F1_Score(y_pred  =Rmerge$`Predicted `, y_true = Rmerge$`Observed `, positive = NULL)
```

[1] 1

```
error <- mean(Rmerge$`Predicted ` != Rmerge$`Observed `)
error
```

 [1] 0

|	Predicted	Observed
1	3	3
2	2	2
3	3	3
4	3	3
5	2	2
6	2	2

K- Nearest Neighbors method is adopted to predict the respondents belong to which income groups generated by K- Mean clustering. After executing the unsupervised machine learning, we divided respondents into three major income groups which are high-income group, middle-income group and low-income group respectively. Then we hope to handle other regression and classification problems to predict and classify new points with similar characteristics. Thus, we make use of the KNN algorithm to try to predict whether someone belongs to different income groups based on their monthly

employment income, occupation and number of rooms. The result shows that

it gets a remarkable prediction. We first divided the data into a training dataset

and test dataset. Then we use 5929 observations in total to predict the

respondents' groups. We use the F-1 score to measure our model's accuracy

on the dataset as it is a way of combining the precision and recall of the model.

Finally, the F-1 score is equal to 1, and the error is 0. It means that the

respondents' groups are well predicted in this method.

Unsupervised Learning

Table 4

Below figure generated three k-mean cluster groups of the income, occupation

and room.
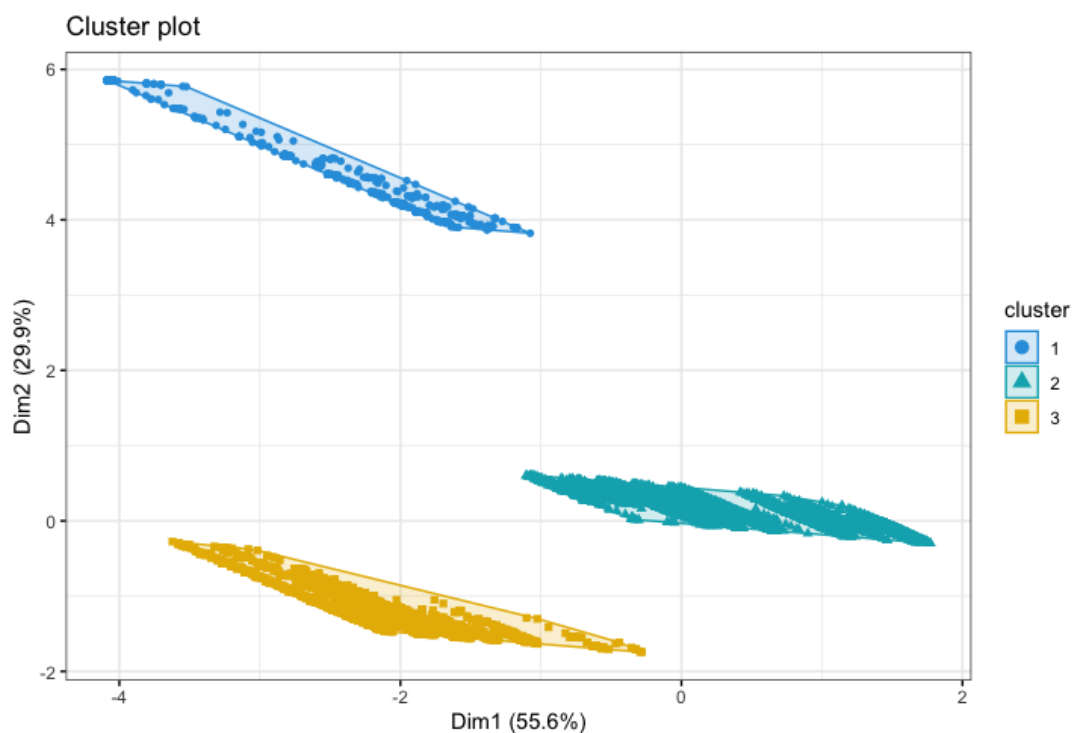
cluster <int>	HKC2011$mearn <dbl>	HKC2011$occup <dbl>	HKC2011$room <dbl>
1	106021.76	158.9765	4.605544
2	10236.82	590.8678	3.403152
3	36800.38	267.4975	3.996668

```
# A tibble: 6 x 4
  `HKC2011$mearn` `HKC2011$occup` `HKC2011$room` cluster
  <labelled>      <labelled>      <labelled>       <int>
1 20000           330             3                    2
2  3740           910             2                    2
3  9000           410             3                    2
4 30000           310             5                    3
5 50000           120             3                    3
6 10000           830             4                    2
```

The table shows that 3 cluster groups are generated by k-mean clustering. From the above table, we predict three different income groups using respondents' monthly employment income, occupation and number of rooms. The centroid of cluster 1 represents that social level group whose respondents' monthly employment income is around $106021.76HKD, with approximately 4.6 number of rooms. The occupation level is about 158.9, which means respondents' kind of work performed belongs to managers and administrators such as chief executive, commercial manager, etc. The centroid of cluster 2 represents that low social level group whose respondents' monthly employment income is around $10236HKD, with approximately 3.4 number of rooms. The occupation level is about 590.8 which means respondents' kind of work

performed belongs to various workers and assemblers such as salespersons, metal, machinery and related trades workers, cleaners, helpers and related workers, etc. The centroid of cluster 3 represents that middle social level group whose respondents' monthly employment income is around $36800.3HKD, with approximately 3.9 number of rooms. The occupation level is about 267.4 which means respondents' kind of work performed belongs to professional aspects such as business, administration and related professionals, health associate professionals, legal, social science and cultural professionals, etc. The table provides information for the public to understand how much income they are likely to obtain in different occupations and to know which income group they similarly belong to in Hong Kong's situation.

Discussion

General, our studies establish the relationship between education and employment income, and the relationship between the place of living and employment income by using the factor of the number of rooms. To answer our hypothesis 1, the result of regression analysis shows that higher education levels can be of benefit to citizens' employment income so that it alerts citizens to attain higher education such as postgraduate to increase their income situation. The regression analysis also proves that hypothesis 2 is statistically

significant. The results show that citizens who are working in management, and professional aspects can receive higher employment income than citizens who are working for low-skilled jobs in general. It can explain why the recruitment requirement for such high-income jobs is more restricted than other jobs.

Furthermore, our studies illustrate how employment income, occupation, and the number of rooms can be grouped to define their income level. The reason is to generalize information to the public to understand which income group they belong to. Alternatively, it provides information to citizens who hope to have a better income situation to understand which conditions they should fit in to reach their target. Different from traditional social science methods, apart from reviewing past studies, we adopted machine learning skills by obtaining big dataset from the government to predict future trends and identifying and grouping similar data points in the dataset without concern for the specific outcome. However, none of these approaches to date holds the information which cannot be quantified, and error might exist because of these unquantified elements. Our studies serve as a proof-of-concept that attained education in Hong Kong affects employment income and explains the place of living can be related to the employment income with the combination of the factors of occupation and number of rooms of the respondents.

On the other hand, limitations exist in the study. An accurate statistical measure of a large population can only be obtained when a full list of the entire population to be studied is available. For now, only 5% of the sample size is used to the study so that the result generalized might not be able to reach entirely accurate. Income alone cannot accurately predict the experience of forced exclusion from new ways of living. A focus on resources cannot preclude consideration of actual material conditions of life.

Moreover, since the data obtained by the government emphasize anonymity, due to the social desirability effect, respondents might provide false information which causes the result imprecise. The study also has a lack of investigation on how different districts citizens living in pose influence on their employment income situation and hasn't investigated the relationship between education level and occupation. Further research can generalize knowledge of which areas citizens in Hong Kong can choose to live to receive higher occupational income to prove whether the result of Bauer, Breitwieser, Nunn, & Shambaugh (2018) can be applied to Hong Kong society.

Conclusion

Using data from the Hong Kong Census and Statistics Department, we examined the association between educational level and employment income situation, and the association between occupation and employment income situation. With the use of both supervised and unsupervised learning, different respondents identified as three other income groups containing high-income group, middle-income group and low-income group. The information can use for the public and the government to recognize their income situation to formulate appropriate career planning and policy to reach various targets.

References

Adamkiewicz, G., Spengler, J., Harley, A., Stoddard, A., Yang, M., Alvarez-

Reeves, M., & Sorensen, G. (2014, September). Environmental

Conditions in Low-Income Urban Housing: Clustering and Associations With

Self-Reported Health. Retrieved from

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3954449/

Easterly, W., Robelo, S. (1993). Fiscal policy and economic growth: an

empirical investigation. Journal of Monetary Economics, 32, 417–455.

Hallerod, B., Larsson, E. (2008). Poverty, welfare problems and social

exclusion. International Journal of Social Welfare, 17, 15– 25.

Krueger, A., & Lindahl, M. (2001). Education for Growth: Why and for Whom?

Retrieved December 21, 2020, from

https://www.aeaweb.org/articles?id=10.1257%2Fjel.39.4.1101

Littlewood, A., & Munro, M. (1997). POVERTY AND POOR HOUSING IN

SCOTLAND: THE RELATIONSHIP BETWEEN INCOME AND HOUSING

CONDITION. *Netherlands Journal of Housing and the Built Environment,*

12(2), 199-220. Retrieved from http://www.jstor.org/stable/41107693

McKnight, Abigail and Cowell, Frank (2014) Social impacts: health, housing and intergenerational mobility. In: Salverda, Wiemer, Nolan, Brian, Checchi, Daniele, Marx, Ive, McKnight, Abigail, Tóth, István György and Werfhorst, Herman van de, (eds.) Changing Inequalities in Rich Countries: Analytical and Comparative Perspectives. Oxford University Press, Oxford, UK. ISBN 9780199687435

OXFam Hong Kong. (2018, May 8). Hong Kong Inequality Report. Retrieved from

https://www.oxfam.org.hk/f/news_and_publication/16372/Oxfam_inequality%2 0report_Eng_FINAL.pdf

Shawna, K. (2020, January 20). Hong Kong Homes Remain World's Least Affordable for 10th Year. Retrieved from

https://www.bloomberg.com/news/articles/2020-01-20/hong-kong-homes-remain-world-s-least-affordable-for-10th-year

Sylwester, K. (2002). Can education expenditures reduce income inequality? Economics of Education Review, 21, 43–52.

Turčínková, J., & Stávková, J. (2012, November 10). Does the Attained Level of Education Affect the Income Situation of Households? Retrieved December

21, 2020, from

https://www.sciencedirect.com/science/article/pii/S1877042812040578

Westfall, P. (2020, October 26). Gini Index Definition. Retrieved from

https://www.investopedia.com/terms/g/gini-index.asp