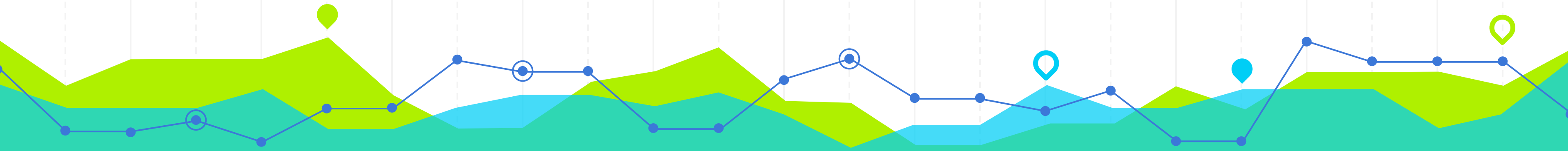


Energy Economy

Is there an Economic Case for Energy-Efficient Dwellings in the UK Private Rental Market?

“

Without Data you are just another person with an opinion

-W Edwards Deming

Split incentive problem

- In the private rental sector, one party, the landlord, makes capital investments in energy efficiency and does not gain immediately, but the advantages are reaped by another, the tenant, who benefits from lower utility costs and greater thermal comfort.
- This affects landlords' investment decisions and creates a barrier to achieving improved energy efficiency.
- "Does improved house energy efficiency contribute to higher home sales prices?"

- In addition to that, we relate **the housing prices to the socio-economic environment of the area**, and then we address another important question **‘how does the sale prices of houses fluctuate with time in years?**’ using repeat sales method.

Multiple Regression Model

- Suppose we would like to predict variable y using $x_1, \dots, x_k$. We have the data:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + e_i = \mathbf{x_i^T \beta} + e_i \quad (i = 1,2,\dots,n)$$

- Write $X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ and $\mathbf{e} = \begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \\ \vdots \\ \boldsymbol{e_n} \end{bmatrix}$.

- A model of the form

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

is to be fit to data, i.e. $\widehat{\boldsymbol{Y}} = X\widehat{\boldsymbol{\beta}}$.

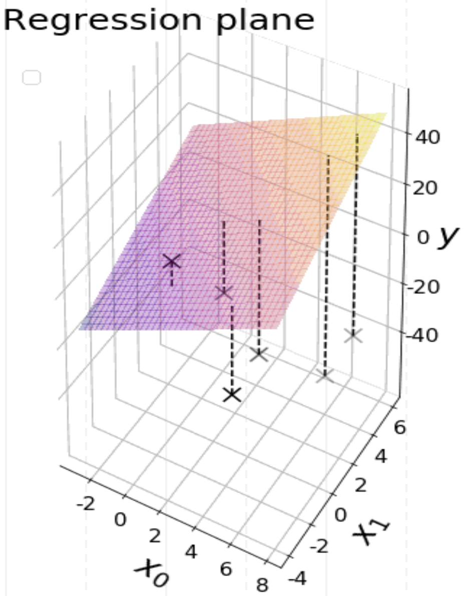
- $e_i = y_i - \hat{y}_i$ is called **residual** for ith observation.

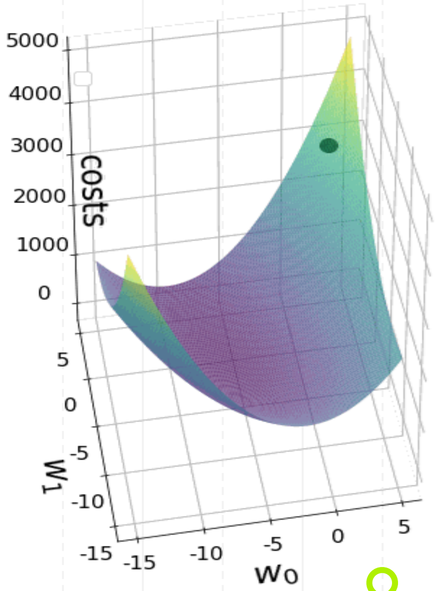
Solving the equations

- We need to find $\hat{\beta}$ that minimizes $SSE = S(\widehat{\boldsymbol{\beta}}) = \left\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\right\|^2$.
- Equate $\frac{\partial S}{\partial \beta_i} = 0$ and we obtain: $(X^T X)\widehat{\boldsymbol{\beta}} = X^T \mathbf{Y}$.
- A formal solution is $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}$.
- The fitted values are $\widehat{\mathbf{Y}} = X\widehat{\boldsymbol{\beta}} = X(X^T X)^{-1} X\mathbf{Y} = \mathrm{H}\mathbf{Y}$ where H is called the projection matrix.

Multiple Regression Model

- Intercept β_0 predicts where the regression *plane* crosses the Y axis
- Slope for variable X_1 (β_1) predicts the change in Y per unit x_1 holding x_2 constant
- The slope for variable x_2 (β_2) predicts the change in y per unit x_2 holding x_1 constant





Important terms

- We make usual four **assumptions** on **e:** $E[\mathbf{e}] = 0$, $\mathbf{e} \sim \mathcal{N}_n(\vec{0}, \sigma^2 I_n)$.
- The **covariance matrix** of random vector $\widehat{\boldsymbol{\beta}}$ is $\Sigma_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}} = \sigma^2 (X^T X)^{-1}$
- The **standard error** for $\hat{\beta}_i$ is given by $\sigma_{\widehat{\beta}_i} = \sigma^2 [(X^T X)^{-1}]_{ii}$.
- The **residual standard error**, which is an unbiased estimator of σ, is given by

$$s = \sqrt{\frac{SSE}{n - (k+1)}}$$

R-squared

- Multiple R-squared is defined as

$$R = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$$

- Adjusted R-squared is defined as

$$\bar{R}^2 = 1 - \frac{n-1}{n-(k+1)}(1 - R^2)$$

The value of R may be large due to the excess number of regressors, which may not add to there regression's explanatory power. This is penalized by adjusted R-squared.

Diagnostic plots

- After running a regression, one has to check if the assumptions $E[\mathbf{e}] = 0, \mathbf{e} \sim \mathcal{N}_n(\vec{0}, \sigma^2 I_n)$ are satisfied.
- To check normality assumption, one can simply check the normal QQ plot of residual.
- To check that each e_i has same variance, one can plot scale-location plots. Here fitted values are plotted with standardized residual $\frac{y_i - \hat{y}_i}{\sqrt{y_i}}$ to check the variability of residual.

Influential points: Leverage

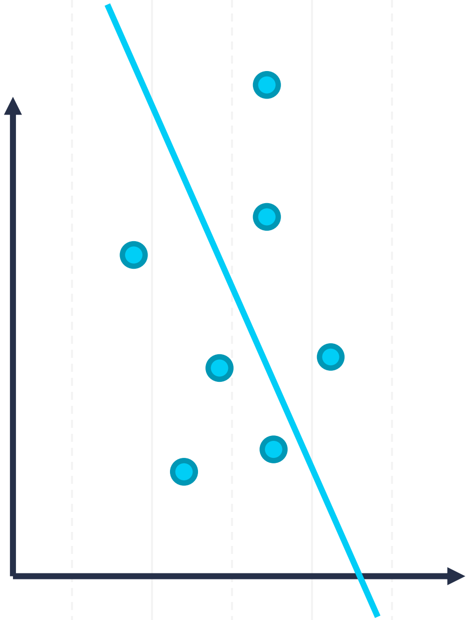
- A leverage score is given to each observation: $h_{ii} = [H]_{ii} = \mathbf{x_i^T}(X^T X)^{-1}\mathbf{x_i} = \frac{\partial \hat{y}_i}{\partial y_i}$.
- It is the degree by which ith measured value influences the ith fitted value.
- Standardized residuals can be plotted against leverage to check for outliers.

Influential points: Cook's distance

- The Cook's distance statistic for every observation measures the extent of change in model estimates when that particular observation is omitted.

$$D_i = \frac{1}{(k+1)s^2} \sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)} \right)$$

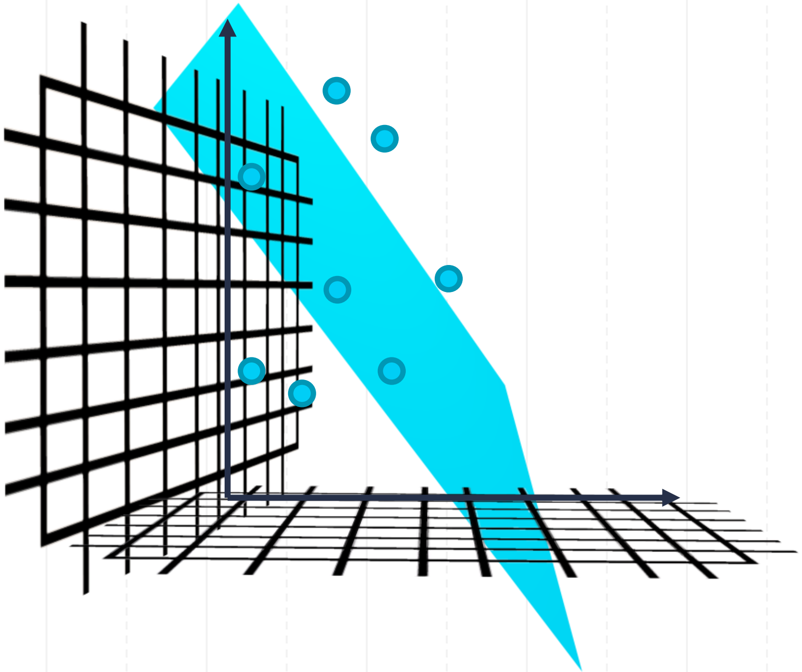
- Cook's distance can be plotted for each observation to see if deleting an observation makes a lot of change. Cook's distance can also be plotted against leverage.



Partial residue plots

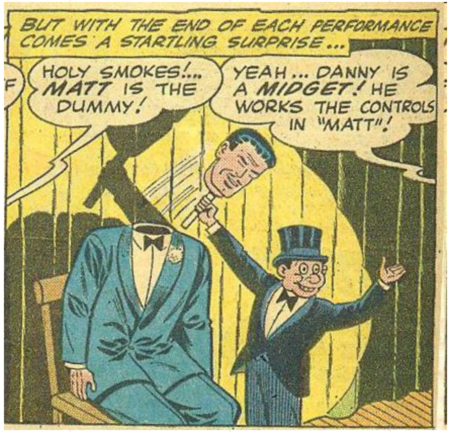
- A partial residual plot is a scatterplot to show the relationship between a given independent variable and the response variable.
- In partial residual plot, partial-residue=$e_i + \hat{\beta}_i x_i$ is plotted versus x_i. It is a graphical way of checking linear relationship.

Partial Residual Plots



Categorical Explanatory Variables

- Categorical independent variables can be incorporated into a regression model by converting them into 0/1 ("dummy") variables
- For binary variables, code dummies "0" for "no" and 1 for "yes"



Dummy Variables, More than two levels

For categorical variables with *k* categories, use *k*–1 dummy variables

For example

SMOKE2 has three levels, initially coded

0 = non-smoker

1 = former smoker

2 = current smoker

Use *k* – 1 = 3 – 1 = 2 dummy variables to code this information like this:

SMOKE2	DUMMY1	DUMMY2
0	0	0
1	1	0
2	0	1

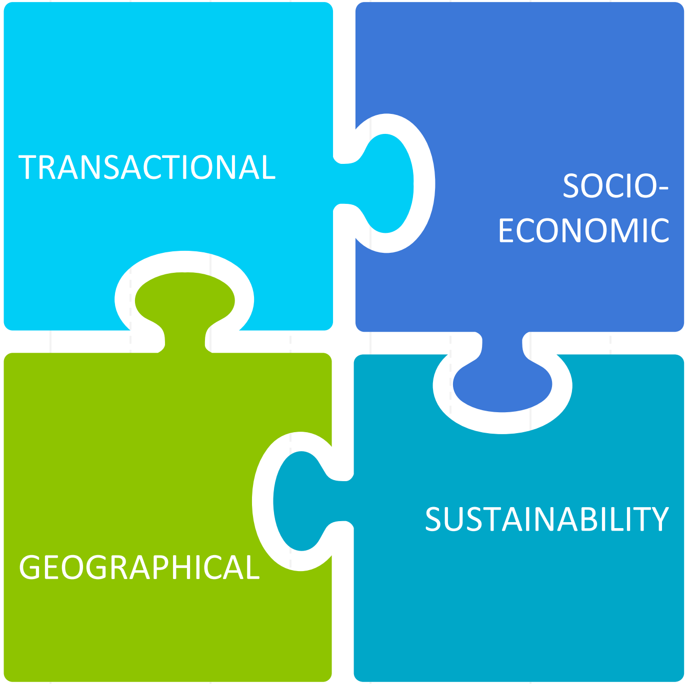
Data Description

1

VARIABLES

- price_1
- date_1
- price_2
- date_2
- perc_change_p2_to_p1
- days_between_sale
- ln_price_1
- ln_price_2

- reg_north_east
- reg_north_west
- reg_yorkshire_and_the_humber
- reg_east_midlands
- reg_west_midlands
- reg_east_of_england
- reg_london
- reg_south_east
- reg_south_west

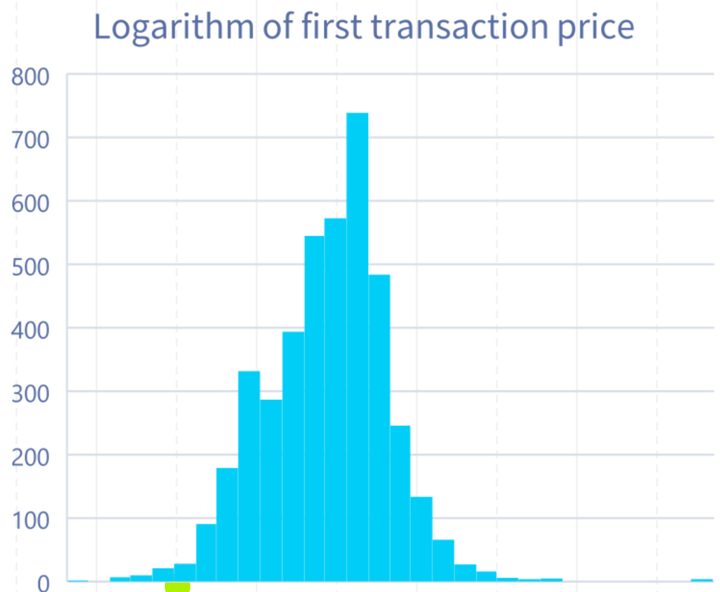


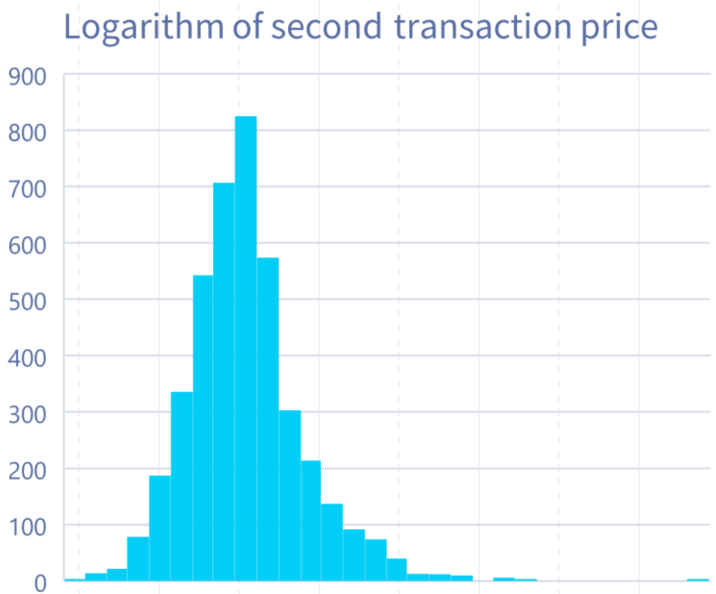
- imd_score
- imd_level
- income_score
- income_level
- emp_score
- emp_level
- educ_score
- educ_level

- health_score
- health_level
- crime_score
- crime_level
- barrier_score
- barrier_level
- living_score
- living_level

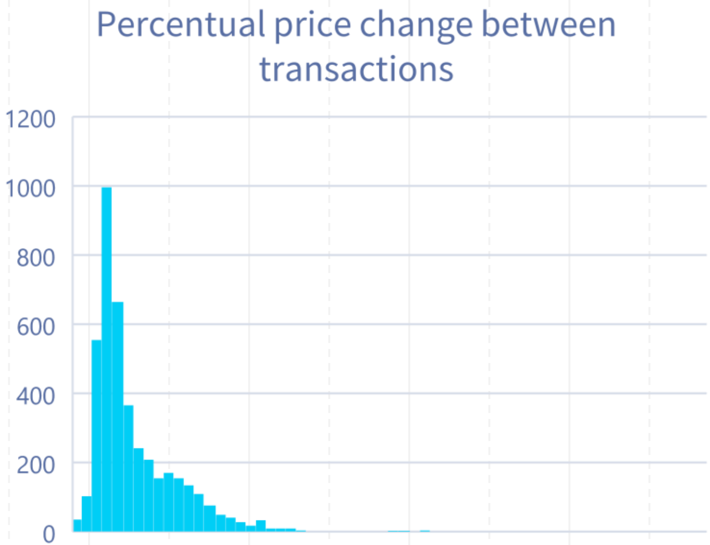
- epc_100
- epc_rating_a
- epc_rating_b
- epc_rating_c
- epc_rating_d
- epc_rating_e
- epc_rating_f
- epc_rating_g
- ln_epc_100

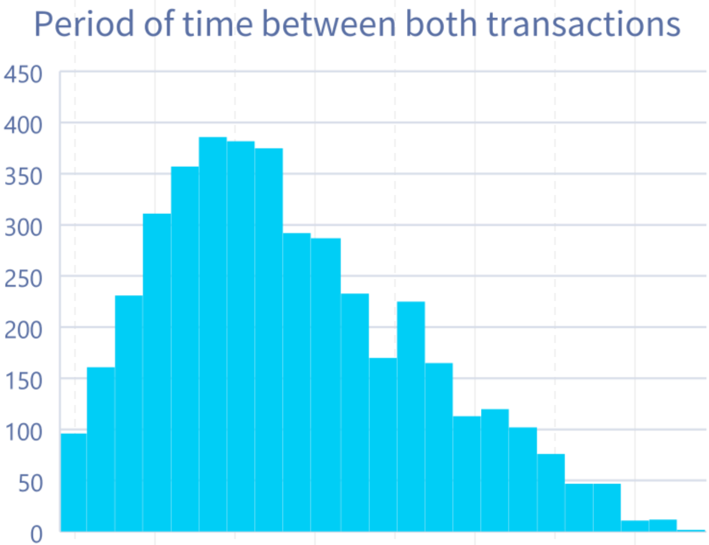
Repeated Sales Prices





Change of Prices and Time between Sales

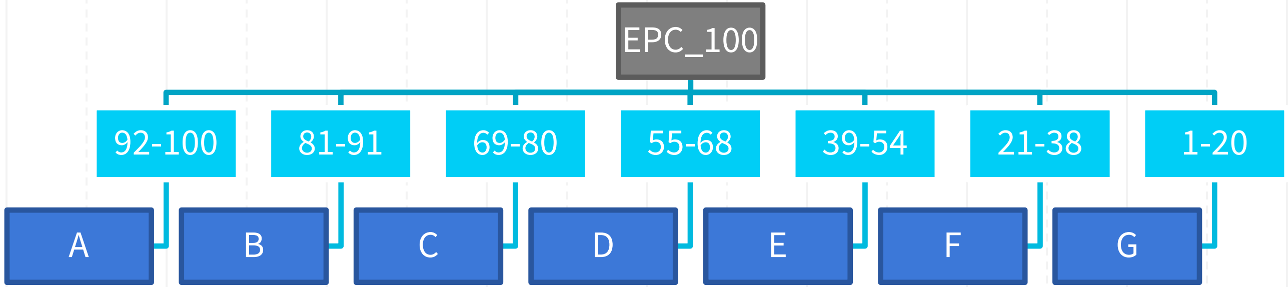




Transactional Prices Summary

Variable	Mean	Median	Std	Skewness	Kurtosis	Smallest	Largest	No. of observations	Normality
price_1	120191	100000	189751	23.96	689.70	6000	5660000	4201	1.30E-80
price_2	154575	120000	263756	24.49	706.57	25000	7900000	4201	8.32E-82
ln_price_1	11.46	11.51	0.65	-0.01	1.64	8.7	15.55	4201	1.62E-21
ln_price_2	11.75	11.7	0.52	1.16	5.01	10.13	15.88	4201	1.73E-35
perc_change_p2_to_p1	0.5	0.2	0.83	3.13	19.92	-0.62	10.42	4201	8.46E-59
days_between_sale	2400.1	2196	1236.9	0.56	-0.32	187	6156	4201	2.25E-28

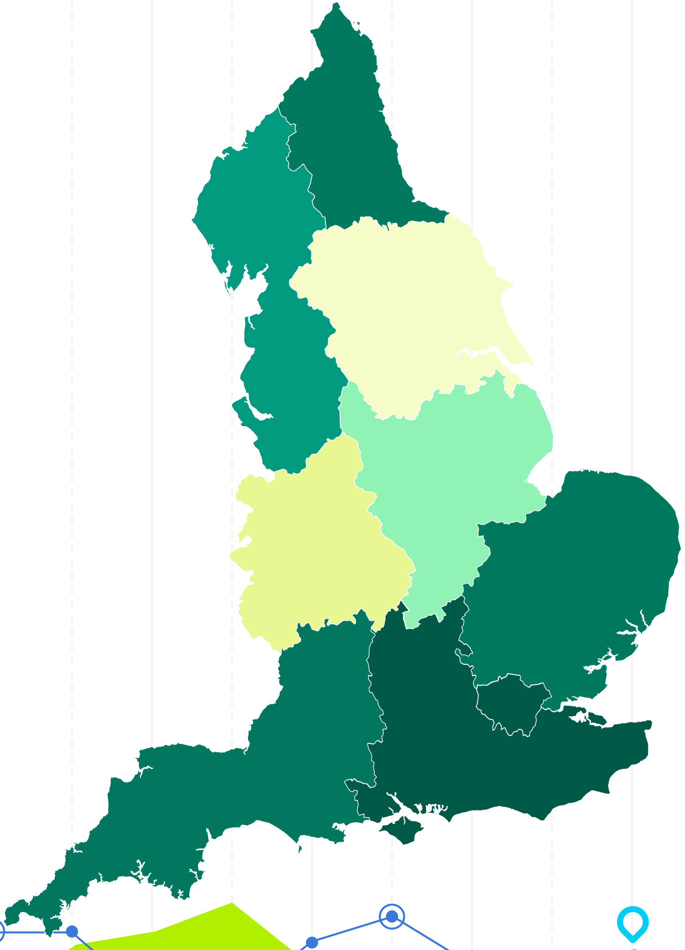
EPC Categorization

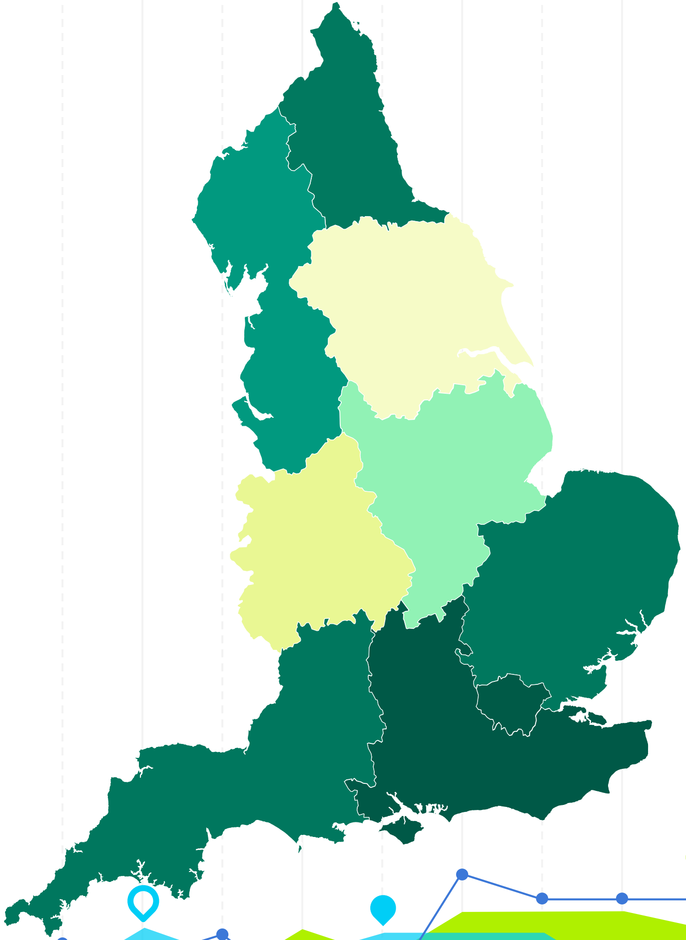


EPC Summary

EPC Band	Frequency	Fraction
EPC a	0	0.000000000
EPC b	379	0.090216615
EPC c	1442	0.343251607
EPC d	1480	0.352297072
EPC e	699	0.166388955
EPC f	162	0.038562247
EPC g	39	0.009283504

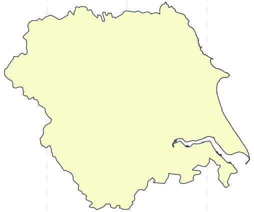
ENGLAND



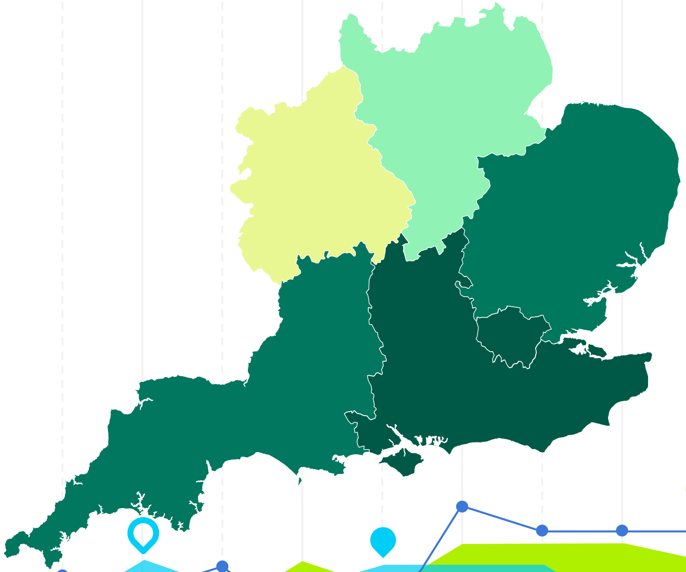


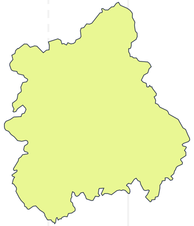
19.99%
North west

3.66%
North east

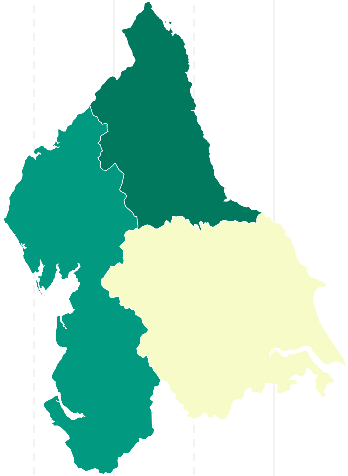


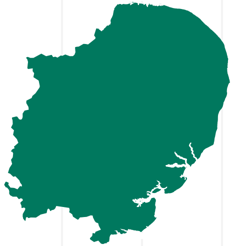
19.97%
Yorkshire and the Humber





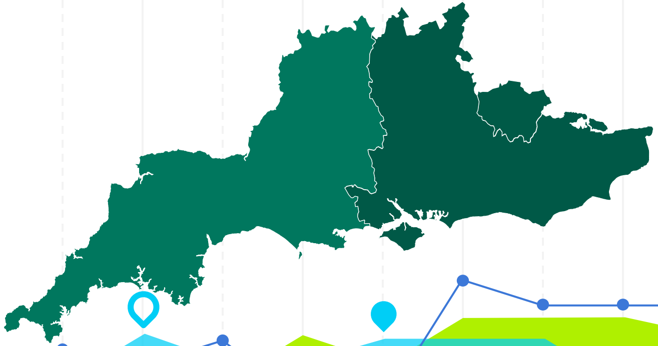
14.25%
West midlands



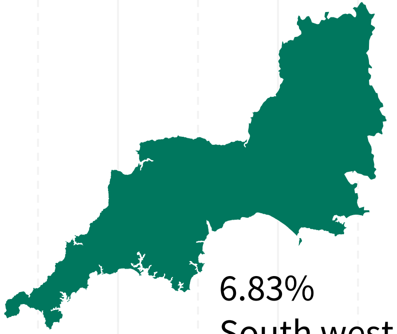


6.64%
East of England

10.35%
East midlands

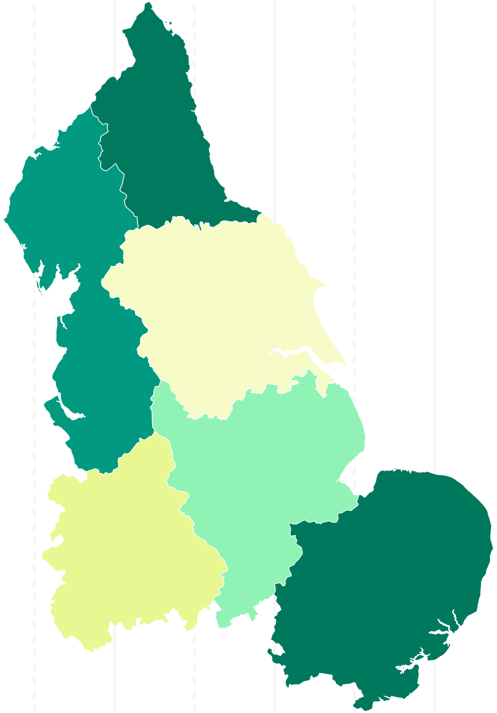


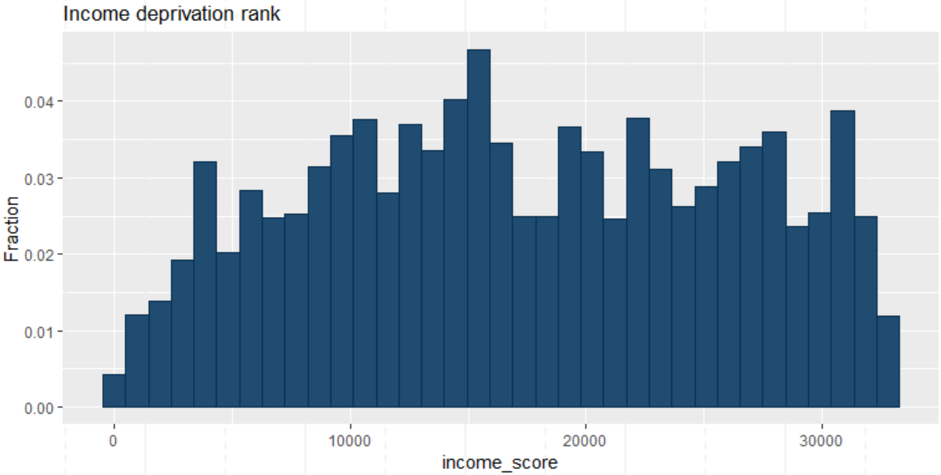
8.59%
London

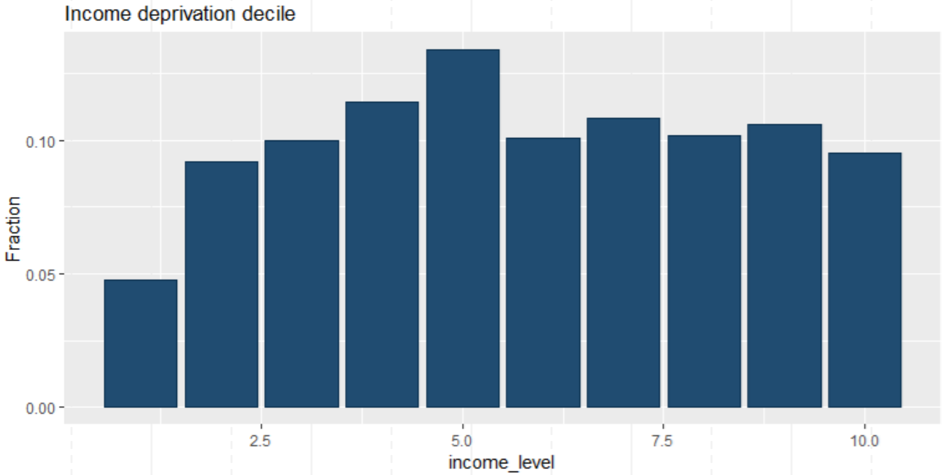


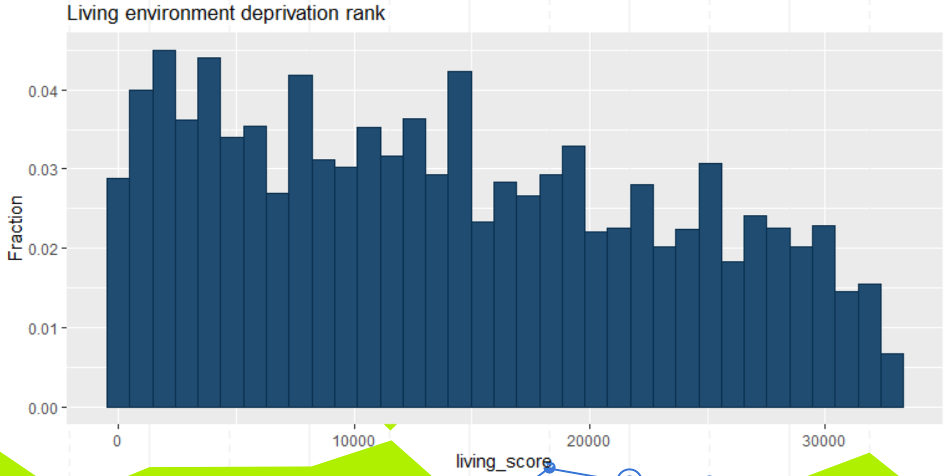
6.83%
South west

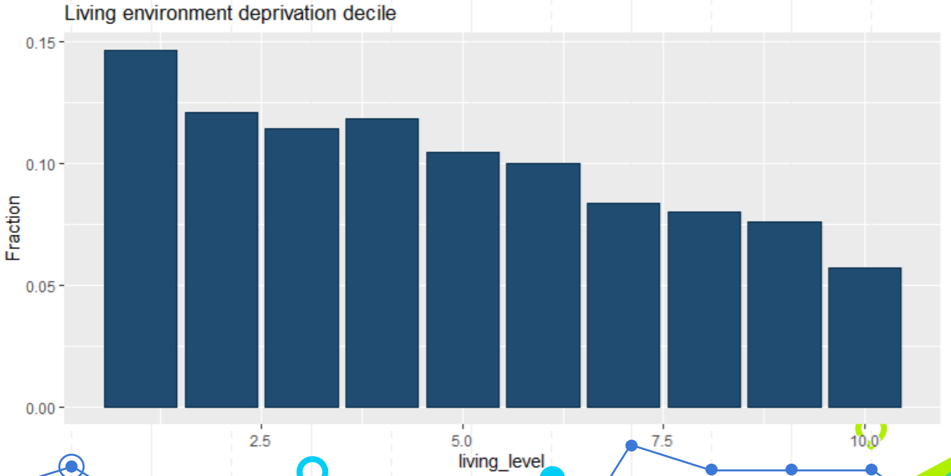
9.68%
South east

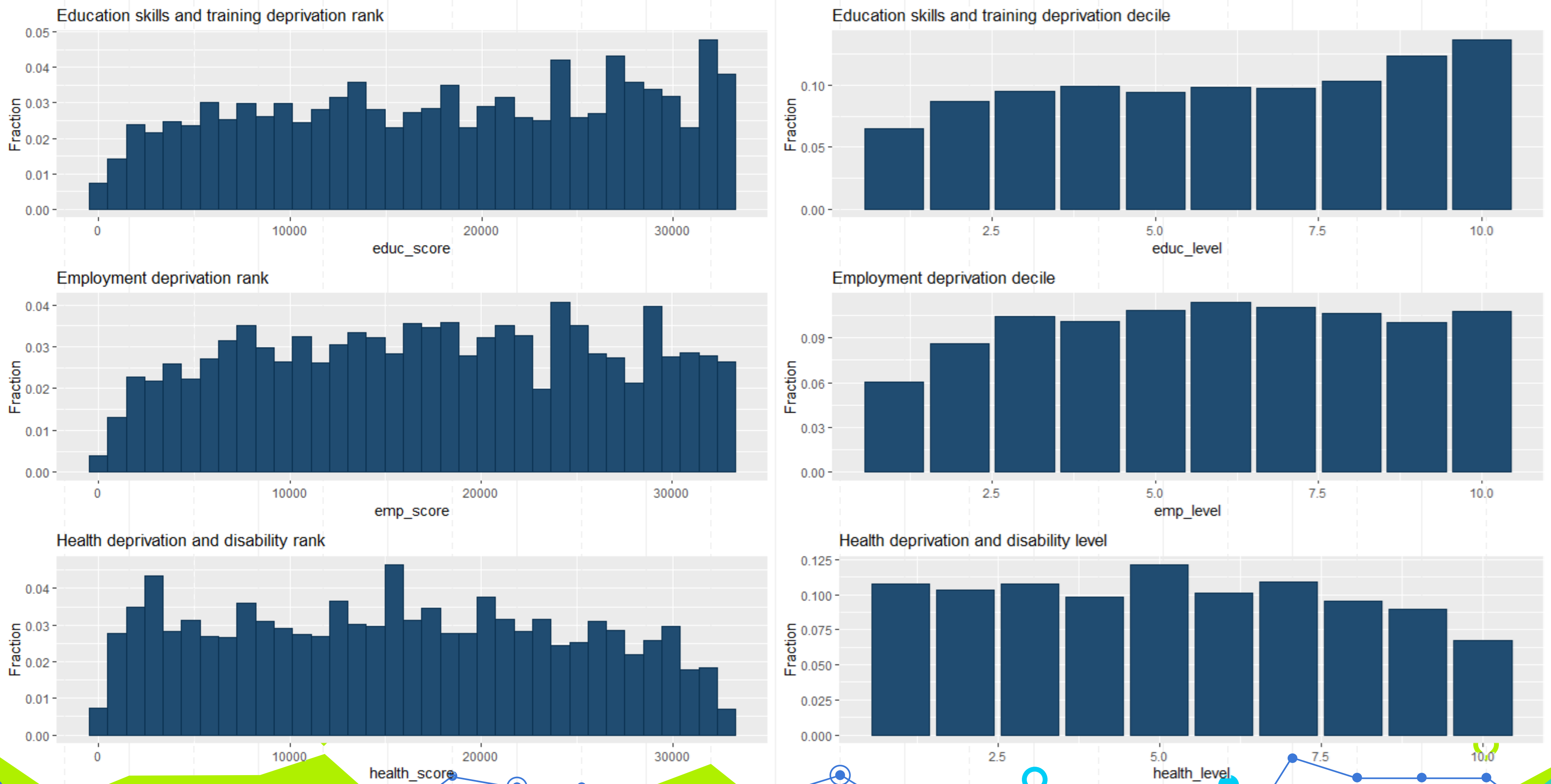



Income deprivation rank
Fraction
0.00
0.01
0.02
0.03
0.04
0
10000
20000
30000
income_score

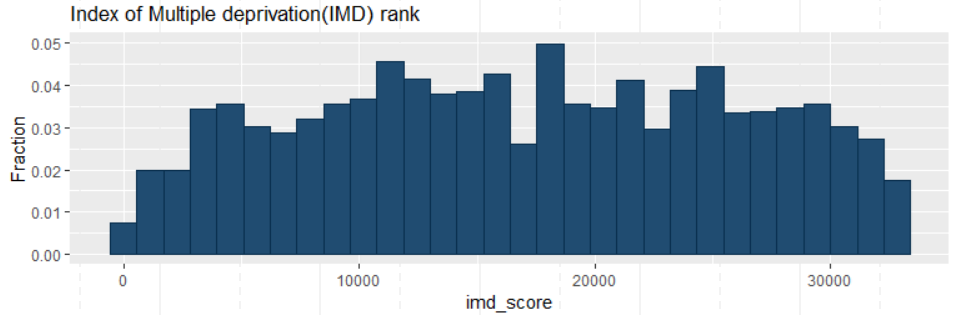

Income deprivation decile
Fraction
0.00
0.05
0.10
2.5
5.0
7.5
10.0
income_level

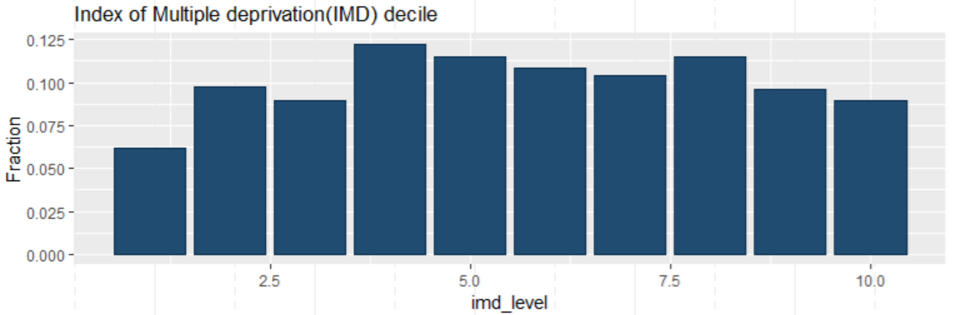

Living environment deprivation rank
Fraction
0.00
0.01
0.02
0.03
0.04
0
10000
20000
30000
living_score

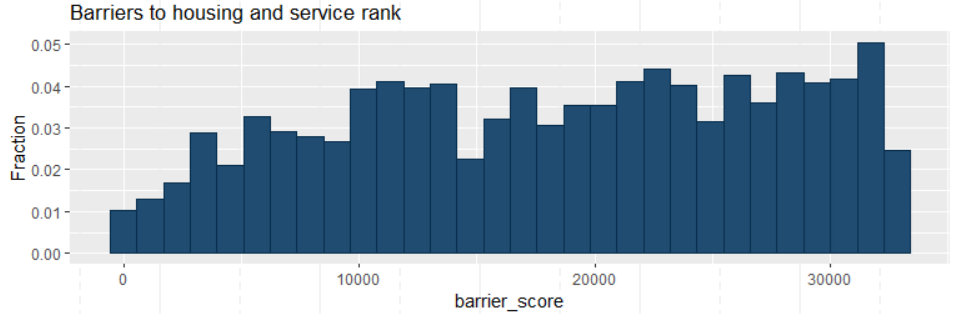

Living environment deprivation decile
Fraction
0.00
0.05
0.10
0.15
2.5
5.0
7.5
10.0
living_level

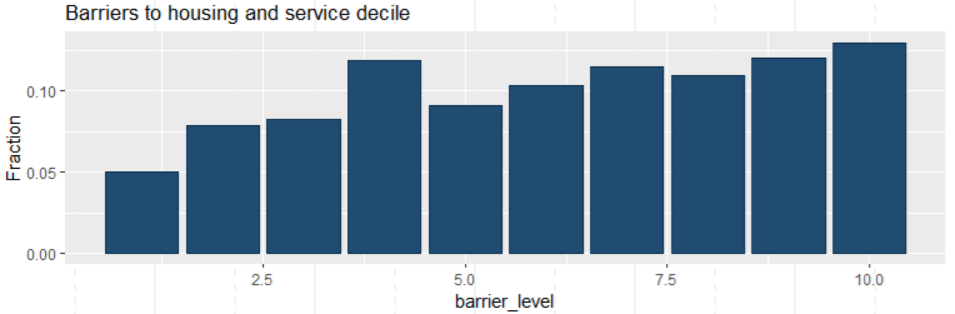


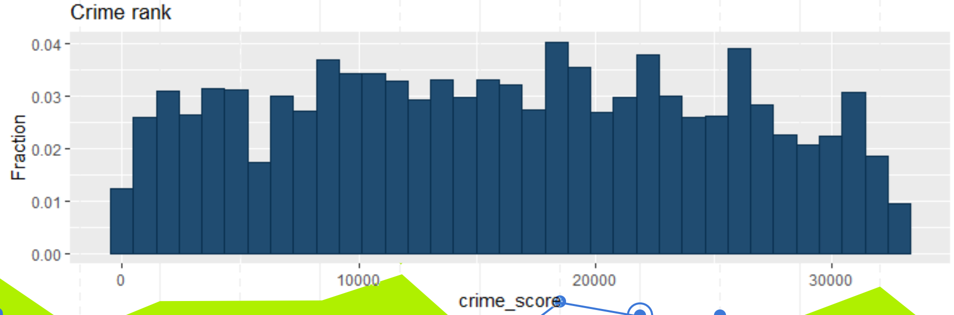
Education skills and training deprivation rank
Fraction
educ_score
Education skills and training deprivation decile
Fraction
educ_level
Employment deprivation rank
Fraction
emp_score
Employment deprivation decile
Fraction
emp_level
Health deprivation and disability rank
Fraction
health_score
Health deprivation and disability level
Fraction
health_level

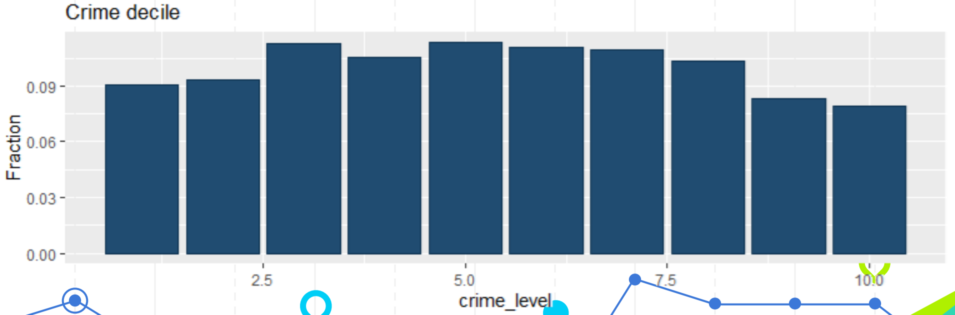

Index of Multiple deprivation(IMD) rank
Fraction
imd_score


Index of Multiple deprivation(IMD) decile
Fraction
imd_level


Barriers to housing and service rank
Fraction
barrier_score


Barriers to housing and service decile
Fraction
barrier_level

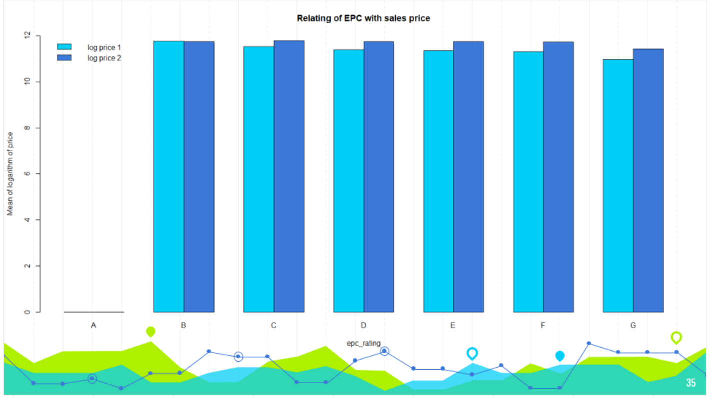

Crime rank
Fraction
crime_score

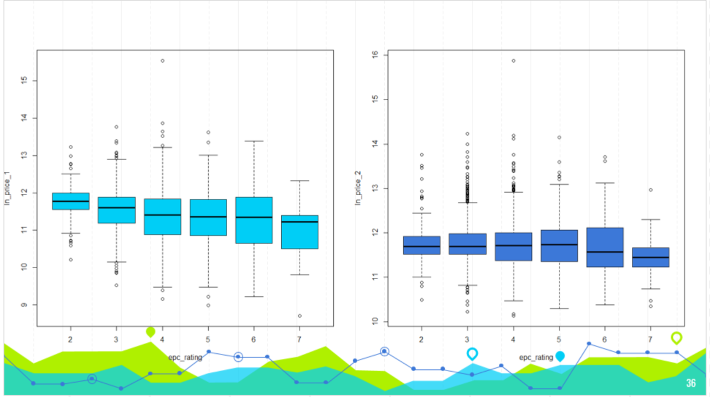

Crime decile
Fraction
crime_level

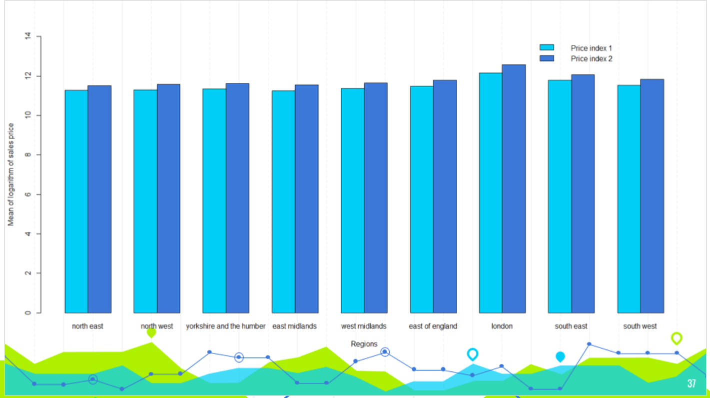
Hedonic Regression

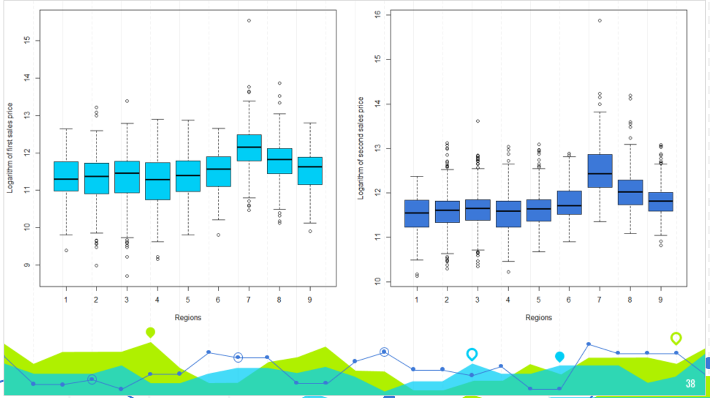
2

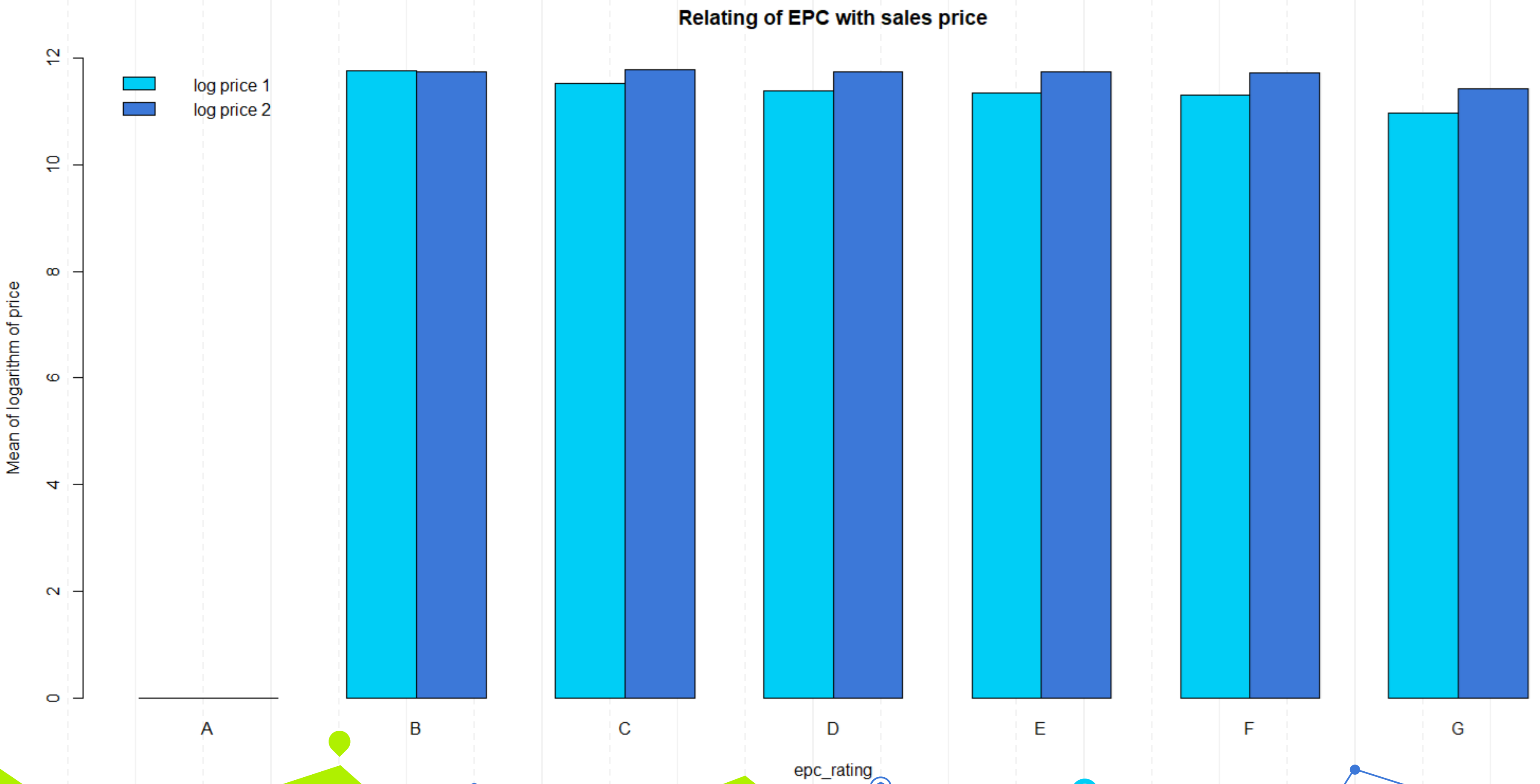
Preliminary Analysis



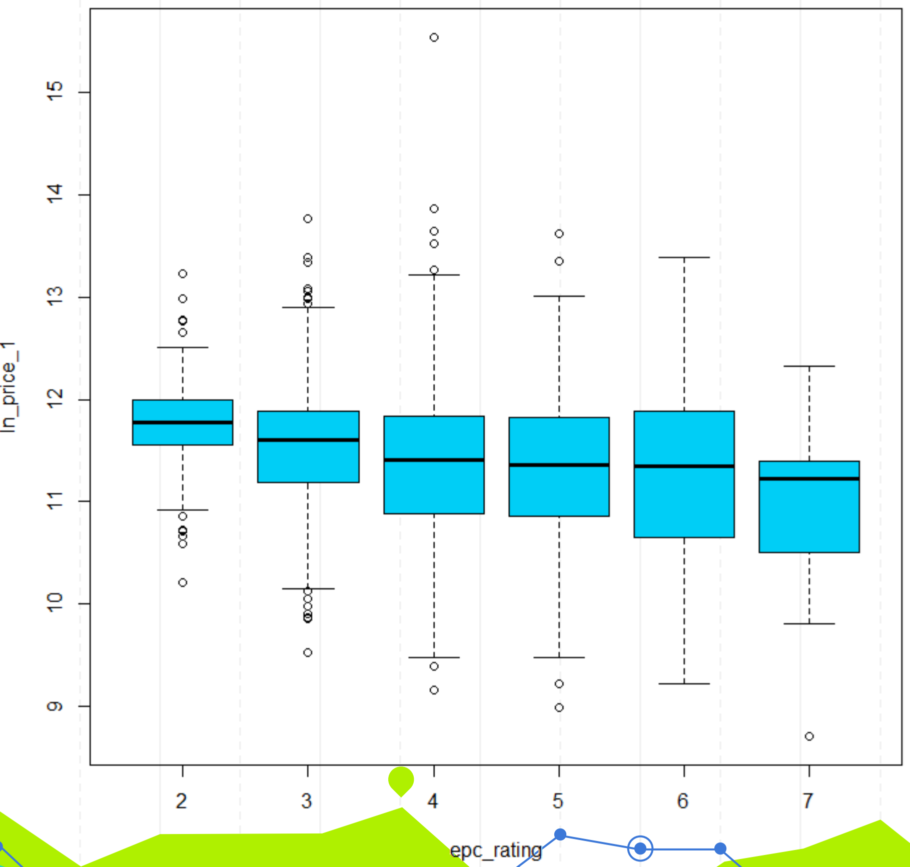




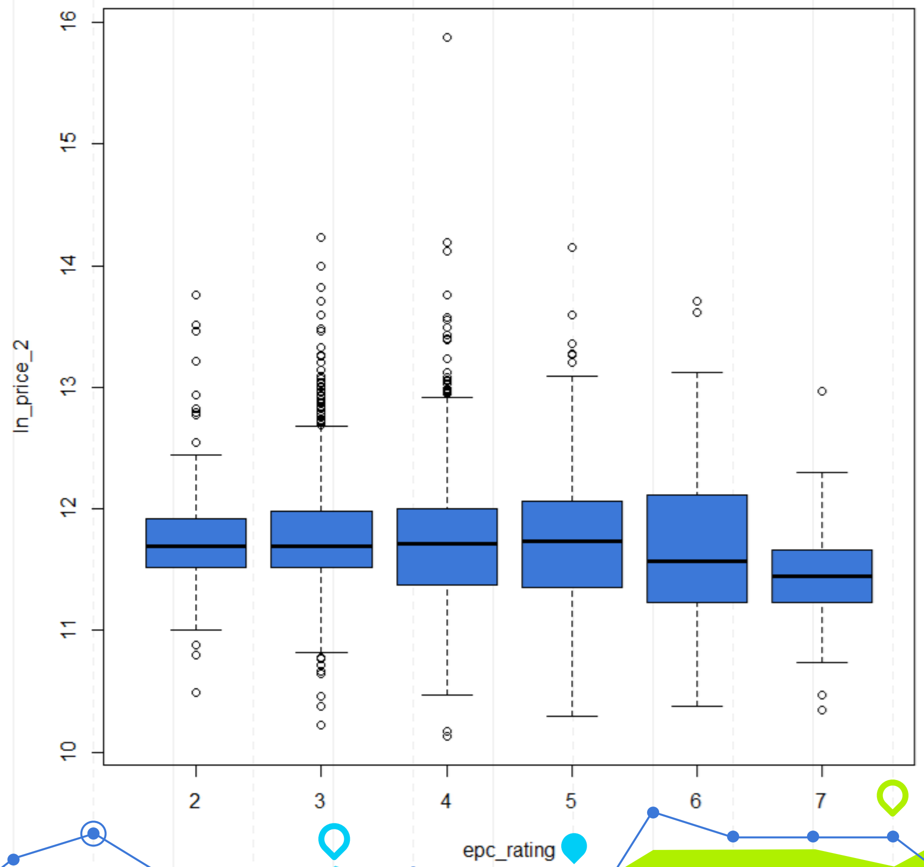




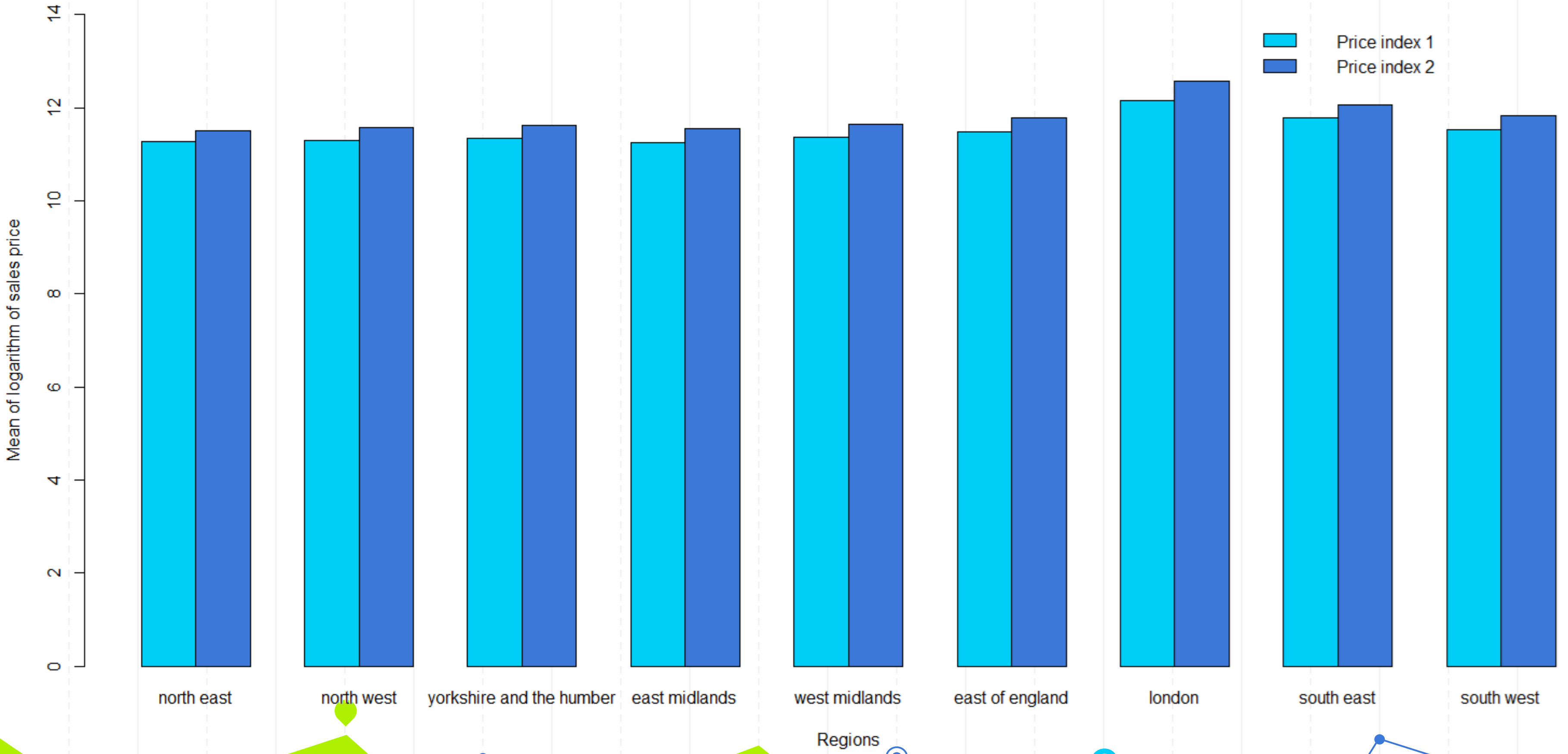
Relating of EPC with sales price
log price 1
log price 2
Mean of logarithm of price
0
2
4
6
8
10
12
A
B
C
D
E
F
G
epc_rating



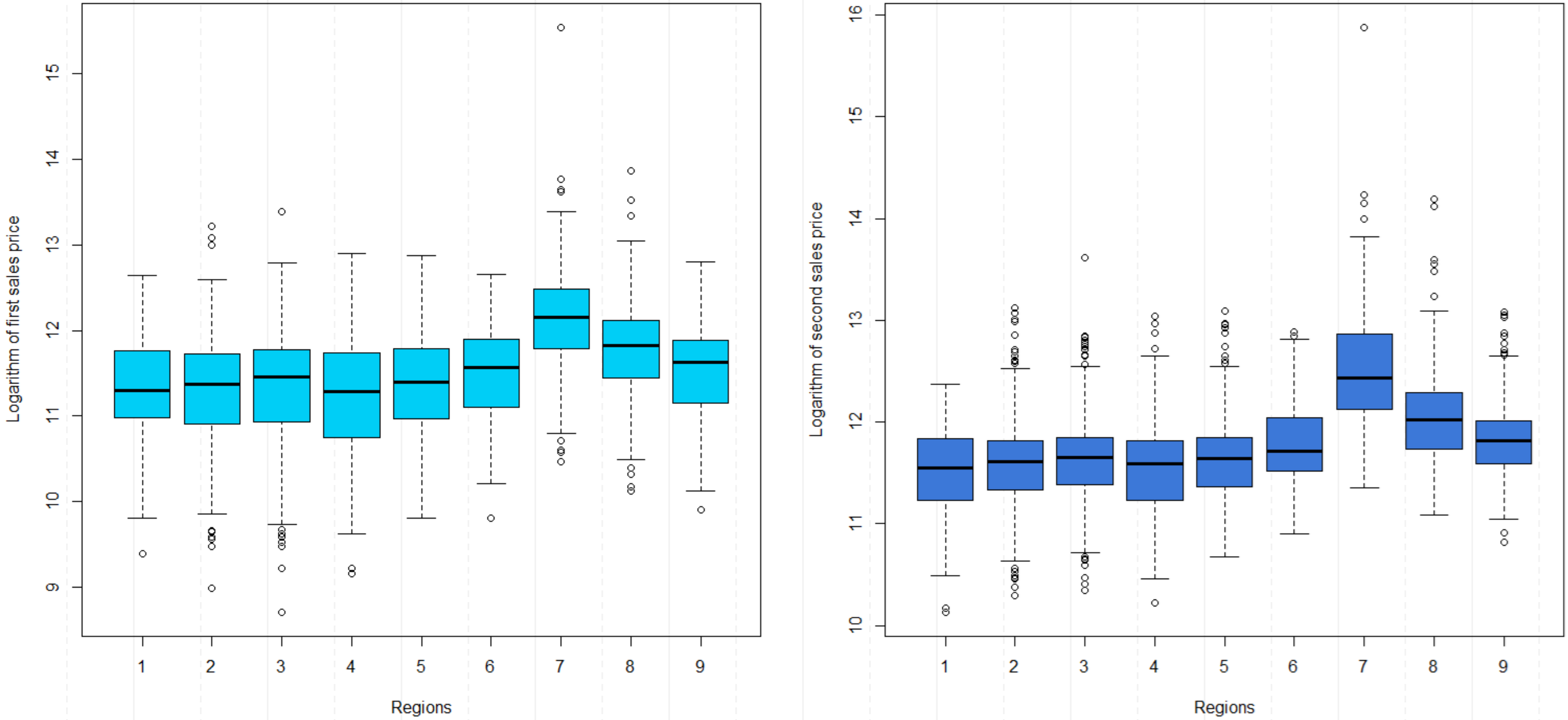
ln_price_1
15
14
13
12
11
10
9
2
3
4
5
6
7
epc_rating



ln_price_2
16
15
14
13
12
11
10
2
3
4
5
6
7
epc_rating



Mean of logarithm of sales price
0
2
4
6
8
10
12
14
Price index 1
Price index 2
north east
north west
yorkshire and the humber
east midlands
west midlands
east of england
london
south east
south west
Regions


Logarithm of first sales price
9
10
11
12
13
14
15
1
2
3
4
5
6
7
8
9
Regions
Logarithm of second sales price
10
11
12
13
14
15
16
1
2
3
4
5
6
7
8
9
Regions

BRIEF INTRODUCTION TO HEDONIC REGRESSION

- **Hedonic regression** is a revealed preference method of estimating the demand for a good, or equivalently its value to consumers
- An attribute vector, which may be a dummy variable, is assigned to each characteristic or group of characteristics.

$[EPC, Regional\ Variables]$

Hedonic Model

$$\ln(P_{it}) = \beta_{0t} + \sum_{j=1}^{K} \beta_{jt} x_{ijt} + \varepsilon_i$$

R

$$\ln(\widehat{P}_t) = \hat{\beta}_{0t} + \sum_{j=1}^{K} \hat{\beta}_{jt} x_{jt}$$

Partial Changes

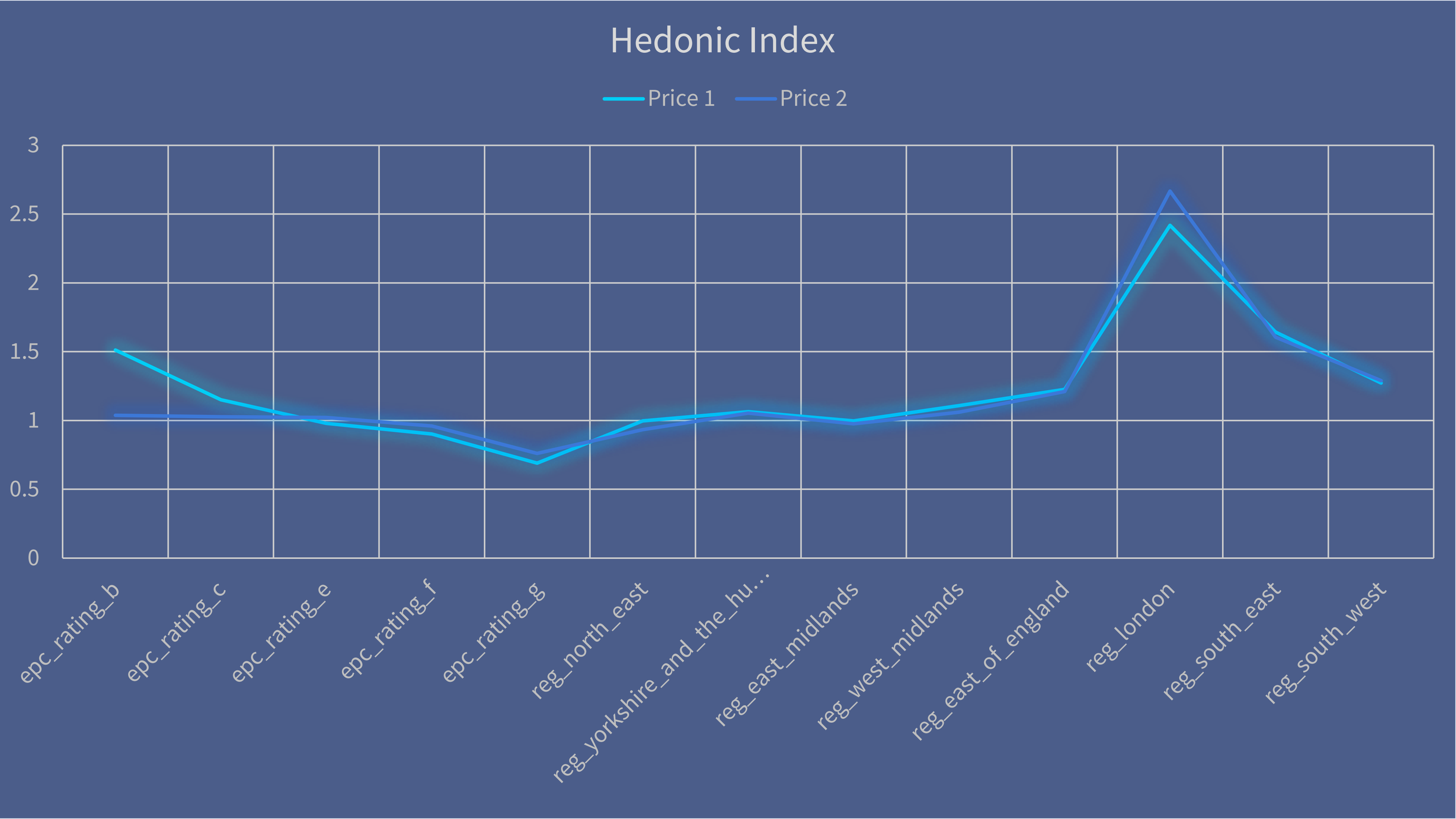
$$\ln \hat{P}_t - \ln \hat{P}'_t = \hat{\beta}_{jt} \Delta x_{jt}$$

$$\Longrightarrow \hat{P}'_t = \hat{P}_t e^{\hat{\beta}_{jt} \Delta x_{jt}}$$

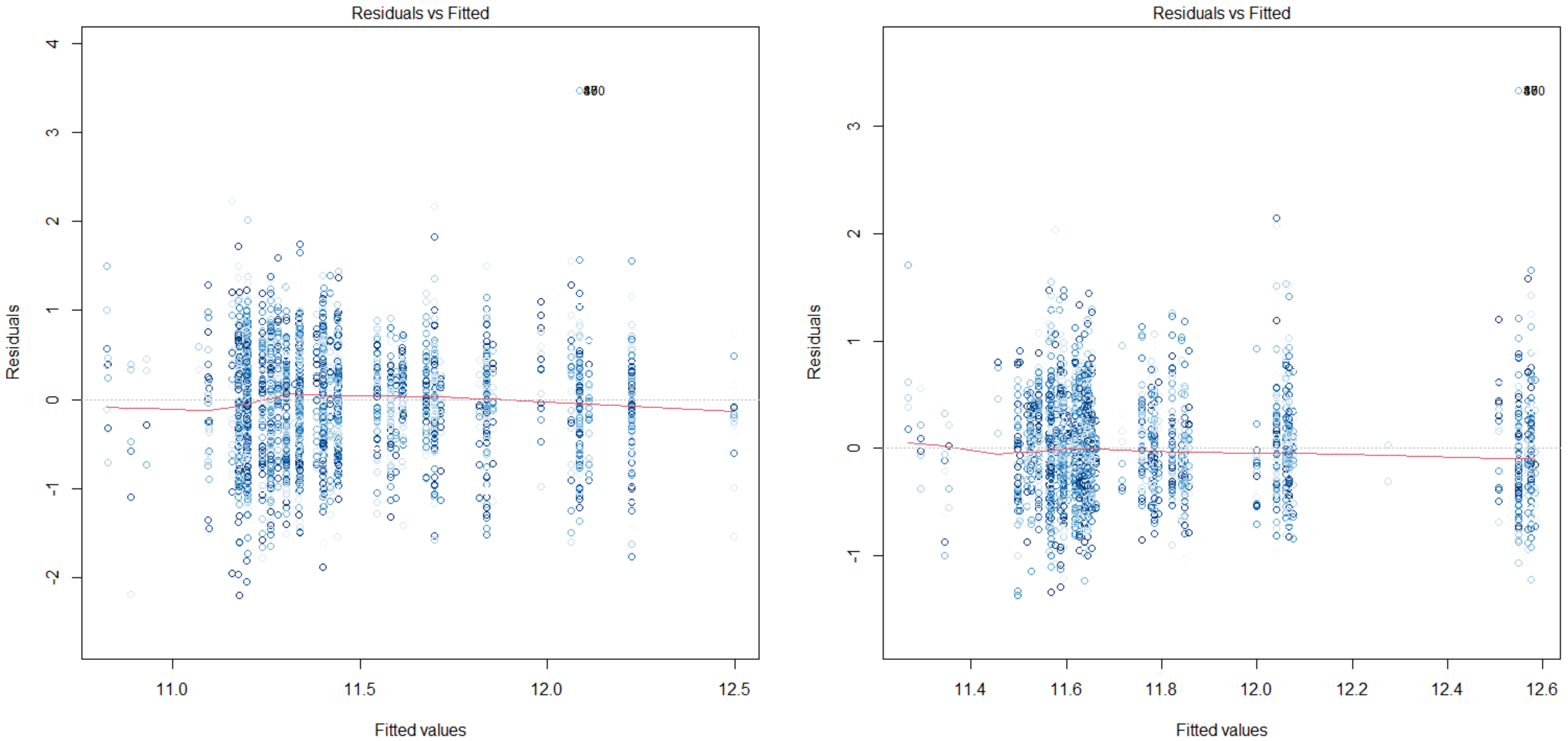
$$\Longrightarrow \hat{P}'_t = \hat{P}_t e^{\hat{\beta}_{jt}}$$

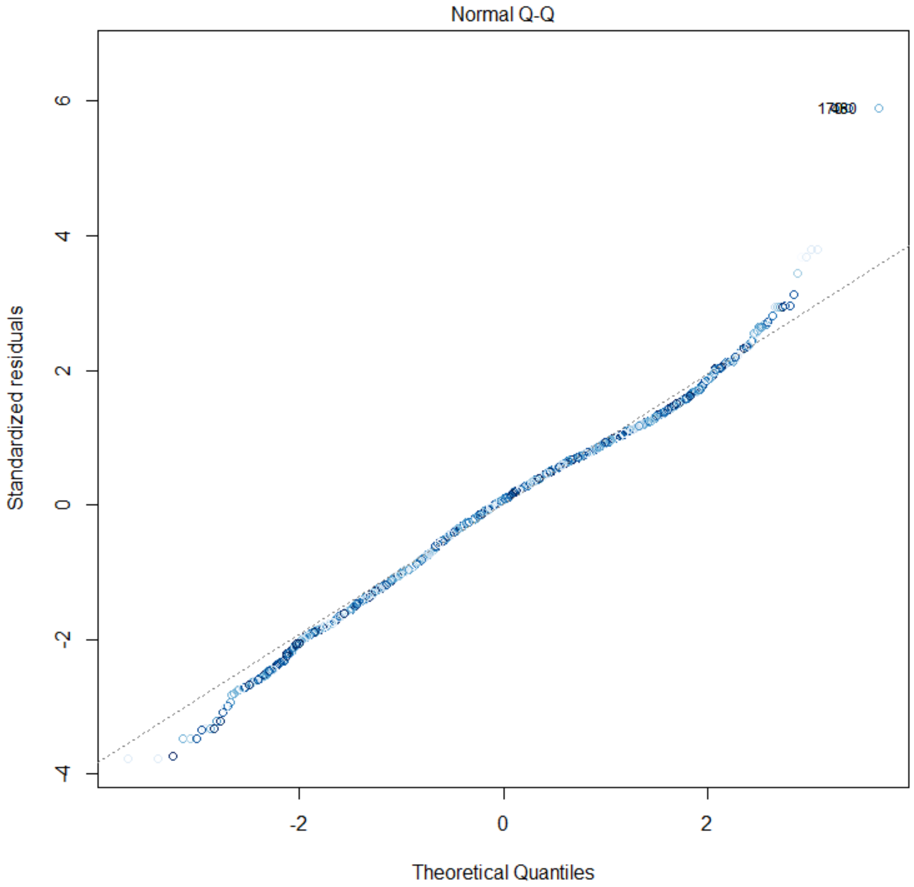
Model Choices

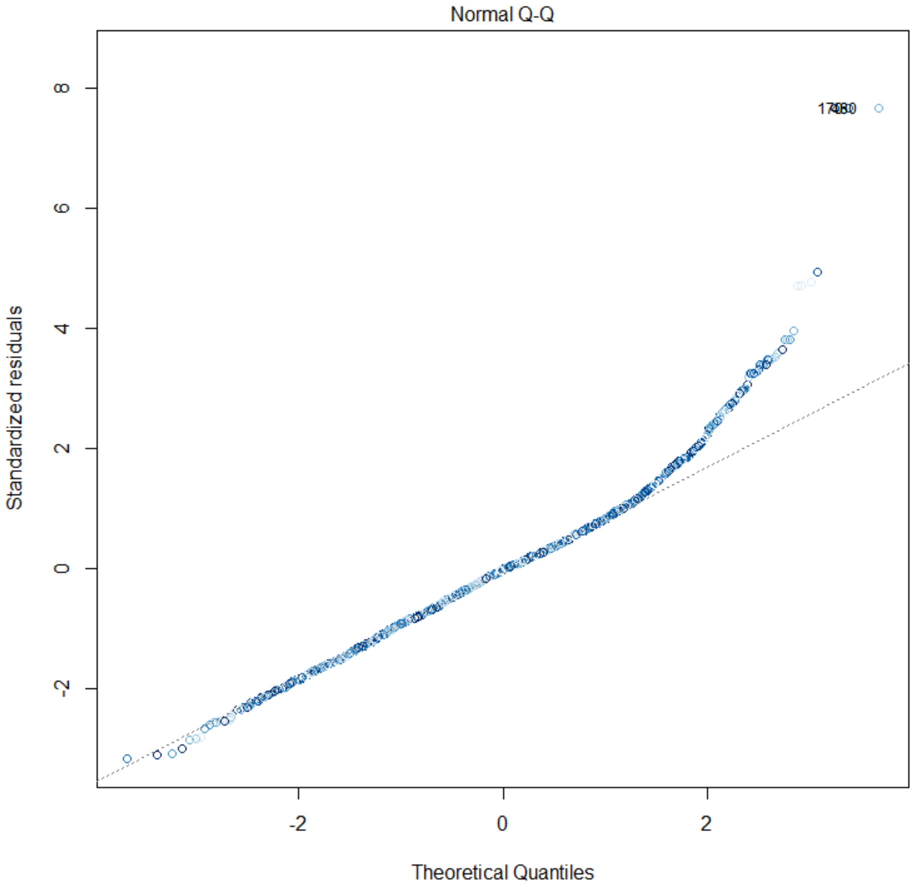
- Referenced on D [∵ EPC D has most number of values]
- Referenced on North-west[∵ North-west has most number of values]

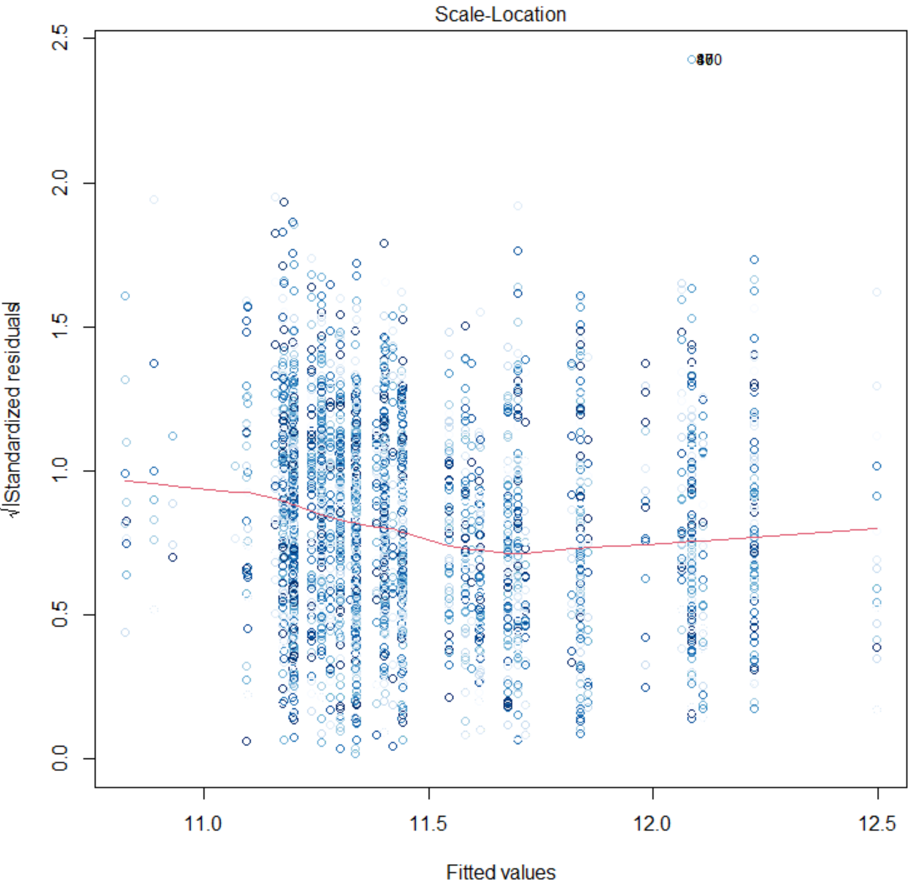


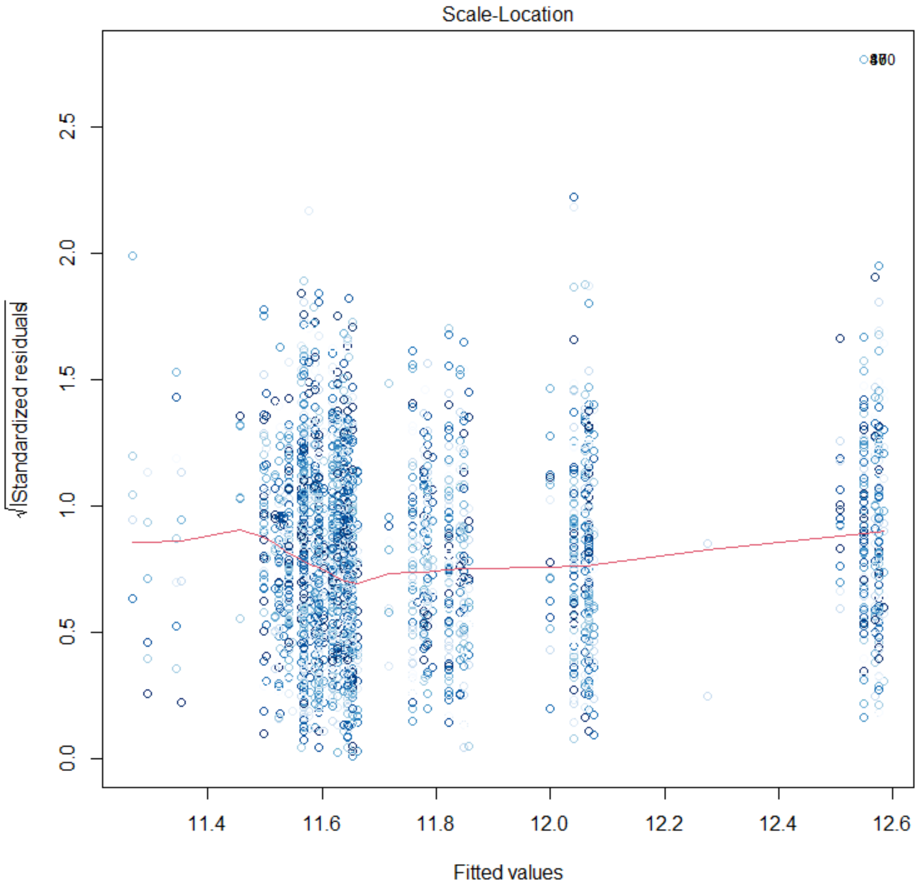
Hedonic Index
Price 1
Price 2
3
2.5
2
1.5
1
0.5
0
epc_rating_b
epc_rating_c
epc_rating_e
epc_rating_f
epc_rating_g
reg_north_east
reg_yorkshire_and_the_hu...
reg_east_midlands
reg_west_midlands
reg_east_of_england
reg_london
reg_south_east
reg_south_west

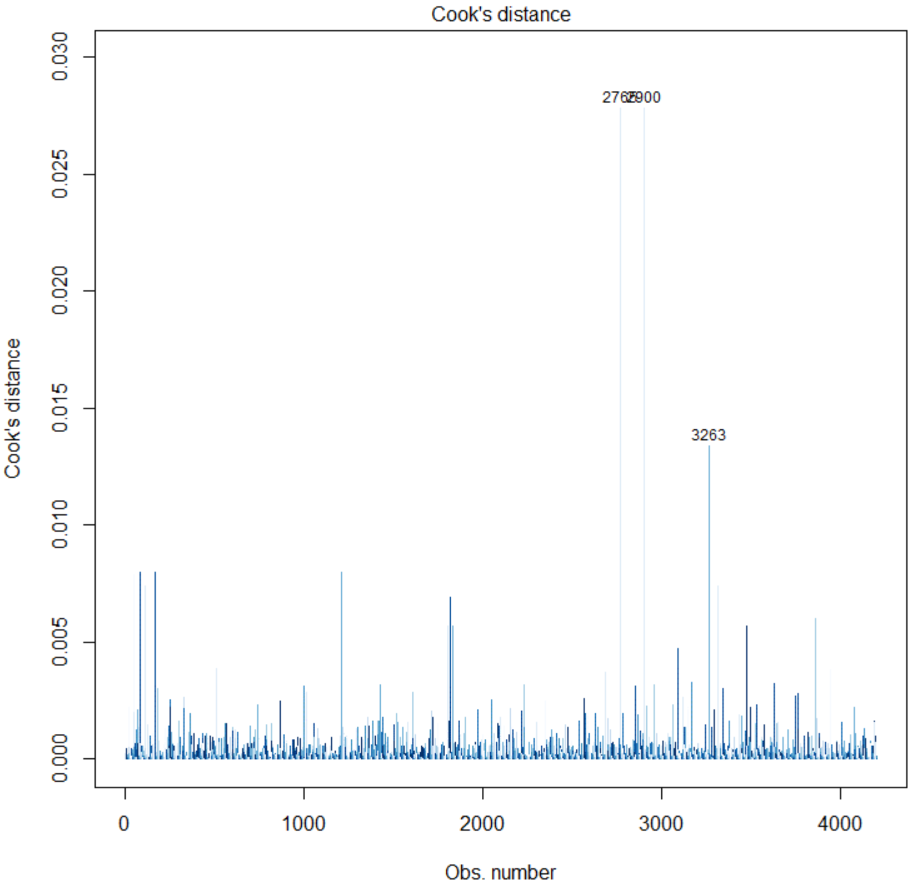
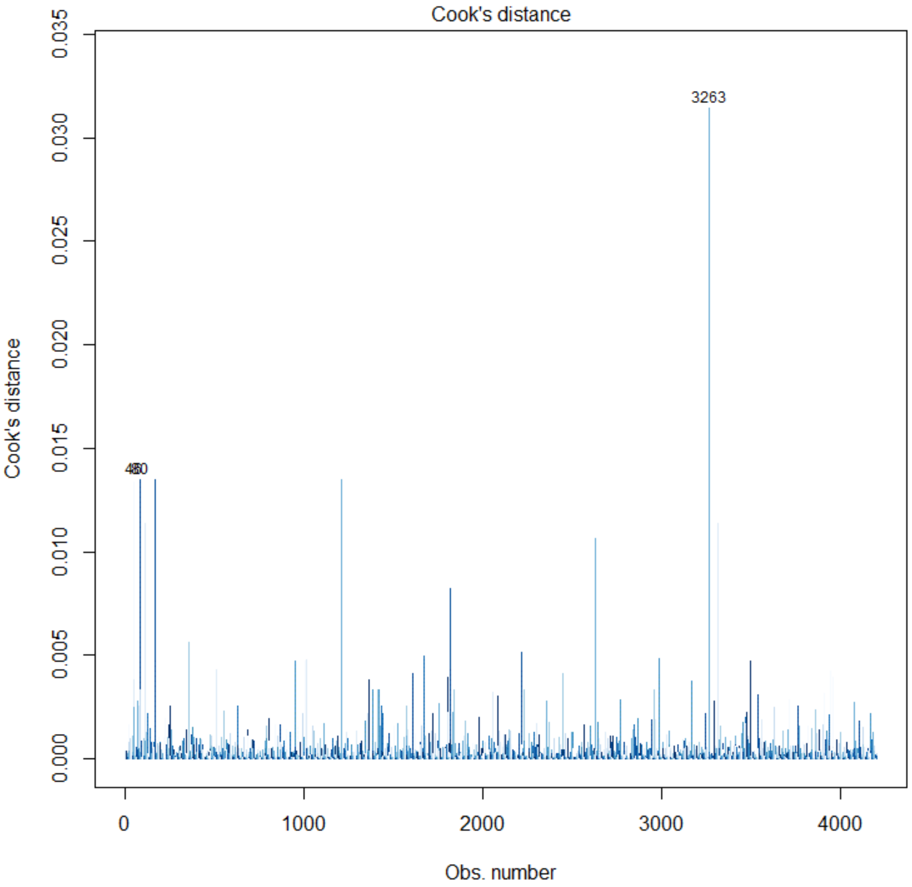

Residuals vs Fitted
Residuals
Fitted values
Residuals vs Fitted
Residuals
Fitted values

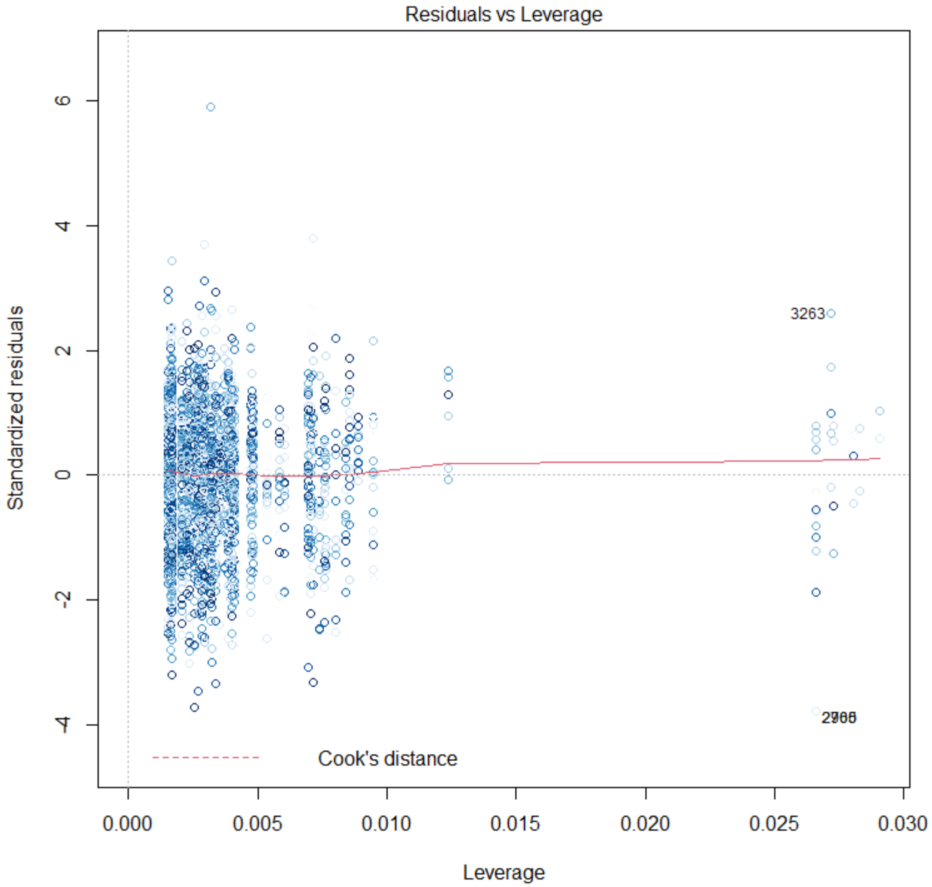

Normal Q-Q
Standardized residuals
Theoretical Quantiles
-4
-2
0
2
4
6
-2
0
2
17080

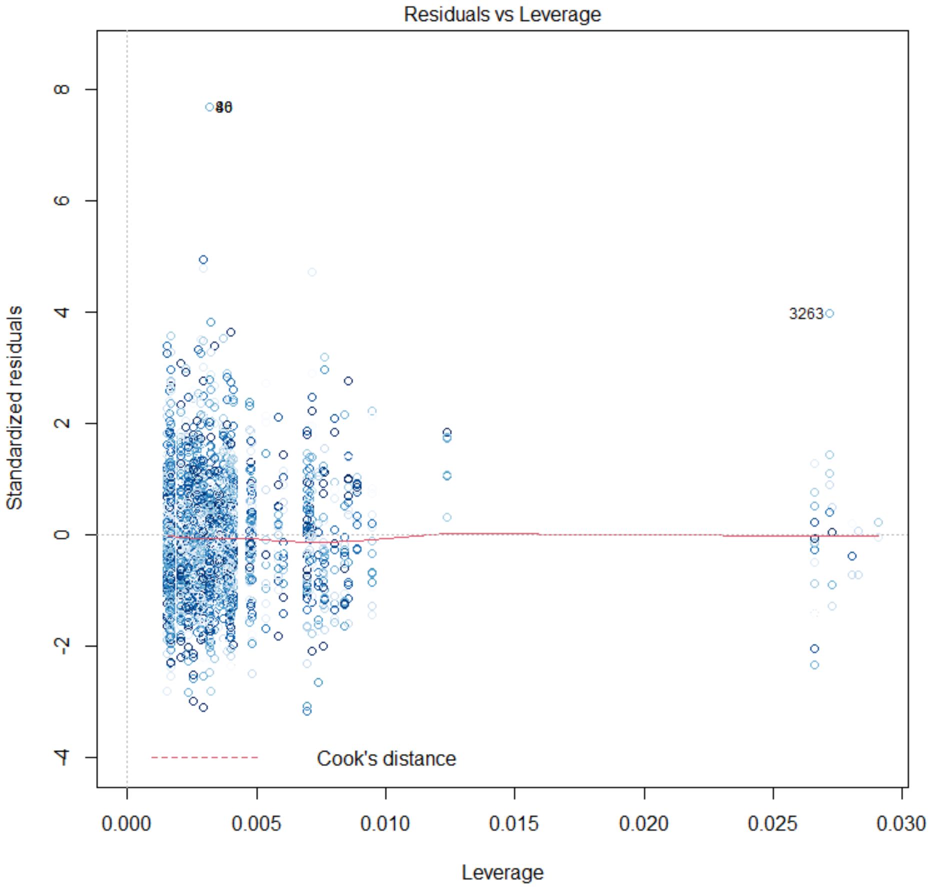

Normal Q-Q
Standardized residuals
Theoretical Quantiles
-2
0
2
4
6
8
-2
0
2
17080

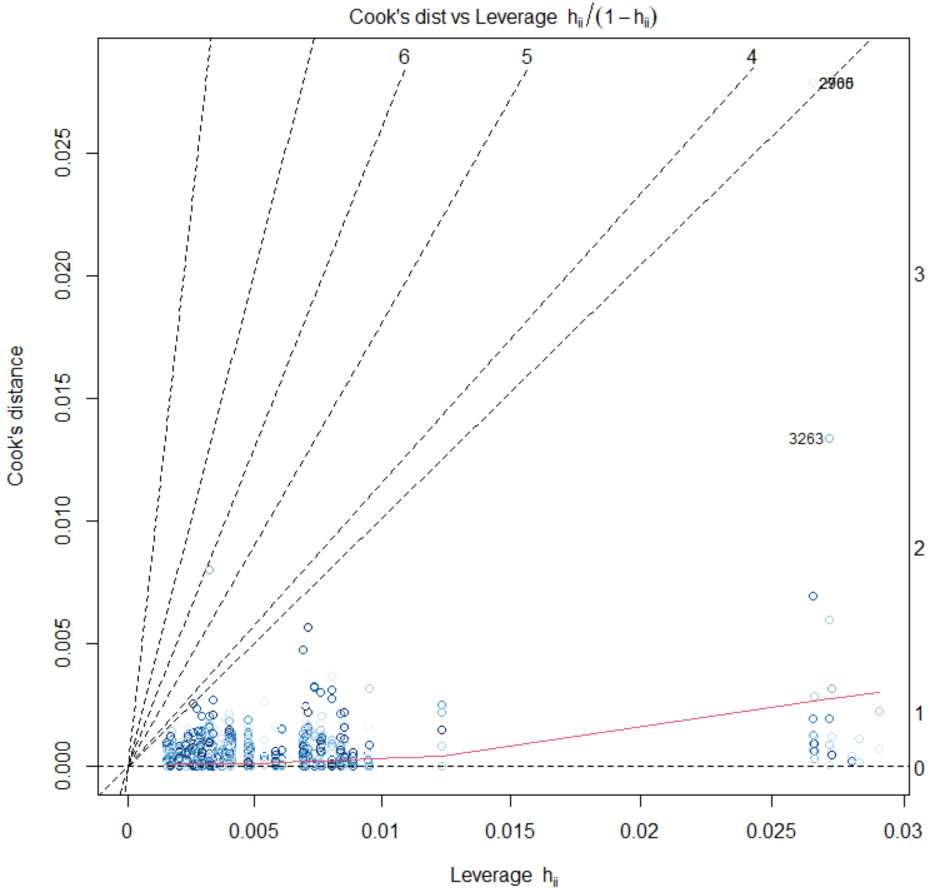

Scale-Location
√|Standardized residuals|
Fitted values
0.0
0.5
1.0
1.5
2.0
2.5
11.0
11.5
12.0
12.5


Scale-Location
√|Standardized residuals|
Fitted values
0.0
0.5
1.0
1.5
2.0
2.5
11.4
11.6
11.8
12.0
12.2
12.4
12.6

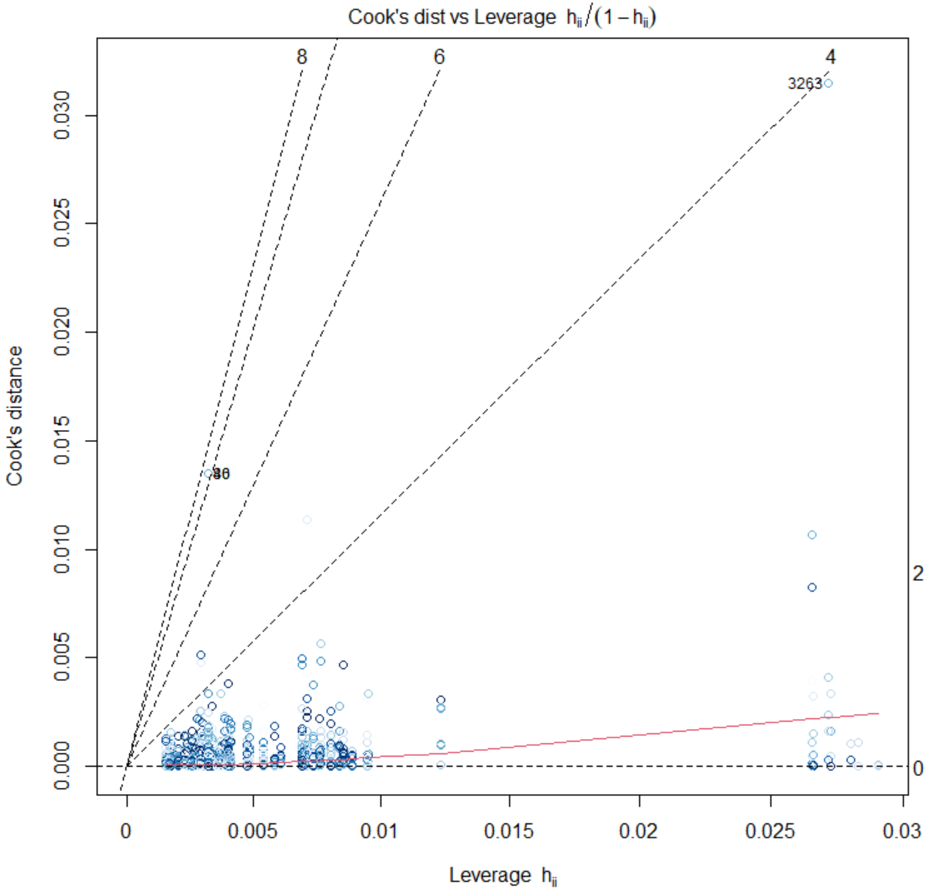

Cook's distance
Cook's distance
Obs. number
2765
2900
3263
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0
1000
2000
3000
4000

Cook's distance
Cook's distance
Obs. number
3263
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0
1000
2000
3000
4000


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
3263
2766
-4
-2
0
2
4
6
0.000
0.005
0.010
0.015
0.020
0.025
0.030


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
3263
80
-4
-2
0
2
4
6
8
0.000
0.005
0.010
0.015
0.020
0.025
0.030



Cook's dist vs Leverage $h_{ii}/(1-h_{ii})$
Cook's distance
Leverage h_{ii}
0.000
0.005
0.010
0.015
0.020
0.025
0
0.005
0.01
0.015
0.02
0.025
0.03
6
5
4
3
2
1
0
3263



Cook's dist vs Leverage $h_{ii}/(1-h_{ii})$
Cook's distance
Leverage h_{ii}
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0
0.005
0.01
0.015
0.02
0.025
0.03
8
6
4
2
0
3263

Price vs Socio-Economic

3

Normality

- An assumption is made that the errors are normal.
- We also know that the predicted value has a normal distribution under a fixed value of explanatory variables

Box Cox Transformations

- So, we search for transformations which may make it normal.
- $$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda} \; if \; \lambda \neq 0 \\ \ln(y) \; if \; \lambda = 0 \end{cases}$$

Then choose the lambda which makes it closest to the normal distribution curve.

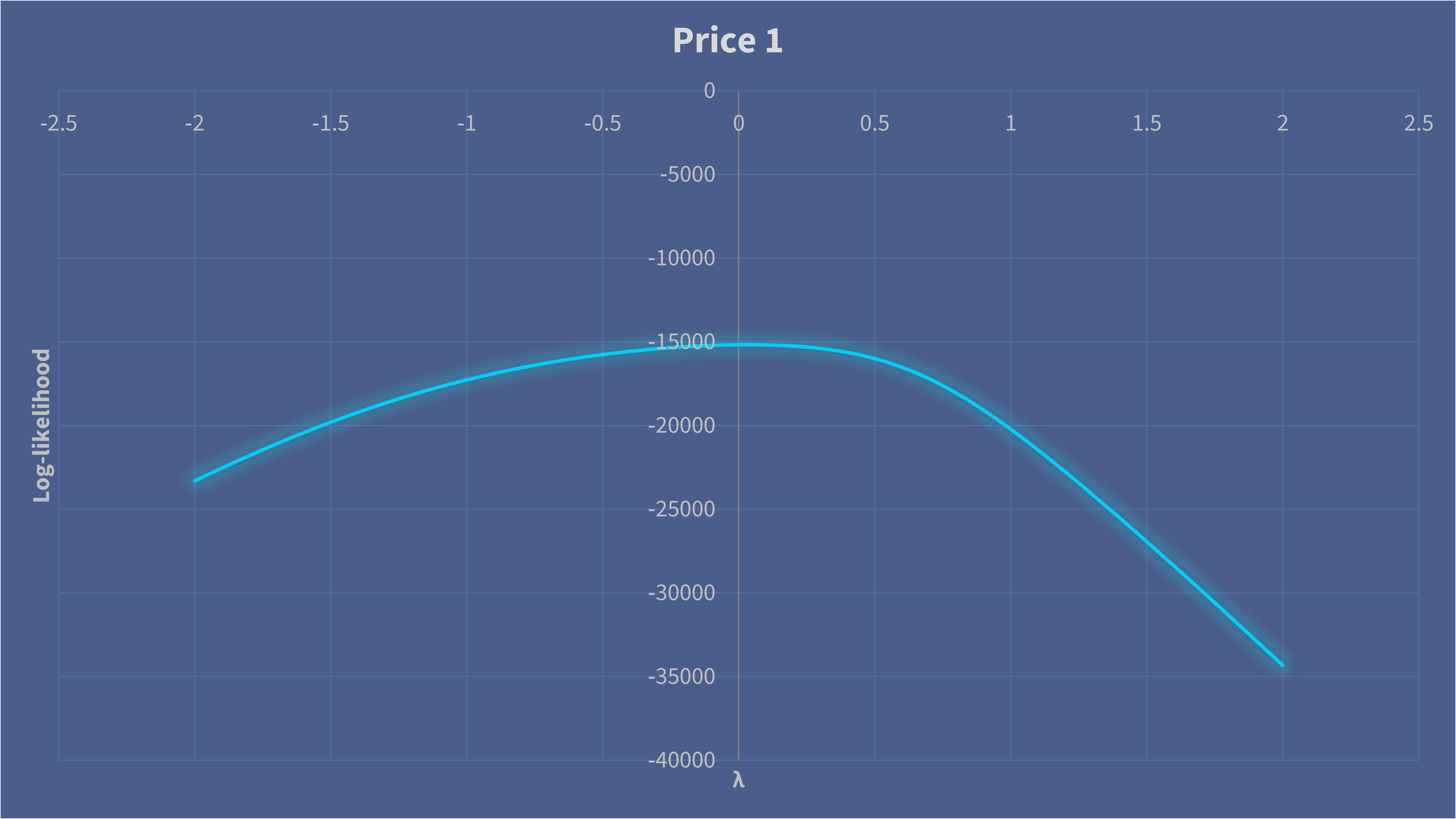
OUR PROCESS IS EASY

y

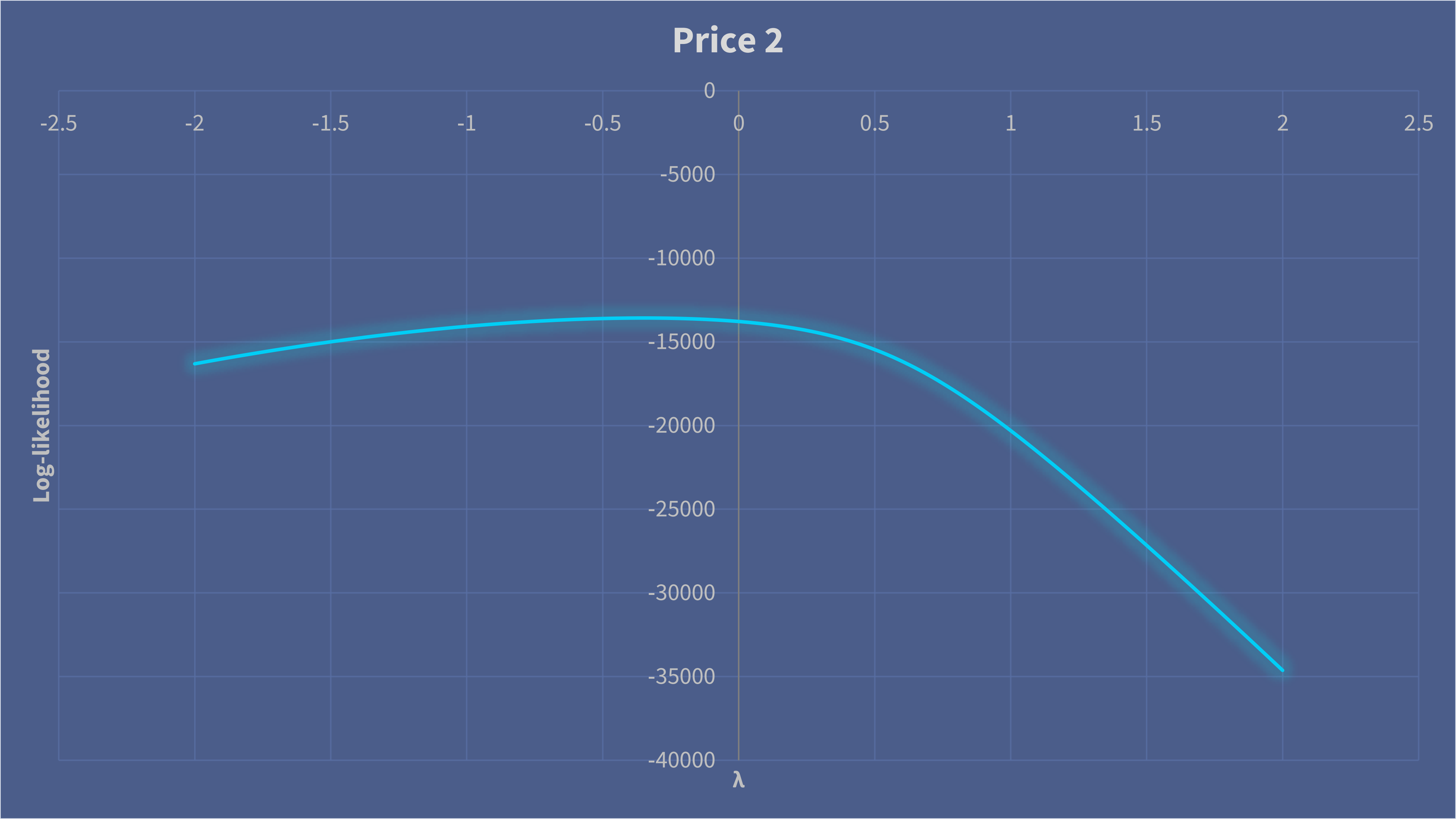
$f(y)$

Perform Regression

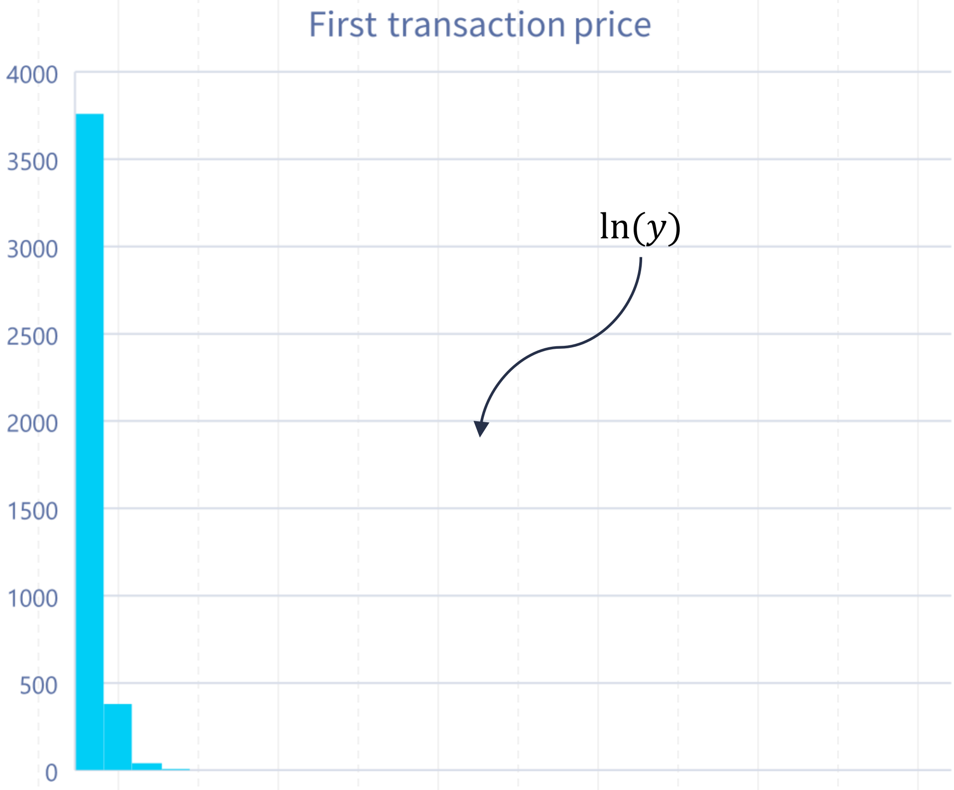
$$\hat{f}(y) = \sum_j \hat{\beta}_j x_j + \hat{\beta}_0$$



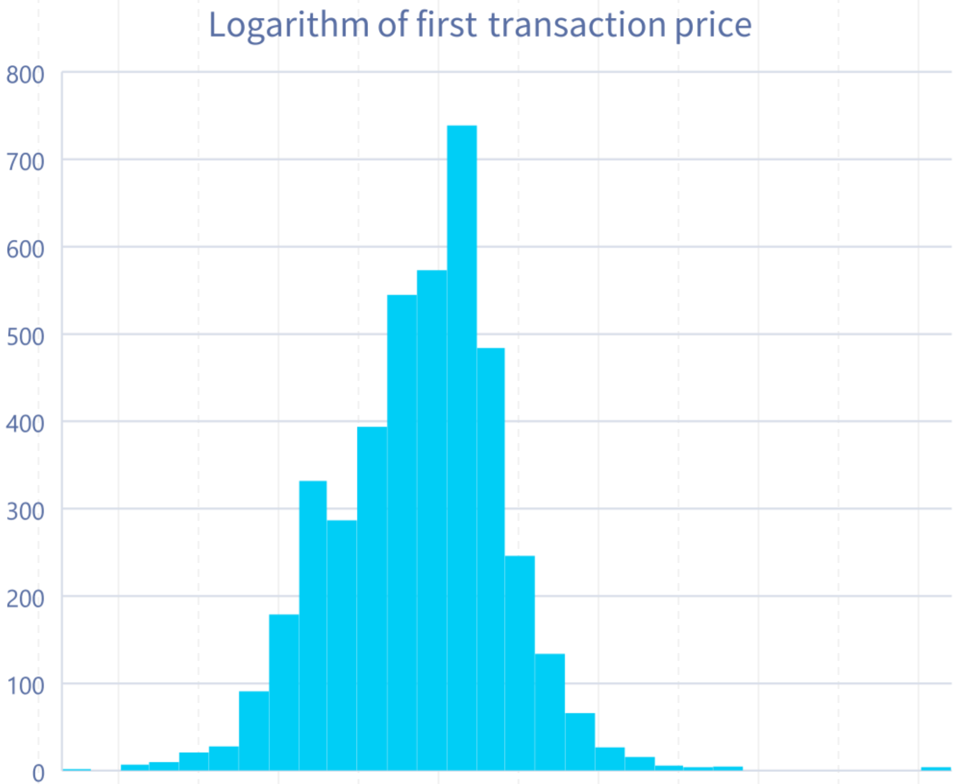
Price 1
Log-likelihood
λ
0
-5000
-10000
-15000
-20000
-25000
-30000
-35000
-40000
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5



Price 2
Log-likelihood
λ
0
-5000
-10000
-15000
-20000
-25000
-30000
-35000
-40000
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5

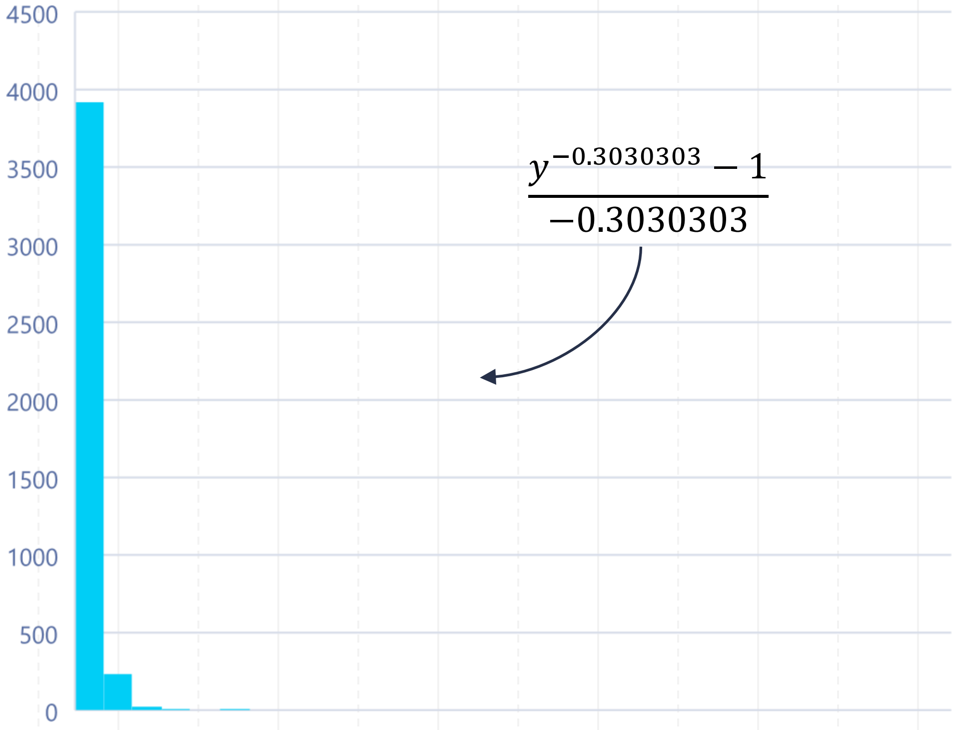


First transaction price
4000
3500
3000
2500
2000
1500
1000
500
0
$\ln(y)$

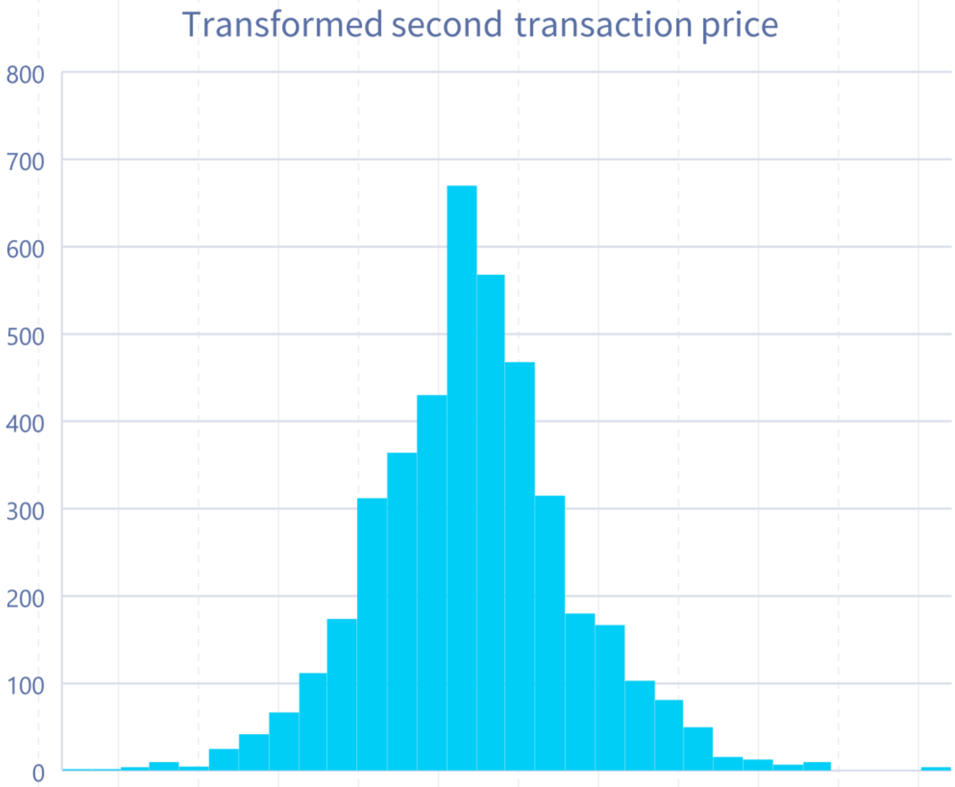


Logarithm of first transaction price
800
700
600
500
400
300
200
100
0

Second transaction price



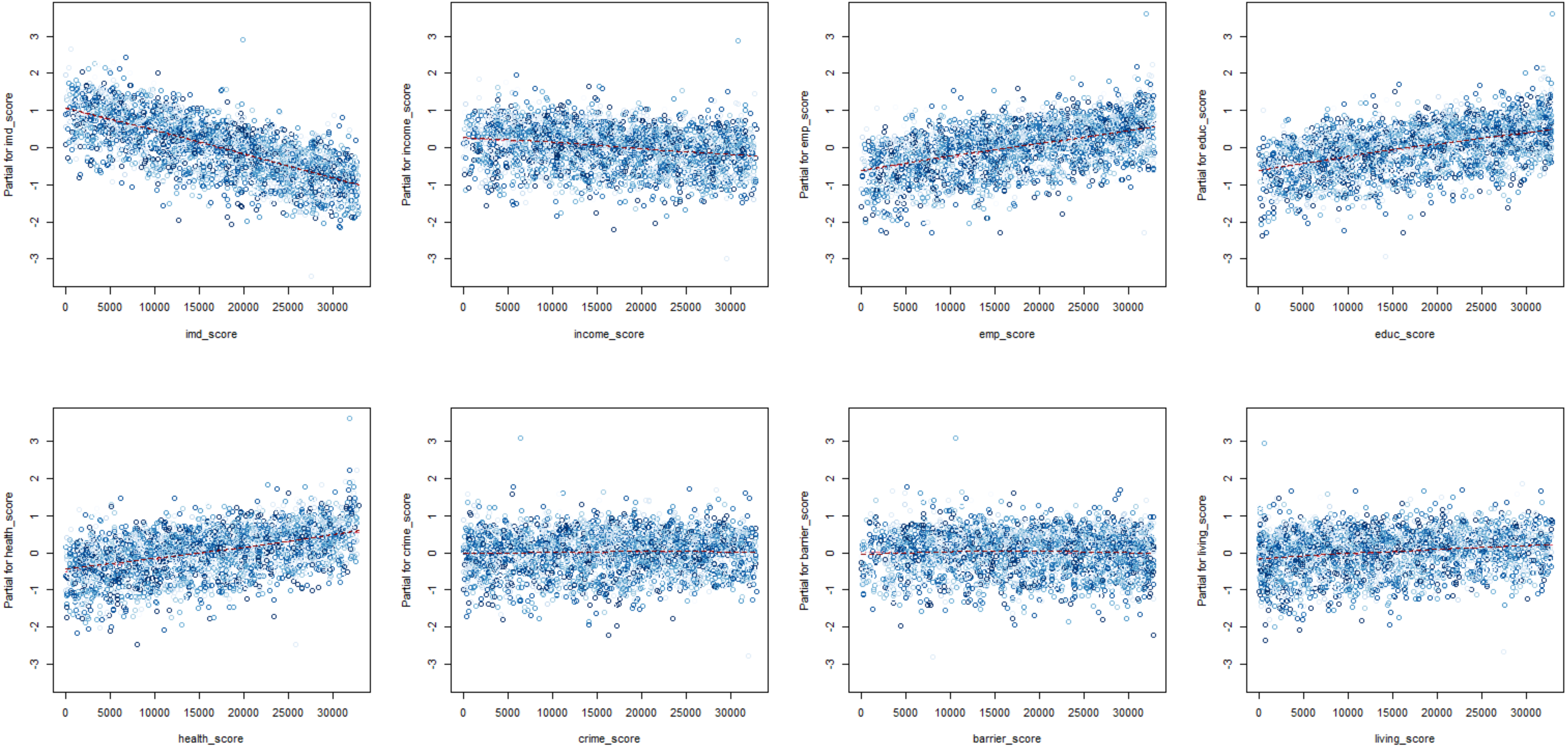
$$\frac{y^{-0.3030303} - 1}{-0.3030303}$$


Transformed second transaction price
800
700
600
500
400
300
200
100
0

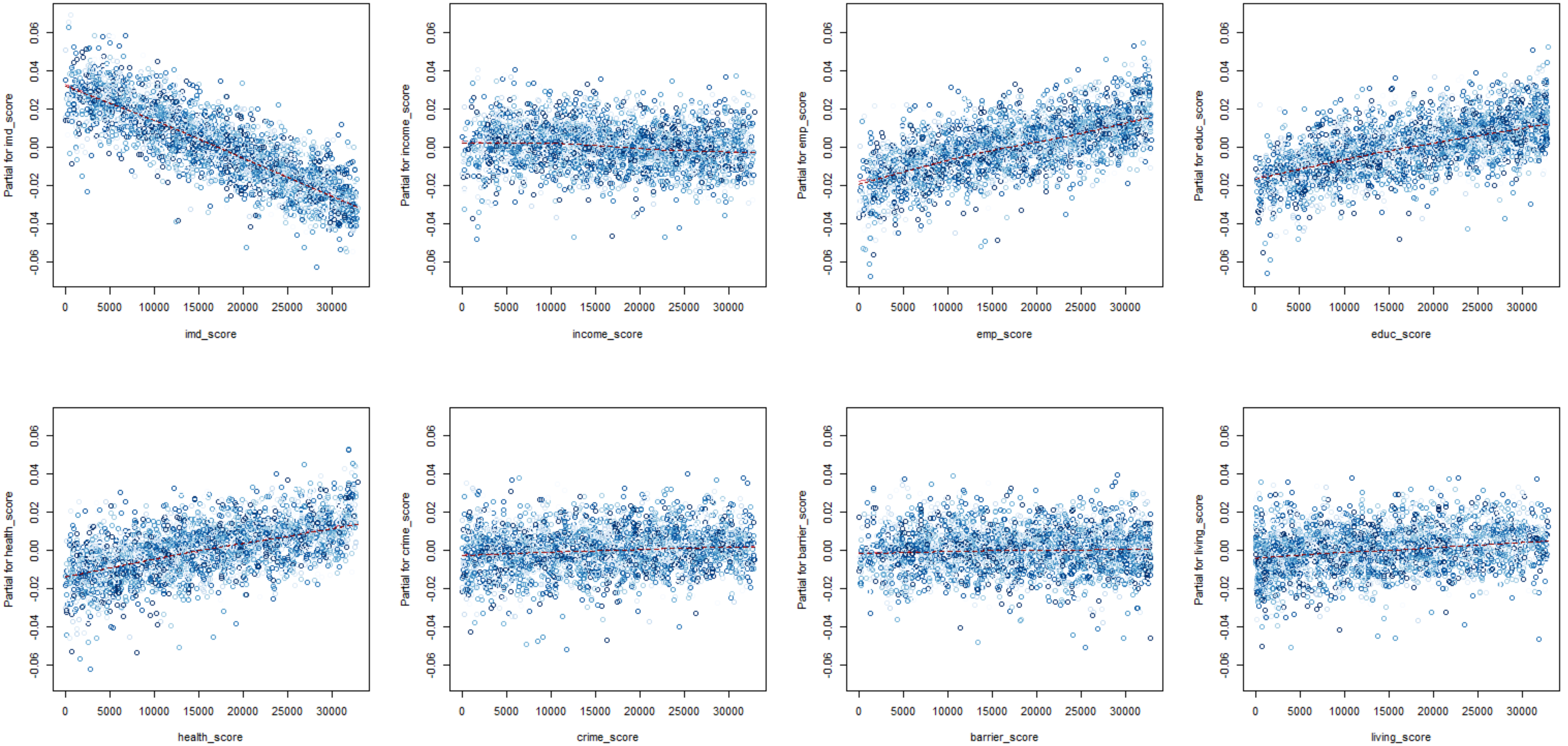

Coefficients in OLS(Continuous) estimations
Price 1
0.00006
0.00004
0.00002
0
-0.00002
-0.00004
-0.00006
-0.00008
imd_score
income_score
emp_score
educ_score
health_score
crime_score
barrier_score
living_score

Coefficients in OLS(Continuous) estimations
Price 2
0.0000015
0.000001
0.0000005
0
-5E-07
-0.000001
-1.5E-06
-0.000002
-2.5E-06
imd_score
income_score
emp_score
educ_score
health_score
crime_score
barrier_score
living_score

Partial Residue Plots for Price 1



Partial Residue Plots for Price 2



4 Repeated Sales Index

Key Ideas

- Was developed by MARTIN J. BAILEY RICHARD F. MUTH AND HUGH O. NOURSE in 1963
- An index to measure the change of prices over the years
- The goal was to make sure that these indices are close to the change in prices of houses.

Theory

→ $\frac{P_2}{P_1} \cong \frac{I_{year_i}}{I_{year_j}}$

→ $ln(\frac{P_2}{P_1}) \cong ln(\frac{I_{year_i}}{I_{year_j}})$

→ $ln(P_2) - ln(P_1) \cong ln(I_{year_i})$-$ln(I_{year_j})$

→ Now, set the index of base year to 1

→ Define, $\beta_i := ln(I_{year_i})$

→ $y_i := ln(P_{i2}) - ln(P_{i1})$

Procedure

- $y_i = \sum_{j=1} \beta_j x_{ij} + \epsilon_i$
- Where x_{ij} is
  - 1 if year_j is the year of second sale
  - -1 if year_j is the year of first sale
  - 0 otherwise
- Estimating the coefficients with OLS,
- $\hat{y} = \sum_j \hat{\beta}_j x_j$

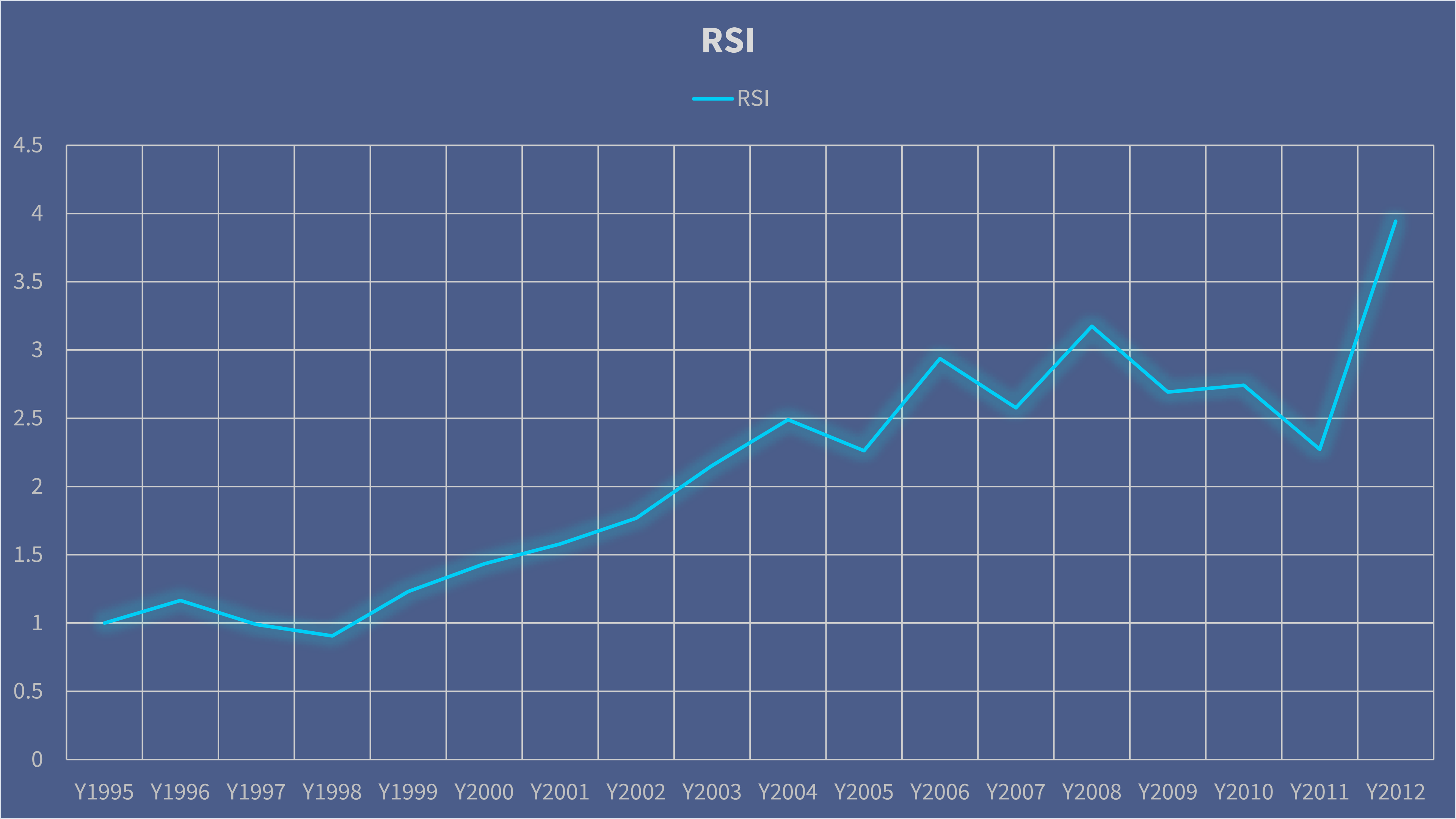
Data Processing

Example

Year	Log Price_1	Year	Log Price_2
2007	1	2008	2
2006	3	2007	7
2007	5	2008	9

Data Processing

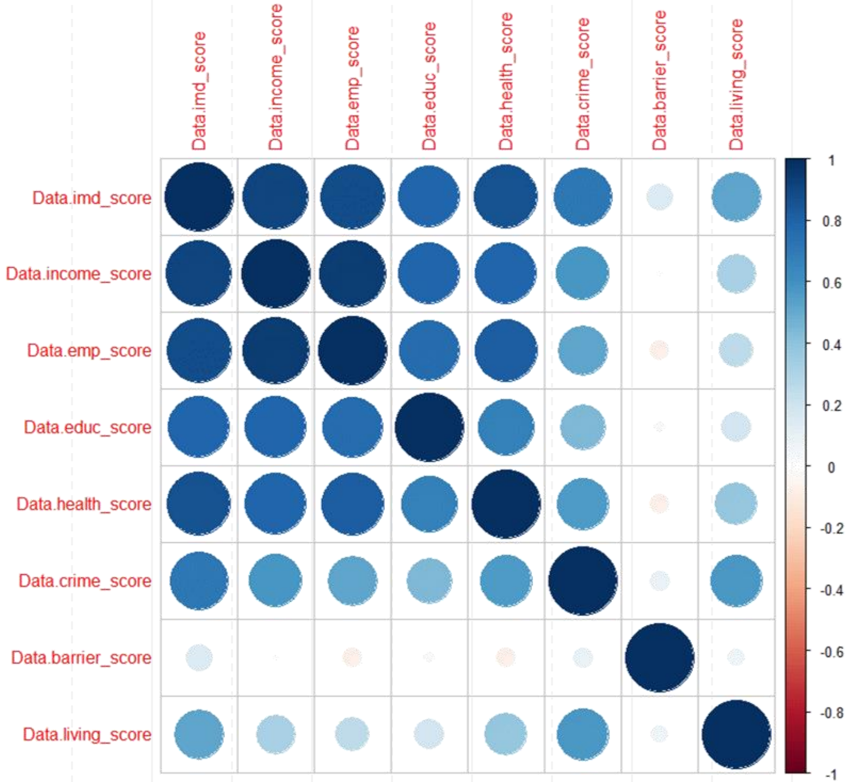
Log Change	2006	2007	2008
1	0	-1	1
4	-1	1	0
4	0	-1	1


RSI
RSI
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
Y1995
Y1996
Y1997
Y1998
Y1999
Y2000
Y2001
Y2002
Y2003
Y2004
Y2005
Y2006
Y2007
Y2008
Y2009
Y2010
Y2011
Y2012

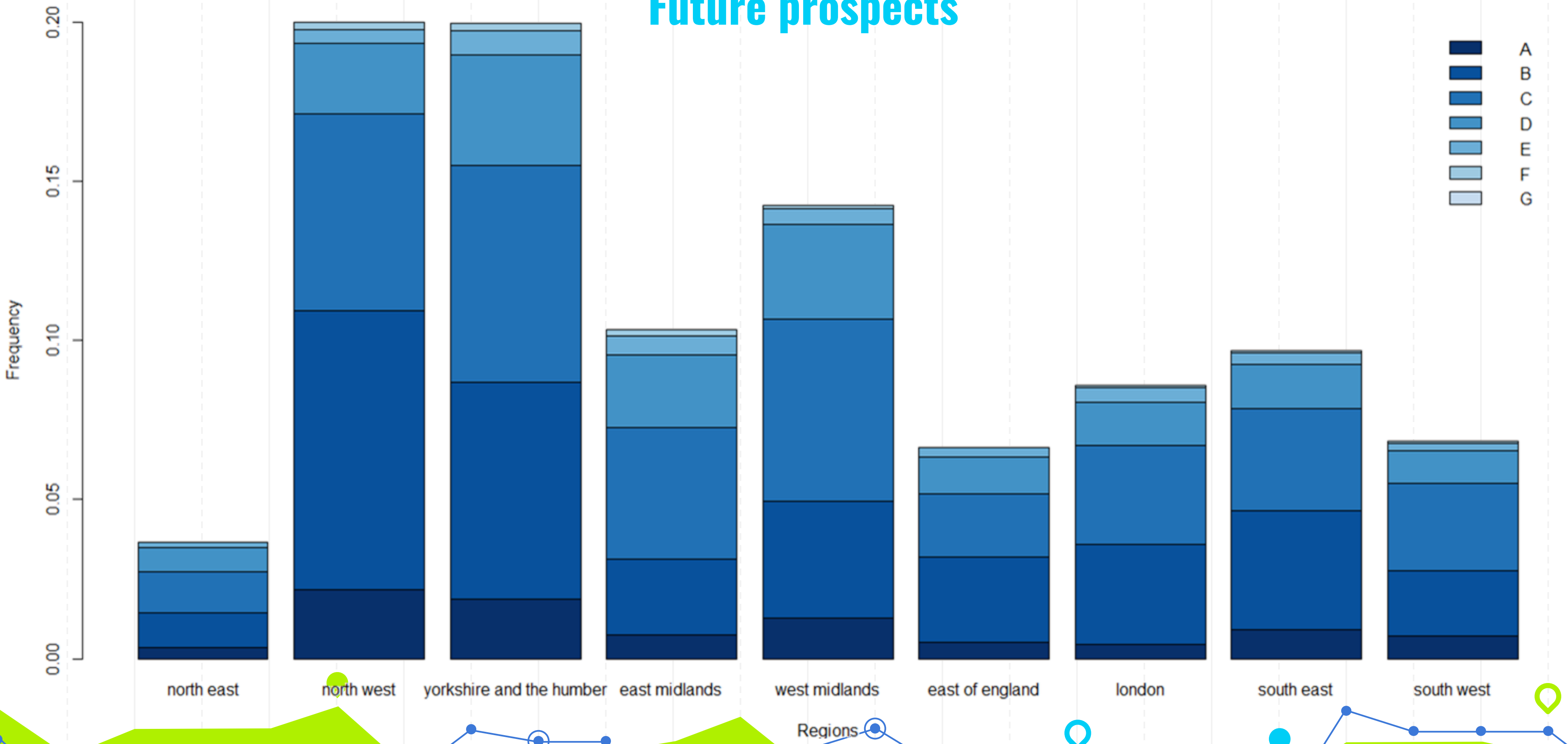
Conclusion

5

Future prospects



Future prospects



LIMITATIONS

- The amount of data that needs to be collected and worked with for hedonic regression needs to be large. However, the used data had only 4201 values
- Hedonic Price Index estimates people's willingness to pay for the supposed variation in environmental qualities and their consequences. However, if the people are unaware of the relation between the environmental qualities, then the value will not be reflected in the price of the property.
- The RSI method is inefficient as it uses only information on units that have sold more than once during the sample period. Hence, a sample selection bias problem is induced.
- If lambda is non-zero for price_2 in Box Cox Transformation so, the target variable is more difficult to interpret than if we simply applied a log transform.

SWOT ANALYSIS

STRENGTHS

Theoretical Analysis of every variable

WEAKNESSES

Many assumptions may not hold for the population

Better statistical methods can be used to get rid of assumptions

OPPORTUNITIES

The whole study was conducted for educational purposes

THREATS

References

- Franz Fuerst, Michel Ferreira Cardia Haddad(2020). “Real estate data to analyse the relationship between property prices, sustainability levels and socio-economic indicators.” Data in Brief. 33 106359.
- F. Fuerst, M.F.C. Haddad, H. Adan(2015). “Is there an economic case for energy-efficient dwellings in the UK private rental market?” Journal for Cleaner Prod. 245 (2020) 118642.
- Bailey, M.J., Muth, R.F., Nourse, H.O. (1963). “A regression method for real estate price index construction. Journal of the American Statistical Association.” 58 933-942 16
- G. E. P. Box and D. R. Cox(1964). “An Analysis of Transformations.” Journal of the Royal Statistical Society. Series B Vol. 26. No. 2.
- Repeat Sales House Price Index Methodology. Journal of Real Estate Literature. Vol. 22, No. 1 (2014), pp. 23-46.
- F. Fuerst , P. McAllister , A. Nanda , P. Wyatt(2015). “Does energy efficiency matter to home-buyers? an investigation of EPC ratings and transaction prices in England?” Energy Economy. 48 145–156 .

Acknowledgements

We would like to thank Dr. Rituparna Sen for her invaluable support as a mentor, and for providing us with an opportunity to delve deeper into the subject. We would also like to thank the authors for their comprehensive work on the subject. Lastly, we would like to thank our some of our classmates, discussing doubts with whom made it easier to navigate whenever we hit roadblocks.

TEAM PRESENTATION

Aprameya G. Hebbar
B Math 2nd Year

Atreya CHoudhury
B Math 2nd Year

Sanchayan Bhowal
B Math 2nd Year

Shreyan Saha
B Math 2nd Year

THANKS!

Any questions?