

Energy Economy

“

*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:

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

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

- A model of the form

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

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

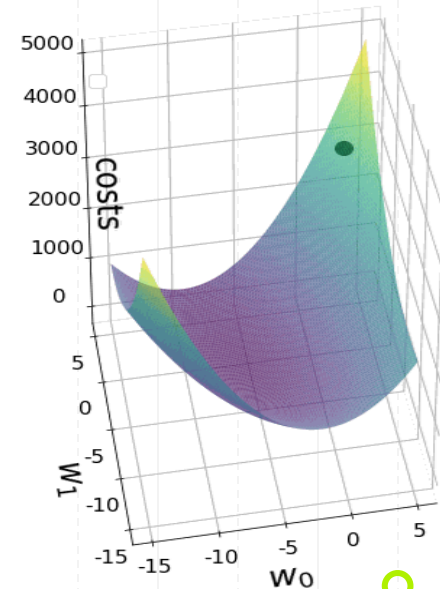
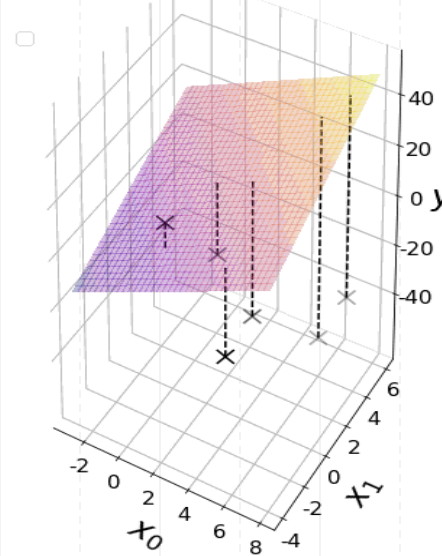
Solving the equations

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

Multiple Regression Model

- Intercept b 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

Regression plane



Multiple Regression Model

A multiple regression model with k independent variables fits a regression “surface” in $k + 1$ dimensional space (cannot be visualized)



Important terms

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

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

Important terms

- 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 $\frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}$ 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 i th measured value influences the i th 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.

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

- 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.

IMPORTANT CONCEPTS



Cook's Distance

Cook's distance is a commonly used estimate of the influence of a data point



Standard Error

Standard Error represents the average distance that the observed values fall from the regression line.



Residual Standard Error

$$\sqrt{\frac{SSE}{n - (K + 1)}}$$



Leverage

Leverage is a measure of how far away the independent variable values of an observation are from those of the other observations. High-leverage points, if any, are outliers with respect to the independent variables.



R^2

$$R^2 = 1 - \frac{SSE}{SS_{yy}}$$



Adjusted R^2

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

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.

Categorical Explanatory Variables in Regression Models

- 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

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

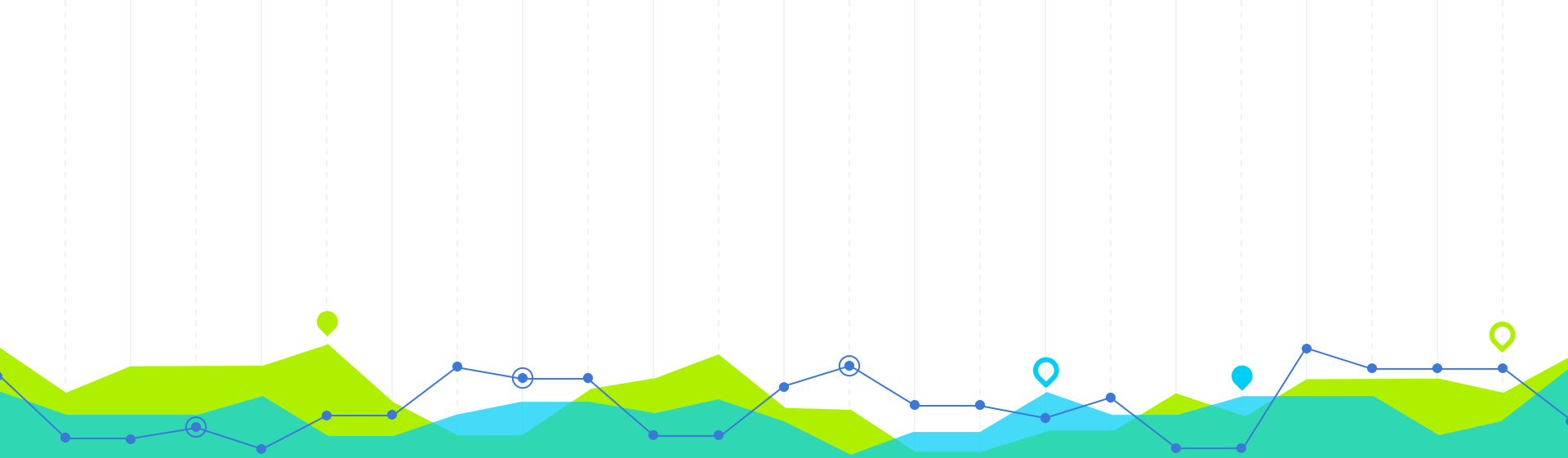
0
1
2

DUMMY1

0
1
0

DUMMY2

0
0
1

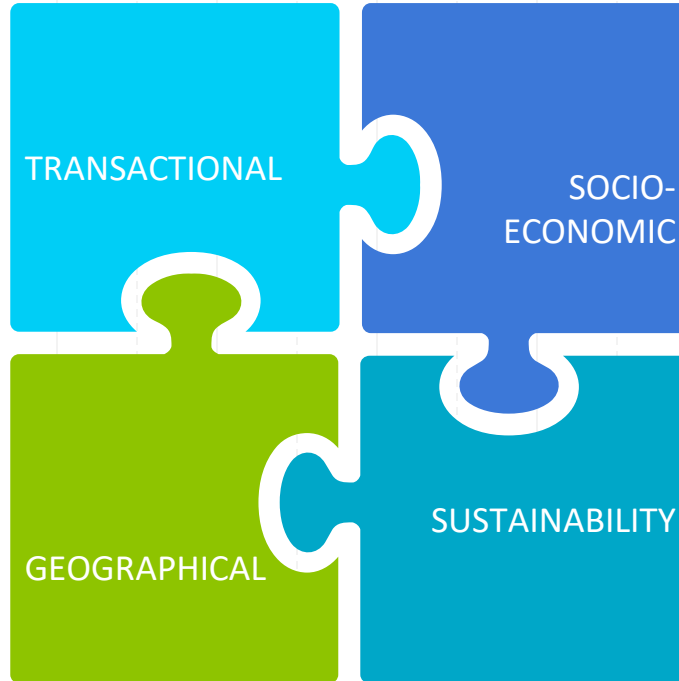


Data Description

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

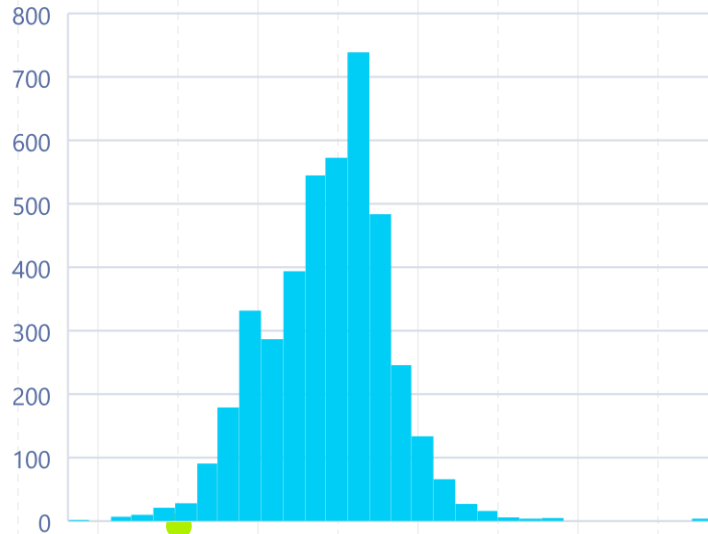


- imd_score
- imd_level
- income_score
- income_level
- emp_score
- emp_level
- educ_score
- educ_level
- health_score
- health_level
- crime_score
- crime_level
- barrier_score
- barrier_level
- living_score
- living_level
- epc_100
- epc_rating_a
- epc_rating_b
- epc_rating_c
- epc_rating_d
- epc_rating_e
- epc_rating_f
- epc_rating_g
- ln_epc_100

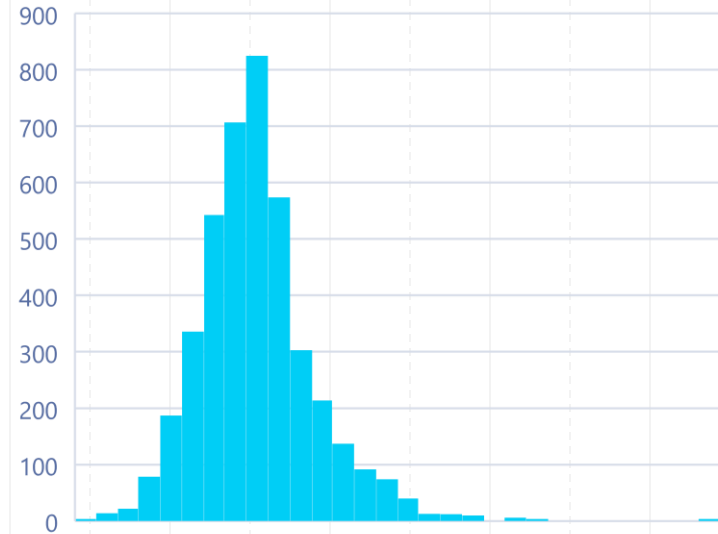


Repeated Sales Prices

Logarithm of first transaction price

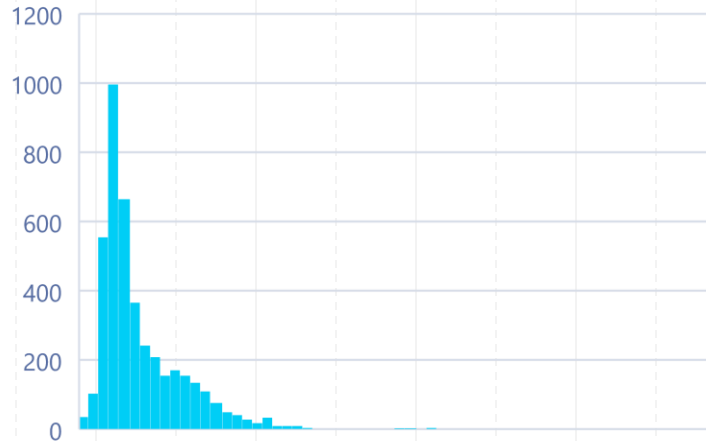


Logarithm of second transaction price

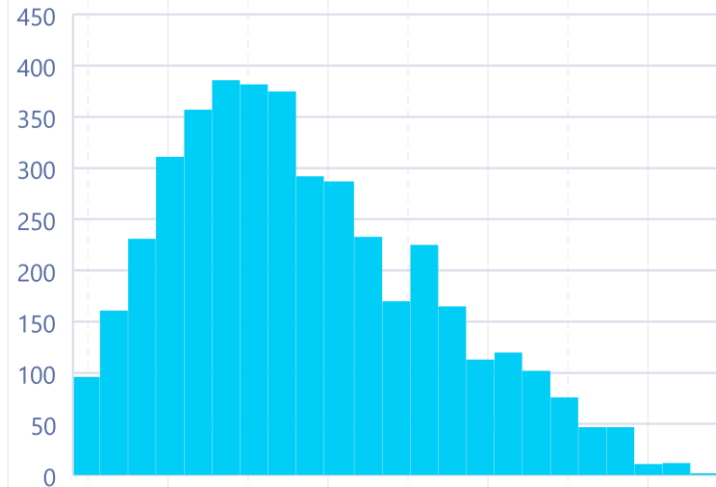


Change of Prices and Time between Sales

Percentual price change between transactions



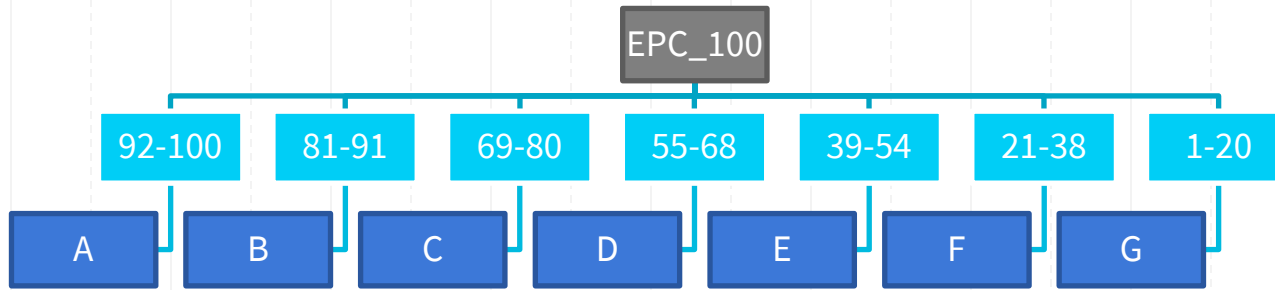
Period of time between both transactions



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

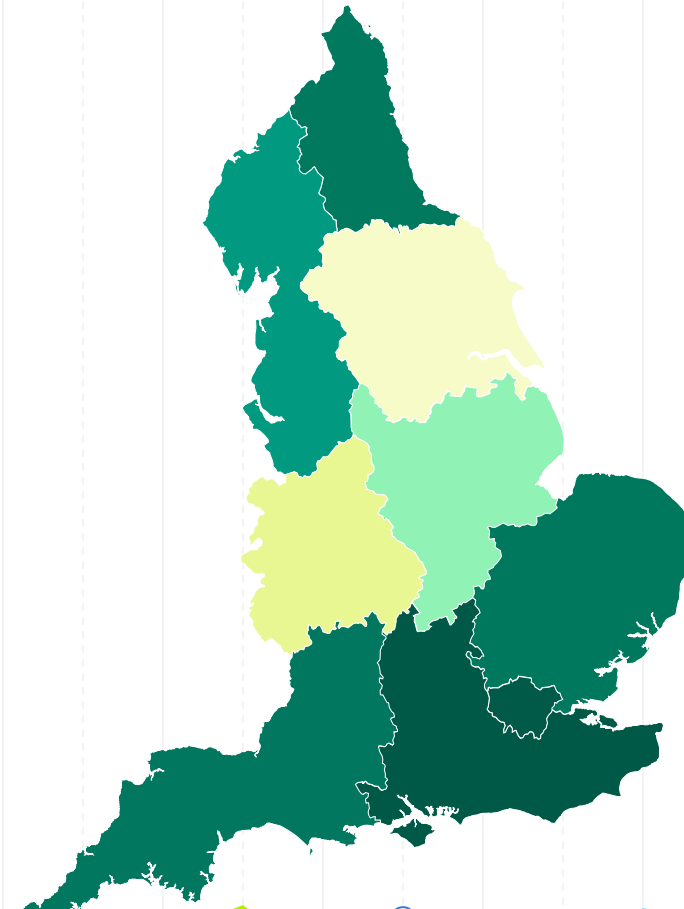
EPC Categorization

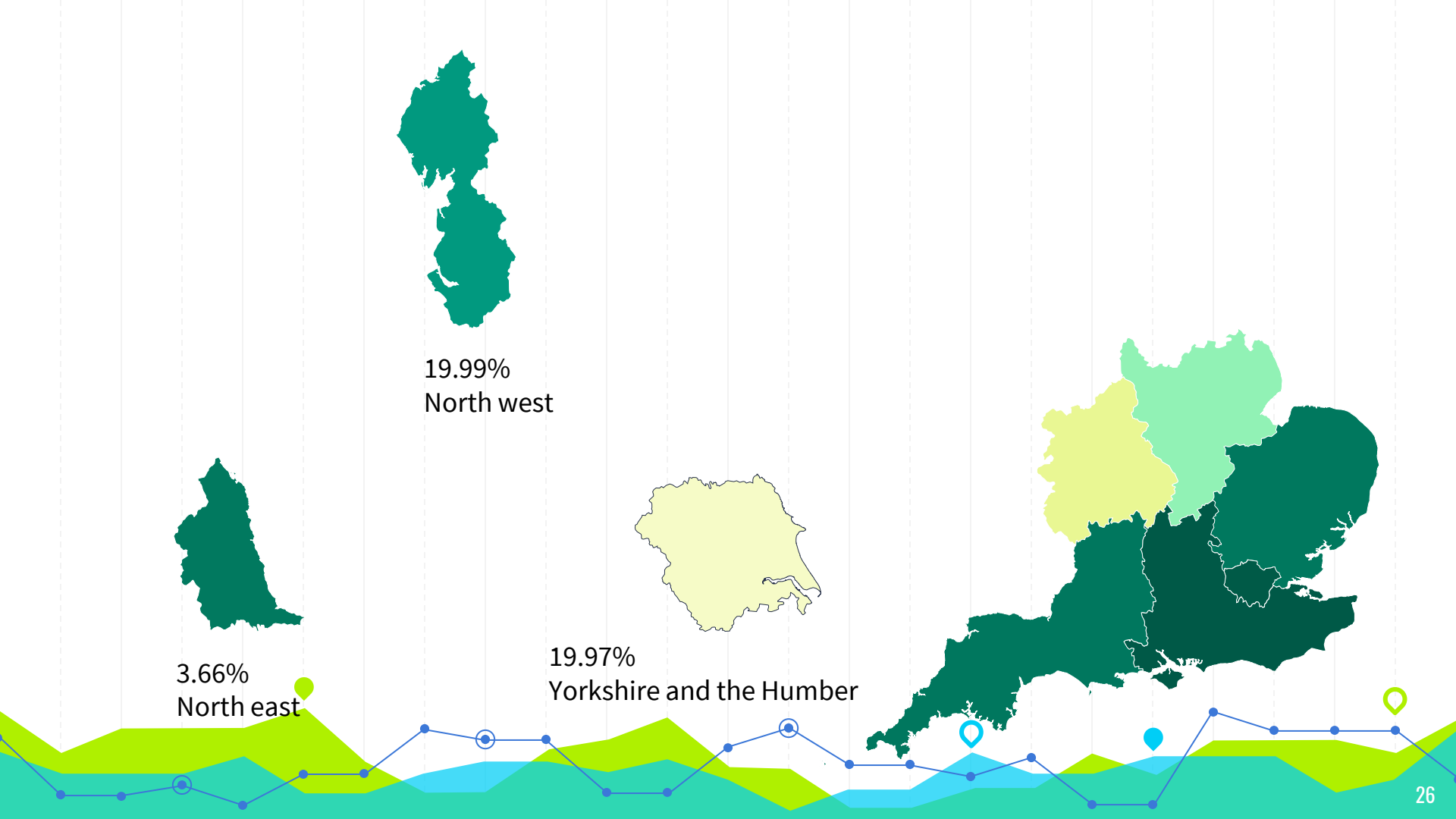


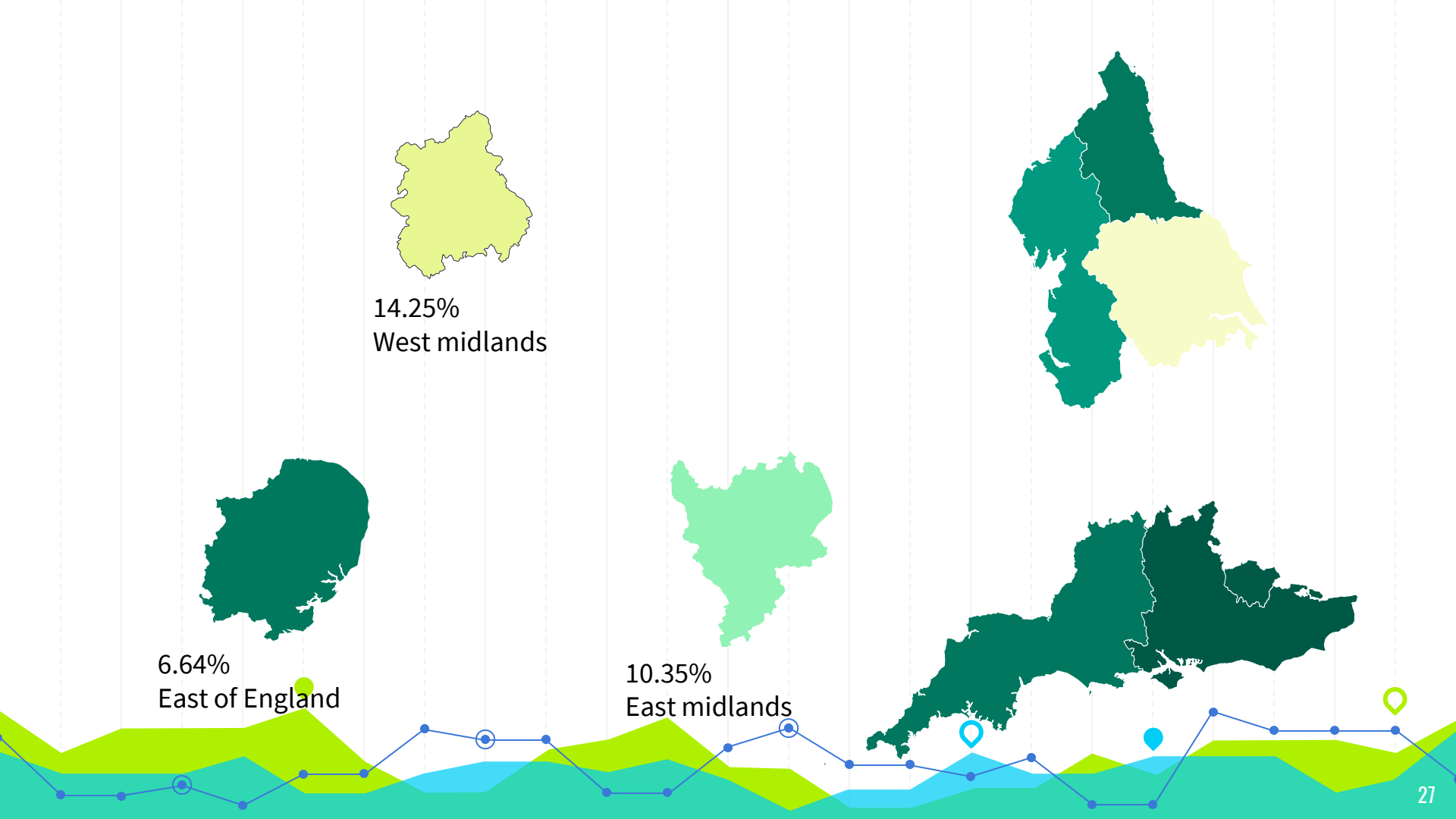
EPC Summary

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



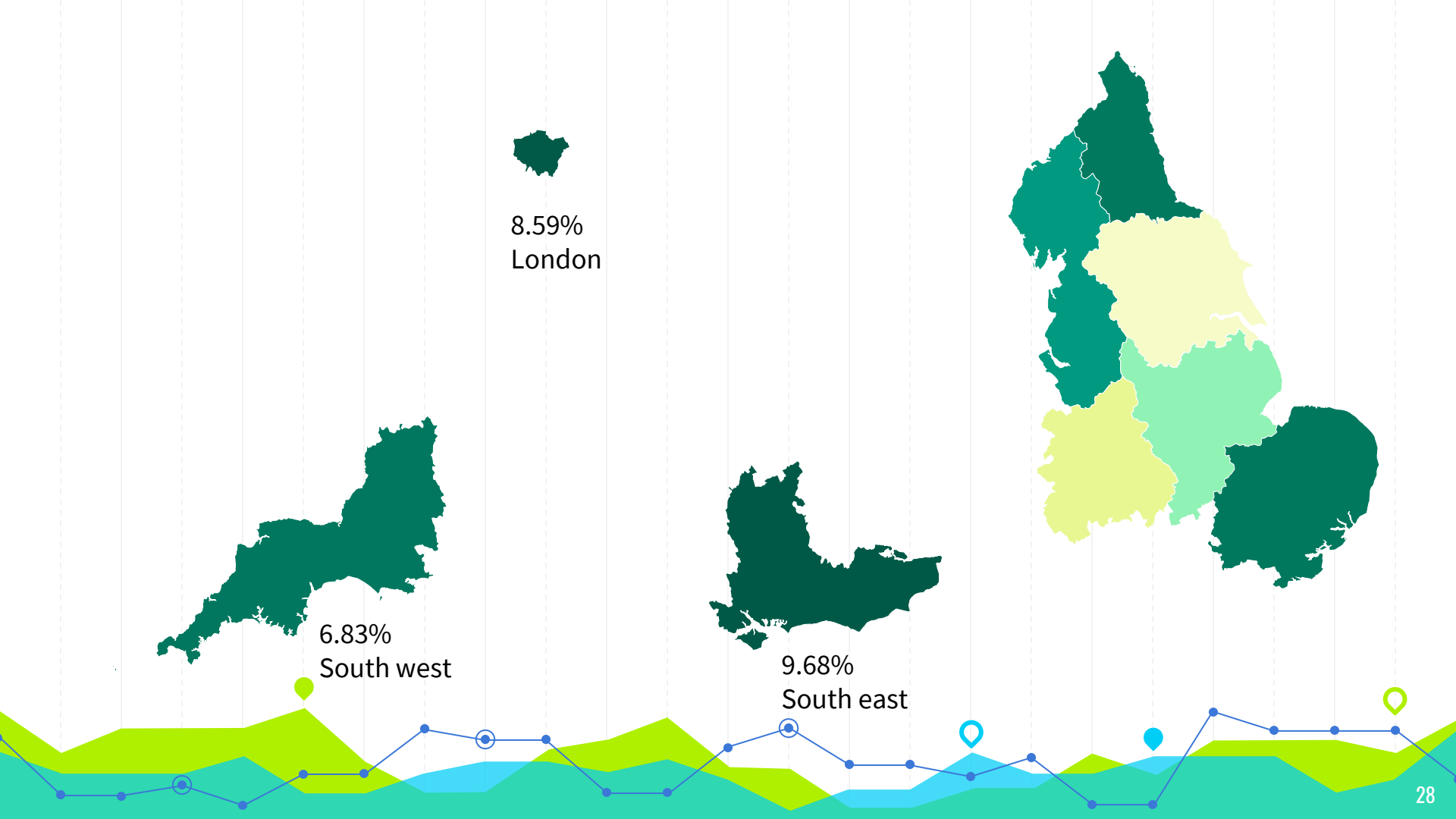


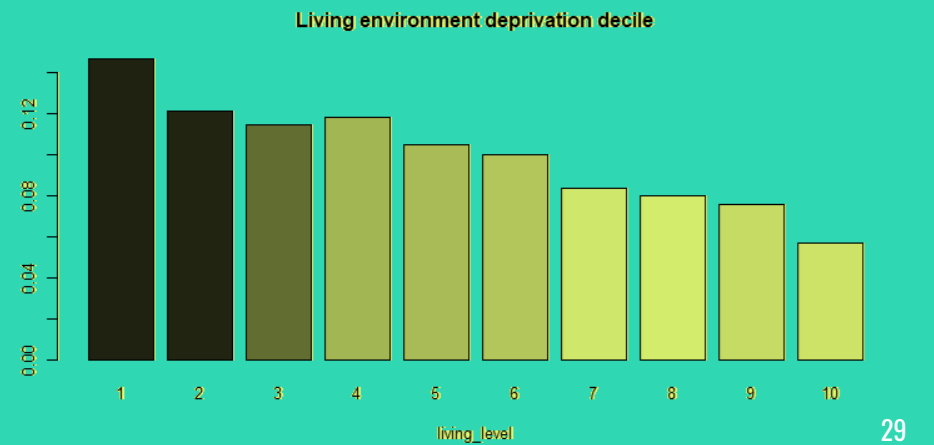
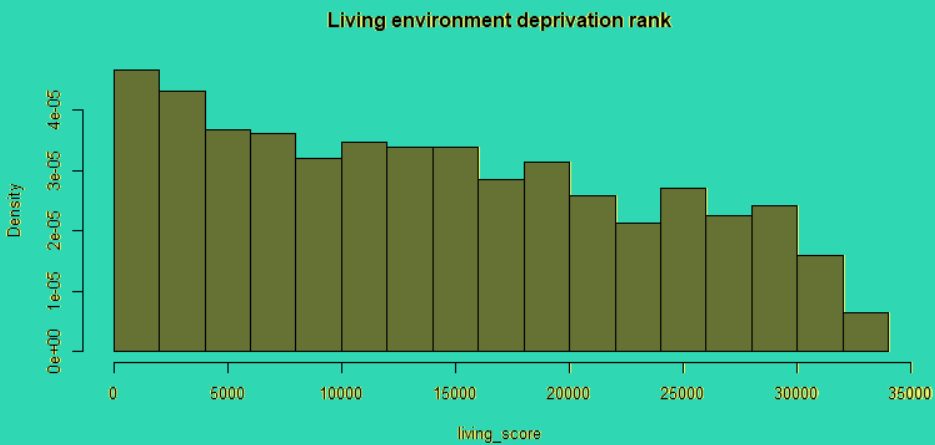
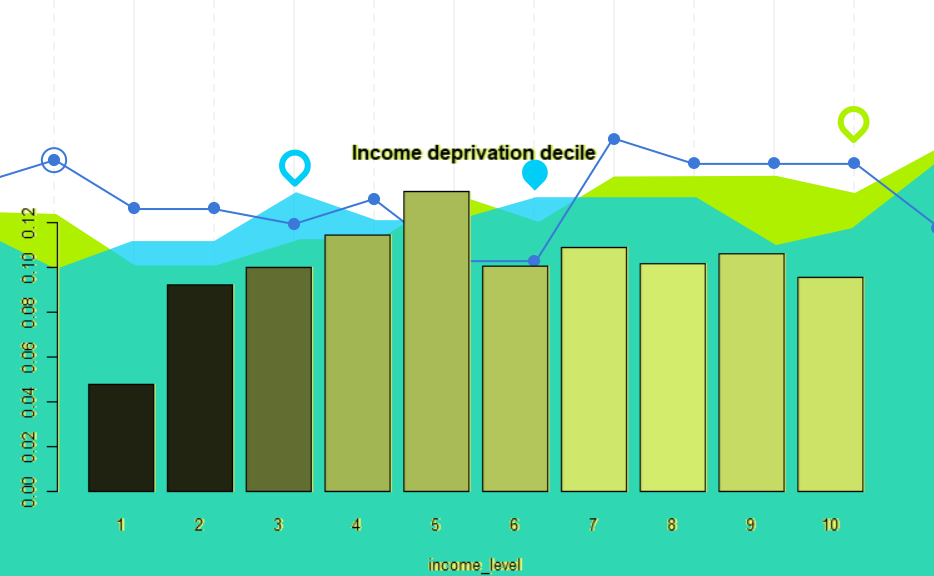
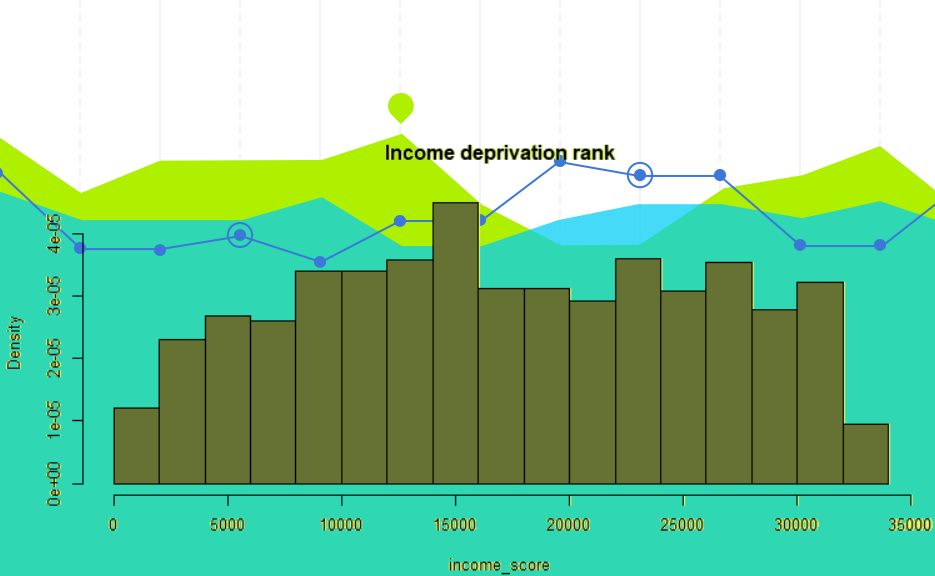


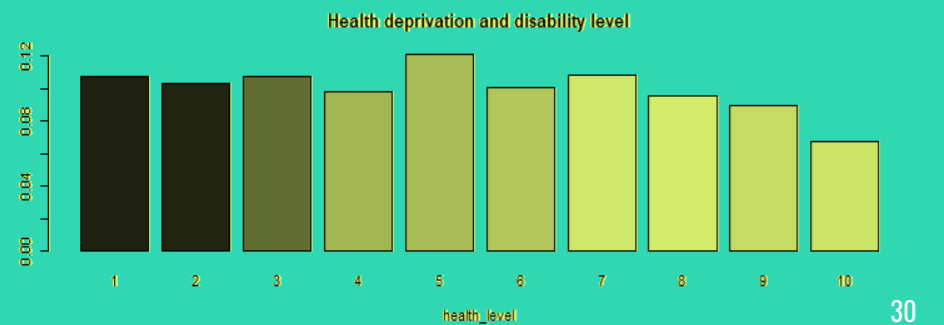
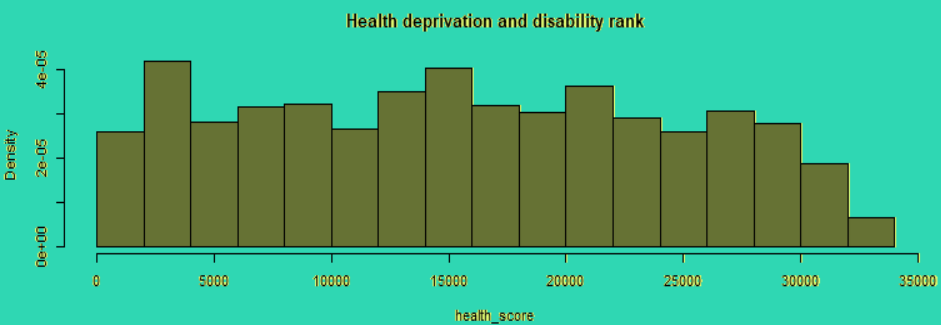
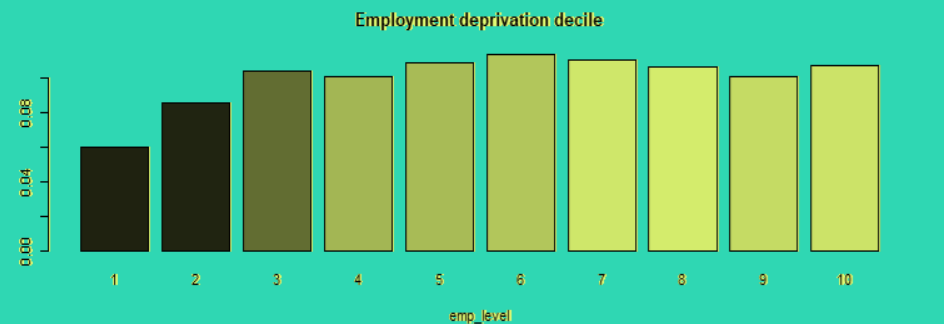
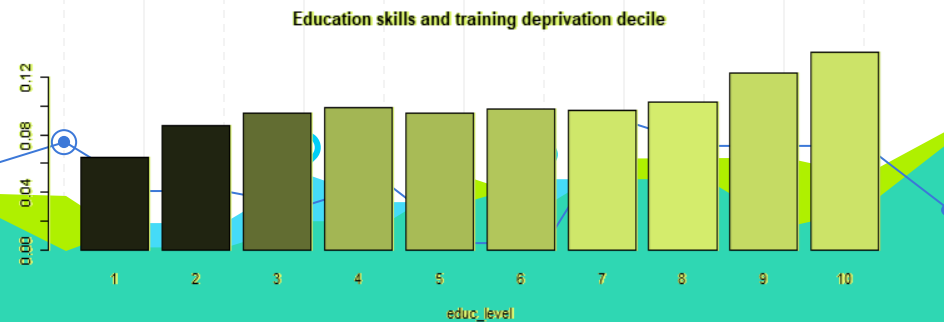
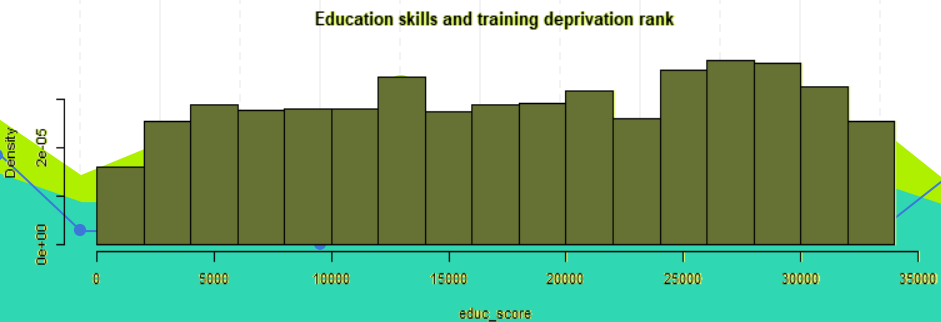
14.25%
West midlands

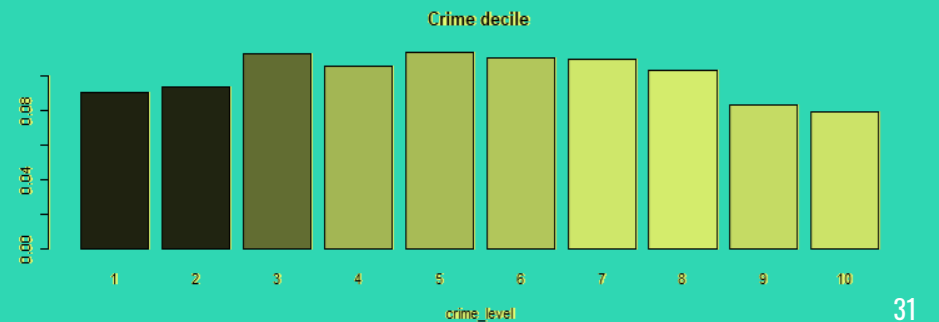
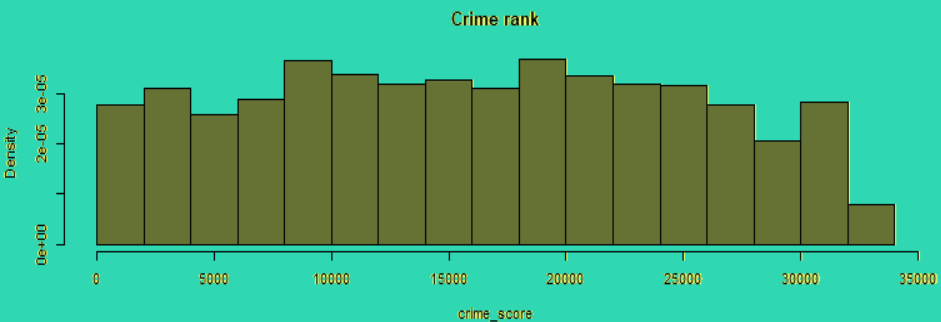
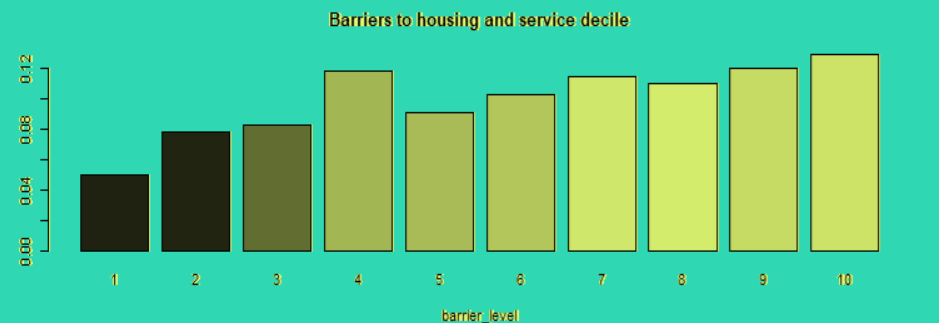
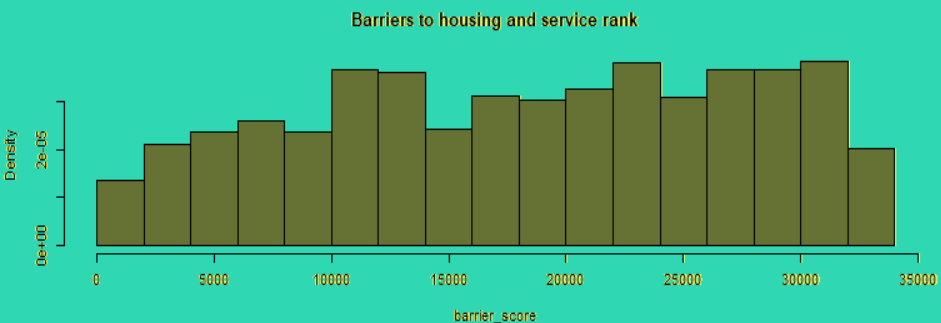
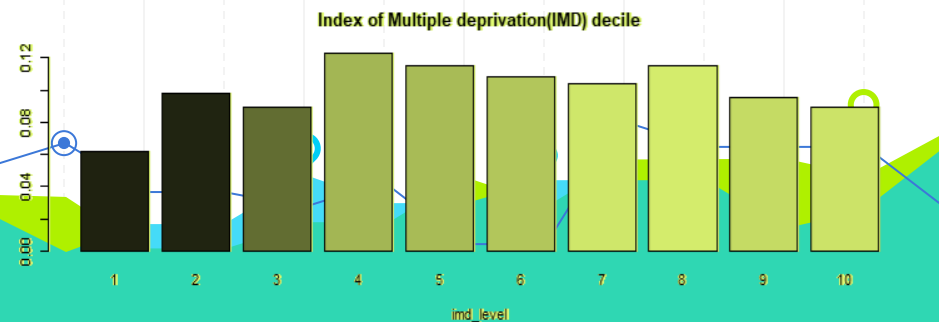
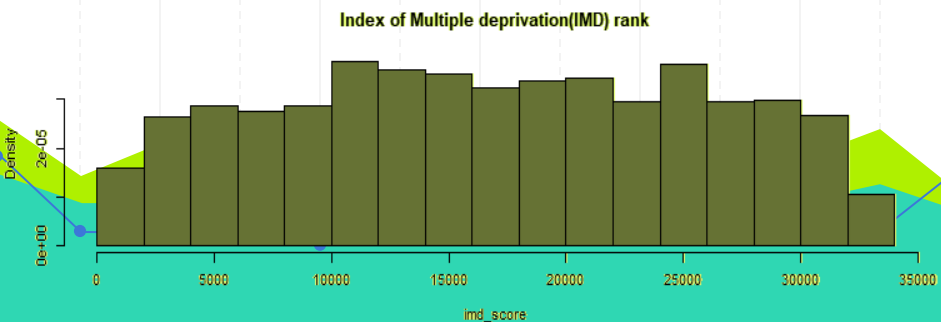
6.64%
East of England

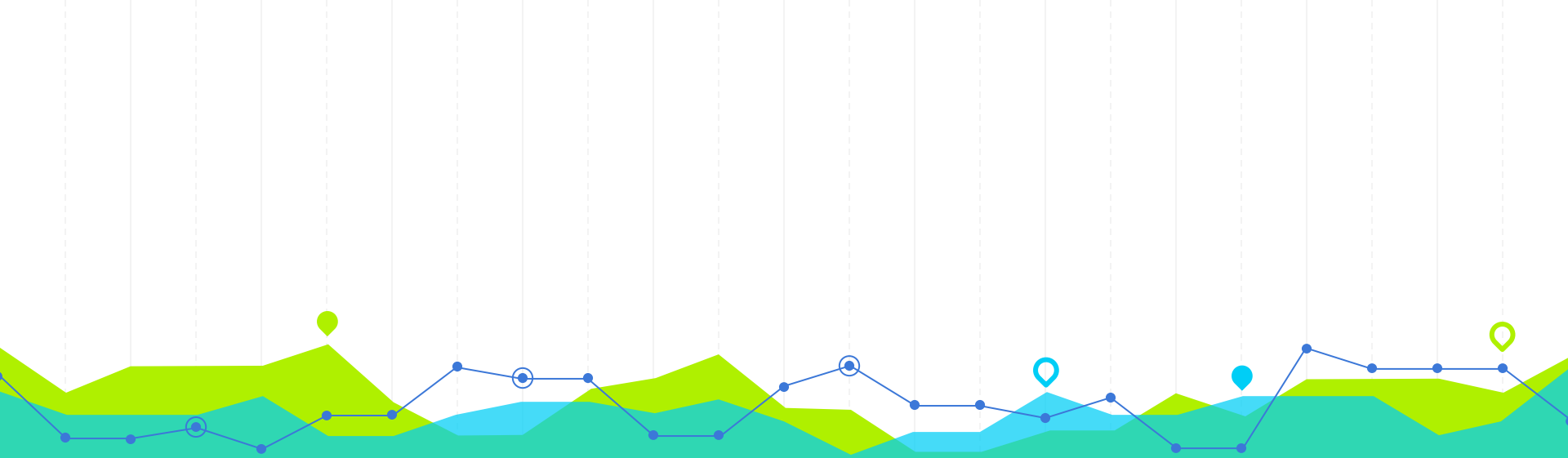
10.35%
East midlands







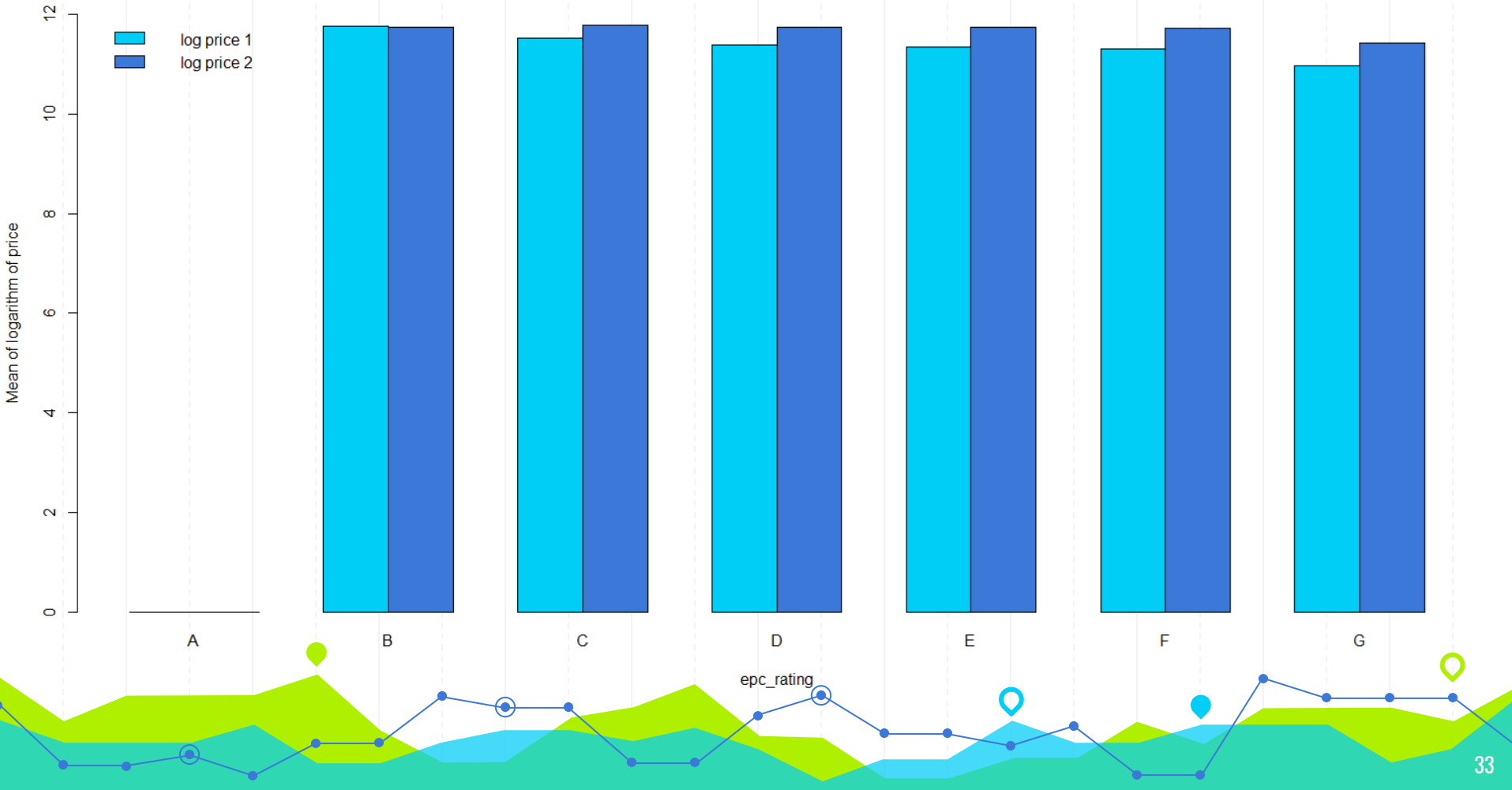


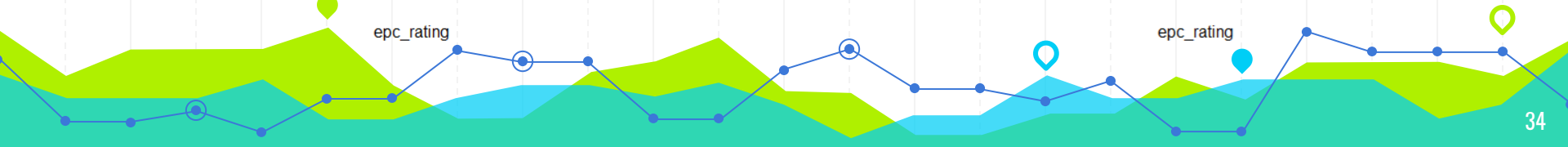
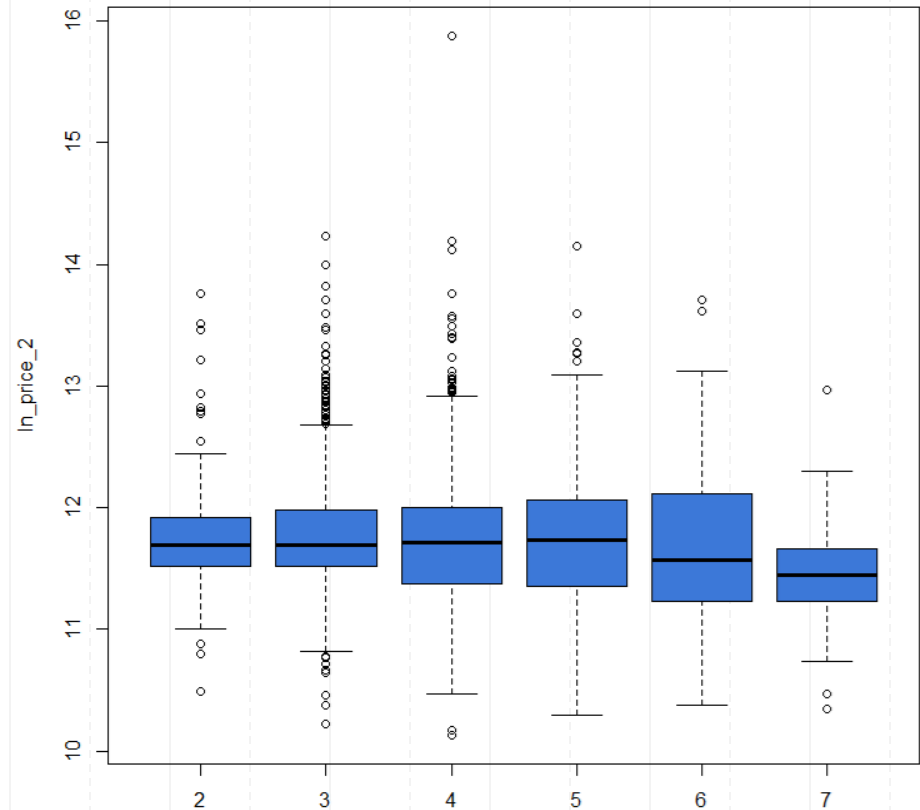
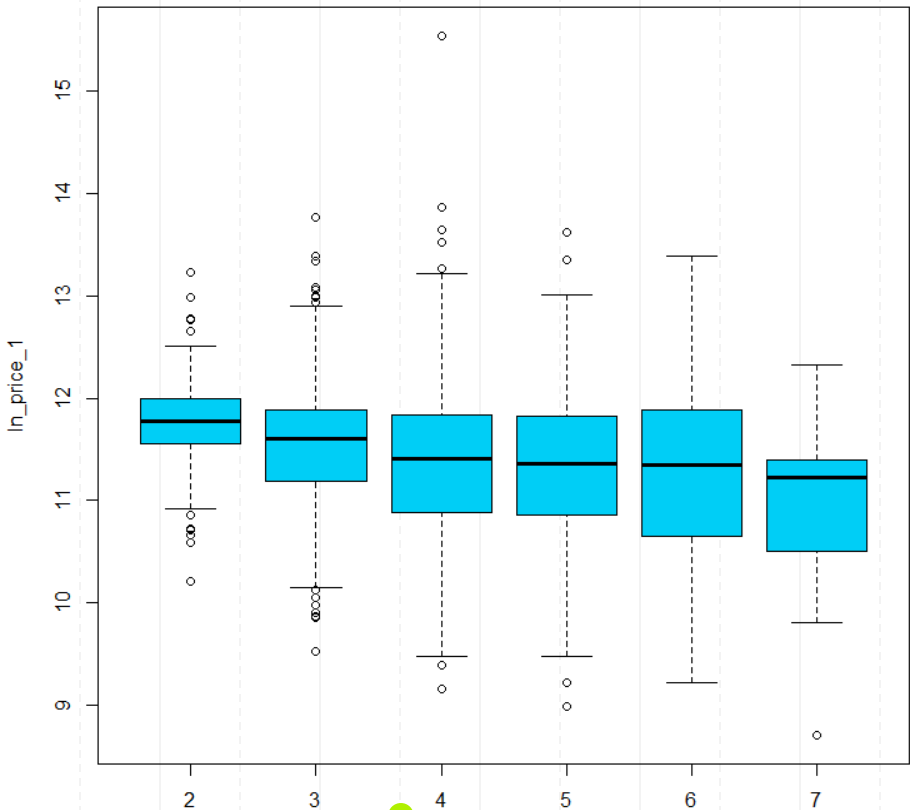


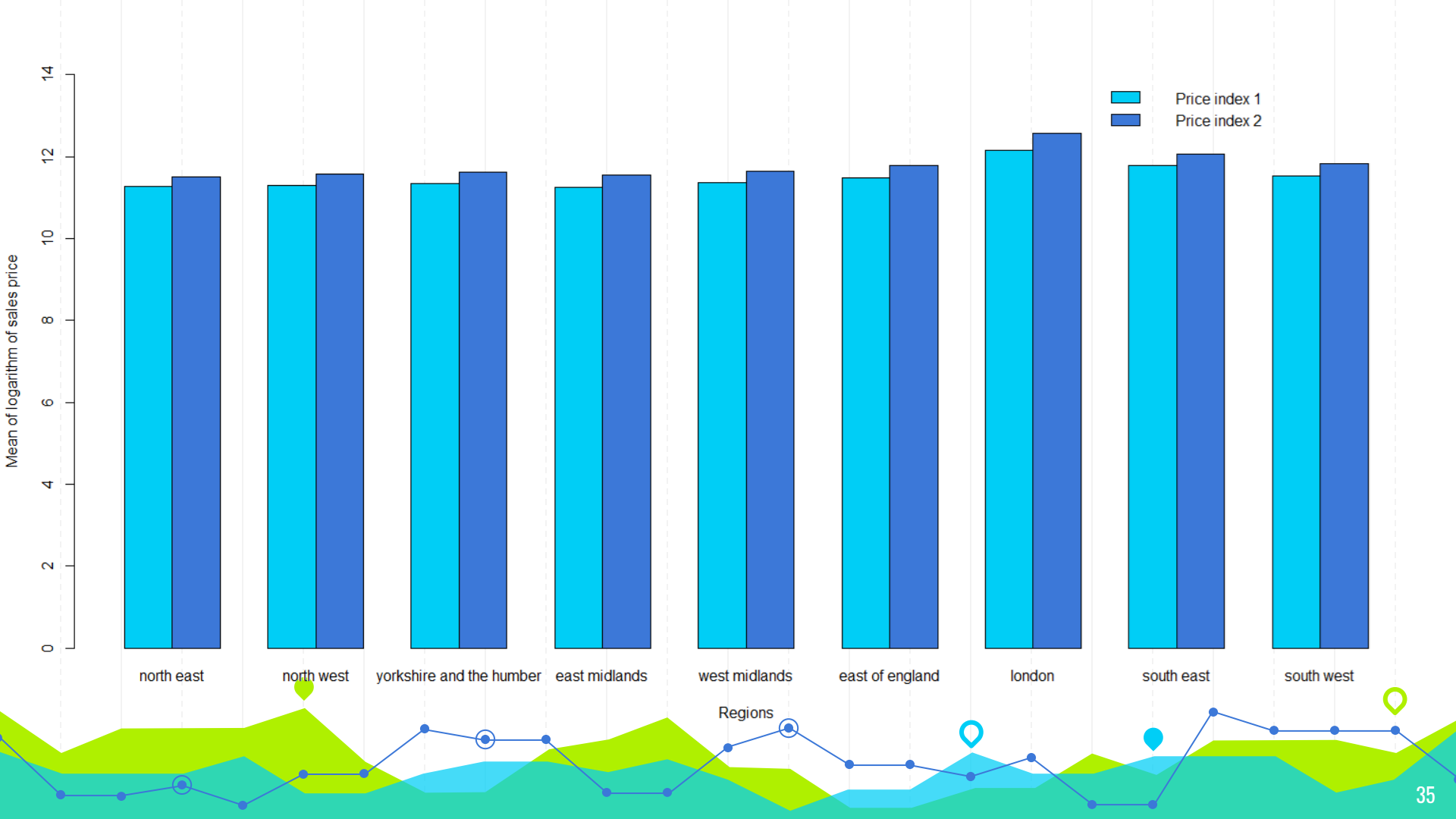
Hedonic Regression

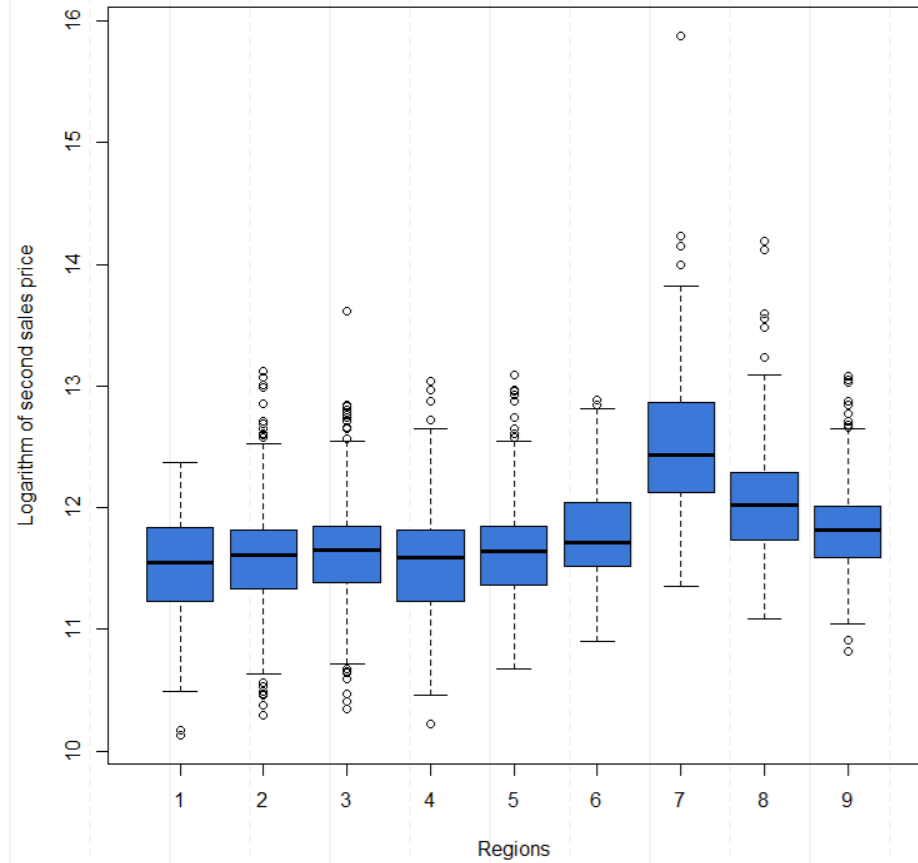
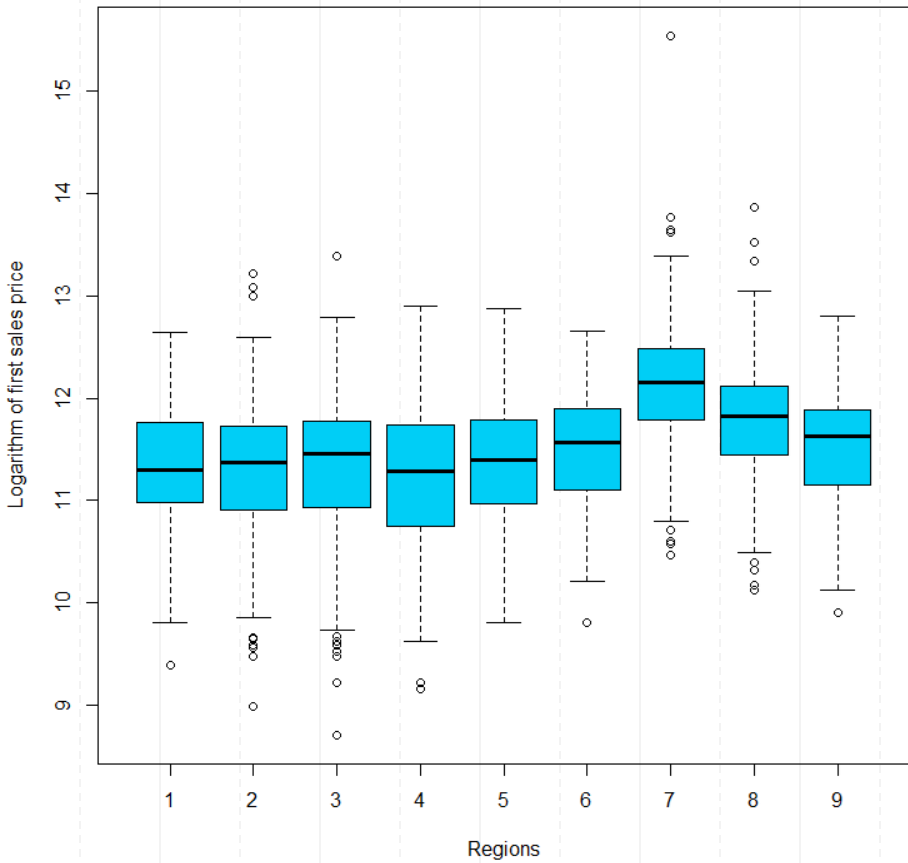
2

Relating of EPC with sales price









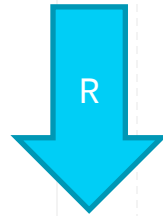
BRIEF INTRODUCTION TO HEDONIC REGRESSION

- **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.



Hedonic Model

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



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

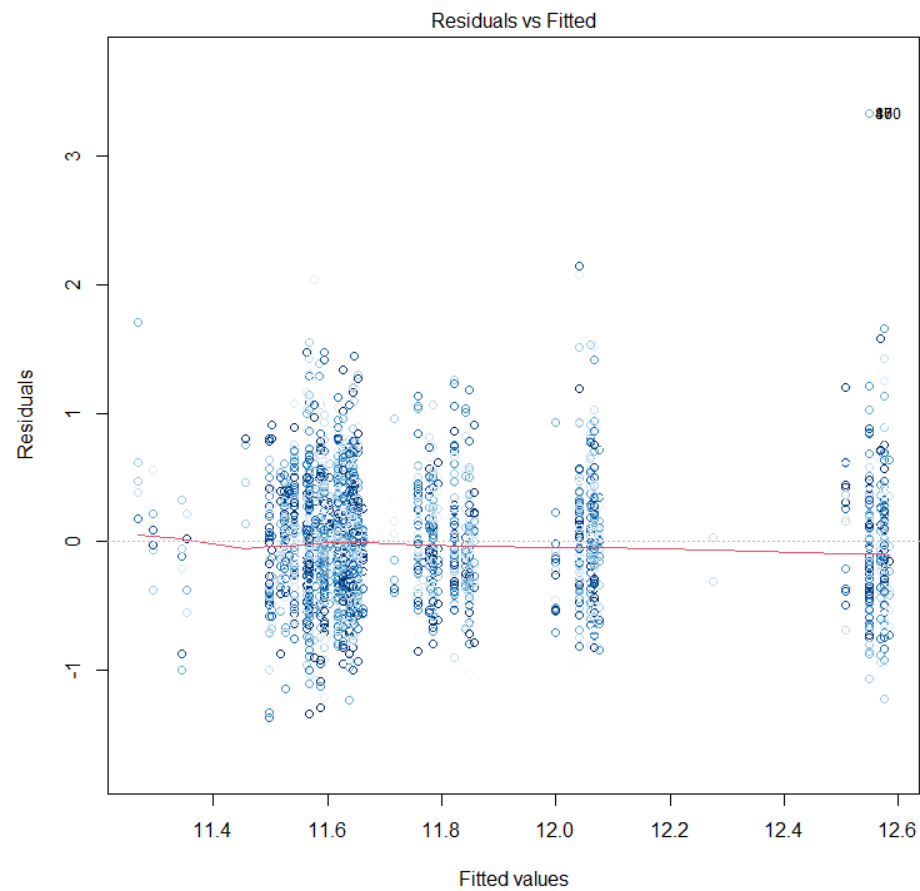
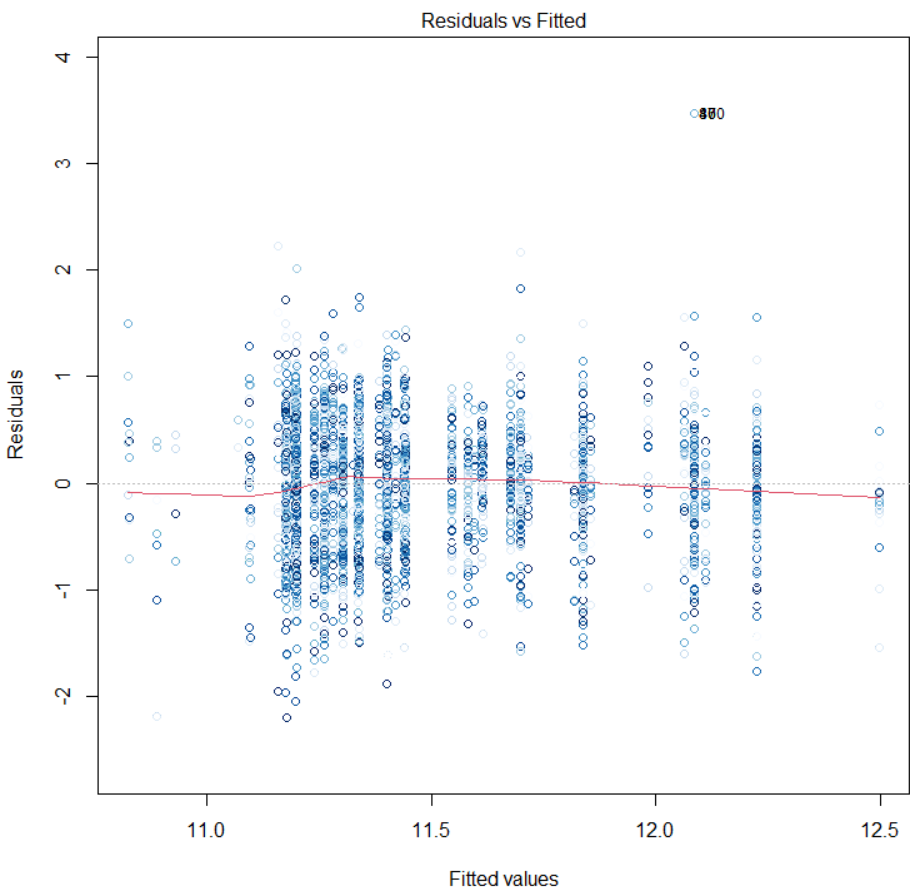
Partial Changes

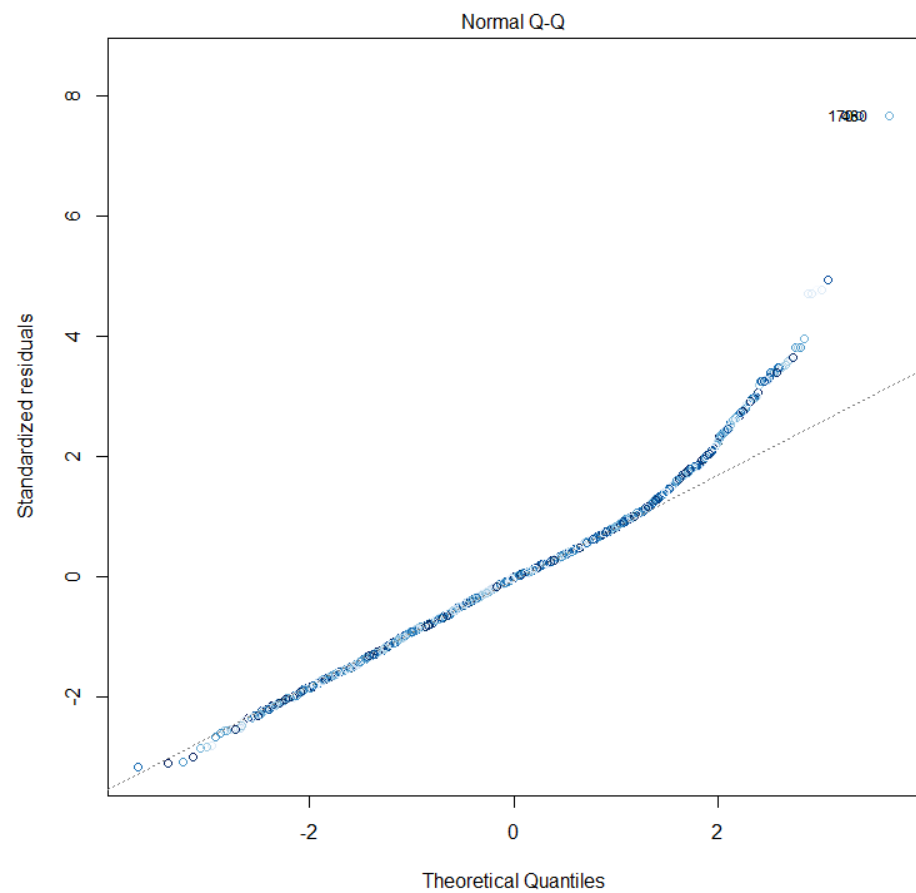
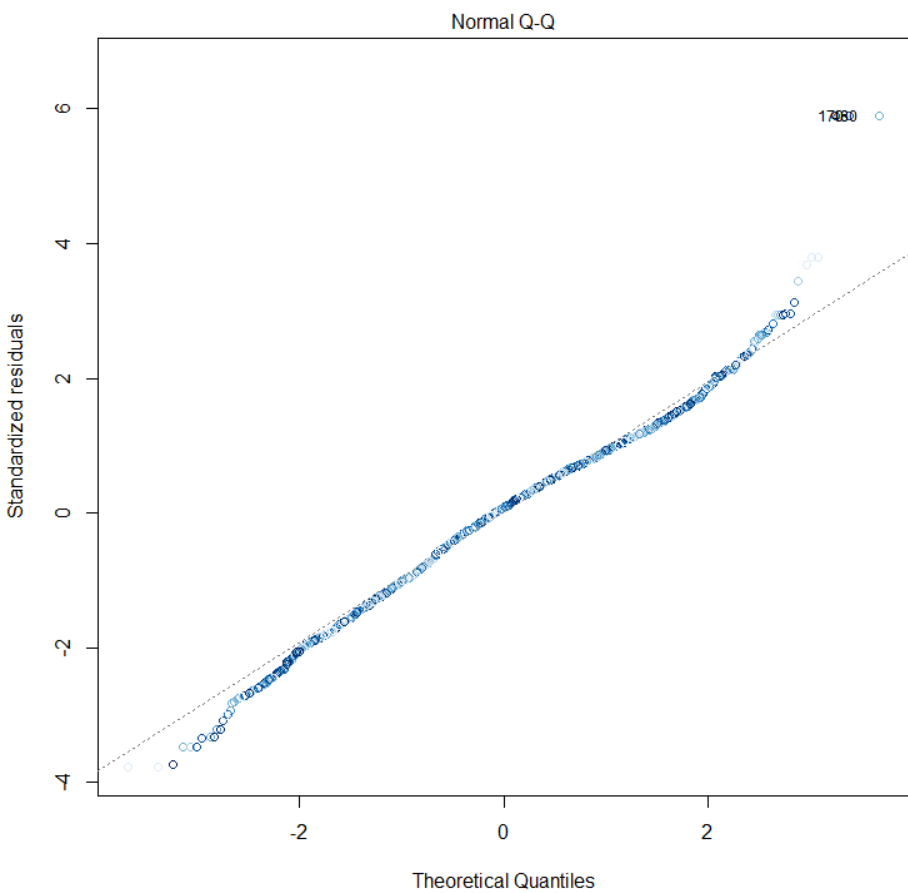
$$\begin{aligned}\ln \hat{P}_t - \ln \hat{P}'_t &= \hat{\beta}_{jt} \Delta x_{jt} \\ \Rightarrow \hat{P}'_t &= \hat{P}_t e^{\hat{\beta}_{jt} \Delta x_{jt}} \\ \Rightarrow \hat{P}'_t &= \hat{P}_t e^{\hat{\beta}_{jt}}\end{aligned}$$

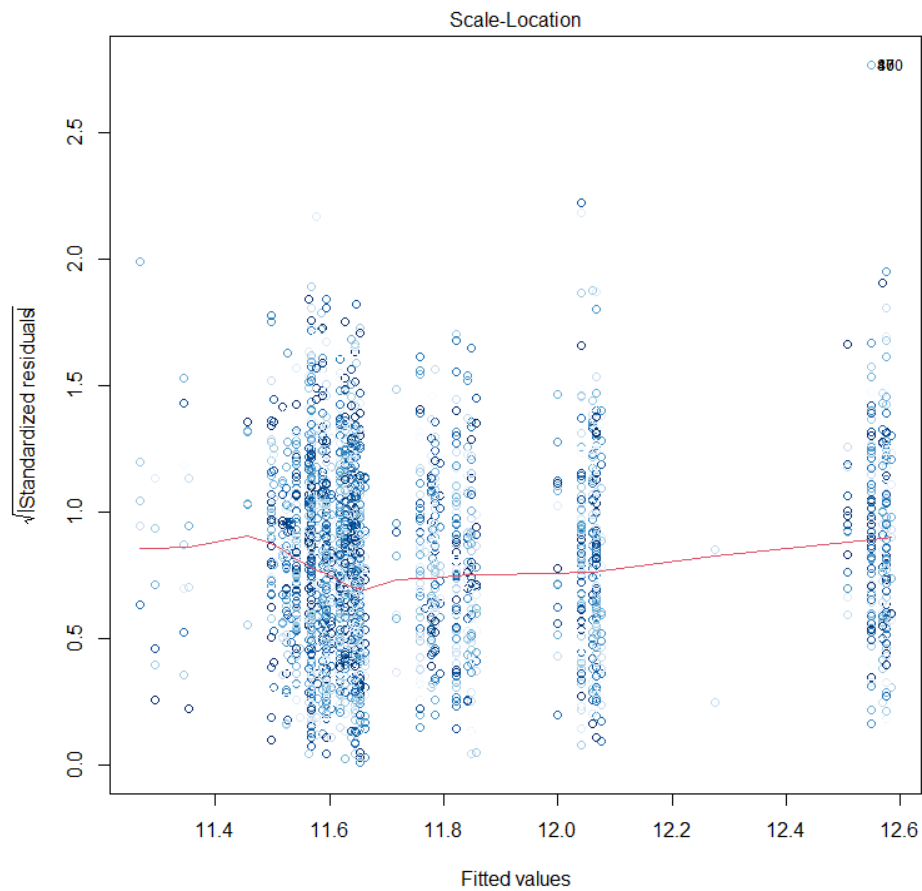
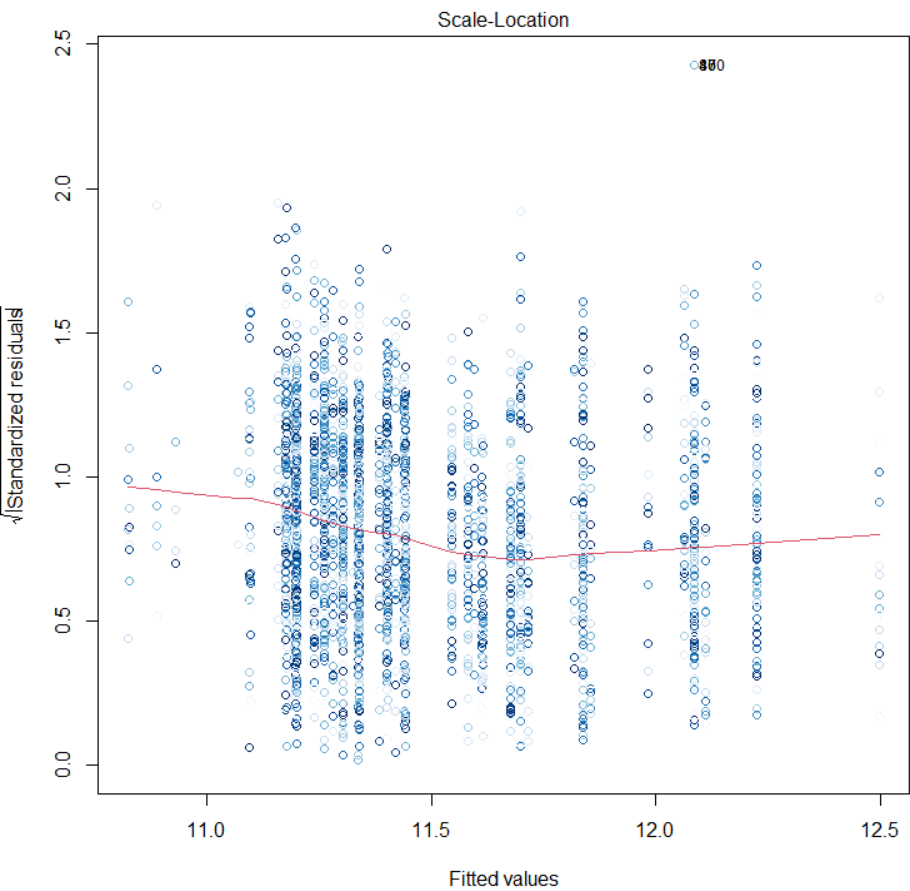
Model Choices

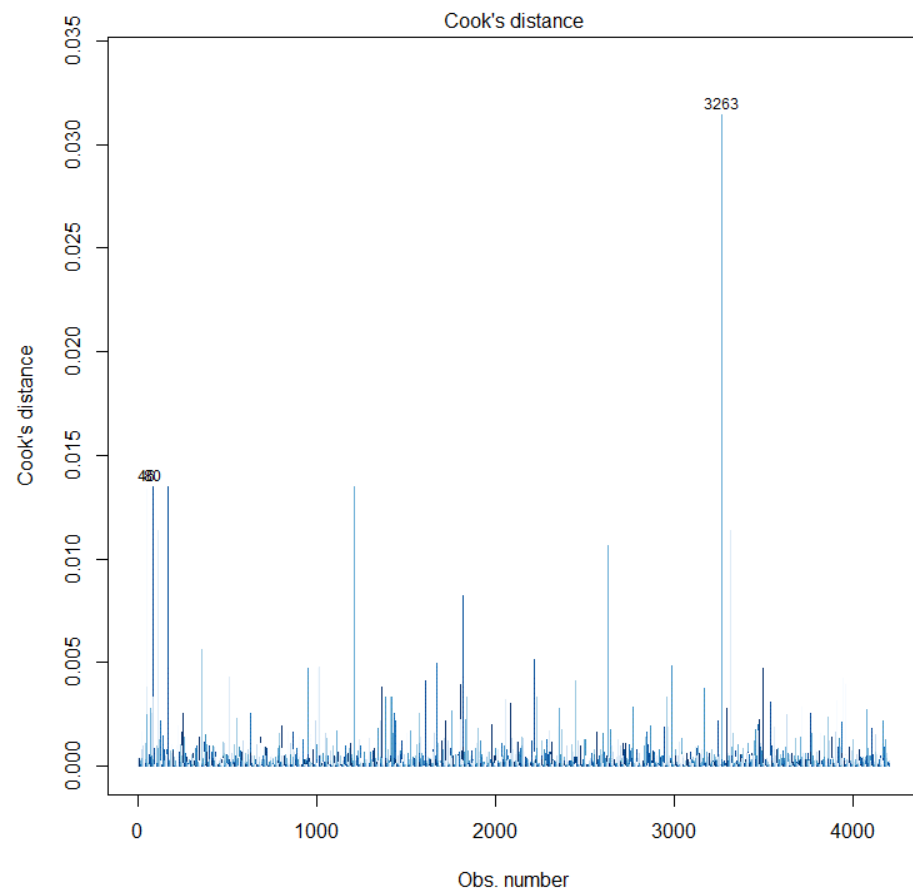
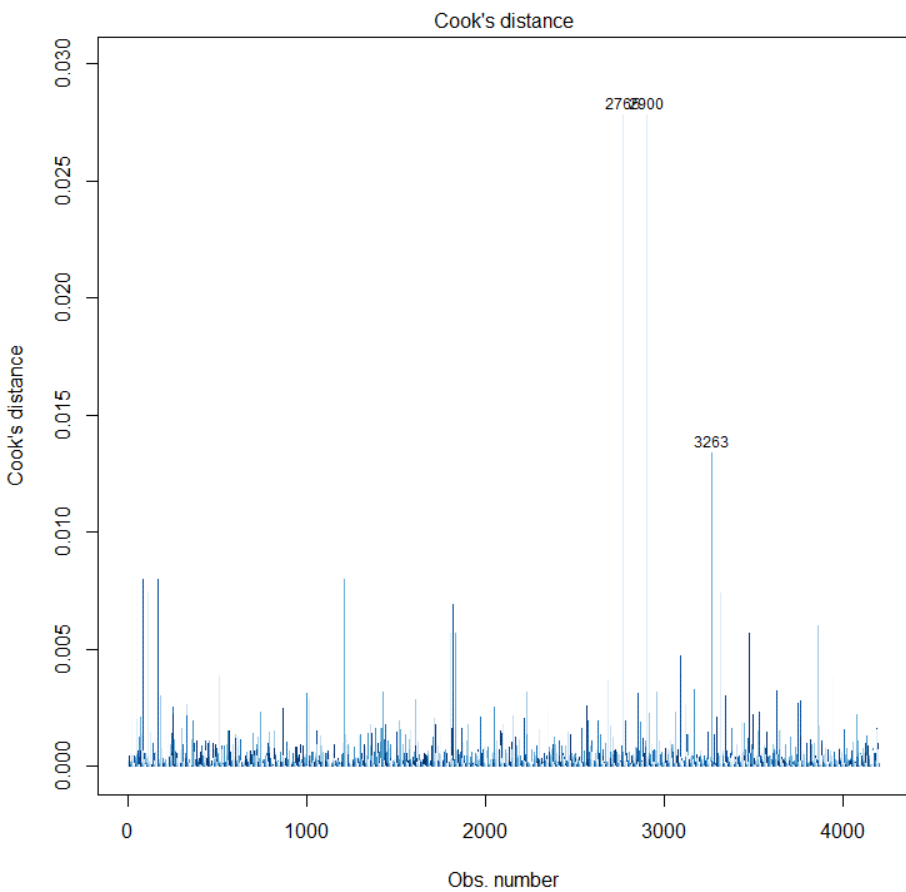
- Referenced on D [\because EPC D has most number of values]
- Referenced on North-west [\because North-west has most number of values]

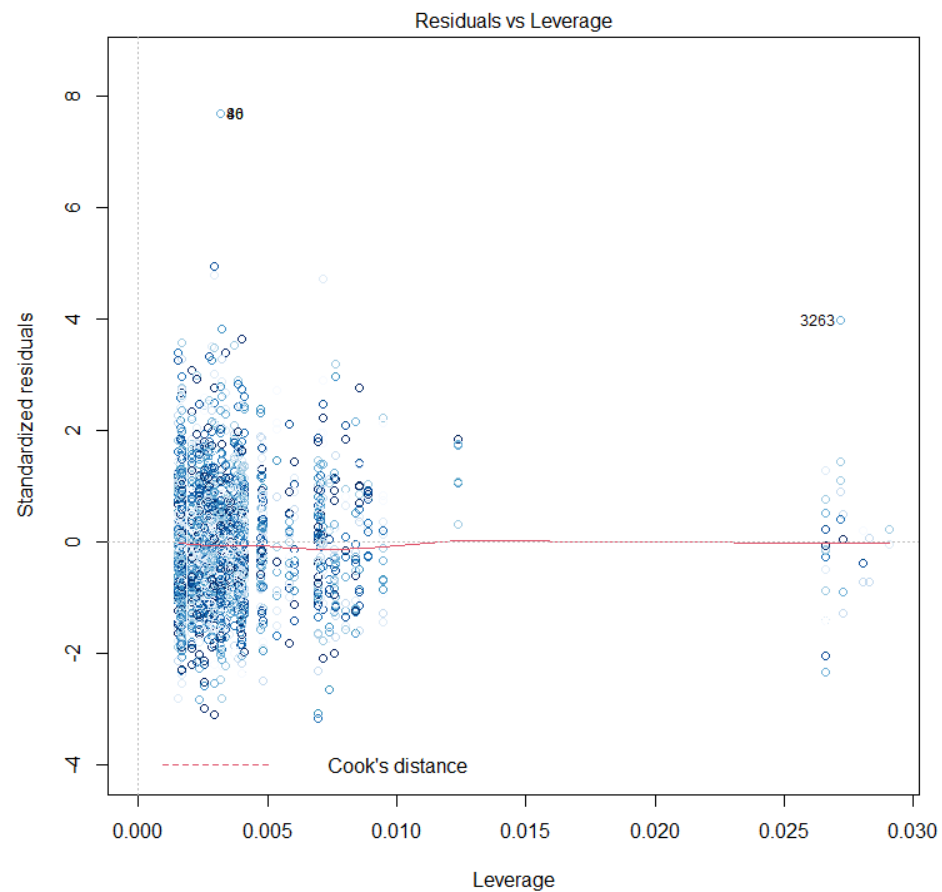
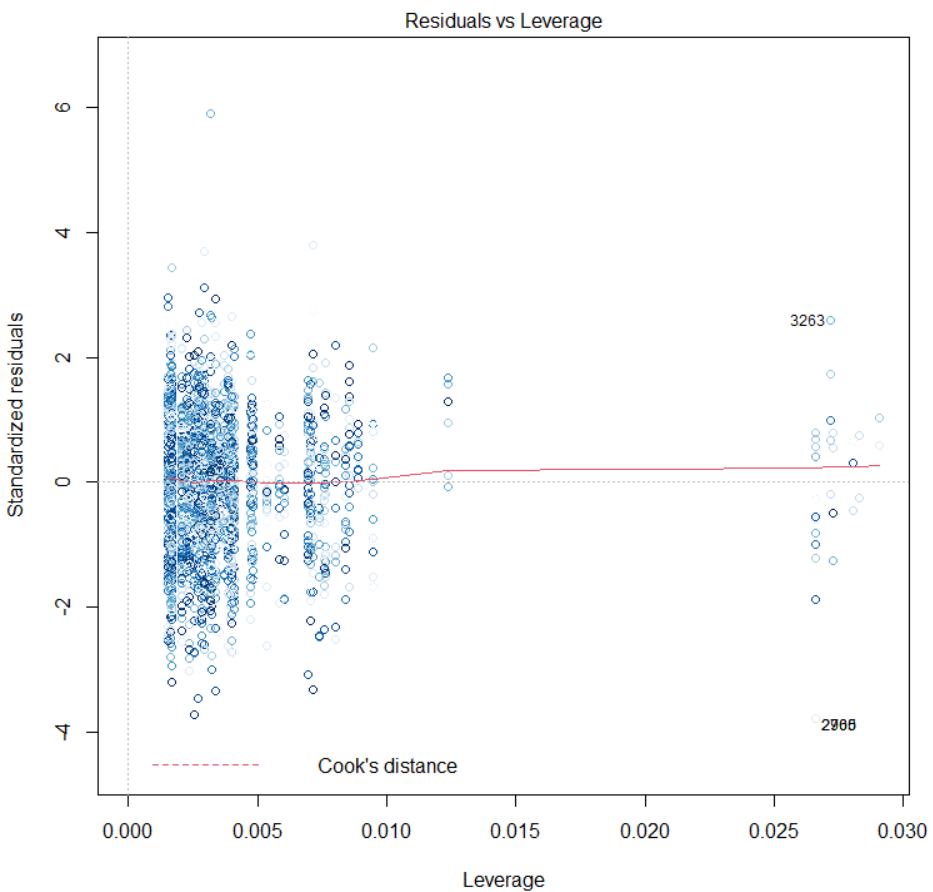


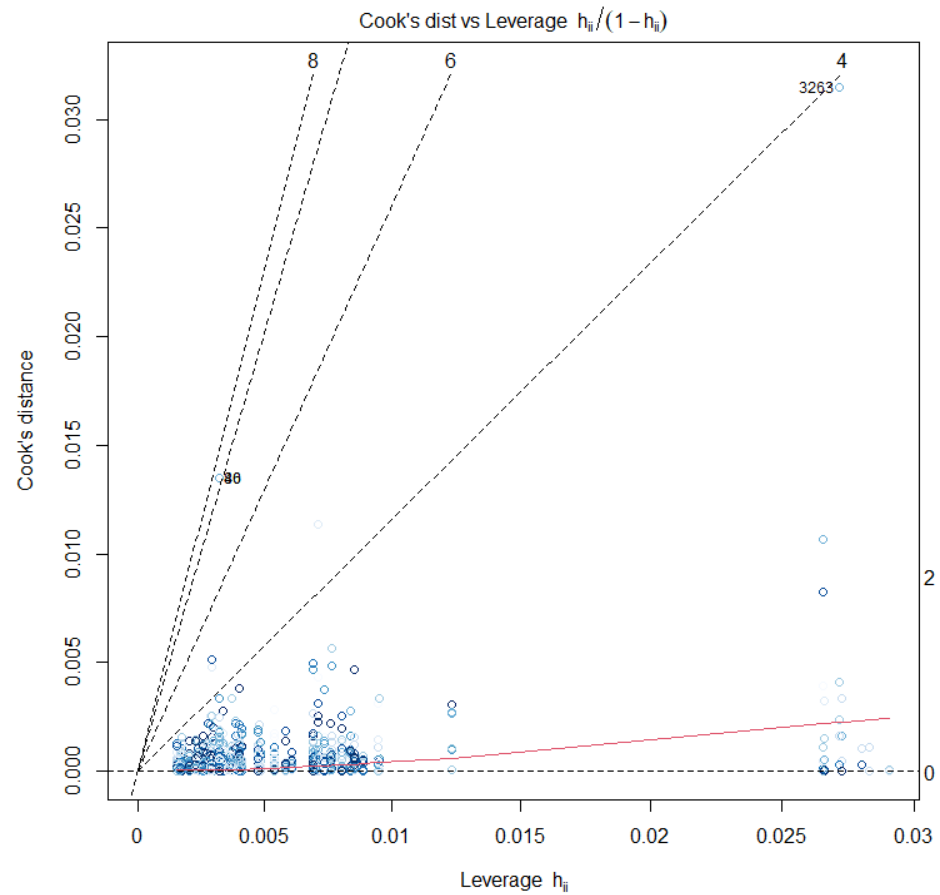
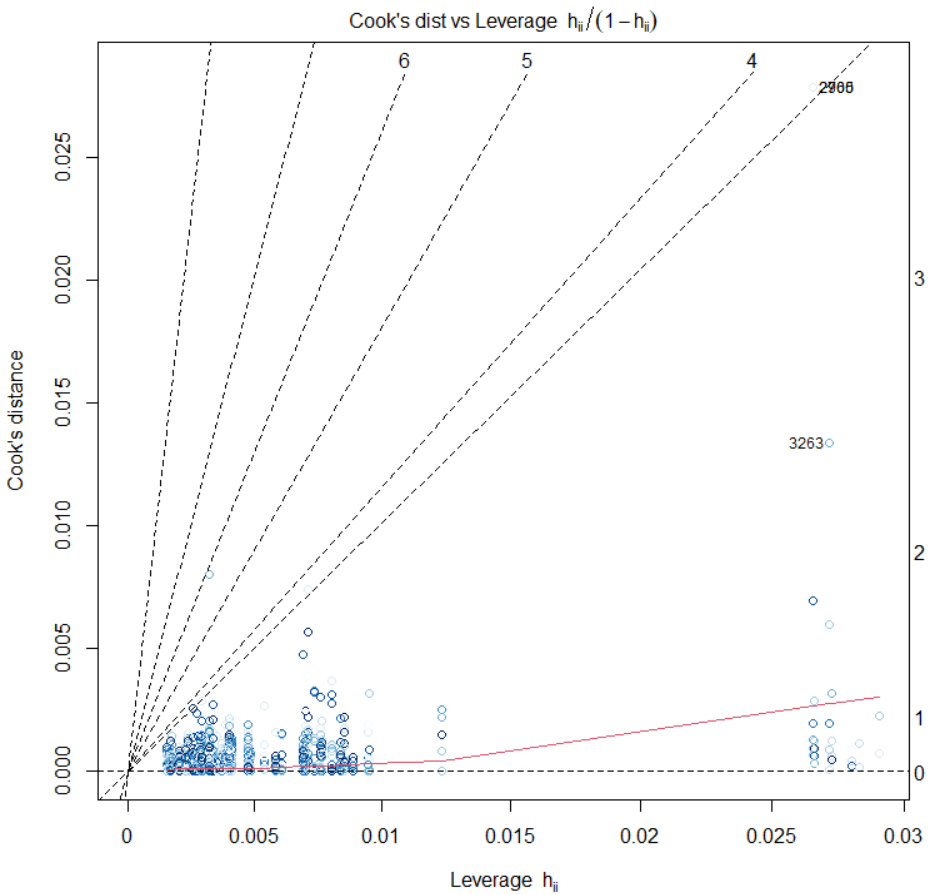


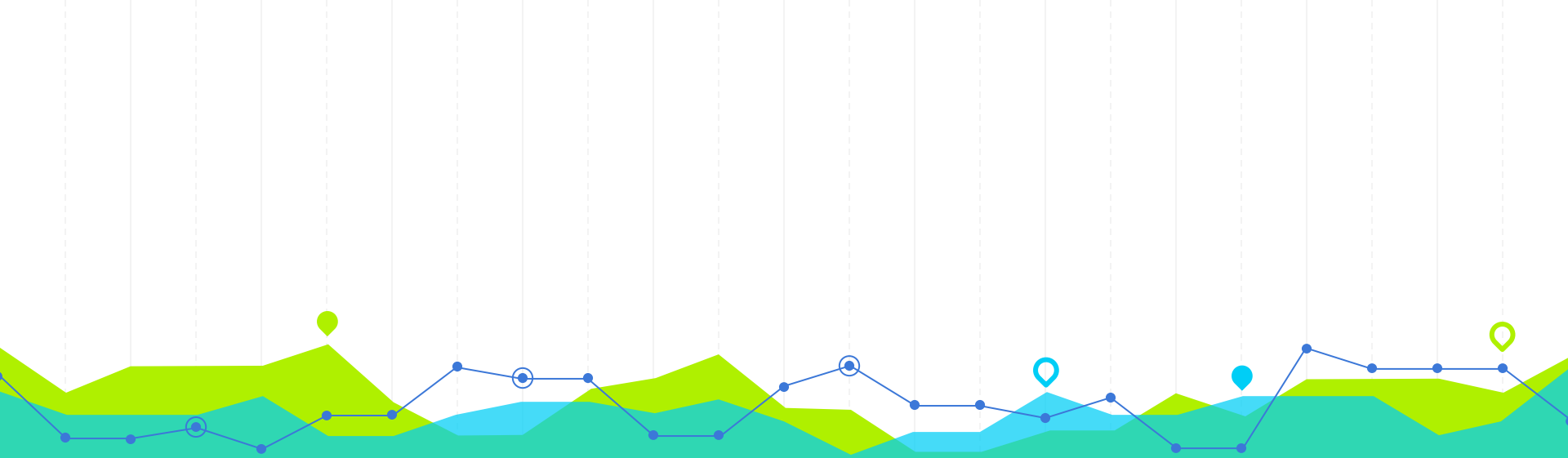












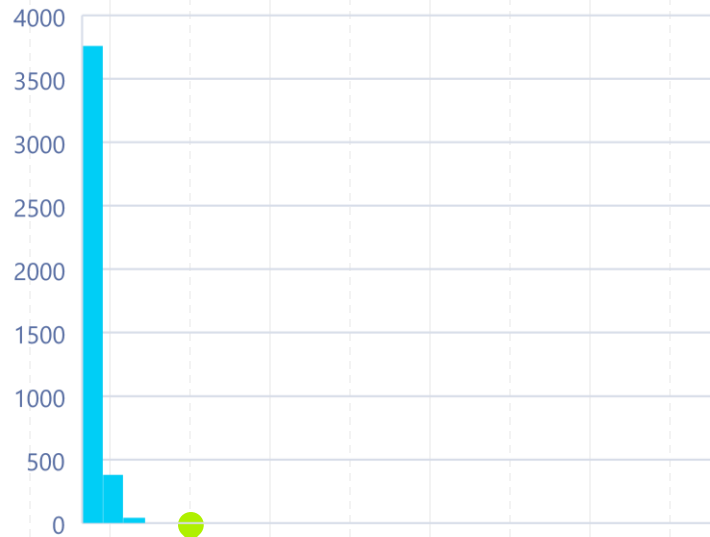
Price vs Socio-Economic

3

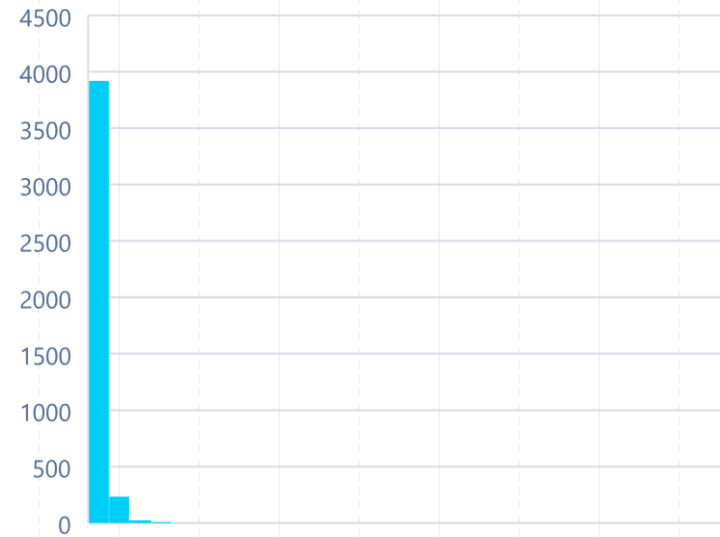
Normality

- 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

First transaction price



Second transaction price



Box Cox Transformations

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

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



OUR PROCESS IS EASY

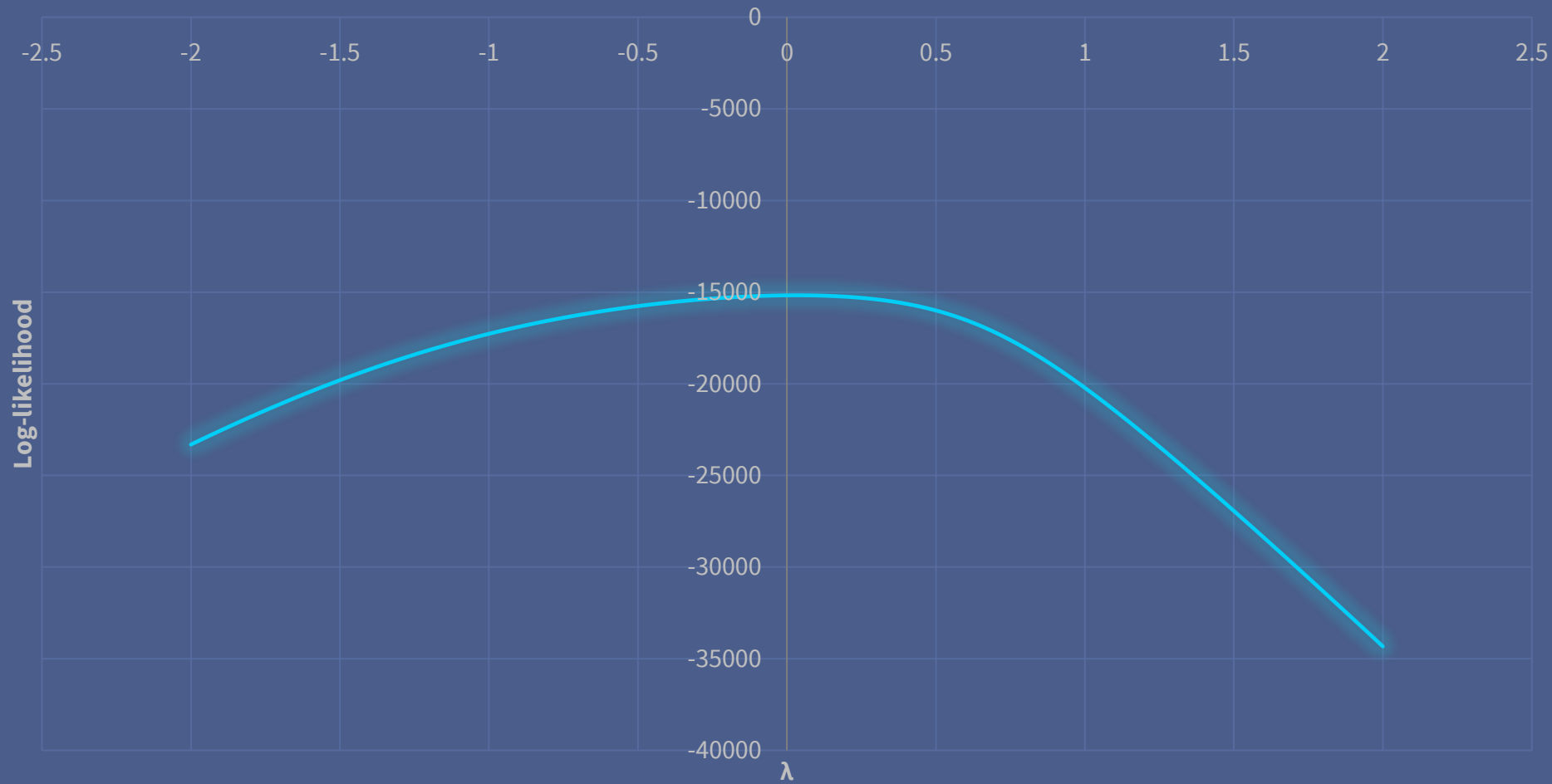
y

$f(y)$

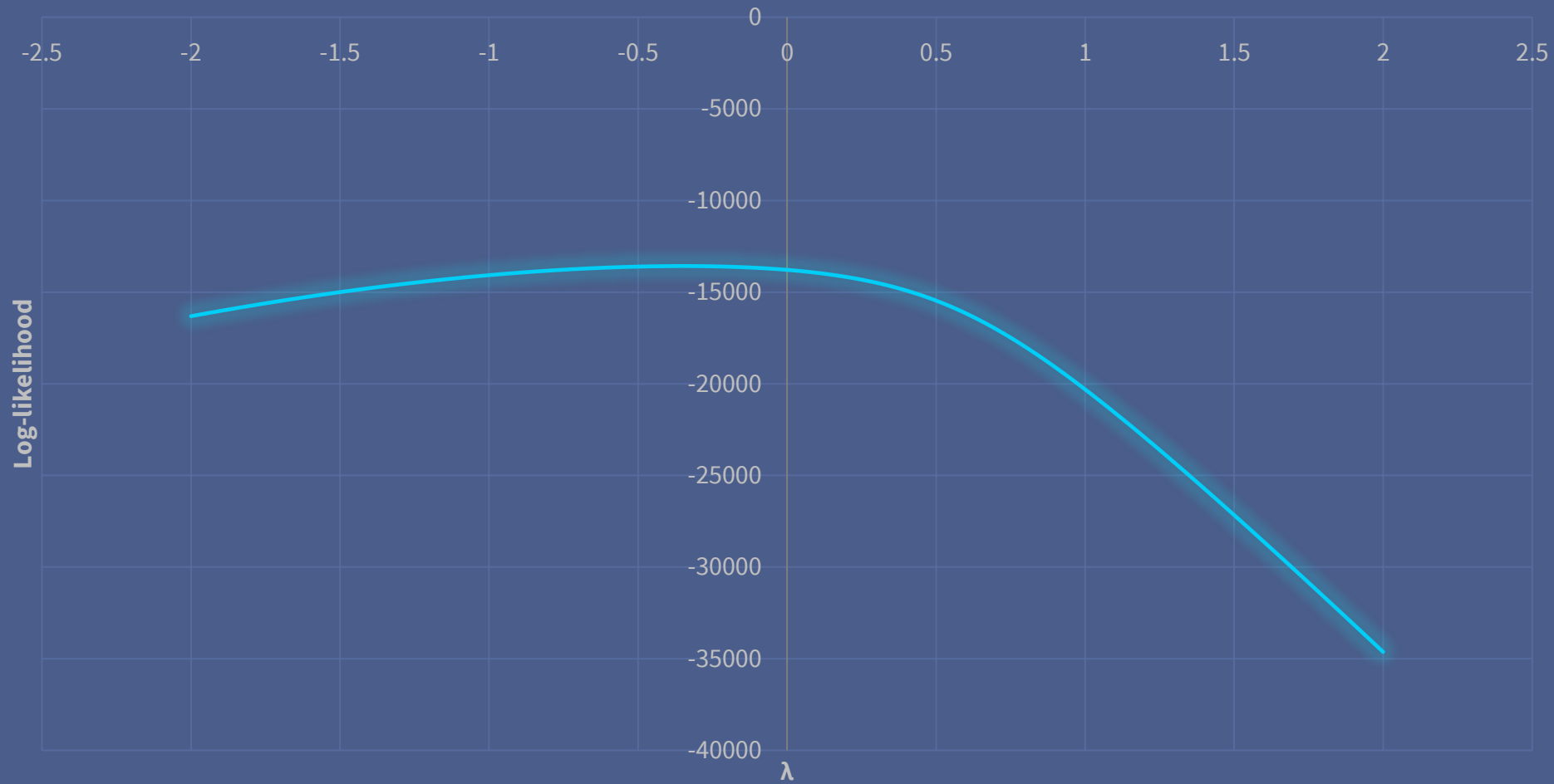
Perform Regression

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

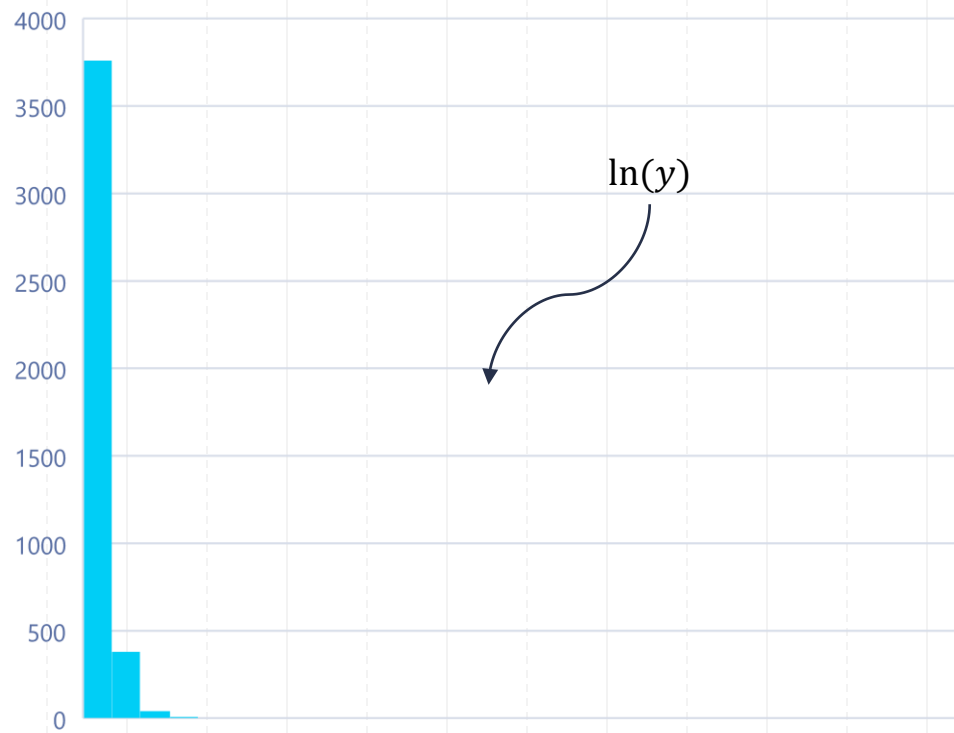
Price 1



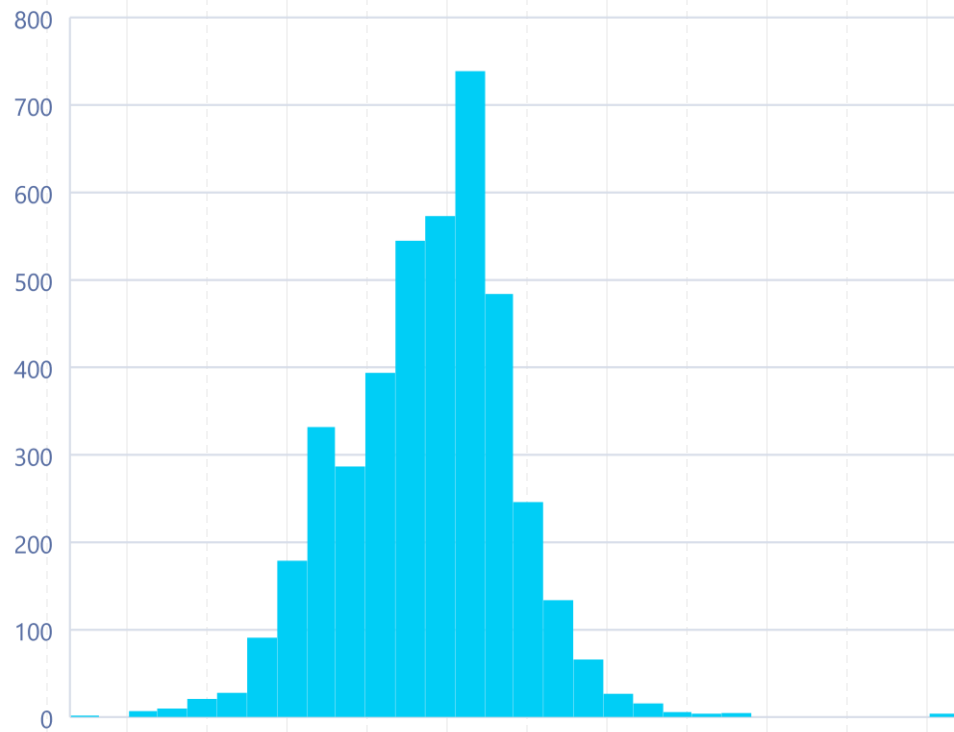
Price 2



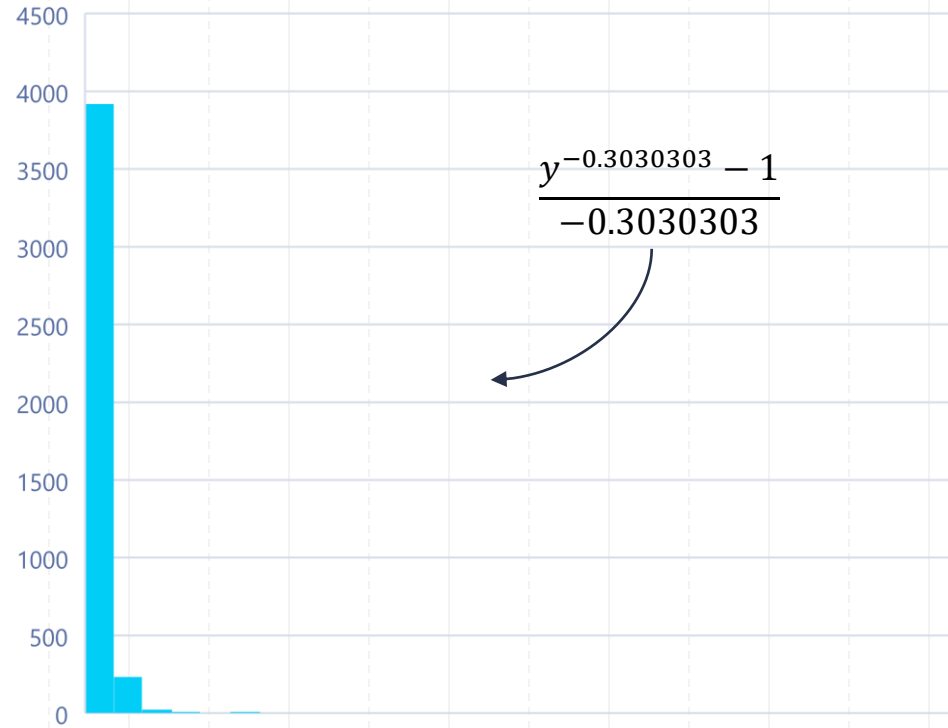
First transaction price



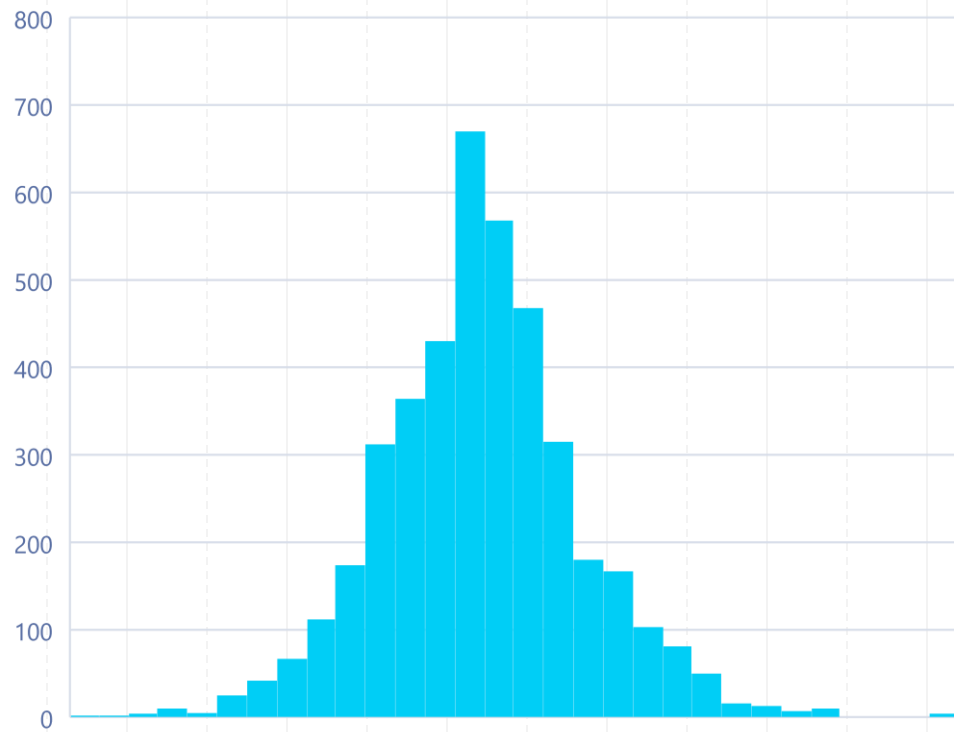
Logarithm of first transaction price



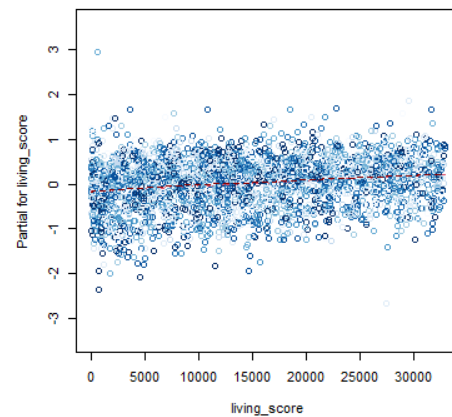
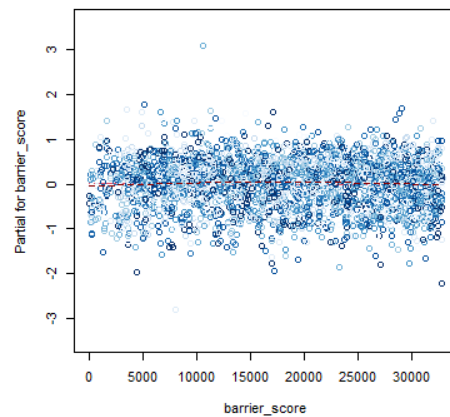
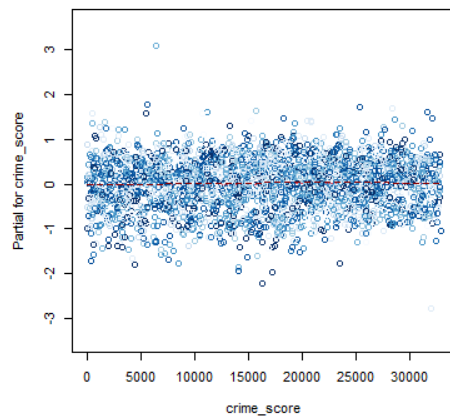
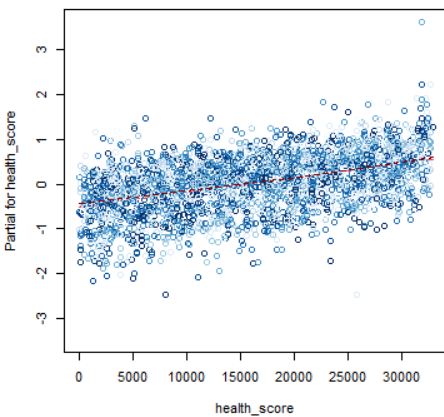
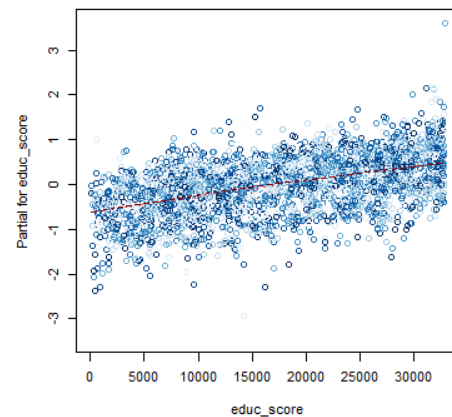
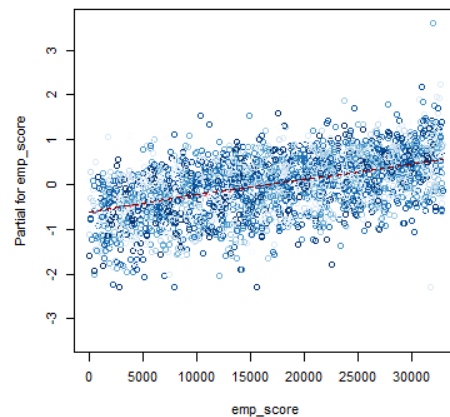
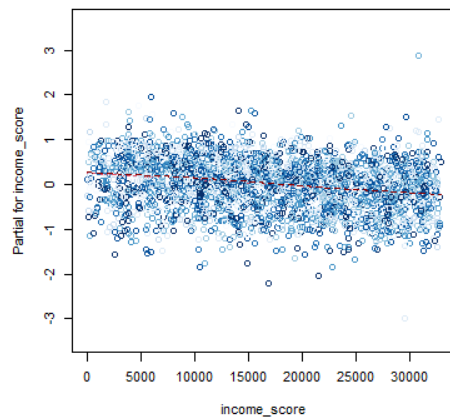
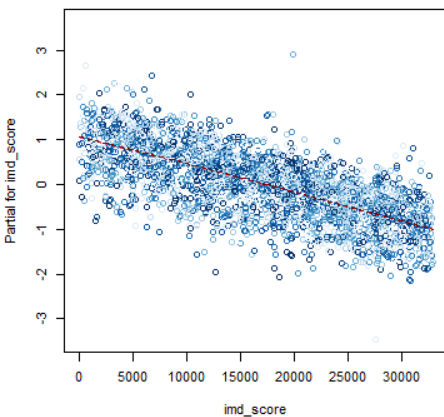
Second transaction price



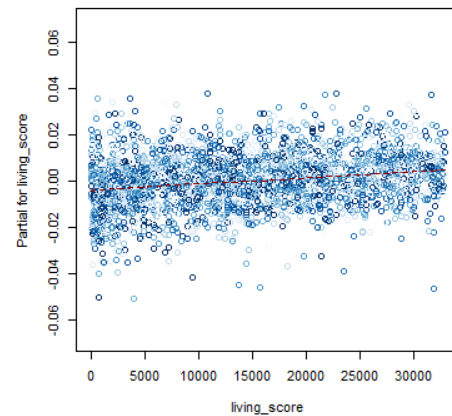
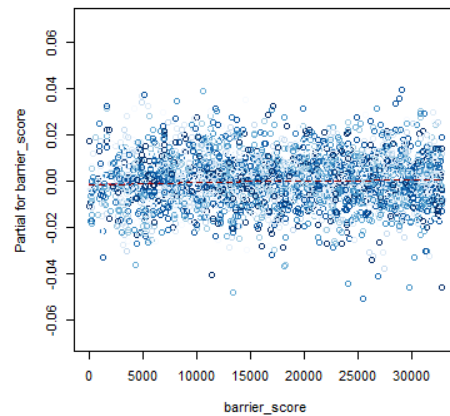
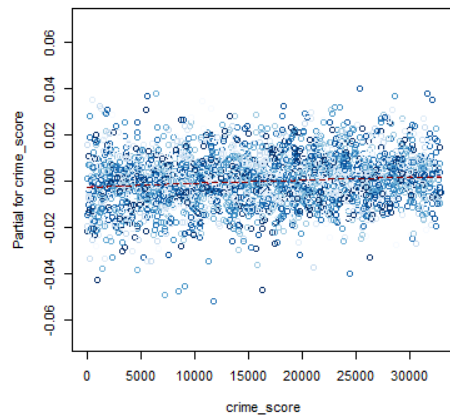
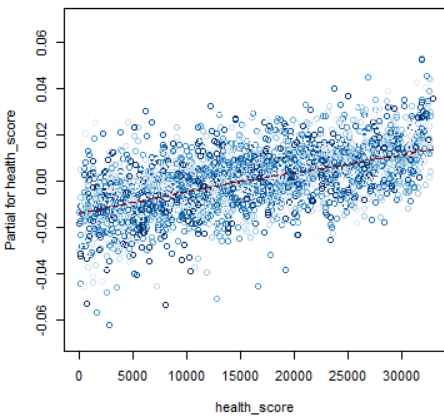
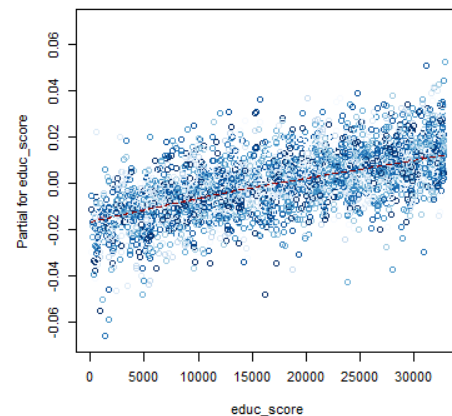
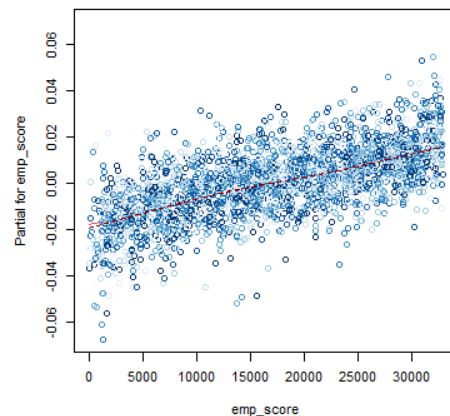
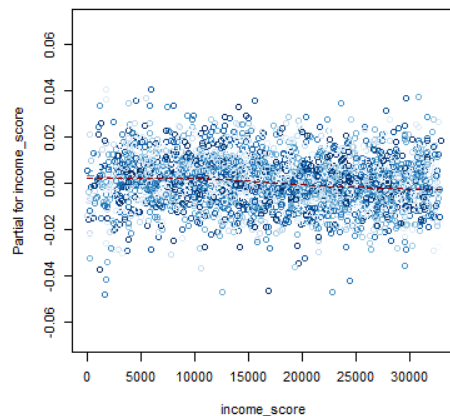
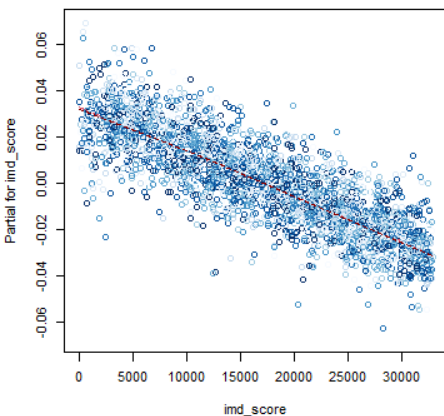
Transformed second transaction price

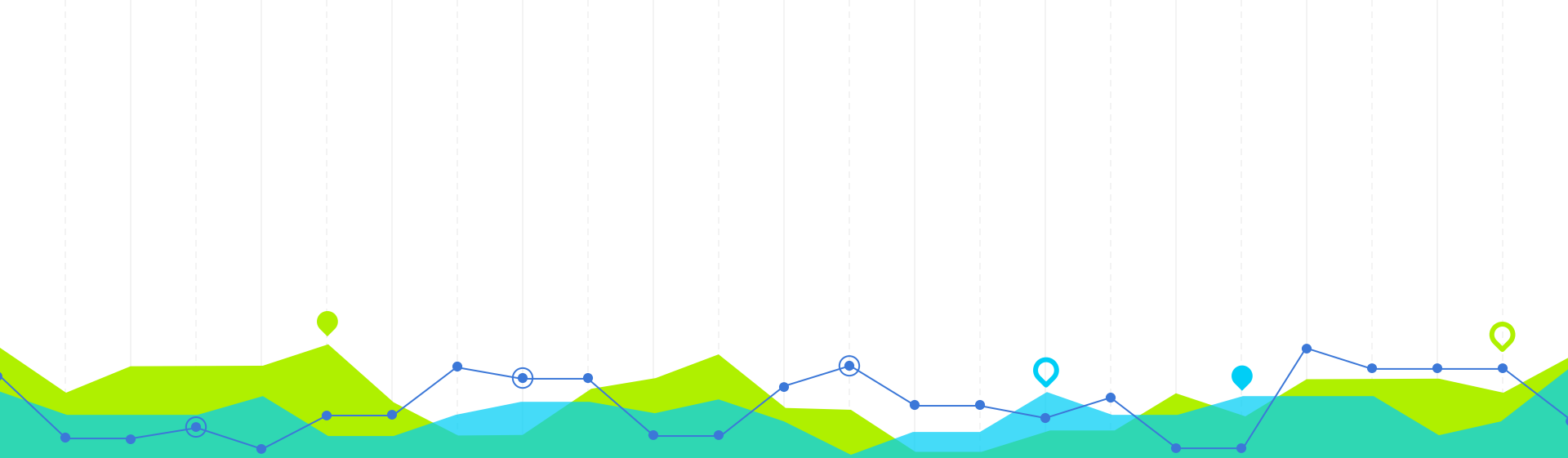


Partial Residue Plots for Price 1



Partial Residue Plots for Price 2





Repeated Sales Index

4

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

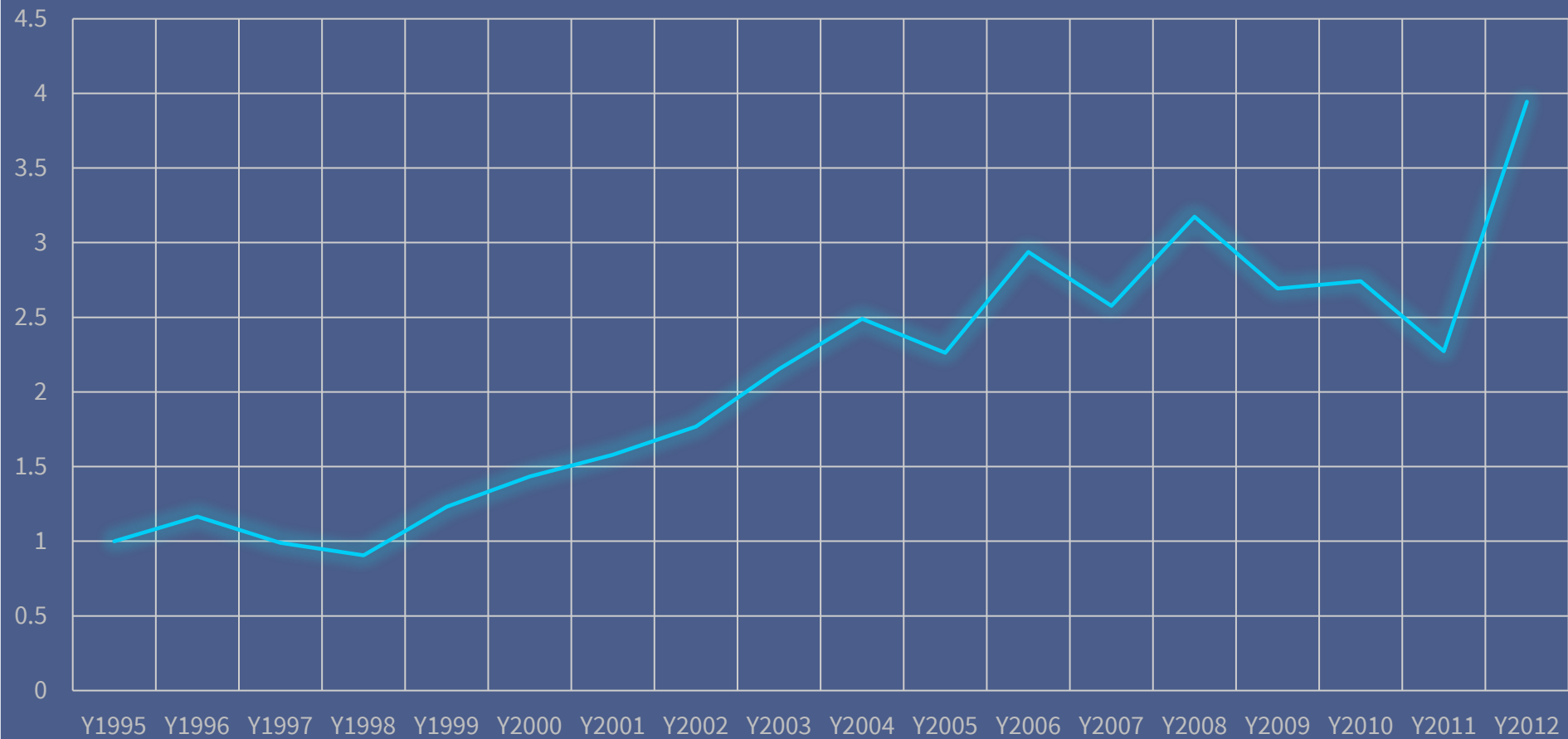
Year	Log Price_1	Year	Log Price_2
2007	1	2008	2
2006	3	2007	7
2007	5	2008	9

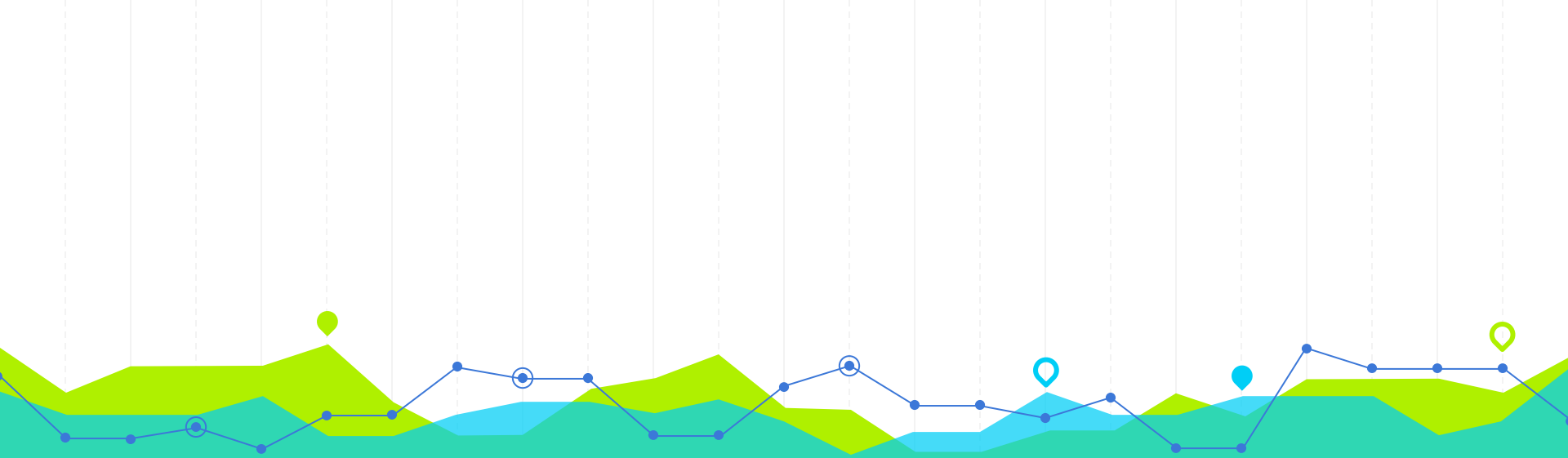
Data Processing

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

RSI

— RSI

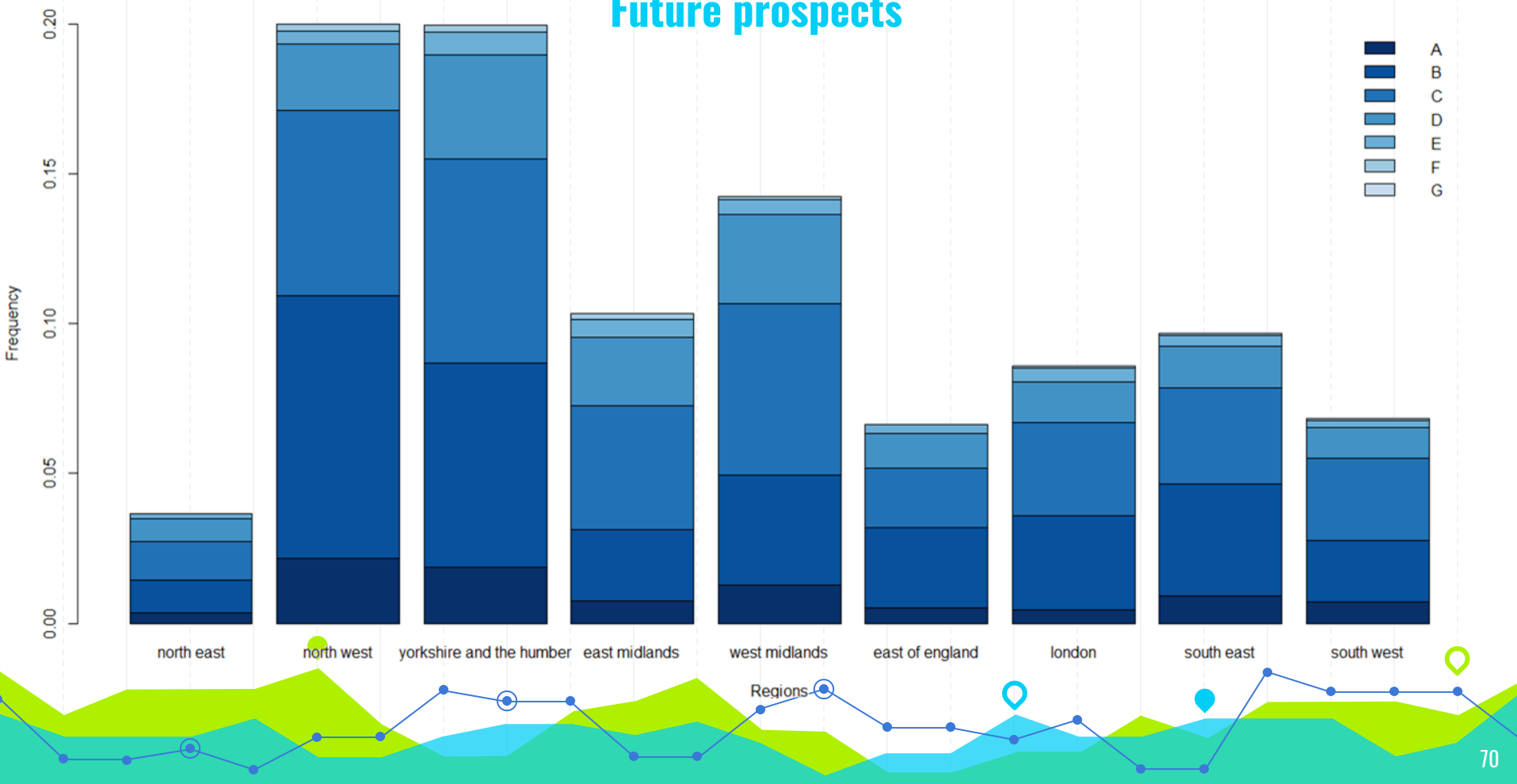




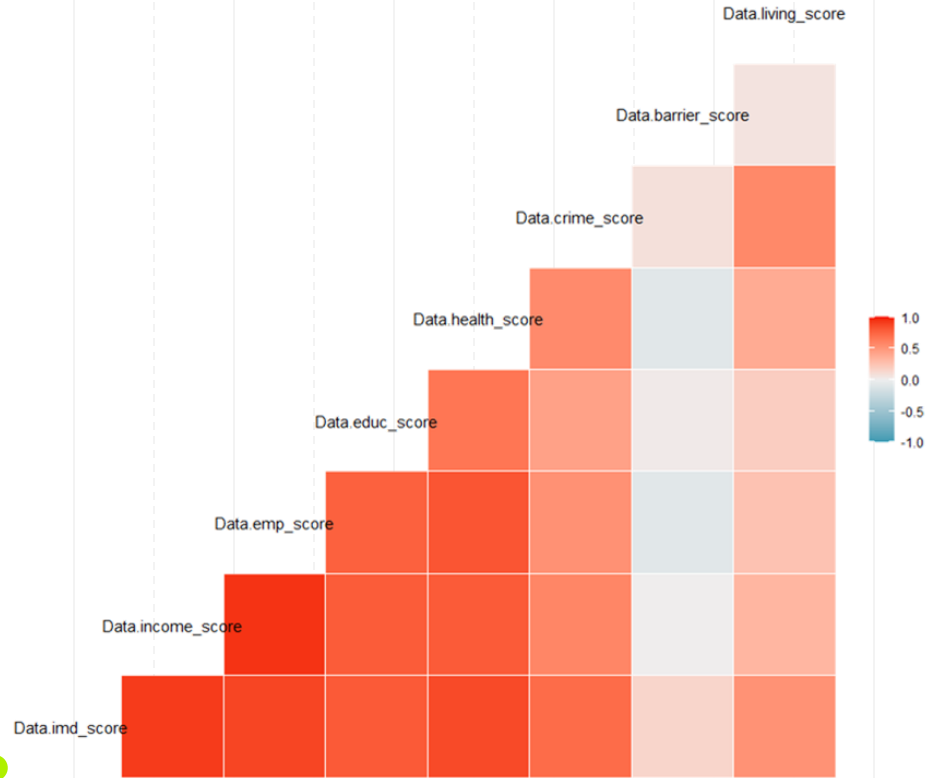
Conclusion

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

S

WEAKNESSES

Many assumptions may not hold for the population

W

Better statistical methods can be used to get rid of assumptions

OPPORTUNITIES

O

T

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.



THANKS!

Any questions?



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