BEST PLACE TO LIVE IN UK

Data Diggers

Yifang Zhou
University Of Southampton
Foundations of Data Science
yz5f17@soton.ac.uk

Richa Ranjan
University Of Southampton
Foundations of Data Science
rr2n17@soton.ac.uk

Anton Okhotnikov
University Of Southampton
Foundations of Data Science
ao2u17@soton.ac.uk

Shihuai Wang
University Of Southampton
Foundations of Data Science
sw1g17@soton.ac.uk

Daniel Pelnar
University Of Southampton
Foundations of Data Science
dp4n17@soton.ac.uk

ABSTRACT

Moving to a new city is always challenging! UK has always been a preferred choice for a wide population in all respects, may it be for higher studies, or work, or relocating in general. We, as a group of students who have moved to Southampton recently, came up with an analysis that helps people to decide which is the most liveable city in the United Kingdom. This project aims to find the most preferred city based on chosen factors that most people find important when moving to a new city.

INTRODUCTION

In order to find the most liveable city in the UK, we first needed to define a model with all the factors that the majority of people find important. After going through various researcher and conducting our own survey, we came up with twelve factors and their weights. Then we found data for these factors. We managed to get about 80 UK cities as the number of observations. The database of our choice was MongoDB where most of the cleaning and merging took place. After that, we run the linear regression and tested our factors for their statistical significance and eliminated the ones that were not significant at 5% significance level. Finally, we combined the estimated coefficients from the regression with our survey weights and got the final weight by which we judge how much each factor is important for determining the liveability of a UK city. The product of our work is an interactive online application which takes inputs like nationality, age, factor preferences of the users, and suggests the best city to live based on those inputs.

MODEL

0.1 DATA COLLECTION

For various factors that determine the liveability quotient of a city, related datasets are required. Various government websites provide data for these factors with respect to the cities in UK. The data has been collected from the following sources:

- UK Government website
- Kaggle.com
- National Health Survey

and some more. These datasets provided the data for various cities in UK. But to understand the weights of each factors affecting the liveability, as well as to ensure that the factors collected are realistic, we additionally relied on the following two sources:

- The Happiness Ranking dataset: This is the dataset from the official government website, which provides the happiness levels of people bases on various factors. The data explains what factors determine people's happiness, and to what level. For example, Weather could be a factor which influences people's happiness. So, the happiness level would be high for this factor in particular. Therefore, the factors containing the maximum happiness level were chosen.
- Survey: A survey was conducted on a sample size of 123 people. They were asked to rank each factor from 1 to 10 (10 being the most important), based on their preferences when moving to a new city. The factors are: Flat/House prices, Population, Road Traffic, GVA per worker, Unemployment rate, Noise level at night, Total Jobs available, Weather, Number of schools, Entertainment, Quality of higher education, Number of hospitals, Connectivity to other cities. Being aware of ambiguity and vagueness of some of these factors, we provided a short explanation for GVA per worker and Entertainment in order to help them make a more informed decision. Additionally, Age, Gender, Home Region and Employment Status were collected as well.

0.2 DATA CLEANSING

For the datasets collected from various government websites, consistency was a major concern. To tackle this problem, we compared all the datasets to confirm if data were available for all the cities, in all the factors. The cities which had no data across all the factors were omitted. As a result, the final count of cities ready for processing was narrowed down to 82. For the datasets obtained after performing general formatting/cleaning, the normalization process was carried out. The normalization techniques used were Z-score and Min-Max. In some cases, the cities had no numerical data. For instance, when we considered Medical facilities as one of the important factors, the dataset to work with was 'The Number of Hospitals each city has'. To normalize this data, the Z-score method was used, as we wanted to preserve the range. Similarly, other datasets were normalized as well, based on the type of data. The Min-Max normalization method was used to normalize the weights from the survey. Further analysis was done on the normalized data.

0.3 DATA MANAGEMENT

The final datasets have been transformed into a different format such as a table-like structure, or a JSON file. Therefore, it was essential to have them all stored in a database for consistent storage. The storage is done on MongoDB, in the form of collections.

0.4 PROCESSING / METHODOLOGY

For the analysis, we used the following methodologies:

- For selecting the most important factors to work upon, a feature selection technique called **Lasso** was used.
- On top of Lasso, we also applied an elimination method that is using F testing and T testing in R.
- The **Linear regression** was used to get the estimates and standard errors, as well as, R squared of the whole model. For this, all the datasets were merged and we regressed happiness rating (the regressand/depedent/explained variable) on all of the 12 chosen factors (the regressor/independent/explanatory variables). The result was estimated coefficients and standard errors for each of the 12 explanatory variables. Certain limitations have to be noted here, namely, some of the assumptions of the linear regression might have been violated. More preciously, the zero conditional mean assumption that says that the unobserved factors in the error term should not be dependent on the explanatory variables. However, we are positive to suspect that there might be some factors which we could not have included in the model but which explain the happiness rating and are dependent on the explanatory variables. For example, there were no usable data for crime rates. However, crime rate can explain our "explained" variable to some extent too. A further problem is that the crime rate is correlated with unemployment rate. We also could not find any data for the Environment (tree cover, the amount of green in the city). This is also an issue. In summary, all this means that the estimated coefficient, might be and probably are slightly biased. We also violated the random sample assumption and we might have measurement errors in the reported data we got from the government websites as people tend to overestimate some things and underestimate others as they are filling out government polls. An example of this might be unemployment rate, as people have incentives to alter a bit the information they are filling in in order to get unemployment benefits.
- For merging the factors and their weights, the **Baye's Decision** rule was used.

0.5 APPLICATION

After performing the analysis, a web application was built, where individual users can interact with the application, and based on the inputs (age, gender etc.) entered by the user, the most preferable city would be recommended. This recommendation is a result of the analysis carried out by our model. The result would look somewhat similar to the visualization shown here:

Another piece of recommendation could also come up from the survey results calculated before. For example, if a user enters a certain age, gender, location, etc., the application could get the data from the survey reports, and would recommend a city that was chosen by people of similar gender/age/location and others.

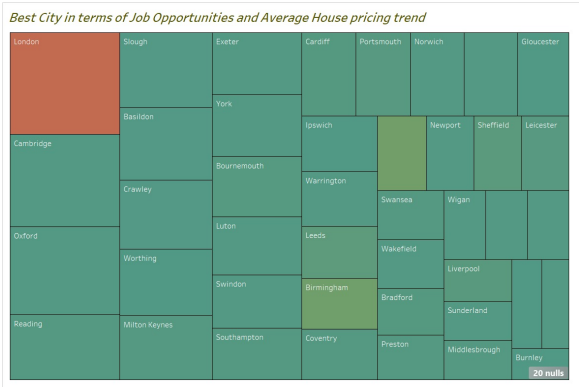

Figure 1: An example of Best City result based on two factors

CONCLUSION / RESULTS

As mentioned under the Application section, the result obtained will be two different recommendations, one based on our analysis, and the other one based on similar parameters from the survey conducted before.

LIMITATIONS AND REFLECTION

There are a few limitations of the model and our work in general:

- Most of the datasets are concentrated during the years 2015 and 2016. Since this is only 1-2 year ago, the data collected on our factors should still be representative to a good extent of what is happing in 2017/2018. However, the interactive application can become outdated in the next 10 years as the factors can change much more due to the slowly and ever changing human culture and what things(factors) people are valuing. For instance, it can happen that in the next 10 years, a new extremely effective way of travel is invented and as a result 'connectivity to other cities' might become redundant. Another example is with the automation of human jobs, as a result a universal basic income might be introduced and people might to value 'total jobs' factor much less in the next years.
- There were insufficient entries for all factors across all cities. Therefore, the list had to be narrowed down to 82 cities only. This might cause a problem with statistical inferences as we do not have consistent estimators asymptotically and testing (F and T test) is out of question as well.
- In order to get unbiased estimated coefficient when running a linear regression, 4 assumptions have to be satisfied. Zero conditional mean assumption was most likely violated due to not having all the factors' data available. For instance, we wanted to collect data for crime rates of each of the city of interest, however, there are no data for this and so the crime is rate is in the error term unobserved, making our

estimates biased. The other factors which we wanted to analyze but were not able to are: Environment(Nature), Views or Transport system. The assumption of random sampling was violated too and finally we suspect that some of the data we got might have measurement errors, for instance the population data.

- For the survey, the sample size for some of the subcategories is too small, and as a result, a bias has to be included in further calculations. Namely, the subcategories that do not have enough observations are:
  a. Age under 18, and over 36
  b. Regions: Africa, Australia, North and South America
  c. The dominant region is Asia (70,7% of all observations), then EU but not UK (16,3%) and UK (6,5%)
  d. For Professional/employment status we have enough observations only for the following three categories: a student (52%), employed for wages (26,8%), self-employed (12,2%). For the others like retired, out of work etc. we do not have enough observations.
- The application does not have the survey responses for the subcategories that do not have sufficient enough observations.

REFERENCES