Investigating Airbnb's in Edinburgh

September 11, 2023

1 Overview

Holiday makers means of tourism have changed, in particular, how holiday makers acquire accommodation. Companies like AirBnb allow user's to make short-term rentals. This report investigates Airbnb's within Edinburgh's context and identify what trends exist when examining **popularity**, **price** and explore the effect of **location** on both of these ideas. To achieve this, a data set containing information on 7389 listings in Edinburgh was wrangled and cleaned. To explore popularity, a measurement variable called **Mean Score** is created and is subject to inferential statistical evaluation and is modelled. This is then applied to the context of location. To measure the effect of location, k-means clustering was performed on latitudinal and longitudinal information which revealed key 'neighbourhoods' of interest. Namely, these were close to city center. Three fields, years_since_first_review, year_since_last_review and reviews_per_month were the highest correlated with 'Mean Score'. Also, certain neighbourhoods were more expensive and was supported with great statistical significance. Further research could investigate alternative interpretations of 'popularity' and more objective measures of 'clusters' close to city center.

2 Introduction

Context and motivation Airbnb is a company which accommodates users who are travelling. People who register for the application as a host can list properties which allow normal users to 'rent' temporarily for a specified duration. This can act as an alternative to staying in Hotels and much of the appeal comes from cheaper prices and no hidden fees like taxes[7]. There is great motivation to understand the complexities of Airbnbs as it can reveal an insight into the business and economics of this sector. This is significant for hosts as there is a great opportunity for business as it was found that travellers likelihood of staying in hotels in the UK had decreased from 79% to 40%.[9] Therefore, realizing patterns in popular Airbnbs for hosts can be advantageous in attempting to replicate similar success with their own properties. This is important because in Edinburgh alone, throughout 2022 there were **117,540** bookings made[1]. Of the 7,389 properties that exist in Edinburgh, 9% of properties are listed to have a review rating of 0[1], presumably due to no bookings.

Previous work Exploring literature work before revealed that a study on Airbnb's in 40 cities in America revealed that Airbnb rentals were in *better* neighbourhoods closer to city center with good public transportation.[8] This finding was replicated in Metro Nashville Tenessee where there was a strong correlation between distance to a key landmark, price and review score. [10].

Objectives In this report, we will be attempting to identify features accounting for popularity in an Airbnb. This will be done by using the review scores in the data set to create a new dependent variable 'Mean Score' which several statistical and visual analyses will be performed. We will then find what variables have the highest correlation with it and attempt to produce a model. A location perspective is achieved by performing k-means clustering to identify popular neighbourhoods for Airbnb's to reside in. Then, price and Mean Score will be looked at in this context in order to identify distributions and it's meaning.

3 Data

Data provenance Data for this exploration was collected from Inside Airbnb's website.[5] The data for which is 'sourced from publicly available information from the Airbnb site'.[4] This report explores the csv file 'listings.csv' which represents a snapshot of the data when it was downloaded (7 March 2023). While Inside Airbnb themselves don't have any restrictions on this data acquisition (as per their Community guidelines)[3], Airbnb themselves state in their Terms of Service Section 12.1 'Do not use bots, crawlers, scrapers or other automated means to access or collect data'[2]. However, since we are not directly performing webscraping on Airbnb themselves, we will continue.

Data description The csv file listings.csv contains a list of properties in Edinburgh with 75 fields on information about each property such as 'id' and 'description' and there are 7,389 different property listings. Going forward, we will mainly be looking at the variables 'price', 'latitude', 'longitude' and 'Mean Score' to further analyze the data set.

Data processing This csv file was wrangled and cleaned with Pandas DataFrames so that statistical analysis can be performed. Firstly, columns which can be immediately discounted where looked at. The following columns where dropped due to being hard to quantify:
id, listing_url, sraped_id, last_scraped, source, name, description, neighborhood_overview, picture_url, host_url, host_name, host_location, host_thumbnail_url, host_picture_url. Also, the following columns were removed due to having all values NaN:
neighbourhood_group_cleansed, bathrooms calendar_updated, calendar_last_scraped.
Then, to ensure that the data set is compatible with calculations, there were some columns that needed adjusting such as columns in a date format.

1. host_since → years_hosting
2. first_review → years_since_first_review
3. last_review → years_since_last_review

Then to account for categorical variables, binary one-hot encoding was used. Then, a separate column was made which contains the Mean Score of a property based on the mean average of the following:

1. review_score_rating
2. review_score_accuracy
3. review_score_cleanliness
4. review_score_checkin
5. review_score_communication
6. review_score_location
7. review_score_value

The column *neighbourhood* was removed as well due to the column *neighbourhood_cleansed* being a tidier representation of what *neighbourhood* was trying to represent.

4 Exploration and analysis

Interpretation of results Figure 1 highlights a correlation heatmap of Mean Score against every other variable in the data set. However, we have removed all constituents which make up 'Mean Score' as well as 'Mean Score' itself since they will all trivially have a high correlation. The results reveal that the following three columns *year_since_first_review, years_since_last_review, and reviews_per_month* all have the highest correlation with 'Mean Score'. Then, multiple linear regression was applied using these three independent variables against the dependent variable 'Mean Score' which yielded the results in

Figure 2. The model produced an R^2 value of 0.336 indicating some positive correlation, however, it isn't incredibly strong. However, our p-values for each of the coefficients are 0 indicating the signifiance and correctness of each.
Figure 3(b) highlights the application of K-means clustering on positional coordinates for each property. The Scree plot helped identify an appropriate value for k. Since there is no sharp elbow, we assume k to be 10. Figure 4 highlights how these clusters are spread across Edinburgh. Table 1 shows that the majority of Airbnb listings are in Cluster 3. It turns out, most listings reside in Clusters 3, 5, 6, and 9.
Figure 5(a) reveals that price varies depending on what region in Edinburgh a property is in. It's revealed that the more expensive listings all reside within the most popular neighbourhoods, with Cluster 3, 6, 9 all being expensive areas. The boxplot of Mean Score in figure 5(b) reveals little variation in Mean Score in different neighbourhoods, hovering around 4.6 to 5.0. There exists variability within each cluster, however, this may be due to there being a small amount of listings in some clusters over others. For example, cluster 2 has the highest average Mean Score and smallest variability, but it is also the cluster with the least listings. To explore this idea further, we can use Hypothesis testing.

Application of Statistical methods To investigate the effect of cluster location, we will explore whether clusters closer to the city are more expensive than clusters further away. While Figure 5(a) highlights a broad overview of how price differs by location, we will explore this further. Figure 6 shows the distribution of the price of listings in Edinburgh. It's clear that the distribution is non-normal and looks skewed. We can use hypothesis testing to see if clusters close to the city center are more expensive. First, we define our null and alternative hypothesis.

H_0=There is no difference in price based on location

H_a=There is a difference in price based on location

Then for a boundary condition, we will set α to 5%. In order to perform statistical analyses, I created two sets P and R such that $P,R \subset Q$ where Q is the set of all clusters. These clusters are arbitrarily chosen with reference to the coordinates 55.945427, -3.188338 (Bristo Square).

$$P = \{Clusters : 0,3,6,8,9|Thecloseclusters\}$$

$$R = \{Clusters : 1,2,4,5,7|Thefarclusters\}$$

Following this, we can apply a statistical test. However, since our data isn't normalized, we don't perform a standard t-test. Instead, the Mann-Whitney U Test was performed due to the non-parameterised shape of our distribution. This is also appropriate since this type of test works with a lower sample size. This is significant because we know that most listings exist in clusters close to the city center. Since our value for α is 5%, any p-value $p < 0.05$ highlights statistical significance and hence strong evidence for rejecting the null hypothesis.

Mann-Whitney U Test	
p-value	**u-stat**
0.000294	78907.5

Table 2

4.3 Interpretation of findings Table 2 highlights the result of the Mann-Whitney U Test. Since our p-value is less than 0.05, we have incredibly strong reason to reject the Null Hypothesis. Therefore, we can conclude the there is some difference in price between listings which are close to the city center against listings which are further out. This is further supported by the u-statistic which is 78907.5 indicating disjoint between the two subsets P and R in relation to price.

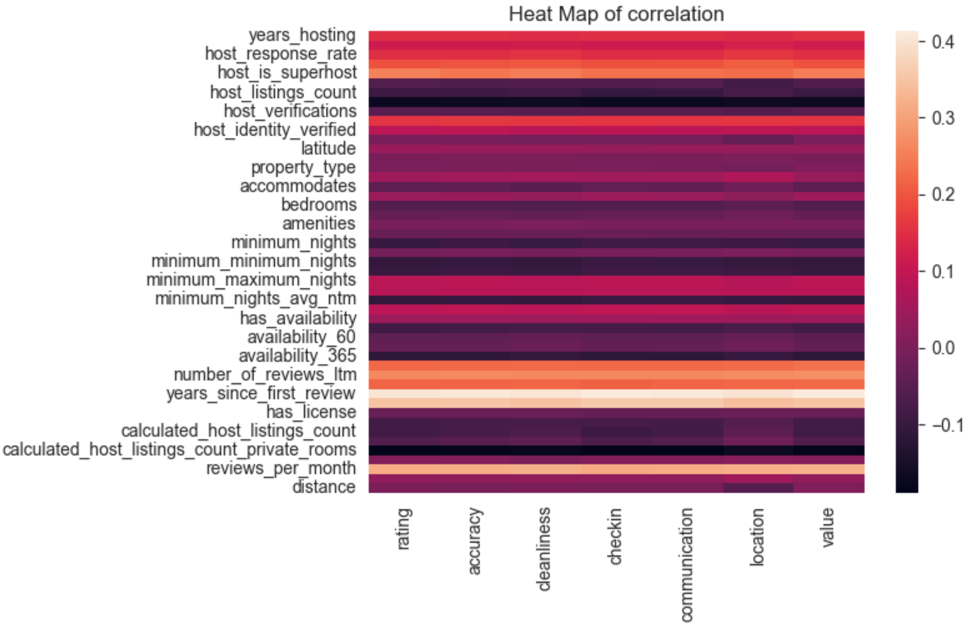
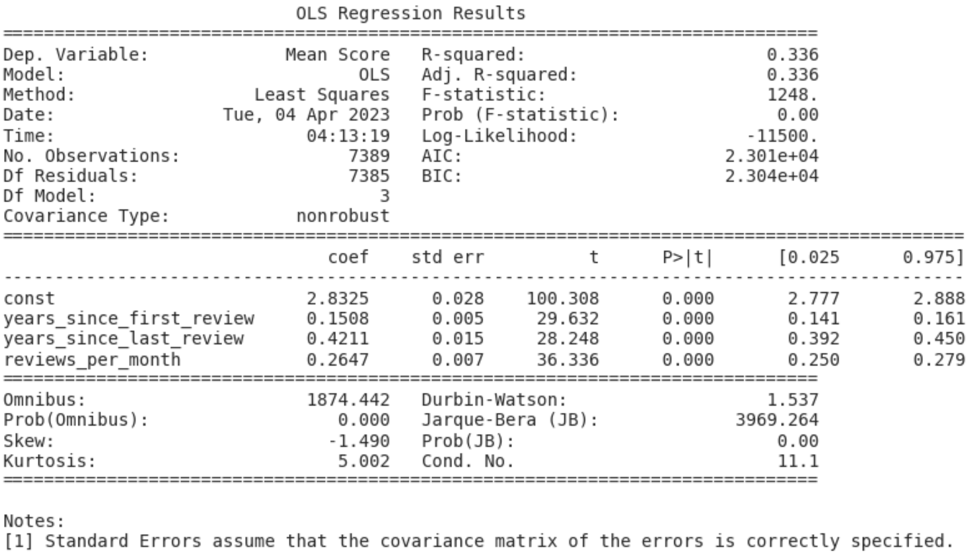


Figure 1 Properties classified by cluster on a map



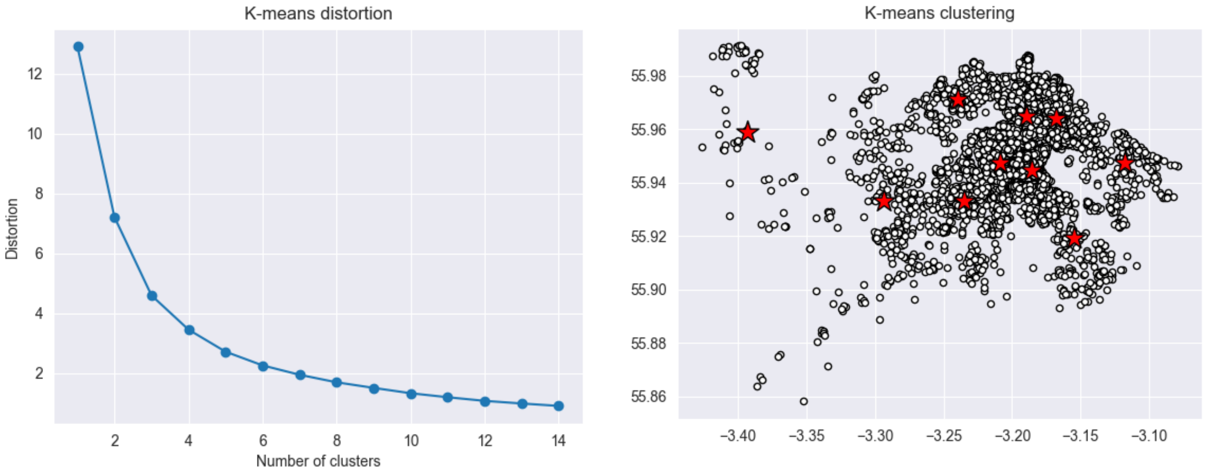
```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Mean Score   R-squared:                       0.336
Model:                            OLS   Adj. R-squared:                  0.336
Method:                 Least Squares   F-statistic:                     1248.
Date:                Tue, 04 Apr 2023   Prob (F-statistic):               0.00
Time:                        04:13:19   Log-Likelihood:                -11500.
No. Observations:                7389   AIC:                         2.301e+04
Df Residuals:                    7385   BIC:                         2.304e+04
Df Model:                           3
Covariance Type:            nonrobust
============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                       2.8325      0.028    100.308      0.000       2.777       2.888
years_since_first_review    0.1508      0.005     29.632      0.000       0.141       0.161
years_since_last_review     0.4211      0.015     28.248      0.000       0.392       0.450
reviews_per_month           0.2647      0.007     36.336      0.000       0.250       0.279
==============================================================================
Omnibus:                     1874.442   Durbin-Watson:                   1.537
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3969.264
Skew:                          -1.490   Prob(JB):                         0.00
Kurtosis:                       5.002   Cond. No.                         11.1
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 2 OLS Model Results for Mean Score

5 Discussion and conclusions

Summary of findings In conclusion, this study has identified key areas in Edinburgh which Airbnb's exist and key variables which account for Mean Score. It was found that the majority of Airbnb's exist close to the city center as our subset P of clusters Q account for 77% of listings in Edinburgh. Each cluster was also located near tourist points of interest in regions with access to public transit. There was also a difference in price depending on where a listing is located. Visually, in Figure 5(a), it's clear that some areas are more expensive than others. This was then verified using the Mann-Whitney U Test which rejected the null hypothesis for the alternative discerning that Clusters further from the city center were different in price than clusters within the city center.



(a) Distortion graph to identify value of k

(b) K-means clustering performed with k=10

Figure 3: K-means clustering on positional data on properties

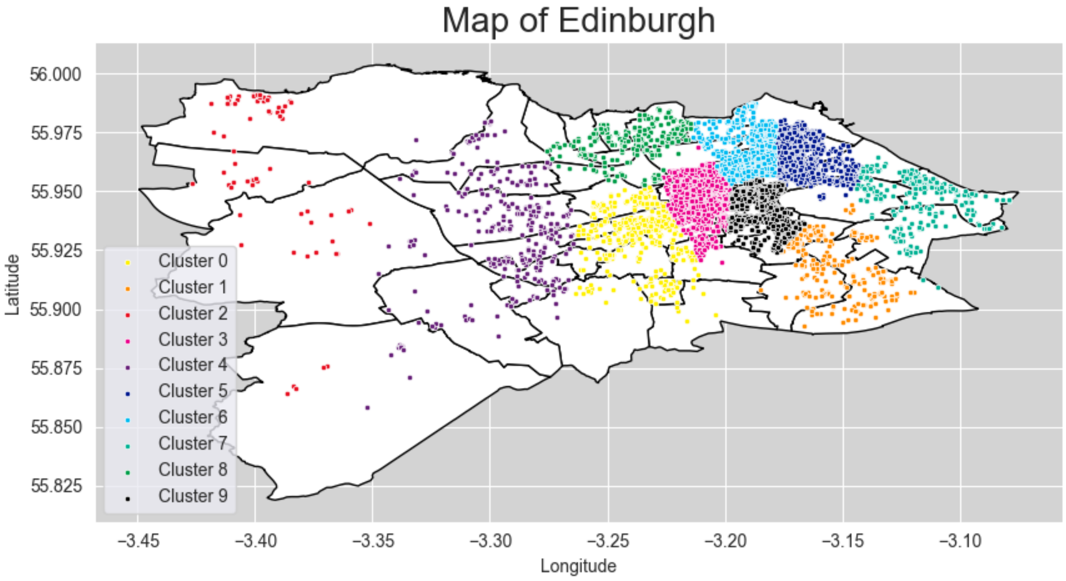
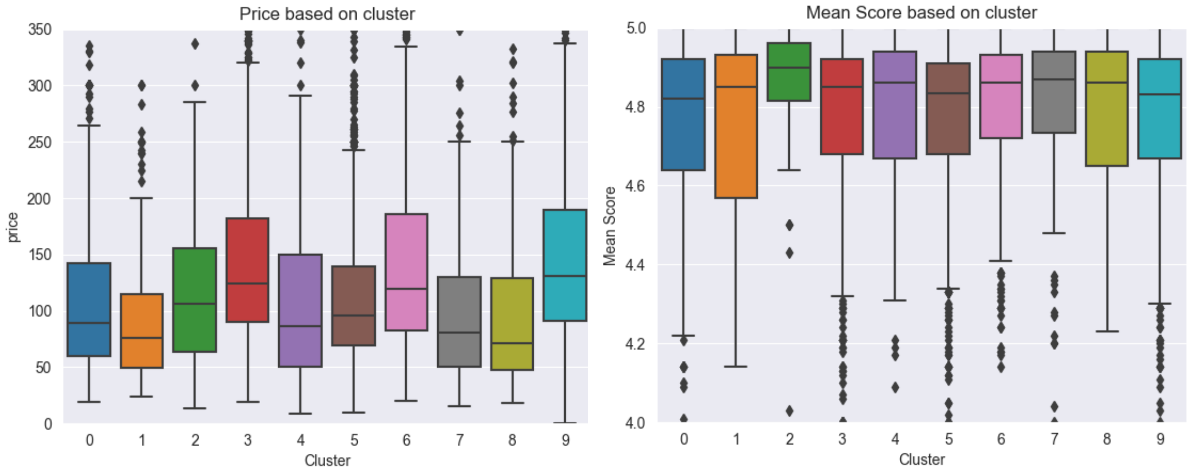


Figure 4 Properties classified by cluster on a map

Cluster (Cl) makeup in total listings (percentage)									
Cl 0	Cl 1	Cl 2	Cl 3	Cl 4	Cl 5	Cl 6	Cl 7	Cl 8	Cl 9
7%	3%	1%	26%	4%	17%	16%	4%	3%	18%

Table 1

Evaluation of own work: strengths and limitations A strength of this report is the cluster analysis performed on location and subsequent visualization. Instead of just performing cluster analysis on a graph with coordinates, a shapefile of the Map of Edinburgh is used to provide a more meaningful representation on how clusters are defined. Another strength from this report was the statistical analyses performed and significance of it's conclusion. A statistically significant value was found in discerning price and location on Airbnb's which reveals an insight into Airbnb's in Edinburgh. However, this reports findings are limited through the limitations of analyses performed. For instance, the Mann-Whitney U test works by ranking observations from both datasets. Therefore, if observed prices are the same, then our calculated p-value may not hold as much significance. Additionally, since 77% of the listings existed close to the city center, the sample size for R is small and could therefore not be an accurate representation of similar listings.



(a) Boxplot of price based on cluster (b) Boxplot of Mean Score based on cluster

Figure 5: Price and Mean Score based on Clusters

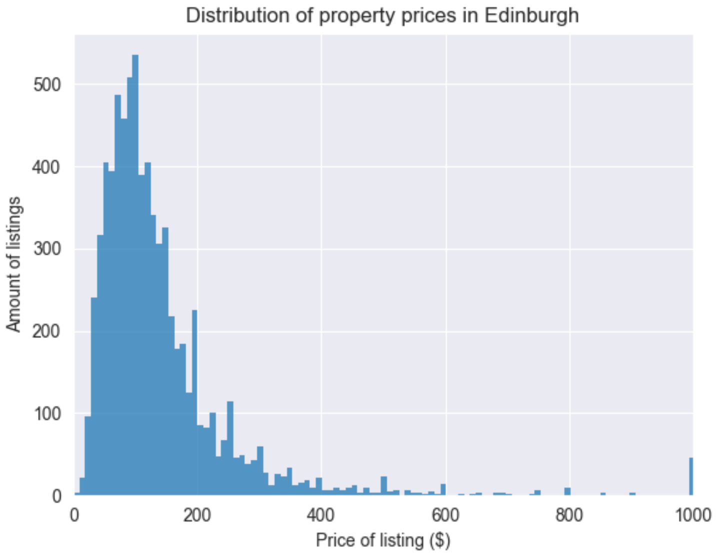


Figure 6: Distribution on Price

Comparison with any other related work A study looking at the determinants of Airbnb prices in European cities concluded that Airbnb prices are ‘spatially dependent’, which is consistent with our findings here.[6]. Furthermore, our findings also coincide with literature examined before conducting the study further supporting the idea between price and distance. [10]

Improvements and Extensions To improve, some variables which were removed for data wrangling could have been significant contributors to variation in the data set. This could then act as a further means of exploration. Moreover, since clusters which were part of P and R were arbitrarily chosen, this also has an impact on our result. A more methodical way, such as distance to a key landmark, could’ve been used to determine which clusters were more ‘central’. Then, a linear regression model could’ve been created to further verify our findings in the conclusion. Furthermore, only one statistical analyses had been performed. Another statistical analysis method could have been performed to reveal further insight into the data. Furthermore, the power of the model predicting Mean Score is weak. A much stronger model could be made with different modelling techniques such as quadratic regression. Alternatively, a different measurement of ‘popularity’ can be used instead which could reveal a different insight into the data set.

References

[1] Airbnb. *Get the data*. last downloaded 7 March 2023. 2023.

[2] Airbnb. *Terms of Service*. Lat accessed 02 April 2023. 2023. URL: `https://www.airbnb.co.uk/help/article/2908`.

[3] Inside Airbnb. *Data Policies*. Last accessed 02 April 2023. 2023. URL: `http://insideairbnb.com/data-policies`.

[4] Inside Airbnb. *How is Airbnb really being used in and affecting the neighbourhoods of your city?* Last accessed 2 April 2023. 2023. URL: `http://insideairbnb.com`.

[5] Inside Airbnb. *Inside Airbnb: Get the Data*. Last downloaded 7 March 2023. 2023. URL: `http://insideairbnb.com/get-the-data`.

[6] Kristóf Gyódi and Łukasz Nawaro. "Determinants of Airbnb prices in European cities: A spatial econometrics approach". In: *Tourism Management* 86 (2021), p. 104319. ISSN: 0261-5177. DOI: `https://doi.org/10.1016/j.tourman.2021.104319`. URL: `https://www.sciencedirect.com/science/article/pii/S0261517721000388`.

[7] EHL Insights. *What motivates travelers to book Airbnb?* Last accessed 28 March 2023. n/a. URL: `https://hospitalityinsights.ehl.edu/travelers-airbnb-study`.

[8] Junfeng Jiao and Shunhua Bai. "An empirical analysis of Airbnb listings in forty American cities". In: *Cities* 99 (2020), p. 102618. ISSN: 0264-2751. DOI: `https://doi.org/10.1016/j.cities.2020.102618`. URL: `https://www.sciencedirect.com/science/article/pii/S0264275119306559`.

[9] Luqi Lu and Saloomeh Tabari. "Impact of Airbnb on customers' behavior in the UK hotel industry". In: *Tourism Analysis* 24.1 (2019), pp. 13–26.

[10] Zhihua Zhang et al. "Key factors affecting the price of Airbnb listings: A geographically weighted approach". In: *Sustainability* 9.9 (2017), p. 1635.