x18110096-CA2-CohortA

by Zainul Abedin Khatik

Submission date: 03-Apr-2019 11:14PM (UTC+0100)

Submission ID: 1105409724

File name: 17305_Khatik_Zainul_Abedin_x18110096-CA2-CohortA_41884_2101207937.pdf (723.21K)

Word count: 2749

Character count: 13896

Finding the hidden patterns of Dublin Airbnb Data using Regression Analysis.

Khatik Zainul Abedin
X18110096
CRM CA2
Master in Data Analytics – Cohort A

Abstract--**Airbnb is an online marketplace which lets people rent out their spare room or their house to guests. Since its inception in 2008 it has been rapid growing. It has been growing alternative to traditional hotel system. Airbnb collects a lot of data about the host, host properties and the reviews given by customer on the host and host property. This study focuses on Dublin, the data is available at the Inside Airbnb official website, we cleaned and transformed the data and conducted a statistical regression analysis like multiple linear regression and logistical regression to find that whether there is an effect of one variable on the other and if that is true then analyse to what extent they are affecting. This study intends to focus on three business queries. The purpose of this study is to investigate to find a) What factors are to be considered to become Superhost. b) What impact does number of reviews has on the host response time and host response rate c) Whether there is a relationship between since the host has been rented out the space and number of reviews.**

We formulated a hypothesis and proved statistically that there exists a relation between the chosen variables selected for this study.

Keywords- **logistic regression, linear regression, superhost, host resposnse time, review score.**

I) Introduction

In this era of technology, most of the companies are leveraging the availability of the data and using to find hidden patterns, relations between people's interest. By using this, companies can target a specific audience in the business environment to gain more profits from the customer needs and demands[1]. Many online marketplace companies leverage this to serve their business and companies like Airbnb built their business model on it. Because of its flexibility and easy profits, the competitions between the private hosts who are listing their properties in Airbnb website are increasing exponentially and so is the data being generated. Due to the higher amount of competitions the generated data needs to be analysed and categorised so that the company, as well as the private hosts, will also get profits, and this helps the company sustained in long run[2]. In this paper, we will concentrate on the hidden patterns and the relations between the categories of the listings with the factors that affect them by using different types of regression analysis.

II) Literature review

Airbnb is an online marketplace where it connects customers and the service providers across the world over 190 countries. The traditional hotel industry is affected by the growth of private home listings in Airbnb[3] and it also affects the housing market tremendously in Dublin[4]. At this point of time the Airbnb has introduced a study on their hosts and started to categorise them and give them Superhost badges based on their reviews and ratings by the customers[5]. This study is also done by many of the researcher and due to the high demand from external users the Airbnb is made their data publicly available regional wise in its official company's website. Dmytro Lakubovskyv has mentioned in the blog that the positive and the negative comments posted by the customers in the Airbnb website and its effects on the hosts reputation and difference between a super and a normal host but he does not provided any statistical evidence that the results which he give are accurate and he also said that factors effecting the price in different cities of USA but not in Ireland [6]

Syuen Loh has also explained in his statistical company website that the factors that are affecting the price in different cities where the company is operating and also at what particulate time of a year, most popular times the rentals is mentioned [5] but he didn't given any particular numeric format to specify the effect of a particular variable on other variable and even his statistical evidence is also missing in his study.

Many of the researchers also predicted the future pricing, availability dates using this Inside Airbnb data which is publicly available and also done analysis like reviews on searchable listings, locations and the host listings metadata[7] but in this stage we are not in need of predicting the variable so first relation will not be considered as an insight to our analysis and the next two relations are done on the variables mostly related to the locations and even though the place of operation is different but the methodology can be applicable on any type of data.

III) Methodology and implementation

We conducted an analysis on the data of Airbnb Dublin where it includes various types of information about the hosts of Airbnb. The process of methodology plays an important role playing as a root map for the researcher to perform the analysis. This analysis is majorly focused on super hosts, since when the host has been with Airbnb and what is his or her response time any new customer enquiring about the place. To find this Regression Analysis is very useful, important and easy concept in finding the hidden relations with in the data between the variables [8]. In this paper, two types of regression analysis techniques are used they are. Multiple linear regression and binary logistic regression. These regression analysis technique has a huge amount of applications in different fields [8].

This research has a keen control over the dependent and the independent variables and this control lead to the higher amount of accuracy in building the model. Before going to the analysis part, we need to prepare the data that suits to the SPSS environment because SPSS cannot calculate the data which is string format. It is mandatory that to represent the fields as numeric, and then we can conduct the research, for that we take help of Microsoft Excel because SPSS cannot do that alone.

IV) Data preparation

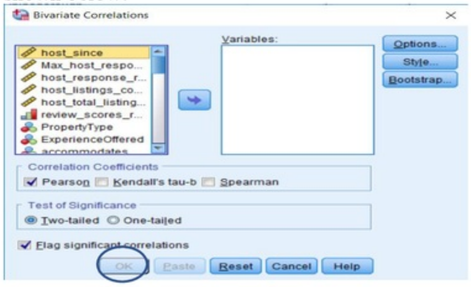
Firstly, we need to code the data in SPSS using values in the variable view of SPSS in such a way that the unique categorical variables are represented as numerical like 1,2,3 etc. once the process is done in the SPSS copy the numbers into excel and then copy the same and again past that in SPSS, so that we can recode the variables in the reverse order as previous. This process will complete once we rename the new variable and delete the old variable in the data. This process is done for every categorical variable which needs to be analysed and most importantly as a final step we need to convert the measure value as nominal in the variable view of spss so that the software can understand that the variable is a categorical variable.

Once the required fields converted to nominal numerical fields then they are ready for any analysis in SPSS

A. *Correlation*

The process of finding correlation between the numeric variables in SPSS will be in a user-friendly manner. In the Correlation process we will consider the variables which are having higher correlations. The process will be like Analyse->correlation->binominal Correlation.

As we go through this process we will get a popup as shown below.



Correlation Coefficients
☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance
● Two-tailed ○ One-tailed

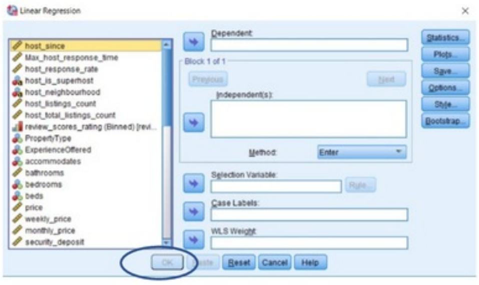
☑ Flag significant correlations

B. *Linear Regression*

In any regression analysis the dependent and independent variables are to be decided before going to the analysis. There are many tools and environments that support these regression models and some of them are IBM-SPSS, R, Python etc but in this analysis the IBM statistical tool called SPSS is used because it is very reliable, easy to use and understand to the user very quickly [9]. Linear Regression can find the relation between one variable and the other variable of any type but it needs to be in a numerical format when it comes to SPSS.

The steps to run the linear regression in SPSS is Analyse-> regression->Linear regression.

A popup will appear as shown below on the screen once you follow the above steps



Method: Enter

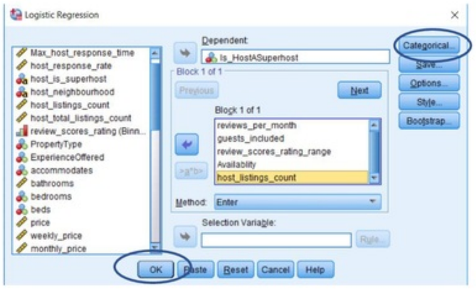
Drag the dependent and the independent variables from the left-hand side to the right- hand side to their respective fields and then click on continue. Answers will be getting on a suppurate window and which are explained in the next section.

C. *Logistic Regression*

The process for the logistic regression is also very much similar to the linear regression but the difference is that we need to select the dependent variable as a binomial variable and that should not consist of more than two variables like yes or no, true or false etc. the process will be like

Analyse->regression-> binary logistic regression

The popup will be as shown



Dependent: Is_HostASuperhost

Block 1 of 1
reviews_per_month
guests_included
review_scores_rating_range
Availablity
host_listings_count

Method: Enter

The dependent and the independent variables should be dragged to their respective fields and declare the categorical variables in the categorical section before running the algorithm. The circle parts will serve as a reference to the readers.

Note: The process of running the algorithm will not change across the dataset.

V) Dataset description:

In the Field of online marketplace and hospitality service which is operated via a website and a mobile application. Customers can get a private homestay, lodging, tourism experience for the people who travel to a new place. The amount of data in that market place is increasing exponentially because of the competition from the hosts.

The data which is generated needs to be analysed because it helps the hosts as well as the company to increase their profits and implement their strategy's Which helps the company to sustain in a long run[10]. The data set has nearly 9521 rows with 102 columns which gives a complete information about the host and their listings with their customer individuals service rating and overall rating. These fields are very much helpful in finding the relations between the variables and improve the quality and customer experience of the host listings in Airbnb. The link to get the data is given below for the reference

Link: http://insideairbnb.com/get-the-data.html

The dataset is available in comma suppurated value format. For this study, we only consider the columns which are useful for the analysis and the remaining are dropped because of huge processing time.

For data transformation of the data we remove the unnecessary columns in Excel and remove null values, remove dollar sign from the price, converted dd/mm/yyyy format of the since_host column to yyyy in R..

VI) Business Query:

1)What factors are to be considered to become a Super host.

2) What impact does number of reviews has on the host response time and host response rate?

3) Whether there is a relationship between since the host has been rented out the space and number of reviews.

VII)Hypothesis

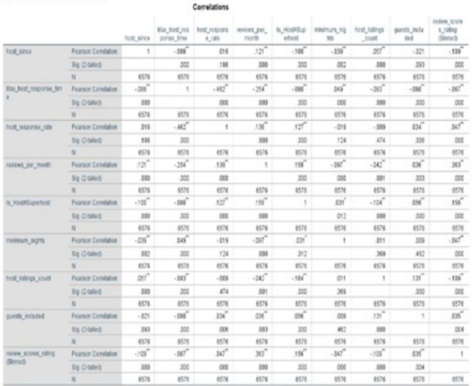
H_0: there is no relation between the variables

H_1: there is a relation between the variables

VIII) Interpretation of Results

As it is explained the correlation process behind the results of SPSS in previous section, in this section the results will be concentrating more.

Correlation:



Correlations

		host_since	Max_host_response_time	host_response_rate	reviews_per_month	Is_HostASuperhost	minimum_nights	host_listings_count	guests_included	review_scores_rating (Binned)
host_since	Pearson Correlation	1	-.088**	.016	.121**	-.100**	-.039**	.057**	-.021	-.109**
	Sig. (2-tailed)		.000	.186	.000	.000	.002	.000	.093	.000
	N	6576	6576	6576	6576	6576	6576	6576	6576	6576
Max_host_response_time	Pearson Correlation	-.088**	1	-.462**	-.254**	-.088**	.049**	-.083**	-.088**	-.067**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000	.000
	N	6576	6576	6576	6576	6576	6576	6576	6576	6576
host_response_rate	Pearson Correlation	.016	-.462**	1	.138**	.127**	-.019	-.009	.034**	.047**
	Sig. (2-tailed)	.186	.000		.000	.000	.124	.474	.006	.000
	N	6576	6576	6576	6576	6576	6576	6576	6576	6576
reviews_per_month	Pearson Correlation	.121**	-.254**	.138**	1	.158**	-.067**	-.042**	.036**	.363**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.001	.003	.000
	N	6576	6576	6576	6576	6576	6576	6576	6576	6576
Is_HostASuperhost	Pearson Correlation	-.100**	-.088**	.127**	.158**	1	.031*	-.104**	.056**	.156**
	Sig. (2-tailed)	.000	.000	.000	.000		.012	.000	.000	.000
	N	6576	6576	6576	6576	6576	6576	6576	6576	6576
minimum_nights	Pearson Correlation	-.039**	.049**	-.019	-.067**	.031*	1	.011	.009	-.047**
	Sig. (2-tailed)	.002	.000	.124	.000	.012		.369	.462	.000
	N	6576	6576	6576	6576	6576	6576	6576	6576	6576
host_listings_count	Pearson Correlation	.057**	-.083**	-.009	-.042**	-.104**	.011	1	.131**	-.109**
	Sig. (2-tailed)	.000	.000	.474	.001	.000	.369		.000	.000
	N	6576	6576	6576	6576	6576	6576	6576	6576	6576
guests_included	Pearson Correlation	-.021	-.088**	.034**	.036**	.056**	.009	.131**	1	.035**
	Sig. (2-tailed)	.093	.000	.006	.003	.000	.462	.000		.004
	N	6576	6576	6576	6576	6576	6576	6576	6576	6576
review_scores_rating (Binned)	Pearson Correlation	-.109**	-.067**	.047**	.363**	.156**	-.047**	-.109**	.035**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.004	
	N	6576	6576	6576	6576	6576	6576	6576	6576	6576

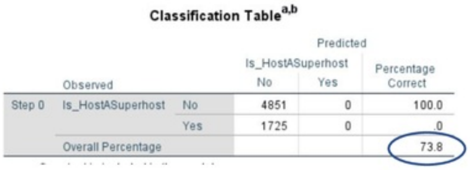
There is week relation exists between most of the variables in this table. This evident that there are still other variables which impact more on the dependent variables.

When we consider independent variables like Host since, maximum host response time and is host a super host they have theoretically low correlation but virtually they effect the model to considerable amount. There correlations range from 10% to 25% which needs to be considered.

Business Query 1:

For solving the first business query which is what factors are to be considered to become a Super host. Binary logistic regression is used to solve this query and the process to run the analysis is explained in the methodology section. Now the answers will be explained in this section.

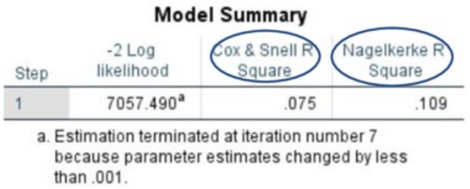
The factors are to be considered to become a super host according to this data as shown below with their percentage and statistical evidences.



Classification Table^{a,b}

	Observed		Predicted		
			Is_HostASuperhost		Percentage Correct
			No	Yes	
Step 0	Is_HostASuperhost	No	4851	0	100.0
		Yes	1725	0	.0
	Overall Percentage				73.8

The clasification table tells us that the model build is 74% accurate and the model built on the data provided without any changes. The accuracy of the model is very much important to know that the values we got are accurate and got correct results.

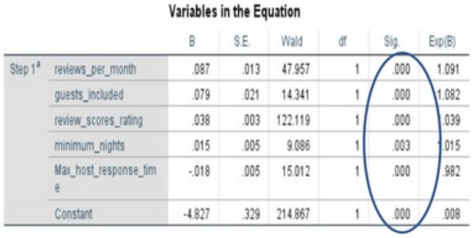


Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	7057.490^{a}	.075	.109

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

The model summary table is considered in such a way that it provides a statistical evidence that the model built on the right variables. the cox and Snell r square and Nagelkerke are called the pseudo square values and if they are more than 5% or 0.05 then the model is statically significant.



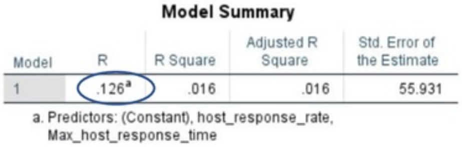
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	reviews_per_month	.087	.013	47.957	1	.000	1.091
	guests_included	.079	.021	14.341	1	.000	1.082
	review_scores_rating	.038	.003	122.119	1	.000	.039
	minimum_nights	.015	.005	9.086	1	.003	.015
	Max_host_response_time	-.018	.005	15.012	1	.000	.982
	Constant	-4.827	.329	214.867	1	.000	.008

The above table shows the percentage of effect of independent variables effect on the dependent variable individually. When we come to the reviews per month, guests included, review score rating should be high and the minimum nights, max host response time should be less for a host to be supper host in Airbnb marketplace

Business query:2

To answer the second business query, multiple linear regression is used. The process to run the algorithm is explains in the previous section and the answers are shown below with the statistical evidences.

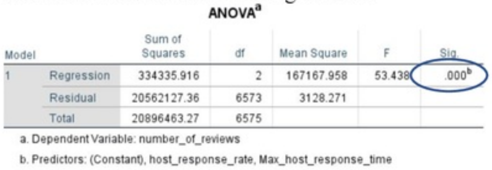


Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.126[a]	.016	.016	55.931

a. Predictors: (Constant), host_response_rate, Max_host_response_time

The model summary will explain about the overall effect of independent variables on a dependent variable. Here approximately 13% of effect on number of reviews due to host response time and host response rate and this can be known using R value.



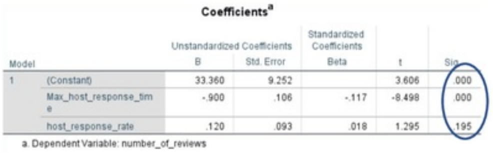
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	334335.916	2	167167.958	53.438	.000[b]
	Residual	20562127.36	6573	3128.271		
	Total	20896463.27	6575			

a. Dependent Variable: number_of_reviews

b. Predictors: (Constant), host_response_rate, Max_host_response_time

The Anova table provides the proof that the model is statistically significant. If the sig value is less than 0.05 then we can say that the model is significant.



Coefficients[a]

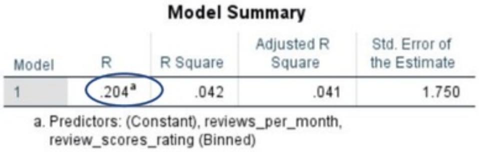
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	33.360	9.252		3.606	.000
	Max_host_response_time	-.900	.106	-.117	-8.498	.000
	host_response_rate	.120	.093	.018	1.295	.195

a. Dependent Variable: number_of_reviews

The above coefficients table says that the individual effect of each independent variable on dependent variable and if the reviews per month needs to be more then the response time of the host should be low and response rate should be high that for one unit increase of host response time then there is a probability that the monthly reviews will increase by 90%.and the response rate will increase to 13%.

Business query:3

The third business query also uses the same algorithm as second query because it has a dependent variable as continuous variable.

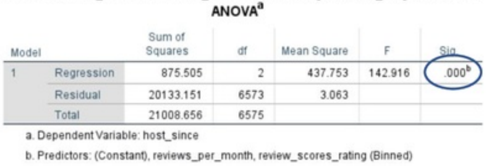


Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.204[a]	.042	.041	1.750

a. Predictors: (Constant), reviews_per_month, review_scores_rating (Binned)

The combined effect of reviews per month and reviews score rating is affecting the seniority rating by 20.5%



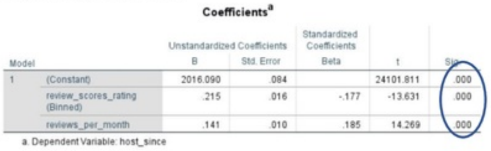
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	875.505	2	437.753	142.916	.000[b]
	Residual	20133.151	6573	3.063		
	Total	21008.656	6575			

a. Dependent Variable: host_since

b. Predictors: (Constant), reviews_per_month, review_scores_rating (Binned)

As explained that the sig value is less than 0.05 then we can say that the model is significant and the values which we got are also having a significant effect on the dependent variables.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2016.090	.084		24101.811	.000
	review_scores_rating (Binned)	.215	.016	-.177	-13.631	.000
	reviews_per_month	.141	.010	.185	14.269	.000

a. Dependent Variable: host_since

The coefficients table says that the effect of reviews will be high and the reviews per month will also high for a senior host who lists their properties frequently in Airbnb marketplace.

IX) Conclusion

There is a significant effect of independent variables on the dependent variables so that there is no statistical evidence to prove that the H_1 is false and finally, H_0 is rejected for all the three business queries.

We found the hidden affects of dependents variables and the independent variables which are useful in the business environment to achieve more profits for the hosts and these also helps the sustainability of the company in a long run.

X) Bibliography

[1] L. Gallagher, "HOW AIRBNB FOUND A MISSION—AND A BRAND. (cover story)," *Fortune*, vol. 175, no. 1, pp. 56–62, 1/1/2017 2017.

[2] "One Size Does Not Fit All: Predicting Product Returns in E-Commerce Platforms," *2018 IEEEACM Int. Conf. Adv. Soc. Netw. Anal. Min. ASONAM Adv. Soc. Netw. Anal. Min. ASONAM 2018 IEEEACM Int. Conf. On*, p. 926, 2018.

[3] J. B. Garau-Vadell, D. Gutiérrez-Taño, and R. Díaz-Armas, "Residents' Support for P2P Accommodation in Mass Tourism Destinations," *J. Travel Res.*, vol. 58, no. 4, pp. 549–565, Apr. 2019.

[4] V. Lima, "Towards an understanding of the regional impact of Airbnb in Ireland," *Reg. Stud. Reg. Sci.*, vol. 6, no. 1, pp. 78–91, Jan. 2019.

[5] S. Loh, "Using data to understand the market for AirBnB rentals in Seattle," *Medium*, 10-Sep-2018. .

[6] D. Iakubovskyi, "Digging into Airbnb data: reviews sentiments, superhosts, and prices prediction (part1)," *Towards Data Science*, 01-Oct-2018. [Online]. Available: https://towardsdatascience.com/digging-into-airbnb-data-reviews-sentiments-superhosts-and-prices-prediction-part1-6c80ccb26c6a. [Accessed: 03-Apr-2019].

[7] H. Rechavia, "Statistical Overview of Barcelona's Airbnb Market," *Towards Data Science*, 14-Aug-2018. [Online]. Available: https://towardsdatascience.com/statistical-overview-of-barcelonas-airbnb-market-83dc7d6be648. [Accessed: 03-Apr-2019].

[8] S. Chatterjee and A. S. Hadi, *Regression Analysis by Example*. John Wiley & Sons, 2015.

[9] S. J. Bae, H. Lee, E.-K. Suh, and K.-S. Suh, "Shared experience in pretrip and experience sharing in posttrip: A survey of Airbnb users," *Inf. Manage.*, Dec. 2016.

[10] B. Mihalčová and M. Pružinský, "Category Management – Project Implementation in E-Shop," *Procedia Econ. Finance*, vol. 23, pp. 267–275, Jan. 2015.

x18110096-CA2-CohortA

GRADEMARK REPORT

FINAL GRADE

/100

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5