Statistical Analysis of real-time usage of Dublinbikes

Rahul Dhande
18182852
MSc in Cloud Computing

18th June 2020

Abstract

Nowadays, Increase demand and usage of renting a bike is growing day by day in cities. It popular and productive every day helps to reduce the usage of car journeys and improve green transportation. Besides, One major challenge is some bike stations(more than 10) is empty in morning commuting period. This statistical report aims to provide a detailed analysis of real-time Dublinbikes based on usage data. Moreover, It includes a data structure of bike stations, available bike stands as well as available bikes. However, The report shows analysis using IBM SPSS tool with different tests such as Independent sample t,Mann-Whitney U and Chi-Square for independence. This analysis helps to improve usage of dublinbikes as per commuter needs such as reduce waiting time for every commuter as well as business revenue in future.

1 Introduction

In many cities, public bike renting popular every day. It saves cost, environment from pollution and significant benefit it helps to stay healthy for users by physical exercise. One of the cost-effective public bike renting is Dublinbikes, bike-sharing scheme founded in 2009. Moreover, It has over 64,000 subscribers and 16,3 million journeys across Dublin city increasing every year. Shyram Ravichandran ([1],2019)

However, Some researchers previously show analysis on Dublin bikes.([1],2019) describes statistical clustering analysis of dublinbikes on the basis of busiest and the quietest stations, including check-in and check-out date.The significant purpose of this analysis is to Improve understanding of user behaviours leading to insights that facilitate planning, such as rebalancing bikes across stations on demand using K-nearest algorithm aims to divide usage data of bikes into 4 different clusters. Overall the analysis results show a major difference in bicycles usage Stations on weekends and weekdays, during business hours and after work.As compared to,Enda Murphy and Joe Usher ([2],2015) presented a unique analysis of Irish bicycle-sharing experience based on the human age group. Overall analysis results show the issues of a male and female group of this bike scheme.

The Wu ([3],2019) explored and analysed oral squamous cell carcinoma data using SPSS as well as calculate measure the transfection efficiency and aggression results of mir-150 using Independent Sample t-test. Besides, Baran([4],2019) performed a statistical analysis of lymphoblastic leukaemia data using SPSS with a Mann-Whitney U test to determine significance for the same. Also, John Nutor ([5],2019) analysed HIV data

between males and females in SPSS and performed a chi-square test of independence to test the relation between the HIV status and the predictor variables.

In this statistical report, I plan to use Dublinbikes dataset, i.e. dublinbikes.sav and show detailed analysis of available bike stations, bikes and bike status. For the same, I use IBM SPSS tool.Pallant([6],2016) The analysis includes three different tests.

- Independent Sample t-test: acts like a parametric test aims compares the mean value between two separate or unrelated groups. It contains dependent and Independent variables. To find the significant mean score, we consider available bikes as a grouping variable and available bikes stand as a continuous variable for Independent sample t-test.

- Mann-Whitney U test : referred to as non-parametric tets aims to compares the means value between two Independent groups if the dependent variable is either ordinal or continuous. In this report, we consider Station ID as an Independent variable, and available bikes as a dependent variable.

- Chi-Square test for Independence: referred to as a Chi-Square test of association aims to find statistical relationships between correlated variables such as Independent variable or related variables. In this report, we consider two categorical variables, such as available bikes and Address of bikes stations.

1.1 Data Collection

The Data collected from Dublin city Council Republic of Ireland.Dublinbikes ([7],2020)The data is collected in CSV format and import into SPSS that becomes sav file.The Data contains information from first week of January 2020 until the middle of February 2020 show in Table 1.

Column	Description
1	Station ID
2	Time
3	Name
4	Bike Stands
5	Available Bike Stands
6	Available Bikes
7	Status
8	Address
9	Longitude
10	Latitude

Table 1: Structure of Dublinbikes Data

2 Independent Sample t Test

2.1 Research Question

What is an exact significant difference in the mean value of dependent variable Available bikes for the independent variable, i.e. Available bikes stands per day?

2.1.1 Hypothesis

The Hypothesis of Dublinbikes data of morning period is following.
H0: Does available bikes more than Available bike stands
H1: Available Bike stands not more than available bikes

2.2 Statistical results analysis

T-Test

Group Statistics

	AVAILABLEBIKES	N	Mean	Std. Deviation	Std. Error Mean
AVAILABLEBIKESTANDS	7	79573	24.11	7.987	.028
	10	78230	21.55	7.893	.028

Figure 1:

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
AVAILABLEBIKESTANDS	Equal variances assumed	.091	.763	64.146	157801	.000	2.565	.040	2.486	2.643
	Equal variances not assumed			64.153	157796.814	.000	2.565	.040	2.486	2.643

Figure 2:

The above figure1,Evidently show that **Group Statistics** provides difference on Mean and Standard deviation of Available bikes stands (Independent Variable) per day as per Available bikes during the morning commuting period.Additionally, the Standard deviation for seven available bikes is 7.987 and for ten is 7.893

Simultaneously, The figure 2 shows **Independent Samples Test** including assumptions and results of Leven's test and t-test. Firstly, The Equality of variances shows significance value p= 0.763, which is greater than 0.05, so the value of t is 64.146. Secondly, the t-test for equality of means has Significance value (2-failed) 0 which is overall show the difference between the mean value of dependent variable Available bikes for the independent variable, i.e. available bikes stands per day.

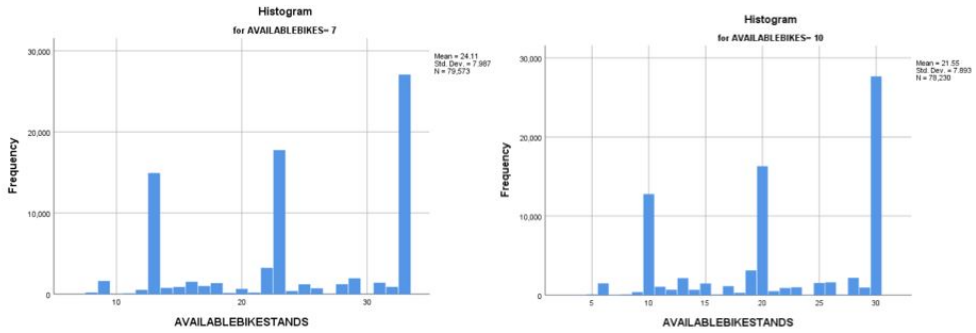

Figure 3:

The figure 3 shows the histogram for seven and ten available bikes in different available bike stands proves the mean value difference between dependent and independent variable.

In addition, the calculation for the effect size for the independence sample t-test using Eta squared formula, which gives exact value. For this, we consider t=64.146 ,N1=79573,N2=78230

Eta-Squared $= t^2/t^2 + (N1 + N1 - 2) = (64.146)^2/(64.146)^2 + (79573 + 78230 - 2) =$ **0.025**

2.3 Benefits

- Statistical significant mean difference will helps to reduce the waiting time for commuter.
- It will help to improve business revenue of Dublin Bikes.

2.4 Conclusion

Overall,According to significant value(0.05), the Null hypothesis is impossible as compared to alternative hypothesis accepted as well as the Leven's test for variance clearly shows 0.025 would be 2.5% of available bikes stands of Available bikes(grouping variable).

3 Mann-Whitney U test

3.1 Research Question

Does the available bikes more than Bikes Station ID per day during morning commuting period?

3.1.1 Hypothesis

The Hypothesis of Dublinbikes dataset is following.

H0: No correlation between the Station ID and Available bikes per day

H1: Correlation between Available bikes per day and Station ID

Variables for test:

- Station ID (Independent-Categorical)
- Available bikes (Dependent-Continuous)

3.2 Results of analysis

NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles 25th	50th (Median)	75th
STATIONID	499	13.96	29.824	2	101	2.00	3.00	4.00
AVAILABLEBIKES	498	4.00	2.998	0	10	2.00	3.00	7.00

Figure 4:

The figure 4 evidently shows descriptive statistics of Comparison for Mean value, Standard Deviation between Station ID and Available Bikes.The Mean value of Station ID 13.96 and for Available bikes is 4.00 per day during morning commuting period.

Mann-Whitney Test

Ranks

	AVAILABLEBIKES	N	Mean Rank	Sum of Ranks
STATIONID	7	74	67.97	5030.00
	10	41	40.00	1640.00
	Total	115		

Figure 5:

Subsequently, The figure 5 depicts "Mean Rank" for the group of available bikes seven (67.97) is greater than Ten bikes (40.00) as per Station ID per day. Moreover, The Sum rank for a group of seven bikes is also greater than a group of ten bikes.

Additionally, the below figure 6 referred to as 'Test Statistics' aims to provides two values one is Z-value and second is Asymp Sig. (2-tailed) value. The Z-value is -5.829 and Significance value p=0. However, the probability value is less than 0.05, which concludes group variable (Available bikes) is statistically different as per Station ID per day during morning commuting period.

The figure 7 shows the two histograms for available bikes as per Station ID proves the mean value and standard deviation for group of seven bikes is also greater than a group of ten bikes. Overall, There is a statistical difference between Independent and dependent variables.

In addition, the calculation for the effect size for the Mann – Whitney U Test using Eta squared formula, which gives exact value. For this, we consider Z=5.289, N= 115 **Eta-Squared** $= Z^2/(N-1) = (5.289)2/(114) =$ **0.2453**

Test Statistics[a]

	STATIONID
Mann-Whitney U	779.000
Wilcoxon W	1640.000
Z	-5.289
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: AVAILABLEBIKES

Figure 6:

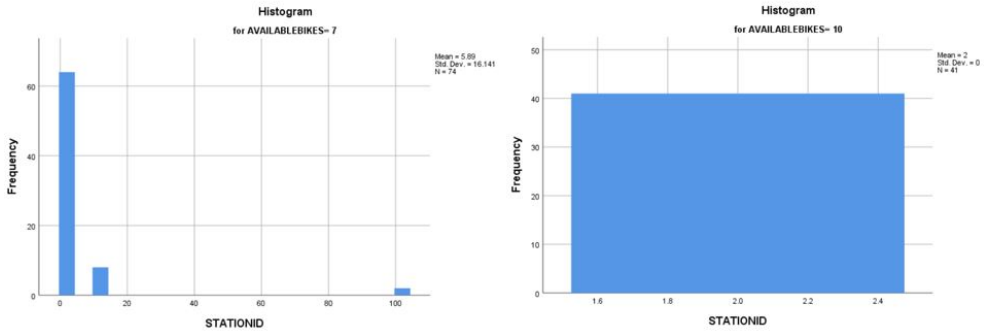

Figure 7:

3.3 Benefits

- Statistical significant mean difference will helps to reduce empty bike stations as per station ID.
- It will help to improve business revenue of Dublin Bikes.

3.4 Conclusion

To Summarise, according to the research question, it is evident that the 2.4% variability of a station ID as per Seven and Ten bikes per day during morning commuting period.Also, the significance value p=0 concludes that Null hypothesis is accepted.

4 Chi-square test for independence

4.1 Research Question

Does the proportion of available bikes more than in Address(Streets)?

4.1.1 Hypothesis

The Hypothesis of Dublinbikes data of morning period is following.
H0: No correlation between the Address(Streets) and Available bikes per day
H1: Correlation between Available bikes per day and Address(Streets)

4.2 Results of analysis

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
ADDRESS * AVAILABLEBIKES	699	100.0%	0	0.0%	699	100.0%

Figure 8:

The above figure 8 shows the Case Processing summary used for statistical analysis. It includes valid cases which are 100%, and Missing cases are 0 for both address and available bikes.

ADDRESS * AVAILABLEBIKES Crosstabulation

Count

		AVAILABLEBIKES											
		0	1	2	3	4	5	6	7	8	9	10	Total
ADDRESS	Blessington Street	17	3	3	0	5	18	15	46	14	6	41	168
	Bolton Street	34	42	15	12	22	5	9	28	0	0	0	167
	Charlemont Street	12	1	8	22	8	0	0	0	0	0	117	168
	Christchurch Place	0	0	0	0	0	0	0	0	0	0	28	28
	Greek Street	18	1	54	69	26	0	0	0	0	0	0	168
Total		81	47	80	103	61	23	24	74	14	6	186	699

Figure 9:

The above figure 9 Crosstabulation Table depicts detailed statistics of available bikes as per on a various Address(Streets). As per Statistics analysis, On a Charlemont street 117% group of Ten bikes available whereas on Blessington street 41% of Ten bikes per day. Overall, The total average of available bikes for a group of ten is 186% and for a group of Five is 23% for each Address (Streets) per day.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	839.902[a]	40	.000
Likelihood Ratio	878.733	40	.000
N of Valid Cases	699		

a. 18 cells (32.7%) have expected count less than 5. The minimum expected count is .24.

Figure 10:

Simultaneously, the figure 10 referred to as Chi-Square test Analysis which evidently shows the Pearson chi-square value X(40)=839.902,N=699 and p=0 which is less than alpha value 0.05. these conclude there is a statistically significant association between available bikes and Address group.

Additionally, The below figure 11 of Effect Size calculation Statistics includes Phi and Cramer's V tests. It clearly shows the Phi tests has the strength of association between available bikes, and Address(Streets) is good enough, whereas Cramer's tests have very low with approximate significance is 0.

Symmetric Measures[c]

		Value	Approximate Significance
Nominal by Nominal	Phi	1.096	.000
	Cramer's V	.548	.000
N of Valid Cases		699	

c. Correlation statistics are available for numeric data only.

Figure 11:

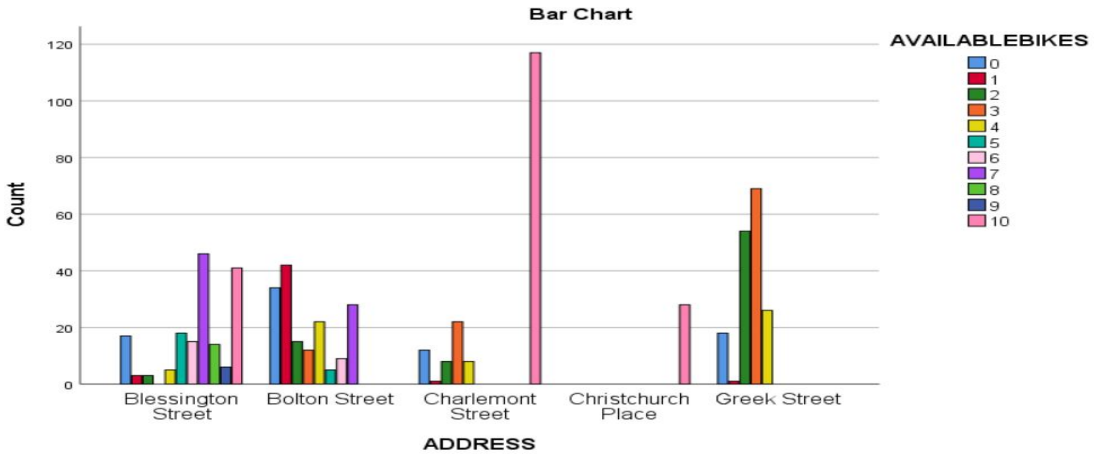

Figure 12:

The above figure 12 generated a clustered bar chart shows highlights the Available bikes groups as per Address(Streets) groups.

4.3 Benefits

- Statistical relationship between two categorical variables will helps to reduce improve availability of bikes as per Address(Streets).
- It will help to improve business revenue of Dublin Bikes.

4.4 Conclusion

Overall, It is evident that Significance value p=0 which is less than alpha value 0.05 which also concludes that Null hypothesis is accepted as compared to no alternative hypothesis accepted.

4.5 Final Conclusion and Discussion

The report aims to focus on providing a reliable solution to commuters during the morning period using Dublinbikes real-time data statistical analysis. For the same, we performed three various hypothesis tests using IBM SPSS. The following table 2 shows final results show Independent sample t-test accept alternative hypothesis whereas Mann-Whitney U and ChiSquare accept the null hypothesis. The statistical results help to improve the availability of bikes and reduce waiting time for commuters.

Tests	Null Hypothesis	Alternative Hypothesis
Independent Sample t test	—	✓
Mann-whitney U test	✓	—
Chi-Square for Independence	✓	—

Table 2: Hypothesis Results

References

[1] T. T. P. Thi, J. Timoney, S. Ravichandran, P. Mooney, and A. C. Winstanley, "Bike renting data analysis: The case of dublin city," *CoRR*, vol. abs/1704.06802, 2017.

[2] E. Murphy and J. Usher, "The role of bicycle-sharing in the city: Analysis of the irish experience," *International Journal of Sustainable Transportation*, vol. 9, no. 2, pp. 116–125, 2015.

[3] C. Wu, M. Yang, and H. Chen, "Inhibition effect of mir-150 on the progression of oral squamous cell carcinoma by data analysis model based on independent sample t-test," *Saudi Journal of Biological Sciences*, vol. 27, 11 2019.

[4] G. Baran, H. Arda Sürücü, and V. Üzel, "Resilience, life satisfaction, care burden and social support of mothers with a child with acute lymphoblastic leukaemia: a comparative study: Resilience, life satisfaction and care burden in mothers," *Scandinavian Journal of Caring Sciences*, vol. 34, 06 2019.

[5] J. John Nutor, P. A. Duodu, P. Agbadi, H. O. Duah, K. E. Oladimeji, and K. W. Gondwe, "Predictors of high hiv+ prevalence in mozambique: A complex samples logistic regression modeling and spatial mapping approaches," *PLOS ONE*, vol. 15, pp. 1–21, 06 2020.

[6] J. Pallant, *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using SPSS for Windows Version 15*. USA: Open University Press, 3rd ed., 2007.

[7] "Dublinbikes real time api, near real time api and historical data."
URL: https://data.gov.ie/dataset/dublinbikes-api.