Analysis of the Distribution of Bike Rentals during Summer and Spring

Introduction

In recent years, bike-sharing schemes have received significant attention due to their ability to tackle a wide range of problems within cities [1]. One of the benefits these schemes offer is providing low-cost transportation to households. Due to the sharp increase in the cost of living in the UK [2], public transport has become even less affordable [3]. According to a report conducted by the CCC (Clean Cities Campaign), Birmingham, London and Manchester ranked as cities with the most expensive transport among 36 European cities [3]. The report indicated that people who live in these UK cities spent almost 10 percent of their monthly budget on public transportation [3]. Additionally, sharing a bike has enormous environmental benefits since it enables sustainable transportation [4]. It allows individuals to travel without polluting the air, compared to fossil-fuel based transportation. On top of the economic and environmental benefits[5], bike sharing also reduces transport congestion[6]. Given all these benefits, bike sharing schemes should be encouraged throughout the country.

Although the bicycle-sharing system has evolved significantly since the 1960s [7], it needs more development to secure more users. Therefore, in this report, data will be analysed to understand the demand of bike usage. By anticipating demands, suitable insights for better strategies can be discovered. Allocating bikes can be optimised considering the imbalances of the demand, making it easier for riders to hire and return the bikes.

Discussion

In this report, the data is pre-processed by visualising it to acquire initial insights. Then, the data is filtered and more concentration is put closely on the quantity and time factors to define evening peak commuting hours, which is the keen interest of the client.

The first part of the statistical analysis is two “goodness-of-fit tests”, which are the Shapiro-Wilk Test (S-W test) and Anderson-Darling Test (A-D test). Although chi-squared test is widely used for normality test, its non-composite design requires a specific hypothesis which need to reuse the data for both estimating and testing. Moreover, it may lead to binning problem. Therefore, S-W test and A-D test are selected. A-D test is a modification of Kolmogorov-Smirnov test (K-S test) as it takes specified distributions into consideration and places more weights on the tails. It increases the complexity but provides a more precise outcome. These tests are used to check whether the distribution of bike hires follows a normal distribution during evening peak hours in spring and summer. The results of these tests help our clients to precisely adjust the number of spare bikes during peak commuting times. Two “two-sample tests”, Wilcoxon rank sum test and K-S test, are conducted in the second part to see whether the distribution of bike usage varies between spring and summer. If bike hires are equally distributed in both seasons, then the client only needs to prepare one set of bicycle quantity allocation plans, otherwise they need to formulate different plans for different seasons.

Following statistical analysis, detailed evaluation is done and some problem-solving recommendations based on the findings are offered to the client. The constraints placed on the analysis are also considered and clearly stated.

Presentation of the dataset

As a first step to the analysis, the data is pre-rpocessed and organized. To evaluate the demand this report uses the total count of bike users (both registered and causal) as its goal is to evaluate the overall demand of bikes. Furthermore, the data is filtered out such that only working days are considered and holidays and weekends are excluded. This decision came after inspecting the data and realizing that non-working days tend to have a higher variation compared to working days. The next step includes deciding on the time range to be considered when examining the peak evening hours. The Transport for London (TfL) web page suggested that the peak evening hours range from 4 pm to 7 pm [8] so the data is further filter out to only include entries for these times. Next, the data is filtered based on the season such to keep the summer and spring seasons. Lastly, the number of users in the peak hours for each day is aggregated to ensure independence.

After organizing the data, the next step is to observe the distribution of bike rentals within the selected peak hours (4 pm to 7 pm) for both seasons to better understand the demand for bike hires in the evening. As can be seen in Figure 1, the distribution of bike rentals appears normal in spring, and bimodal in summer. One noticeable difference between the summer and spring distribution is that the average demand (the blue dashed line) for bikes in the summer is higher. One assumption for this could be that this difference in demand is a result of the weather. Warmer weather may incentivize more people to use the bike. This trend can be further investigated by plotting a scatter plot of the number of bikes rented against the temperature. This relationship can be observed by looking at Figure 2. This graph indicates that the demand for bikes seems to increase with an increase in temperature.

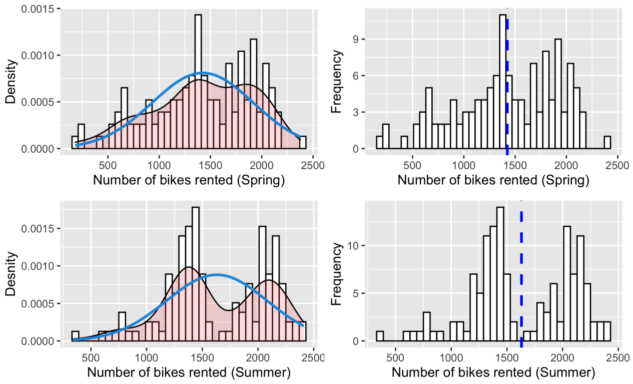

Figure 1: Distribution of bike rentals

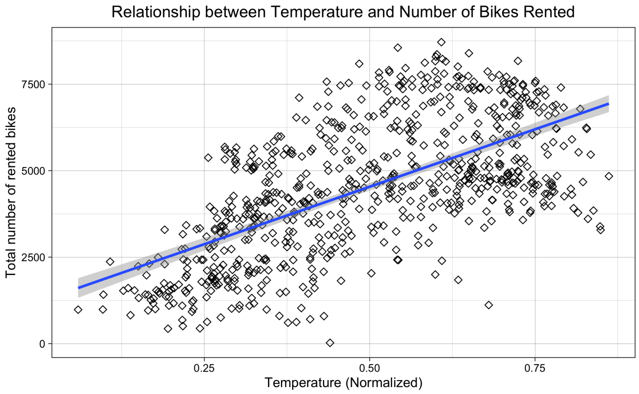

Figure 2: Relationship between number of bikes rented and temperature

Goodness-of-fit Tests

The "goodness-of-fit test" is a method to measure how well a sample data fits a distribution. In this case, it is used to determine whether the total count of bikes hired during evening peak commuting times in spring and summer follows a normal distribution. To examine the normality of the distributions, two commonly used techniques are used in this report, the S-W Test and A-D Test.

In this report, the count of total rental bikes during evening peak time in spring is denoted as X_i, $i = 1, ..., n$, where $n = 128$, and in summer as Y_j, $j = 1, ..., m$, where $m = 131$. For convenience, for both S-W Test and A-D test, the samples of the two seasons are ordered such that $X_1 \leq X_2 \leq ... \leq X_n$ and $Y_1 \leq Y_2 \leq ... \leq Y_m$. The assumptions for the two methods are that X_i and Y_j are both identical independent distributed. The null and alternative hypothesis are as follows:

$$H_0 : \text{The distribution is normally distributed in spring (summer)}$$

$$H_1 : \text{The distribution is not normally distributed in spring (summer)}$$

The p-value is a measure of the evidence against H_0. The smaller the p-value the stronger the evidence against H_0. In this report, the levels of test are all set to 0.05.

For the S-W test, under the null hypothesis, the original test statistic is $W = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, where x_i is an ordered sample to be tested and $\bar{x}$ is the sample mean, the coefficient $\mathbf{a}$ is defined as $\mathbf{a} = (\mathbf{m^T V^{-1} V^{-1} m})^{\mathbf{-1/2}} \mathbf{m^T V^{-1}}$, where $\mathbf{V}$ is the covariance matrix of the normal order statistic with expected vector $\mathbf{m}$. However, for large sample size, it is difficult to compute the inverse of the covariance matrix. Verrill and Johnson provide an approximate W^* test statistic which substitutes the a_i by $b_i = (\mathbf{m^T m})^{-1/2} m_i$.[9]

As for the A-D test, to obtain the test statistic, initially, the order variable X is supposed to be standardised to create a new variable Z in which $Z_i = \frac{X_i - \hat{\mu}}{\hat{\sigma}}$, where $\hat{\mu}$ and $\hat{\sigma}$ are correspondingly the sample mean and sample standard deviation. Under the null hypothesis, the original test statistic A^2 is to measure the distance between the theoretical normal distribution and the samples, which is calculated by

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[ln\phi(Z_i) + ln\phi(1 - Z_{n+1-i})]$$

In this formula, ϕ stands for the standard normal CDF. Since both the mean and variance are not specified under the null, a modified test statistic for unknown mean and variance is used such that $A^{*2} = A^2\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right)$. [10]

The results for the test statistics, critical region and p-value for both tests are given as below.

Method	***Season***	***Test Statistic***	***Critical Region***	***p-value***	***Result***
S-W Test	spring	0.9677	$W^* < W_{128,0.05}{}^{(1)} = 0.9795$	0.003768	Reject H_0
	summer	0.9442	$W^* < W_{131,0.05} = 0.9800$	3.867e-5	Reject H_0
A-D Test	spring	1.1352	$A^{*2^{(2)}} > 0.752$	0.005506	Reject H_0
	summer	3.0401	$A^{*2} > 0.752$	1.137e-07	Reject H_0

(1): The denominator of W^* is the sum of square errors, thus, the more extreme case for H_1 is smaller W^* with larger denominator, i.e. the critical region is $W^* < c$. (2): To some extent, A^{*2} is the distance of the sample distribution and theoretical normal, therefore, the more extreme case for H_1 is larger A^{*2}. i.e. the critical region is $A^{*2} > c$. Moreover, A^{*2} has taken the number of samples into consideration thus the critical region is fixed for different sample numbers in same level.

In conclusion, both S-W test and A-D test reject the null hypothesis for spring and summer. The p-values for summer are closer to zero, which means that the distribution of the number of rented bikes in summer is more unlikely to be normal. In comparison, the S-W test produces a smaller p-value for the spring data while A-D test produces a smaller p-value for summer data. One of the reasons is that the S-W test is strong against short-tailed and skewed distributions but weak against symmetric moderately long tailed distributions.[11] It is shown in Figure 1 that the spring data is more skewed while summer data is more symmetric, to some degree. For another perspective, the S-W test is not as affected by ties as the A-D test although it does get affected. Thus, for this dataset with all variables being integers, a limited precise measurement in S-W test and A-D test will be generated. In most cases, found by Monte-Carlo experiment, S-W test has highest power among all normality tests.[11]

Two-sample Tests

Without the assumption of normality of the underlying data, non-parametric tests are preferred when determining whether two distributions are the same. After doing this analysis, the result will show whether to allocate an equal number of bicycles during evening peak times in spring and summer. This report tests the consistency in the distributions of total evening peak time rental bikes in spring and summer by performing Wilcoxon rank sum and Kolmogorov-Smirnov test (K-S test). Both tests are classical non-parametric tests used to verify whether differences between two populations exist.

For the Wilcoxon rank sum test (also known as Mann-Whitney U test), the count of total rental bikes during evening peak time in spring and summer correspondingly is denoted as as X_i and Y_j, $i = 1, \ldots, n$ and $j = 1, \ldots, m$ where $n = 128$ and $m = 131$. Additionally, the means of the two distributions are defined as μ_1 and μ_2, respectively. Under Wilcoxon rank sum test, it is assumed that the two samples observations of spring and summer are independent of each other. The null and alternative hypotheses are:

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_1 : \mu_1 \neq \mu_2$$

The significant level is 0.05. Under the null hypothesis, the test statistic W is the sum of Y ranks, i.e. $W = \sum_{j=1}^{m} S_j$ where S_j is the rank of Y_j among all X_i and Y_j. Since here n, m are large enough, the asymptotic distribution of W is preferred,

$$W^* = \frac{W - E_0[W]}{\sqrt{\mathrm{Var}_0(W)}} = \frac{W - \frac{1}{2}m(m+n+1)}{\sqrt{\frac{1}{12}nm(n+m+1)}}$$

where W^* is approximately standard normal distribution for large m and n. The approximation will reduce the complexity for Wilcoxon rank sum test. In this case, the more extreme case for this test is $W^* < c_1$ or $W^* > c_2$. That is, for 95% confidence, the critical region is $W^* < Z_{0.025} = -1.96$ or $W^* > Z_{0.975} = 1.96$. The observed test statistic $W^* = -3.3165 < Z_{0.025}$, so the null hypothesis is rejected. (Or else, the same conclusion by p-value = 0.0009 < 0.05.)

While for the K-S test, according to the given data, the cumulative density functions of total rental bikes during evening peak time in spring and summer is denoted as $F_{1,n}(x)$ and $F_{2,m}(x)$ respectively. The values of n and m are the sample sizes. In this case, they are large enough to test the differences by the K-S test. Then, the null and alternative hypotheses are:

$$H_0 : F_{1,128}(x) = F_{2,131}(x) \text{ versus } H_1 : F_{1,128}(x) \neq F_{2,131}(x)$$

Like the previous tests, the level of the test is 0.05. Under the null hypothesis, the original test statistic $D_{n,m}$ is to measure the maximum distance between two CDFs, which is calculated by $D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$. Based on the conditions above, the observed K-S statistic is 0.2036, and critical region is $D_{n,m} > D_{n,m,0.05} = c(0.05)\sqrt{\frac{n+m}{nm}} = 0.1688$, where $c(x) = \sqrt{-ln(\frac{x}{2})(\frac{1}{2})}$ is the inverse of the Kolmogorov distribution at x, hence H_0 is rejected. (Same conclusion by p-value = 0.0093 < 0.05.)

Both tests draw the same conclusion - the distributions of rental bikes during evening peak time in spring and summer are different. Because of the presence of ties in the data, the p-values are approximated, but this does not make too much difference since the data set is large. A difference to be noted, however, is that the Wilcoxon rank sum test is mainly testing for the location parameter (mean) while the K-S test focuses on the general differences between the two populations. Wilcoxon rank sum test is more detailed and specific. If the result shows that the two populations have the same location, further tests such as Ansari-Bradley rank test will be applied to test the dispersion of scale parameter (variance). In contrast, K-S test is more efficient as it directly provides with a general comparison. Meanwhile, it only uses the maximum difference between two empirical CDFs, without considering the distance in the other domain. In this perspective, the K-S statistic is not a sensible metric.

Conclusion

In the earlier sections it was observed that the distribution for the bike hires in summer and spring does not follow a normal distribution. Additionally, it was seen that the distributions in these two seasons are not the same. In Figure 2, it was also observed that the demand is higher in summer compared to spring. Given these results, the recommendation is for the TfL to deploy different bike allocation strategies for the two seasons. More generally, the suggestion is for the TfL to deploy rebalancing of the bike stations. This method would involve a vehicle that would reposition bikes from full to empty stations during peak hours [12]. Since there is high demand during peak hours in both seasons, this method should be implemented in both summer and spring. However, since the demand is higher in the summer, more of such vehicles should be made available during the summer months. Knowing the demand for bikes in these seasons in advance allows TfL to better understand the costs associated with this method in each of the seasons and offers a better cost management strategy. In the long run, a better allocation strategy could render an increase in demand for bikes which in turn could lead to positive environmental outcomes.

It's worth to note that a limitation of this analysis is that all the tests are only performed on the data in peak evening time during working days and it is assumed that the data in each day is independent of each other, which may not be the case in reality. Moreover, data is available only in the range between 2011 and 2012. If more data was available, analysis such as time series or other implemented methods could have been conducted in order to predict the number of bikes rented in the following years.

References

[1] A. Nikitas. The global bike sharing boom – why cities love a cycling scheme. [Online]. Available: https://theconversation.com/the-global-bike-sharing-boom-why-cities-love-a-cycling-scheme-53895

[2] A. Fleck. How much has the uk's cost of living risen this year? this chart shows you all you need to know. [Online]. Available: https://www.weforum.org/agenda/2022/09/breakdown-of-the-rising-cost-of-living-uk/

[3] S. Spyro. Not fare... uk public transport the most expensive in europe. [Online]. Available: https://www.express.co.uk/news/uk/1571255/UK-public-transport-most-expensive-Europe

[4] D. Freund. How bike sharing can be more efficient. [Online]. Available: https://blogs.scientificamerican.com/observations/how-bike-sharing-can-be-more-efficient

[5] Z. M. Yongping Zhang, "Environmental benefits of bike sharing: A big data-based analysis," *Applied Energy*, vol. 220, no. 296-301, 2018.

[6] Bleeper. Bike sharing reduces congestion - study. [Online]. Available: https://www.bleeperactive.com/blog/bike-sharing-reduces-congestion-study

[7] B. Walker. A brief history of bike sharing. [Online]. Available: https://www.here.com/learn/blog/a-brief-history-of-bikesharing

[8] T. for London. Tube and rail fares. [Online]. Available: https://tfl.gov.uk/fares/find-fares/tube-and-rail-fares

[9] P. Royston, "Approximating the shapiro-wilk w-test for non-normality," *Statistics and Computing*, vol. 2, no. 117-119, 1992.

[10] R. B. D'Agostino and M. A. Stephens, *Goodness-of-fit Techniques.* New York: Marcel Dekker, 2008.

[11] N. M. Razali1 and Y. B. Wah, "Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests," *Journal of Statistical Modeling and Analytics*, vol. 2, no. 21-33, 2011.

[12] P. Leclaire and F. Couffin, "Method for static rebalancing of a bike sharing system," *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 1561–1566, 2018, 16th IFAC Symposium on Information Control Problems in Manufacturing INCOM 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405896318313983