I Want to Ride my Bicycle: How Transport for London Can Promote Year-Round Bike Rentals

Lyn Saunders

Abstract

This paper assesses how Transport for London (TfL) can boost bike rentals year-round by investigating the impact of weather and calendar variables on rental trends. The dataset allowed analysis of important variables (temperature, humidity, windspeed, season, weekends, and holidays) to understand their influence on rental patterns. An exploratory analysis suggested strong seasonal variations, with the most frequent rentals in summer and the least in winter. Statistical tests revealed TfL could expect fewer than 35,000 daily rentals on over 70% of business days, showing significant differences between seasons. An Analysis of Variance (ANOVA) verified rental variability between seasons, while a Welch t-test showed no significant difference between weekends and weekdays. Power analysis suggested a larger dataset to avoid potential Type II errors. To quantify the variables' influence on rental counts, a multiple linear regression model was used. After stepwise selection, the final model accounted for approximately 74% of rental variability, with significant predictors being temperature, humidity, windspeed, holidays, and weekends. Analysis indicated that higher windspeed and humidity correlated with decreased rentals, while higher temperatures were associated with increased usage. These results imply that TfL could improve year-round rental profits by targeting strategies towards weather conditions, such as promotions based on seasonality or advertising cycling gear for adverse weather, to stabilize rentals year-round and maintain steady usage between seasons.

1 Introduction

In modern times, urban areas worldwide have played a key role in the rise of cycling as a mode of transport. The recognised benefits of urban cycling include improved physical health (1) and reduced traffic congestion and journey time (2). Since the introduction of public bike sharing, these systems have grown exponentially in popularity in cities (3). London has embraced bike sharing, with one of its most popular systems, Santander Cycles, expanding over $100km^2$ within 10 years of opening (4). The number of rentals in a large urban area, such as London, may vary due to many factors.

In this report, we identify key calendar and weather variables and analyse their role in the frequency and distribution of bike rentals in London, using a dataset provided by TfL. We aim to use statistical methods to recognise which variables (and which interactions) have the largest impact on bike rentals and create less-profitable days for TfL. To validate our analyses, we assess the goodness-of-fit of our chosen variables and statistical methods. Outlining the results and their implications aims to help TfL target future strategies to increase rental counts. The paper is structured as follows: Section 2 explains our statistical methods, Section 3 relays the results applied to the dataset and specific questions (.1 Exploratory Analysis, .2 Proportion Estimations and Confidence Intervals, .3 ANOVA, t-tests, and Power, and .4 Regression Analysis), and Section 4 presents the discussion and conclusions.

2 Methods

2.1 Exploratory Analysis

We evaluated the summary statistics and visualizations to understand the relationships between key variables (temperature, humidity, windspeed, season, weekends, and holidays). The following methods inform the analyses of how calendar and weather variables affect bike rental trends.

Descriptive statistics (e.g mean and standard deviation) were calculated for each variable across the four seasons (spring, summer, fall, and winter), as seen in (Table 1). To visualize the variability across seasons, we used a box plot (Figure 1b). Further interest in seasonal effects on rental trends is shown in the scatter plot (with a linear regression line) comparing temperature and rentals (Figure 1a). The distribution of rental frequencies is represented by a histogram (Figure 1c) to observe potential skewness or clustering.

2.2 Proportion Estimations and Confidence Intervals

When calculating the expected proportion of days with fewer than y rentals (denoted x) we use the formula: $\hat{p} = \frac{x}{n_{total}}$, where n_{total} represents the total number of days in the dataset. The formula for the 90% upper and lower bounds of the confidence interval of the expected proportion $\hat{p}$ is calculated as follows:

$$\text{Lower bound} = \hat{p} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n_{total}}}$$

$$\text{Upper bound} = \hat{p} + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n_{total}}}$$

where: $\hat{p}$ is the expected proportion, $Z_{1-\frac{\alpha}{2}}$ is the Z-score corresponding to the desired confidence level (for 90%, $\alpha = 10$, hence $Z_{95} = 1.645$) from the standardized normal reference distribution, and n_{total} is as given previously.

To compare whether there exists a difference between the true proportions over different seasons we conduct a two-tailed χ^2 test at the 95% significance level with the following hypotheses: $H_0 : p_i - p_j = 0, H_1 : p_i - p_j \neq 0$. If the returned p-value of the test is less than 0.05, we reject the null hypothesis (H_0), leading us to conclude that there exists a significant difference in the proportions between the seasons i and j.

2.3 ANOVA, t-Tests, and Power

To test whether the expected number of rented bikes varies across seasons, an analysis of variance (ANOVA) was conducted. We compared the variability between the group means (seasons) with the variability within each group, which produces an F-statistic. The F-statistic is computed as:

$$F = \frac{MS_b}{MS_w}$$

where: MS_b is the mean square between groups, and MS_w is the mean square within groups. A p-value is obtained by comparing our F-statistic to the F-distribution. If the p-value is less than 0.05 (at a 95% significance level), we conclude there is strong evidence to reject the null hypothesis, implying that the mean rental count is significantly different for at least one season.

For comparing variability between two groups (e.g., workdays vs. weekends), a Welch t-test is used, comparing the group means without assuming equal variances. The difference is found between the sample means and divided by the standard error of the difference. The standard error is calculated by incorporating the variances of each group weighted by their sample sizes. The degrees of freedom formula is also adjusted to improve the accuracy with unequal variances.

Further accuracy is achieved by conducting a power analysis to assess the likelihood of a Type II error. The magnitude of the difference between the group means (the effect size) is calculated using their standard deviations and observed difference. The effect size is compared with the sample sizes to estimate the test's statistical power, typically at a 5% or 1% significance level. The power calculation output, denoted as $1 - \beta$, indicates the probability of detecting a true difference in the group means.

2.4 Regression Analysis

To understand how the number of bike rentals were affected by calendar and weather variables we fitted a multiple linear regression model, with the response variable being the number of rentals.

The predictor variables were: temperature, humidity, windspeed, season, weekend, and holiday. Considering a potential non-linear relationship between weather parameters, we expanded our model by introducing interaction terms (see temperature $*$ windspeed and humidity $*$ windspeed below). Our new model becomes:

$$\text{count} \sim \text{temperature} * \text{windspeed} + \text{humidity} * \text{windspeed} + \text{season} + \text{weekend} + \text{holiday}.$$

The original model was fit with an ordinary least squares (OLS) regression, but improved the model by applying a stepwise function according to Akaike's Information Criterion (AIC, (5)). This allowed elimination in both directions of predictor variables which did not significantly contribute to the prediction of rentals. The goodness-of-fit of the model was evaluated using the R^2 statistic and residual standard error outputs, while their corresponding p-values allowed us to assess the significance of the predictors.

The new data was input into the `predict()` function in R, which applied the fitted regression model enhanced by the stepwise selection to compute both a point estimate for the number of bikes expected to be rented with a 95% confidence interval. The prediction formula is:

$$\hat{Y} = \beta_0 + \beta_1(\text{temperature}) + \beta_2(\text{windspeed}) + \beta_3(\text{humidity}) + \beta_4(\text{weekend}) + \beta_5(\text{holiday})$$

where $\hat{Y}$ is the predicted bike rental count, and the β's are the estimated coefficients from the model.

3 Results

3.1 Exploratory Analysis

Season	Mean Temp. (°C)	SD Temp. (°C)	Mean Humidity (%)	SD Humidity (%)	Mean Windspeed (km/h)	SD Windspeed (km/h)	Mean Rentals	SD Rentals
Fall	16.54	6.08	73.73	9.43	11.71	4.55	29,927.42	7,088.63
Spring	13.02	4.75	71.82	9.58	14.97	5.06	25,754.24	8,883.32
Summer	22.40	3.46	67.18	9.32	14.95	4.63	35,241.67	6,108.91
Winter	8.87	2.75	78.75	8.65	14.70	6.99	19,086.85	7,219.74

Table 1: Summer yields the highest bike rentals whereas winter yields the lowest. Additionally, the temperature tends to be higher in summer and lower in winter. Thus higher rentals correspond to warmer periods. Humidity also shows its extremities in summer (lowest) and winter (highest), though this factor appears to affect rentals less than temperature.

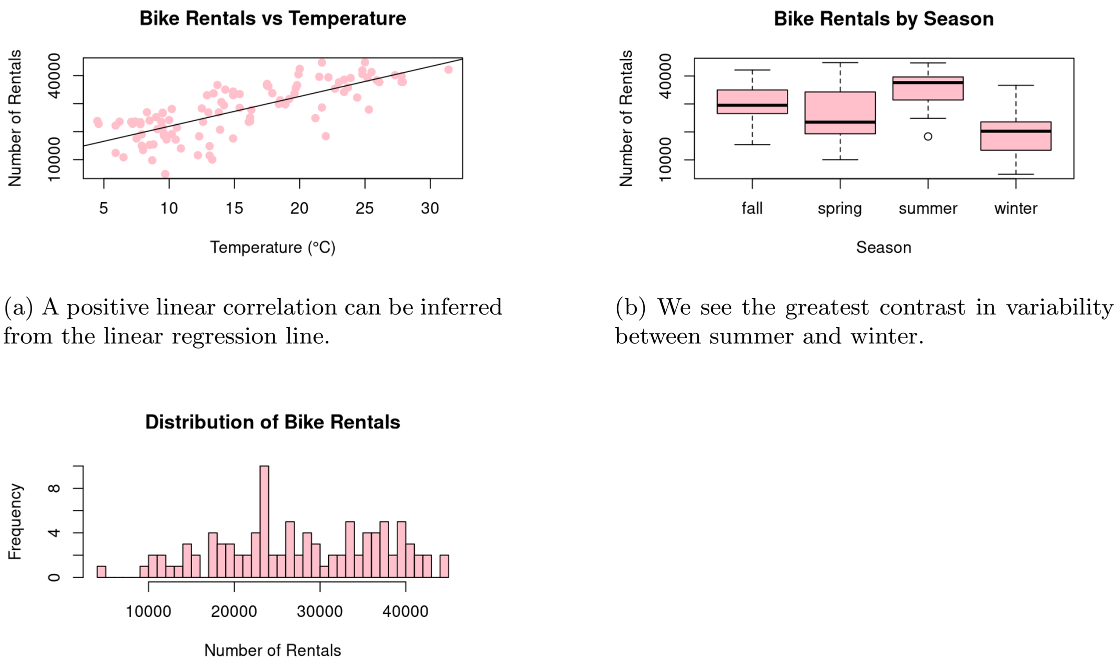
3.2 Proportion Estimations and Confidence Intervals

Using the methods outlined in Section 2.2, we calculated the expected proportion of days with fewer than 35,000 rented bikes (which we define as less-profitable) to be $\hat{p} = 0.71$. Calculations to find the confidence interval suggests that TfL can expect, with 90% confidence, between 63.5% and 78.5% of their business days to be less profitable. The true proportion of less-profitable days will be within this interval 90% of the time.

The proportions for the seasons (spring, summer, fall, and winter) were: $\hat{p_1} = 76.2\%$, $\hat{p_2} = 40.7\%$, $\hat{p_3} = 73.1\%$, and $\hat{p_4} = 96.2\%$. Clearly there exists differences in our observed proportions; more between summer and winter. To assess whether these results represent true differences in the proportions we conducted a two-tailed χ^2 test. The null hypothesis was defined as $H_0 : p_4 - p_1 = 0$. The p-value of 0.110 led us to fail to reject the null hypothesis at the 95% significance level, concluding that there was no significant difference in the proportions of less-profitable days in winter and spring.

3.3 ANOVA, t-Tests, and Power

An ANOVA was conducted to test whether the number of bikes rented per day varied across seasons. The analysis produced an F-statistic of 23.01, with an extremely small p-value of 2.63×10^{-11} which, at the 95% significance level, implies that there is strong evidence to reject the null hypothesis, suggesting that the number of rented bikes varies significantly across the seasons.



(a) A positive linear correlation can be inferred from the linear regression line.

(b) We see the greatest contrast in variability between summer and winter.

(c) Notable peaks can be seen around 24,000 and 35,000 rentals.

Figure 1: A layout of the figures described in Section 3.1.

To test how the rental count varied between working days and weekends a Welch two-sample t-test was performed. The resulting t-statistic of 1.33 and a p-value of 0.189 confirmed that, at the 95% significance level, we could not reject the null hypothesis. Therefore there is no statistically significant difference in the amount rentals between work days and weekends. A power analysis was conducted to measure the risk of a type II error in the t-test. Our calculations returned a low power (0.140), indicating a large probability of a type II error in our t-test and a high risk of not detecting a true difference if one exists. To increase our power to 0.8 we would need to increase our sample size to 156.

3.4 Regression Analysis

We conducted a multiple linear regression analysis to assess how calendar and weather variables affected the rental counts. After adding interaction terms and conducting stepwise selection based on AIC, we were left with the model:

$$\text{count} \sim \text{temperature} + \text{windspeed} + \text{humidity} + \text{weekend} + \text{holiday}.$$

The goodness-of-fit of the resulting model yielded an R^2 value of 0.738. Hence the predictors in the model account for approximately 73.84% of the variability in the rental count. The adjusted R^2 value returned slightly lower at 0.7245. Having both R^2 values be relatively high suggests that the model has substantial predictive power.

4 Discussion

It is important to assess the assumptions underlying our statistical methods and consider potential violations. In Section 2.2 the assumption of a sufficiently large sample size was violated in our χ^2 test. A Fisher test would provide more reliable results due to not relying on the same assumptions of high expected frequencies.

For our ANOVA in Section 2.3 we conducted a Shapiro-Wilk test, implying the normality of residuals from its p-value (0.569). A Levene's test was conducted and returned a significant

Predictor	Estimate	Std. Error	t-value	p-value
Intercept	41027.32	5474.37	7.494	< 0.001
Temperature	848.01	85.62	9.904	< 0.001
Windspeed	-232.05	92.76	-2.501	0.014
Humidity	-294.89	58.53	-5.038	< 0.001
Weekend (1)	-3789.05	1085.43	-3.491	< 0.001
Holiday (1)	-8992.70	2303.82	-3.903	< 0.001

Table 2: All of the predictors were statistically significant at the 5% level and most at the 1% level, except for windspeed. Interpretations of the Estimate column suggest that higher temperatures correlate with an increase in bike rentals, whereas higher windspeed and humidity appear to decrease rentals. The negative estimates imply that holidays and weekends are also associated with fewer rentals.

Scenario	Predicted Rentals	95% CI Lower	95% CI Upper
Working day in spring (16°C, 8% humidity, 12 km/h)	49451.86	41871.62	57032.09
Holiday on summer weekend (26°C, 30% humidity, 7 km/h)	39822.95	32652.06	46993.84
Working day in autumn (10°C, 85% humidity, 30 km/h)	17480.75	13862.74	21098.76
Winter weekend (non-holiday, -1°C, 70% humidity, 16 km/h)	12035.51	8469.46	15601.56

Table 3: Predicted bike rentals for four seasonally specific scenarios. This highlights that weather and seasonal conditions are important factors in bike usage patterns. Colder, more humid conditions could be strong deterrents to cycling.

p-value of 0.252, confirming no violation to the homogeneity of variances assumption.

The multiple linear regression model fitted in Section 2.4 comes with no violations to the linearity or independence assumptions. A Breusch-Pagan test (6) demonstrated no evidence of a violation to the homoscedasticity from a p-value of 0.0646. Additionally, a Variance Inflation Factor (VIF, (6)) test failed to detect evidence of a violation in the multicollinearity among the independent variables. Similarly, no significant evidence of a violation to the autocorrelation assumption was found during the Durbin-Watson test. However, a Shapiro-Wilk test conducted returned a statistically insignificant p-value of 0.00492, suggesting a violation to the normality of residuals assumption.

Implications derived from Figure 1b, supported by our ANOVA (Section 3.3), suggest that bike rentals are strongly influenced by seasonal trends (lower rentals in winter and higher in summer). Figure 1a may suggest that these tendencies are driven by temperature differences. Our table of regression coefficients (Table 2) indicated further dependency on seasonal factors, implying that cyclists behaviour may be heavily influenced by weather conditions. Calendar variables (weekends, holidays) do not appear to substantially affect rentals (Table 2), implying steady daily cycling habits (see Section 3.3 for further reassurance). Our high R^2 values found in Section 3.4 imply that the selected statistical models and chosen variables supply a good predictive framework. Further modifications could strengthen the validity of our framework.

References

de Hartog, J. J., Boogaard, H., Nijland, H., & Hoek, G. (2010). Do the health benefits of cycling outweigh the risks? *Environmental Health Perspectives*, *118*(8), 1109–1116. National Institute of Environmental Health Sciences.

Bullock, C., Brereton, F., & Bailey, S. (2017). The economic contribution of public bike-share to the sustainability and efficient functioning of cities. *Sustainable Cities and Society*, *28*, 76–87. https://doi.org/10.1016/j.scs.2016.08.024.

Eren, E., & Uz, V. E. (2020). A review on bike-sharing: The factors affecting bike-sharing demand. *Sustainable Cities and Society*, *54*, 101882. Elsevier.

Reynolds, L. (2020). London's cycle hire scheme turns 10 this year — Londonist. Retrieved January 22, 2021, from https://londonist.com/london/transport/santaner-cycles-10-years.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression*, Third Edition. Thousand Oaks, CA: Sage. Retrieved from https://socialsciences.mcmaster.ca/jfox/Books/Companion/.