Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Assignment Submission Cover Sheet

Programme Title:	**MSc Business Analytics**
Module Code and Title:	**BU7144 Business Forecasting**
Assessment Title:	**Group Report**
Group Number:	**Team 5**

Student Name and Contribution	**%**		**%**
1. Lim Yue Ying Veronica		**4. Kai Kei Cheung**	
2. Temitayo Coker		**5. Ashley Chew Bo Qing**	
3. Huang Xueyan		**6. Ani Chidera**	

For group work – individual % contributions need to be stated **only** where they **are not equal.**

Please read the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at: http://www.tcd.ie/calendar

Online Tutorial on avoiding plagiarism 'Ready, Steady, Write', is located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write

Capital Bikeshare Forecasting Report

Executive summary

Capital Bikeshare is a company that provides bike-share services around the Metropolitan area of Washington DC. Bike sharing is a new generation of traditional bike rentals, where bike-sharing systems automate the entire process from registration to rental, and return. These systems make it simple for users to hire a bike from one location and return it to another.

Bike-sharing systems are appealing for research due to their data generation qualities and intriguing real-world applications. In contrast to other modes of transportation like the bus or the subway, these systems openly record the distance travelled as well as the location of departure and arrival. This feature transforms the bike-sharing system into a fictitious sensor network that can track urban motion. Capital Bikeshare aims to use the platform to increase the convenience of travel for users as well as to help provide more sustainable modes of transportation for a greener future.

The goal of this project is to provide a forecasting model for the Capital Bikeshare Operations team to accurately predict the supply required to meet the daily demand by users. This will be achieved by forecasting the demand by users daily and allocating the necessary bikes to the bike-share stations.

The data used in this project is a time-series data collected over the first 2 years after the start of Capital Bikeshare in 2010. The data includes the number of users per day, the weather, temperature, holiday, indicator of weekday or weekend, etc.

The key challenge of this project is applying the forecasting model to be implemented - a multivariate linear regression model with additive trend - due to the high level of error percentage at 94.69%. There were attempts to address the high error within the model, but the lack of data makes it difficult to rectify the forecasting model. Recommendation and Implementation suggestions are also presented as part of the findings of the project.

Problem Description

Business problem

Capital Bikeshare is invested in enabling the public to make use of sustainable modes of transport and increase their convenience to get around the city. Consequently, the company investors are always looking for ways to improve the profitability of the company by improving the service provided and ensuring user satisfaction.

Capital Bikeshare does weekly maintenance on the bikes at a central warehouse that stores all bikes that are undergoing maintenance. The maintenance is done in batches to ensure that the bikes are road safe, thus improving their service to users. However, bikes at stations are removed in batches, and this leads to an uneven distribution of bikes at the stations daily.

Thus, the business problem that this report aims to address is: Capital Bikeshare wants to know if it is possible to understand the demand for bikes to ensure that there are sufficient bikes at the stations to meet the user's demands daily.To address the business problem, a clear understanding and outline of the strategy, goal, stakeholders, opportunities, and challenges must be defined.

Strategy: To build a model for the company to forecast the demand for bikes by users.

Goal: To ensure that the supply of bikes meets the demand of users.

Stakeholders:

1. Capital Bikeshare Management Team
2. Capital Bikeshare Operations Team
3. Registered Users, who purchased a monthly or yearly subscription plan.
4. Casual Users, who pay-per-ride or purchase a one-day pass.
5. Competitors of Capital Bikeshare: dockless bike-sharing companies

Opportunities and Challenges:

1. Efficient management of Capital Bikeshare resources.
2. The ability to increase revenue by optimising supply and demand
3. The ability to increase user satisfaction
4. Implementation of the developed forecast model for future use.

Analytics Goal

To develop the forecasting model, the analytics objective needs to be defined. This can be done by translating the business problem into a forecasting problem. That is, to understand the forecasted demand for the bikes for the day ahead. Setting the forecasting problem helps to define the forecasting goal, outcome variables and the models that will be explored in the process of the development of the forecasting model.

Forecasting Goal: To determine the number of bikes needed daily.

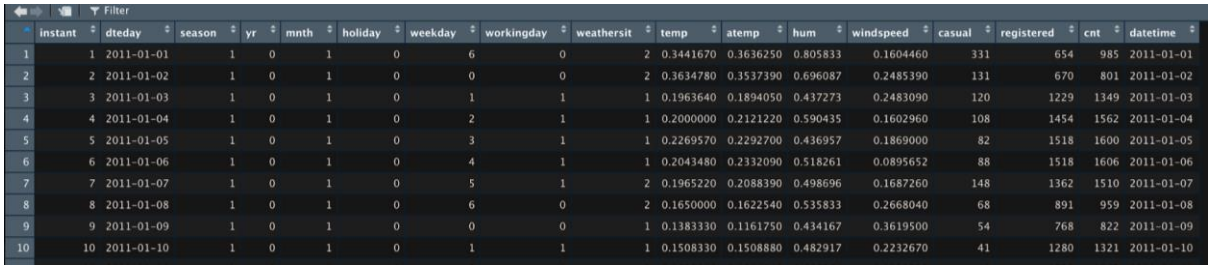
Outcome Variable: The total count of registered and casual users

Data Description

The main dataset for Capital Bikeshare contains the historical data of the hourly and daily count of rental bikes between 2011 and 2012 in the Capital Bikeshare system in Washington DC with the corresponding weather and seasonal information. The focus of the analysis will be based on the daily count of rental bikes between 2011 and 2012.

There are 731 rows and 16 columns within the data set, with each row representing the demand for a single day. The analysis and development of the forecasting model will focus on the following variables to conduct the forecasting objective: the number of users, the period, weather, season, month, holidays, and temperature. These variables were chosen as focal points due to their significant relationship to the number of users, which will be further discussed in the next section.

Raw Data:



	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt	datetime
1	1	2011-01-01	1	0	1	0	6	0	2	0.3441670	0.3636250	0.805833	0.1604460	331	654	985	2011-01-01
2	2	2011-01-02	1	0	1	0	0	0	2	0.3634780	0.3537390	0.696087	0.2485390	131	670	801	2011-01-02
3	3	2011-01-03	1	0	1	0	1	1	1	0.1963640	0.1894050	0.437273	0.2483090	120	1229	1349	2011-01-03
4	4	2011-01-04	1	0	1	0	2	1	1	0.2000000	0.2121220	0.590435	0.1602960	108	1454	1562	2011-01-04
5	5	2011-01-05	1	0	1	0	3	1	1	0.2269570	0.2292700	0.436957	0.1869000	82	1518	1600	2011-01-05
6	6	2011-01-06	1	0	1	0	4	1	1	0.2043480	0.2332090	0.518261	0.0895652	88	1518	1606	2011-01-06
7	7	2011-01-07	1	0	1	0	5	1	2	0.1965220	0.2088390	0.498696	0.1687260	148	1362	1510	2011-01-07
8	8	2011-01-08	1	0	1	0	6	0	2	0.1650000	0.1622540	0.535833	0.2668040	68	891	959	2011-01-08
9	9	2011-01-09	1	0	1	0	0	0	1	0.1383330	0.1161750	0.434167	0.3619500	54	768	822	2011-01-09
10	10	2011-01-10	1	0	1	0	1	1	1	0.1508330	0.1508880	0.482917	0.2232670	41	1280	1321	2011-01-10

Data Preparation and Visualisation

This next section addresses the decisions made regarding using variables, partitioning data, and the period used for the data.

Capital Bikeshare was founded in 2010 and its first year of operation was in 2011, to aid in the formulation of the forecasting model that will be utilised to solve the Capital Bikeshare business problem, the first 4 months of data points/observations – January 1st to April 30th, 2011 – were dropped from the dataset. It is believed that the first few months in operation would not be a true reflection of the user demand and as such, would affect the performance of the forecasting models.

The data has been partitioned into:

- Training period: May 2011 - May 2012 (65%)
- Validation period: June 2012- Dec 2012 (35%)

Before selecting the variables to use for the multivariate linear regression model, an in-depth analysis of each variable was carried out by visualising the data, to get a clear understanding of the relationship between the different variables and the total demand. As the users were defined into 2 groups, registered and casual users, the “cnt” variable is used as an umbrella variable to account for both groups of users and to reduce the complexity of the model.

The visualised plots can be found in the appendix, and the concluded observations resulted in dropping the “workingday” and “weekday” variable as the total number of bikes rented were practically

unaffected by weekdays and weekends. Other variables that will not be used in the development of the model are "normalised feeling temperature", "normalised humidity", and "wind speed" as "temperature" covers all these variables and can be used as a sort of index for all of them.

A time-series object was then created using the date-time variable and the total number of users. The resulting plot is shown in Figure 1.1. After decomposing the data, it is evident that the dataset contains additive trends and multiplicative seasonality.

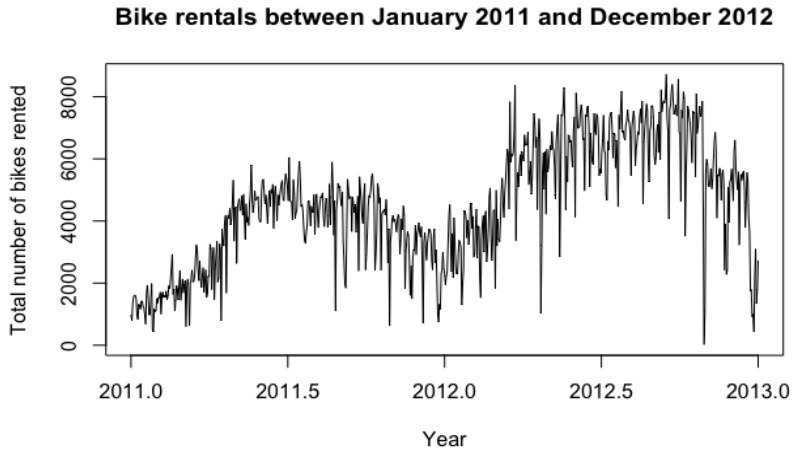


Figure 1.1: Time-series plot of the total number of bike rentals between January 2011 and December 2012

Below are the final variables that would be used to build the multivariate linear regression model.

- "Weathersit"
- "Season"
- "Month"
- "Holiday"
- "Temperature"
- "Year"

There were also external variables to be considered when carrying out the analysis - one such event would be Hurricane Sandy which hit Washington DC at the end of October 2012 (Goldenberg, 2012). This is a possible reason for the abnormal decline in trend at the end of 2012 which likely caused the inaccuracy throughout the forecasting model. Thus, external factors should be considered when forecasting future periods.

Cleaned Data:

Filter

	datetime	seasons	holiday	yr	mnth	weathersit	temp	cnt
1	2011-01-01	1	0	0	1	2	0.3441670	985
2	2011-01-02	1	0	0	1	2	0.3634780	801
3	2011-01-03	1	0	0	1	1	0.1963640	1349
4	2011-01-04	1	0	0	1	1	0.2000000	1562
5	2011-01-05	1	0	0	1	1	0.2269570	1600
6	2011-01-06	1	0	0	1	1	0.2043480	1606
7	2011-01-07	1	0	0	1	2	0.1965220	1510
8	2011-01-08	1	0	0	1	2	0.1650000	959
9	2011-01-09	1	0	0	1	1	0.1383330	822
10	2011-01-10	1	0	0	1	1	0.1508330	1321

Data Forecasting Solution

To determine the number of bikes needed at each station daily to meet user demand, a seasonal naïve model was developed as a benchmark where the user demand forecast from the same season in the previous period is being used to forecast future user demand. The following models were developed and compared against the seasonal naïve model to determine the best model to forecast the demand for bikes:

- Time-series linear regression model
- Holt-Winter's model
- ARIMA model
- Multivariate linear regression model
- Multivariate linear regression model with additive trend

The error metrics are used to evaluate the performance of the different models by comparing the Root Mean Squared Error (RMSE) and the Mean Absolute Percent Error (MAPE) as well as the goodness of fit of the plots. Figure 2.1 provides the summary of all the error metrics, and it can be observed that while the different models perform well in predicting the training set, it performs poorly when predicting the validation set. This is evident from the large MAPE value present in all the validation sets for the different models, which may result from the lack of information within the dataset when building and training the model. However, by just evaluating the RMSE of the different values, the multivariate linear regression model with additive trend has the lowest RMSE values and performs the best in and out of the sample for predicting the demand.

	RMSE	MAPE
SEASONAL NAIVE		
Training set	2249.893	32.10296
Valid set	2564.557	107.20453
TSLM ~ TREND + SEASON		
Training set	233.3712	1.453416
Valid set	1451.2892	138.590252
HOLT-WINTER'S		
Training set	929.3207	22.53235
Valid set	1716.7129	178.43125
ARIMA		
Training set	892.7317	21.07975
Valid set	1699.4703	176.06154
LM ~ EXTERNAL FACTORS		
Training set	693.4603	14.87313
Valid set	1138.4097	98.14325
LM ~ TREND + EXTERNAL FACTORS		
Training set	688.8144	14.66176
Valid set	1133.1127	94.69887

Figure 2.1: Summary of error metrics for each forecasting model

By comparing Figures 3.1 and 4.1, which show the goodness of fit for the seasonal naïve model and multivariate linear regression model with additive trend respectively, it can be observed that the multivariate linear regression model with additive trend has a better fit, thus, supporting its overall better predictive performance than the seasonal naïve model.

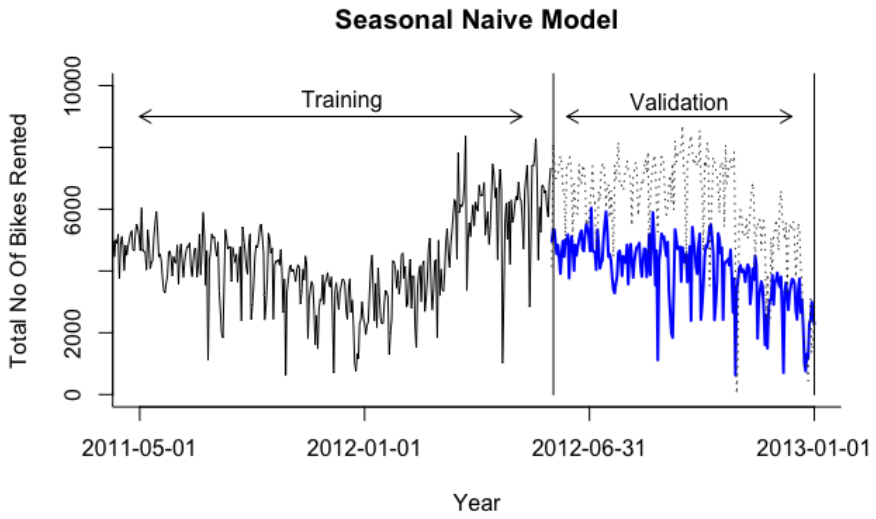


Figure 3.1: Goodness of fit of Seasonal Naïve model

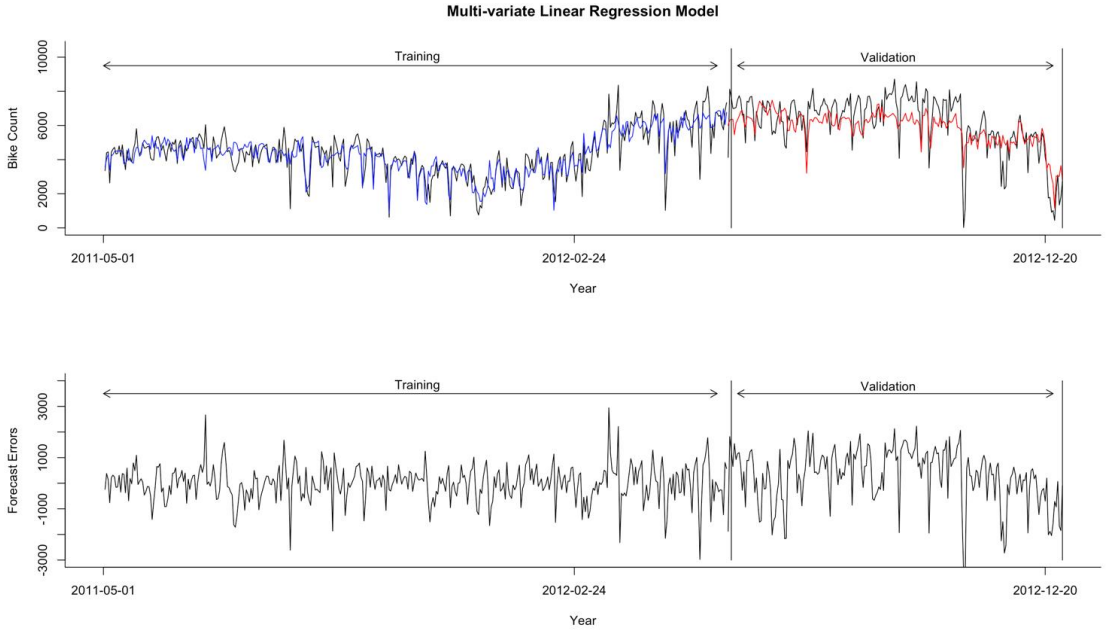


Figure 4.1: Goodness of fit of Multivariate Linear Regression model and Time-plot of Forecast Errors

Upon inspection of the time plot of the forecast errors, there appears to be no particular pattern or additional information in the residuals. An autocorrelation (ACF) plot was generated to prove this assumption and to understand the autocorrelation between the present forecasted bike demand and its past values. This plot hinted at the fact that there may be too many variables present in the model and as such, a Relative Importance of Predictors plot was created in Figure 5.1 to further understand the extent of the impact of each variable.

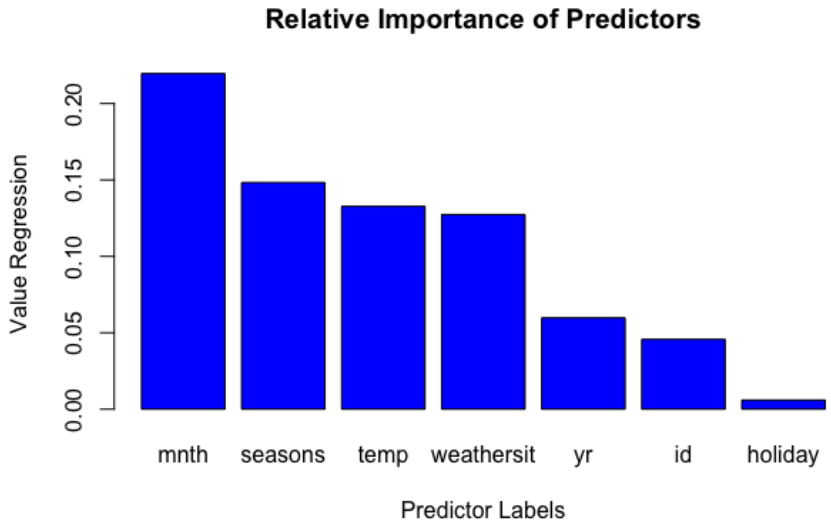


Figure 5.1: Relative importance of variables

From Figure 5.1 above, out of all the variables included in the model, there are only 3-4 main variables that heavily influence the user demand for bike rentals in Washington D.C. Other variables of lower relative importance may cause the model to overfit – which is apparent in the large variation between the training and validation RMSE as seen in the table in Figure 2.1. As a result, there might be serial correlation or additional information that is yet to be captured by the model which can be used to improve the user demand forecasts, although the time plot of the forecast errors does not capture this. Thus, it might be necessary to re-assess the relevant variables that should be included and run the model again.

In this situation, the company can still proceed with using the multivariate linear regression model with additive trend to predict the demand. However, with the high magnitude of the MAPE, there are some considerations which will be addressed in the next section of the report. The model's poor overall performance is likely due to the small number of observations in the daily dataset and methods to improve the model for future use will be addressed in the Recommendation section.

Implementation

With the business problem identified, the forecast model can predict the demand for bikes needed for the day and allow for measures to be taken to ensure that the forecasted demand is met by allocating the necessary bikes to the relevant stations.

Based on the following assumptions:

1. Each bike-sharing station can dock up to 8 bikes at any given time.
2. Most of the bike usage is within the city centre.
3. The bike-sharing stations are located within proximity of each other.
4. All demand is met regardless of the location of the stations, provided that the number of available bikes meets the demand for the bikes.

Although the objective of this analysis is to forecast the demand being met, it is also necessary for Capital Bikeshare to examine the cost matrix of their business model to inform their decision-making on whether to overestimate or underestimate their inventory as well as to understand the extent to which the high error metrics will impact the company financially.

If the company chooses to overestimate the forecasted demand, this increases the cost to the company as it is meeting the predicted demand, but the demand does not meet the supplied bikes - this means that the company will be paying for unnecessary maintenance costs of unused bikes. Assuming that the total cost of maintenance per bike per year is $125, the cost per week will equal $2.40 and the cost per day is $0.34. Due to overcommitting inventory and making superfluous purchases/maintenance in anticipation of surplus demand that does not materialise, overestimating demand will result in a deterioration in the firm's return on assets.

If the company chooses to underestimate the forecasted demand, the demand for bikes is much greater than the supply, then there will be a loss in the company's earning potential. If the loss of profit only comes from casual users, as the registered users have already subscribed for a whole month or year, and so there is no loss of profit from them, the potential loss would be $8 per day per casual user. Fundamentally, underestimating demand will result in higher manufacturing costs, and lower quality standards, followed by lower user satisfaction, leading to a disintegration of the overall reputation of the company.

Whilst comparing the two situations, ultimately it would be more beneficial for Capital Bikeshare to overestimate than to underestimate the demand as the main goal is to meet the demand, satisfy user needs and also generate greater revenue. Another important consideration when estimating forecasted demand would be the competition from other bike-share companies if there is too much competition and the user demand is met, the users may decide to use the services from the competitor companies.

To implement the model with the understanding that the high MAPE value will be the main factor that determines the overestimation or underestimation of the bike demands, the company could take into account the fact that the actual demand for bikes would probably never be as high as twice the predicted demand for bikes. Thus, a good balance of overestimation and underestimation of the predicted bike demand would be to increase the predicted demand by 47.3% (94.6% / 2, from the MAPE value). In doing so, it addresses the high error value of the model, but not sacrificing the ability to meet the user needs while working to increase the revenue for the company.

Recommendation

The forecasting model has its limitations, due to the high MAPE, which makes it a poor predictor of demand. Therefore, to develop a better model, it is recommended that another model is trained with more recent observations or consider using the hourly count of bike demand as this would provide more observations, giving rise to more accurate forecasts.

With Capital Bikeshare moving away from just having traditional bikes to also include electronic bikes (e-bikes), the forecast model can also be used as the basis to develop a more inclusive model to forecast the demand for e-bikes and traditional bikes as well.

Capital Bikeshare can use this model to forecast future expansion for the company itself by including the location of the bikes in the forecast model to ensure that the bikes are being maintained and the allocation of bikes is accurate. This revised model can forecast the demand by location and help the company to evaluate if there are sufficient bikes available for expansion.

Conclusion

This report and project aim to address the business problem of Capital Bikeshare to help them understand the daily demand of users a day in advance. Therefore, the forecast model was evaluated with a focus on a good fit and the lowest percentage of error.

The forecast models developed for this task had high levels of percentage error and therefore, the Multivariate Linear regression model with additive trend is selected as the best model to forecast due to its lower magnitude of errors and better fit to the training data. However, this model does not come without its limitations and the results of the forecasted demand should be implemented carefully.

Advantages and Limitations

By using multivariate linear regression with additive trend model to forecast demand, it takes into consideration many external variables. This is especially important for this specific dataset, as there is more than one variable that influences the users' demand. Thus, accounting for the external variable reflects a more accurate perspective of the forecasted demand. However, the MAPE values for all the forecasting models generated are very high, and even the lowest MAPE value of the final forecast model is 94.69%. This reveals that the demand of the users may not be very predictable and will be difficult to forecast and it poses the challenge to the company of being able to accurately forecast the next day's demand.

Appendix

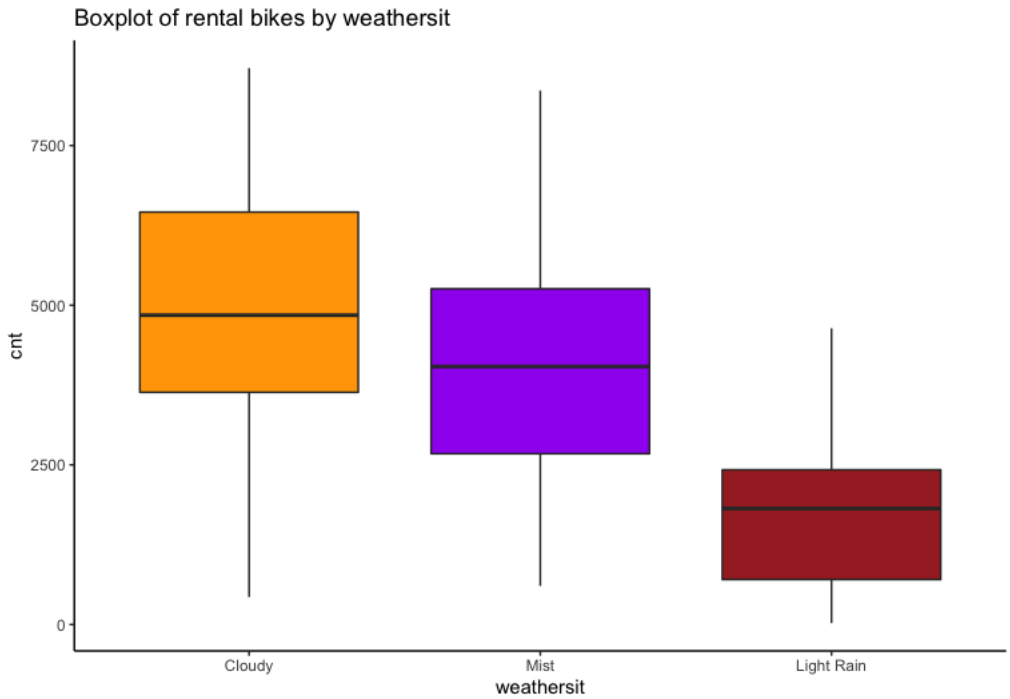


Figure 1.1: Boxplot of count (cnt) and weathersit

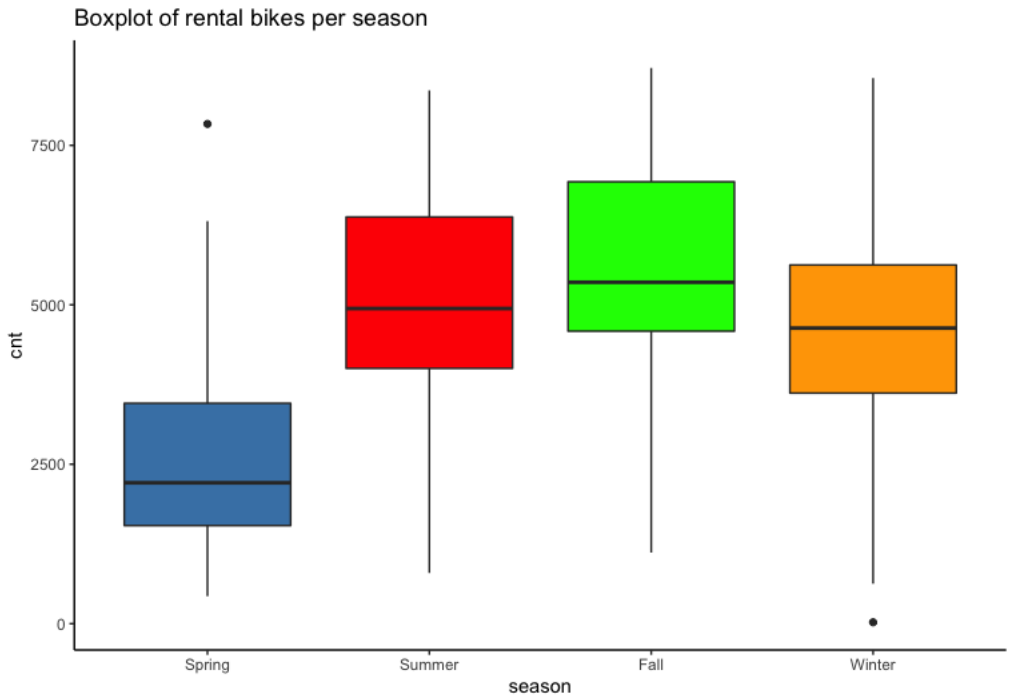


Figure 1.2: Boxplot of count (cnt) and seasons

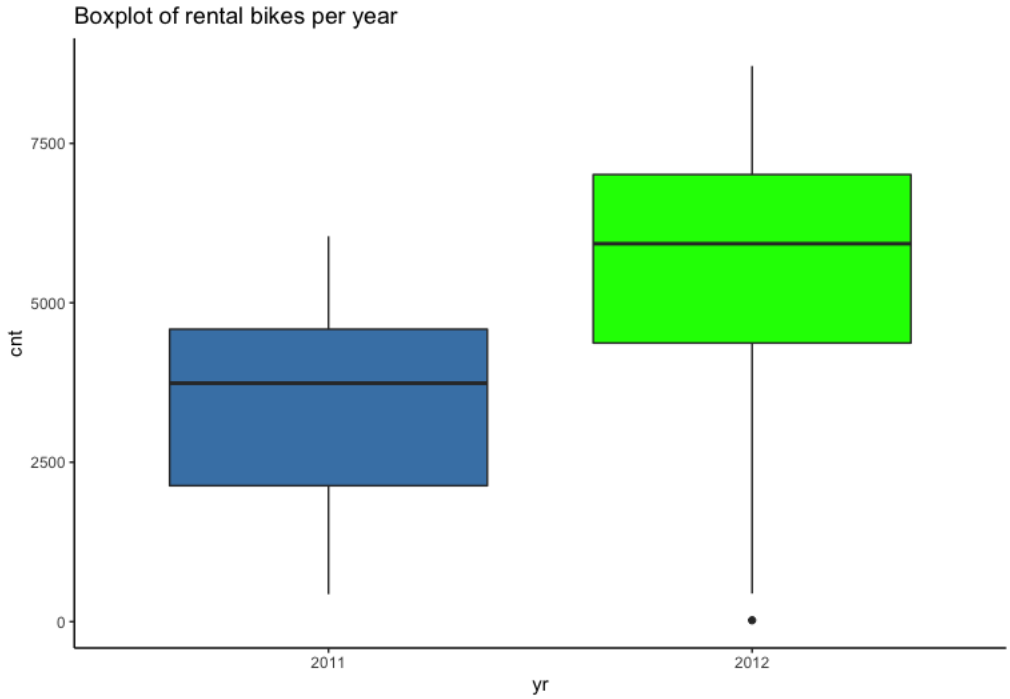


Figure 1.3: Boxplot of count (cnt) and year (yr)

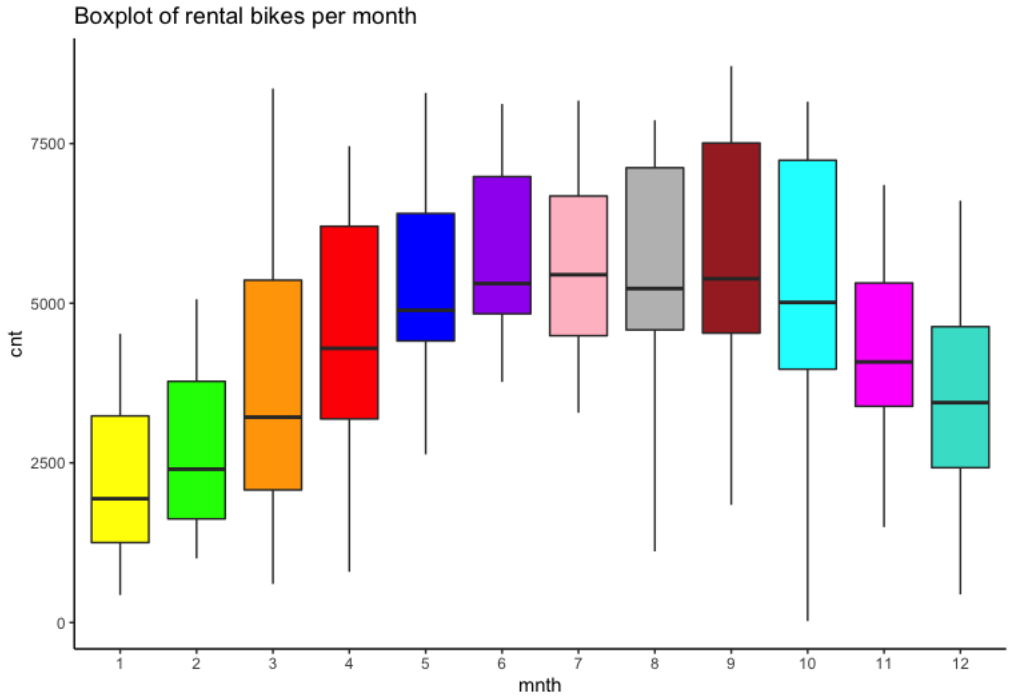


Figure 1.4: Boxplot of count (cnt) and month (mnth)

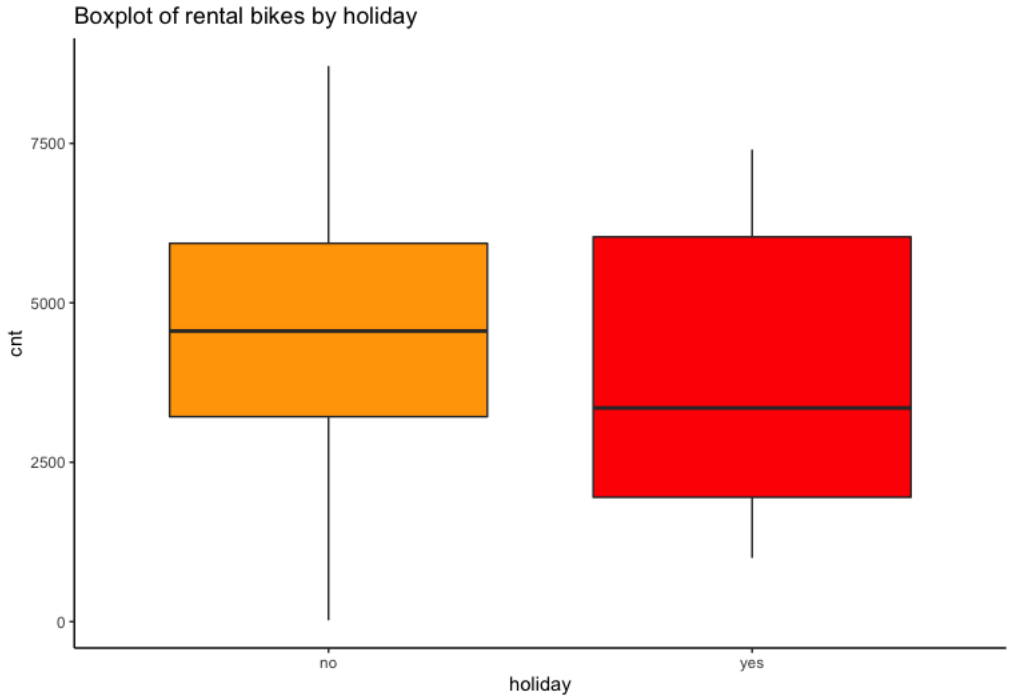


Figure 1.5: Boxplot of count (cnt) and holiday

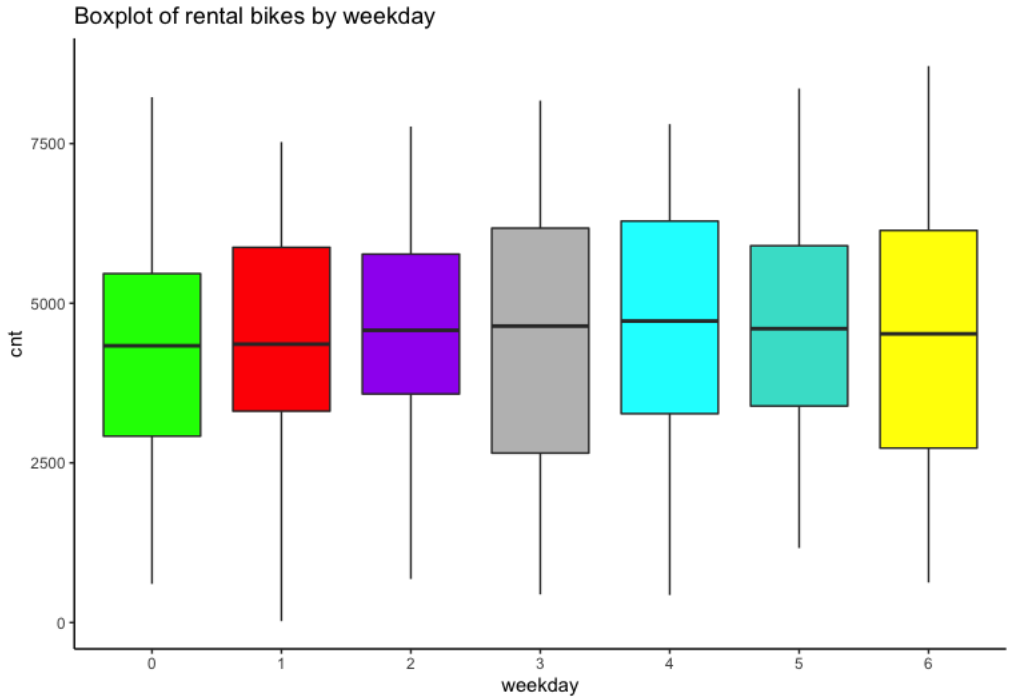


Figure 1.6: Boxplot of count (cnt) and weekday

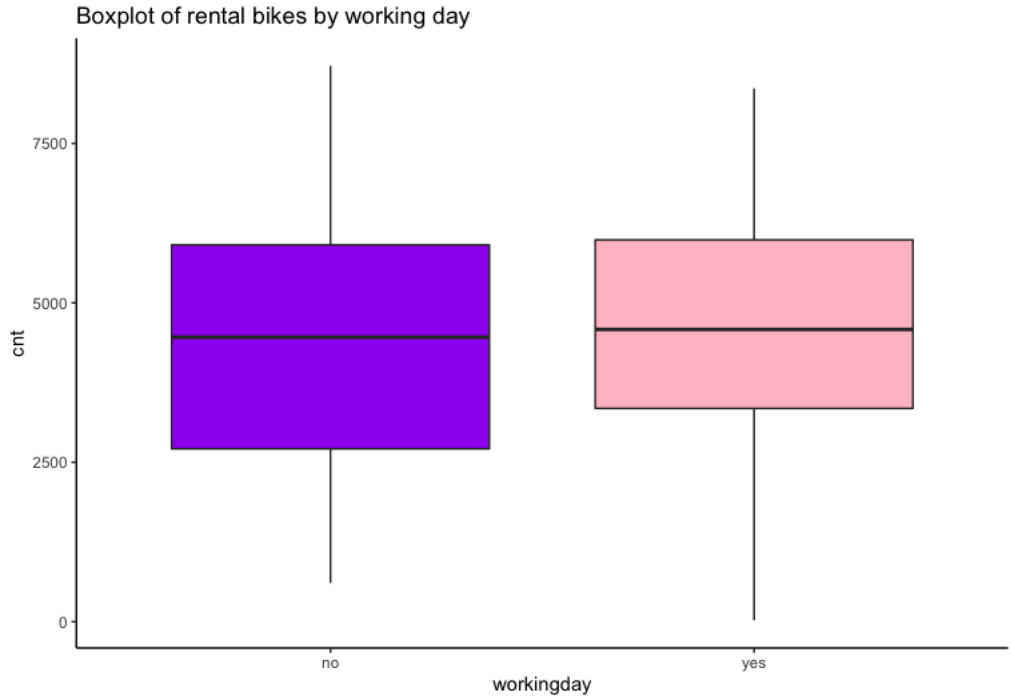


Figure 1.7: Boxplot of count (cnt) and workingday

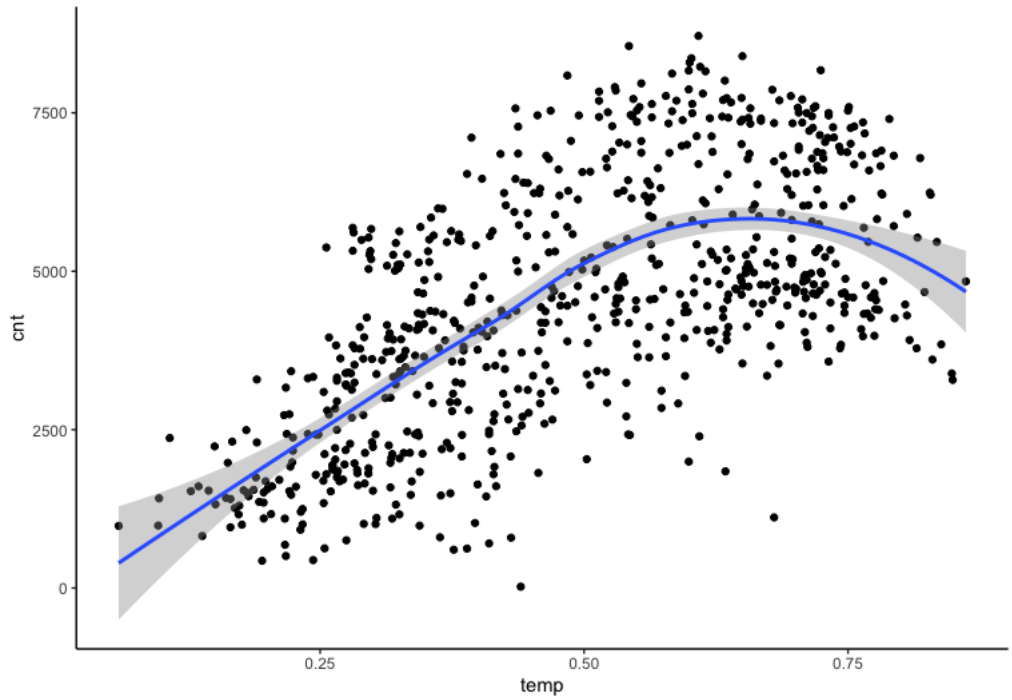


Figure 1.8: Scatterplot of count (cnt) and normalised temperature (temp, °C) with smoothing method = "loess"

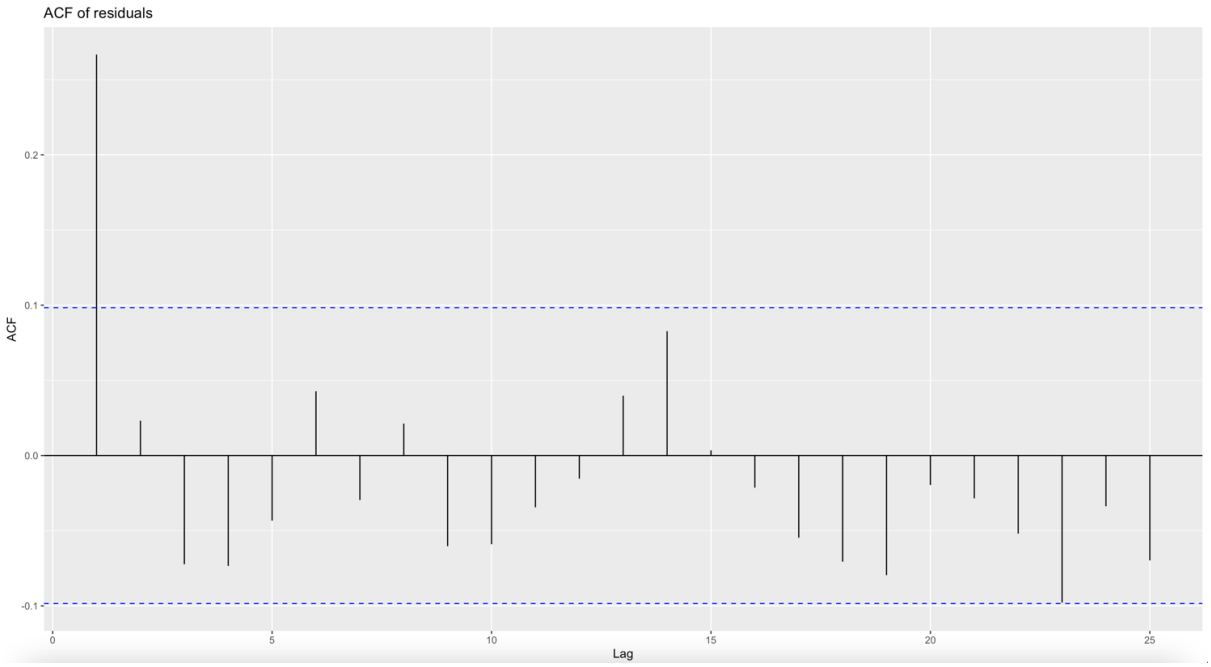


Figure 5.1: Autocorrelation plot (ACF) of residuals

References

Goldenberg, S. (2012, October 29). Washington DC shuts down in preparation for Hurricane Sandy. The Guardian. Retrieved November 11, 2022, from https://www.theguardian.com/world/2012/oct/29/washington-dc-shutdown-hurricane-sandy