FA17 STAT-S 670 EDA - Final Project

Harsh Reddy Gandavarapu, Data Science, Indiana University
Sriram Sridhar, Data Science, Indiana University
Vinoth Aryan Nagabosshanam, Data Science, Indiana University
Vinaysheel Kapgate, Data Science, Indiana University

November 27, 2017

Abstract

Among the various means of transport there are none that are cleaner and economical than the bike (also known as bicycle). We are witnessing the rise of automated bike sharing systems that handle bike rentals and returns with great efficacy. We have with us two years' data (2011 and 2012) from the Capital Bikeshare system in Washington D.C., USA, which is one such system. Using this data we attempt to build a statistical model that can predict the number of bikes rented at a given hour on a given day using information such as weather, day of the week, whether the day is a holiday etc.

Objective

We aim to provide a model that efficiently predicts the number of bike that will be rented at a given part of the day on a given day. We also use the weather variables, which in real life can be easily obtained from weather forecasts. We build a Linear Regression Model that will predict the number of bikes to be stocked by the rental agency. We also use Linear Discriminant Analysis to visualize how well our variables work to predict the number of bicycles. We are not interersted in inferential statistics for our project, our focus lies mainly on prediction. We shall perform Exploratory Data Analysis keeping in mind that it greatly helps us in feature selection.

Data description

Source: **Kaggle/UCI Repository**

Links:

1. https://www.kaggle.com/c/bike-sharing-demand
2. http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

- **dteday**: date: the date of the day on which a particular observation was recorded.
- **season**: factor: One of four seasons of the year.
- **yr, mnth**: factor: attributes extracted from the dteday.
- **hr**: factor: one of twenty four hours of the day.
- **holiday**: factor: whether that particular day was a holiday.
- **weekday**: factor: one of 7 days of the week.
- **workingday**: factor: whether the day was neither a weekend nor a holiday.
- **weathersit**: factor: values from "1" to "4" where "1" indicates clear weather and "4" indicates highly inclement weather.
- **temp**: numeric: normalized temperature.
- **atemp**: numeric: normalized "feels like" temperature.
- **hum**: numeric: normalized humidity.
- **windspeed**: numeric: normalized windspeed.

- **cnt**: numeric: count of total bikes rented.
- **partday**: factor: we bucket the hours of the day for plotting purposes.

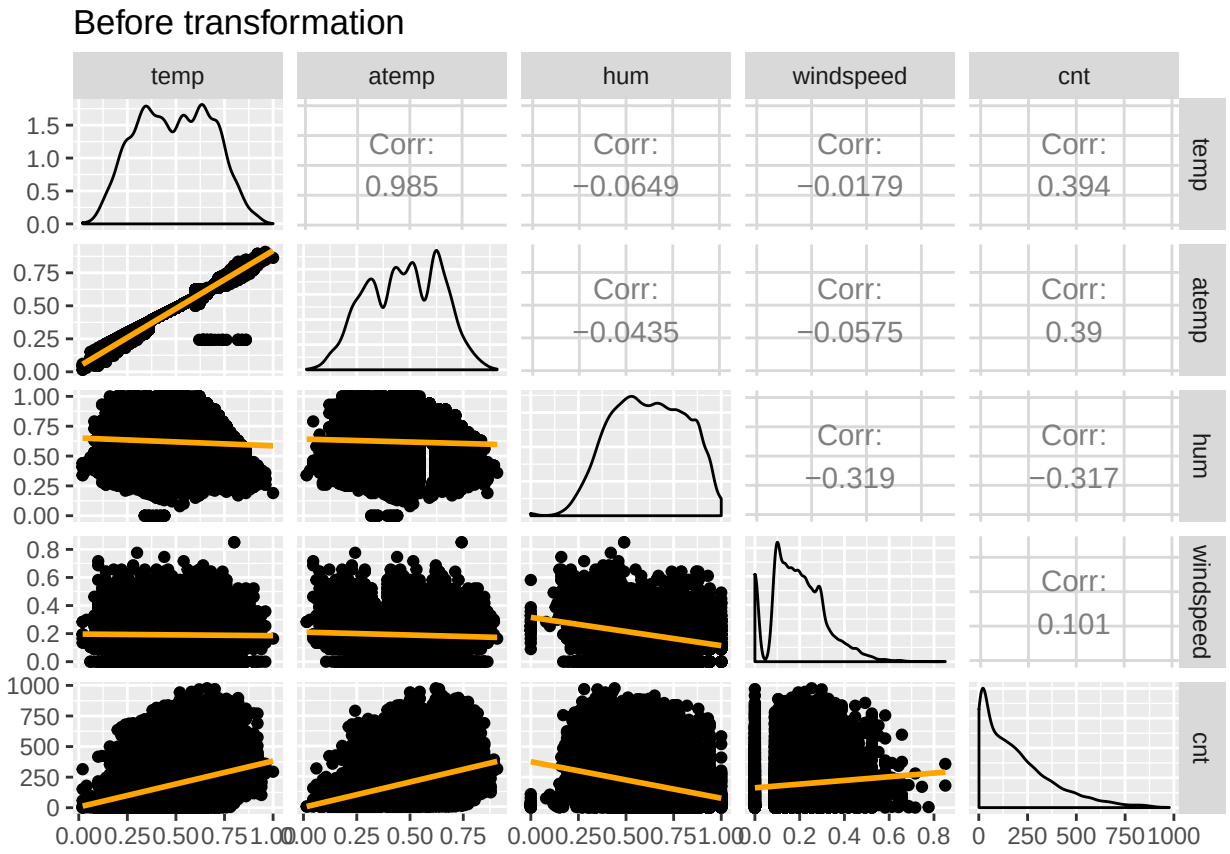
We did not remove outliers as it is most important for the agency to be able to handle situations where bike rentals are highest. We also did not find any missing data.

Data at a glance

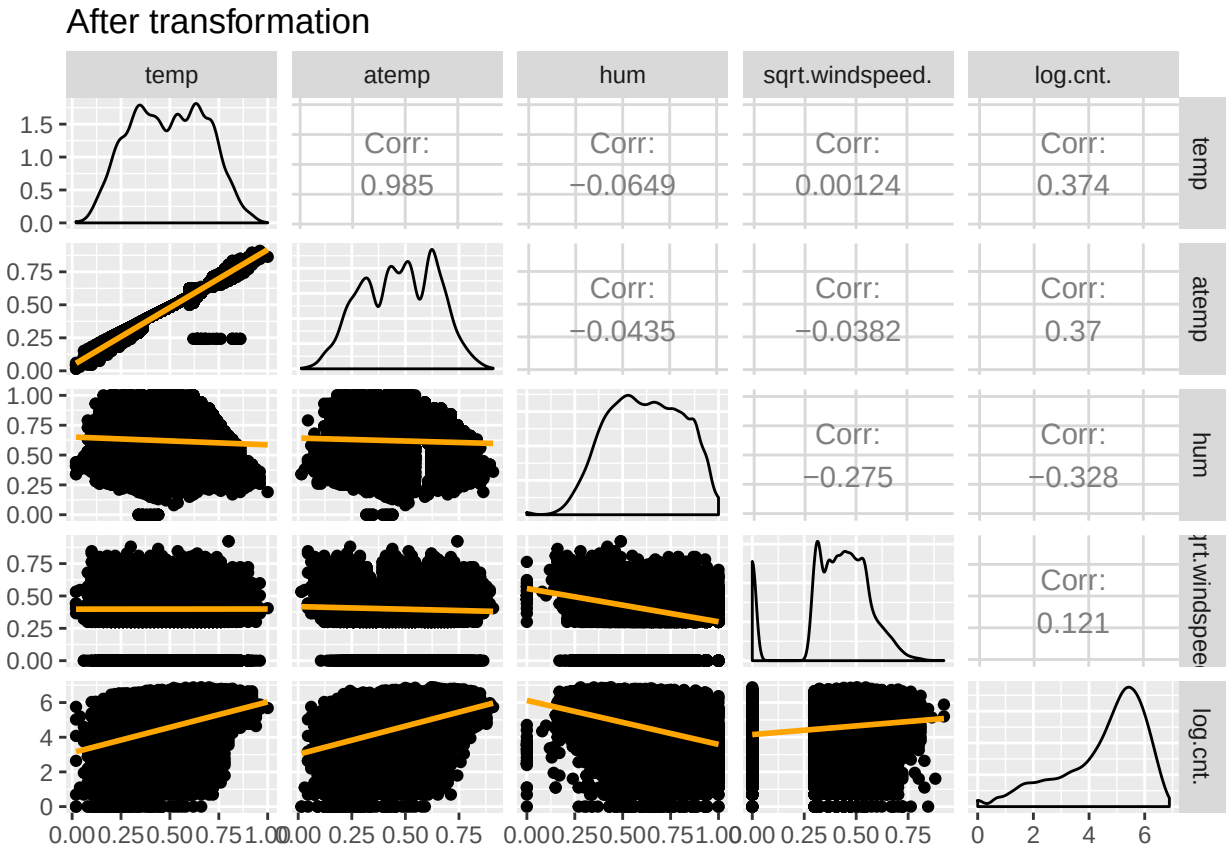
```
##   instant     dteday season yr mnth hr holiday weekday workingday
## 1       1 01-01-2011      1  0    1  0       0       6          0
## 2       2 01-01-2011      1  0    1  1       0       6          0
## 3       3 01-01-2011      1  0    1  2       0       6          0
## 4       4 01-01-2011      1  0    1  3       0       6          0
##   weathersit temp  atemp  hum windspeed casual registered cnt  type
## 1          1 0.24 0.2879 0.81         0      3         13  16 train
## 2          1 0.22 0.2727 0.80         0      8         32  40 train
## 3          1 0.22 0.2727 0.80         0      5         27  32 train
## 4          1 0.24 0.2879 0.75         0      3         10  13 train
```

Univariate/Bivariate Analysis of numerical variables

Let us perform initial univariate and bivariate analysis on the numerical variables to get an idea of how they are distributed and how they vary when taken against the number of bikes rented **cnt**.

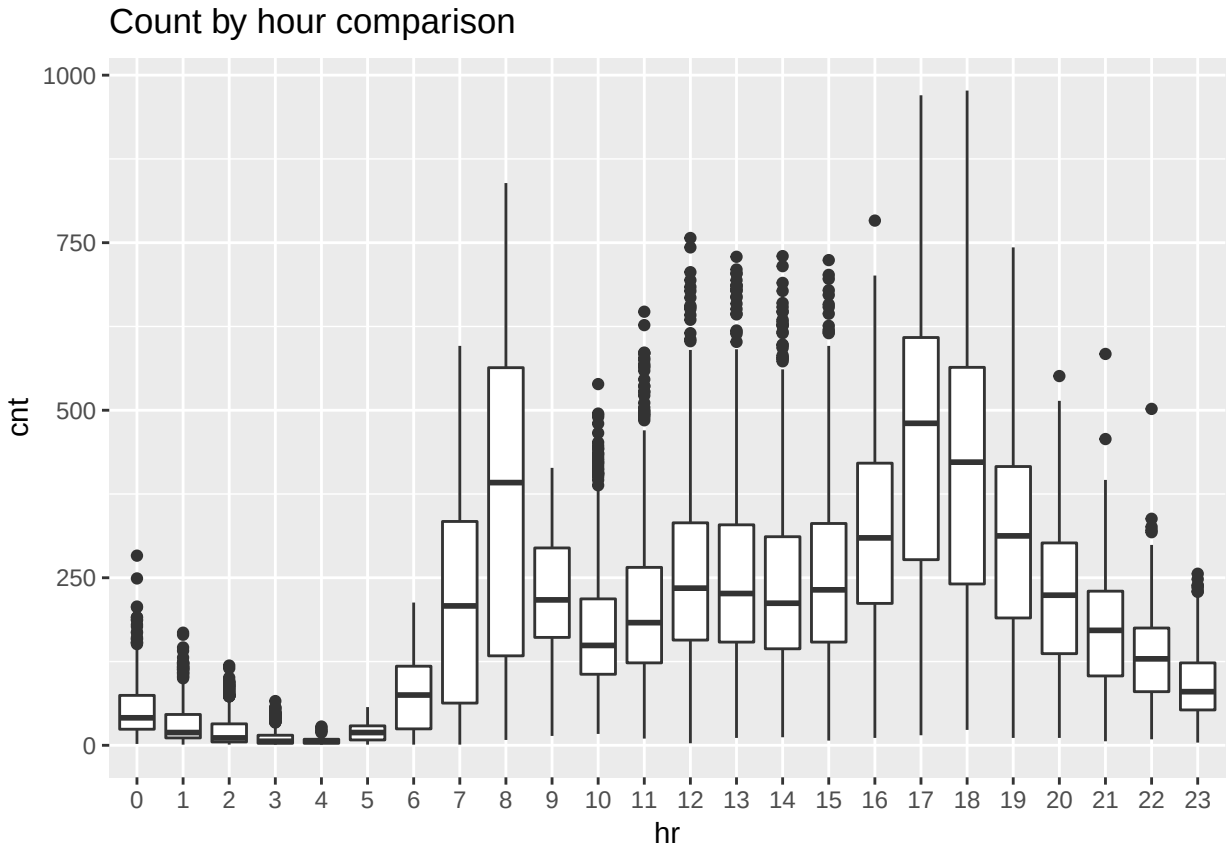


We see that there is significant skew in the density plots of **cnt** and windspeed. We perform a square root transformation on **windspeed** and a log transformation on **cnt**. We see that the linear relationship between the transformed count and the other numerical variables has not improved significantly. But we also see that the transformed variables have become less skewed and help greatly in our linear model as we will see later. We also see that the variables **temp** and **atemp** are very highly correlated and we will choose only one of them for our model.



Bivariate Analysis of count by hour of day

The boxplots below convey the information that we would normally expect by common sense. We see that bicycle rentals are very low during the dead hours of the night. The rentals are quite high around 7-8 AM in the morning and 5-6 PM in the evening, they are relatively low during the afternoon hours and steadily decrease as the day ends.



Trivariate Analysis of categorical variables

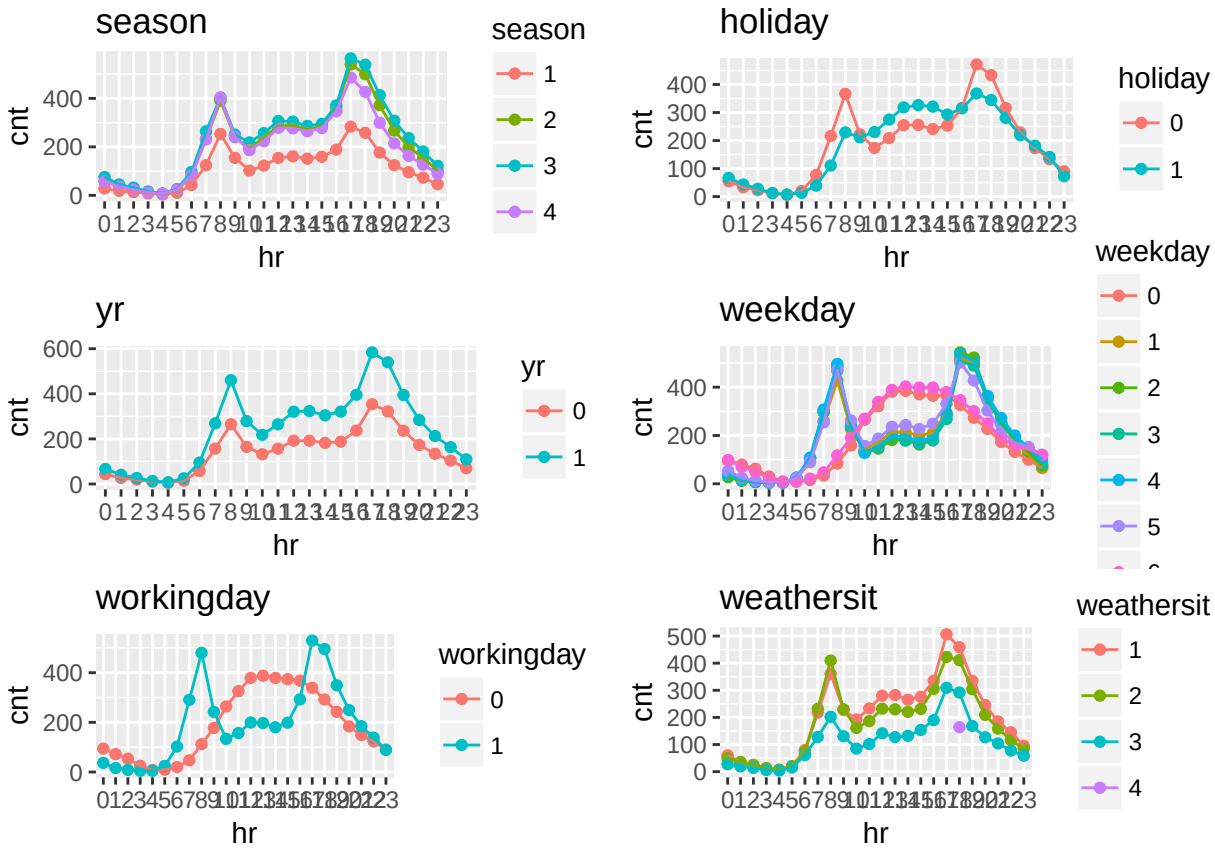
We plot graphs to observe how **cnt** changes averaged by **hr** for the various levels of each categorical variables. We observe that for some of the categorical variables there are significant changes in the plots within the levels of that variable. This tells us that there is some interaction between the hour of the day and the other variables like the season and the day of the week, rather than parallel plots which would occur in the case without interaction. This in addition to the next section greatly helps choose interactions.

Weekday and Hour

From the interaction plot of **weekday** vs **hr** in the trivariate analysis of categorical variables we see that the graphs are not parallel. This is an important interaction that we will add to our model.

Working Day and Hour

From the interaction plot of **workingday** vs **hr** in the trivariate analysis of categorical variables we see that the graphs are far from parallel. This is also a very important interaction that we will add to our model.

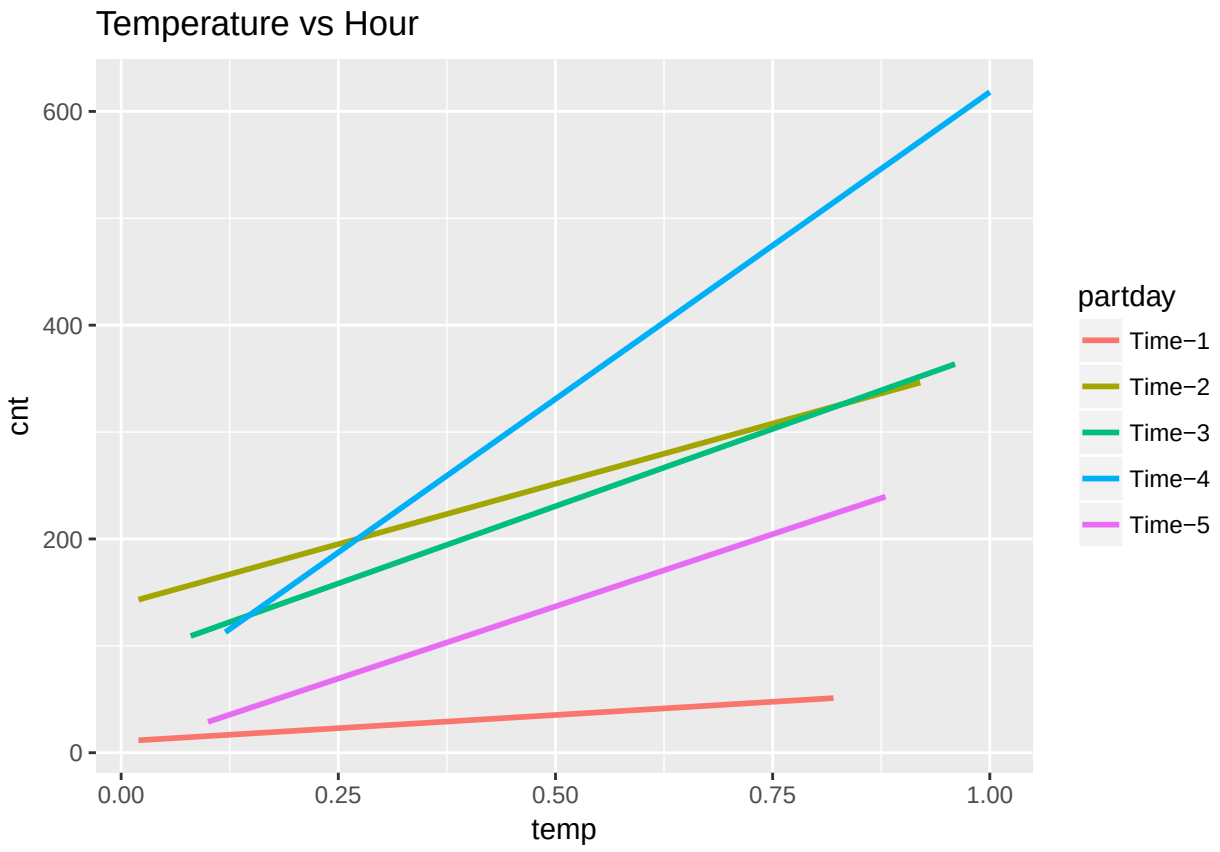


Multivariate Analysis

Here we consider the variation of the response variable **cnt** taken again two or more variables at a time. We show only some of the interesting interaction plots here given that we have a time, energy and space constraint. These plots help us determine if it is worthwhile adding these interaction terms to our linear regression model. We divide the interactions into two types, one containing interactions amongst the weather variables like **temp** and **hum** and the other containing interactions relating to the calendar like **workingday** and **weekday**. We use some intuition and common sense here in choosing the plots though we use a more rigorous method to choose interaction terms for the model.

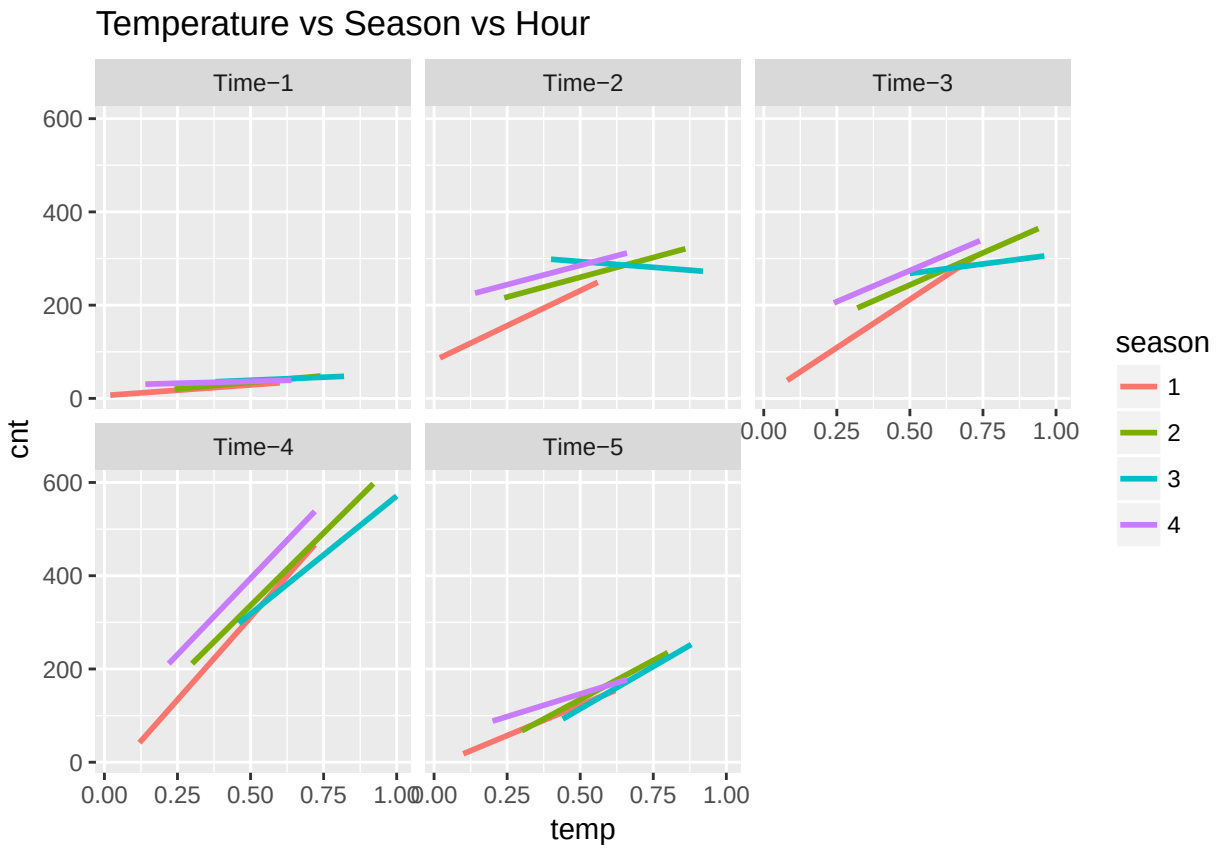
Temperature and Hour

We can clearly see that there are differing slopes across the lines for different **partday**, though not all. On the whole, if any interaction is present, it would be very weak.



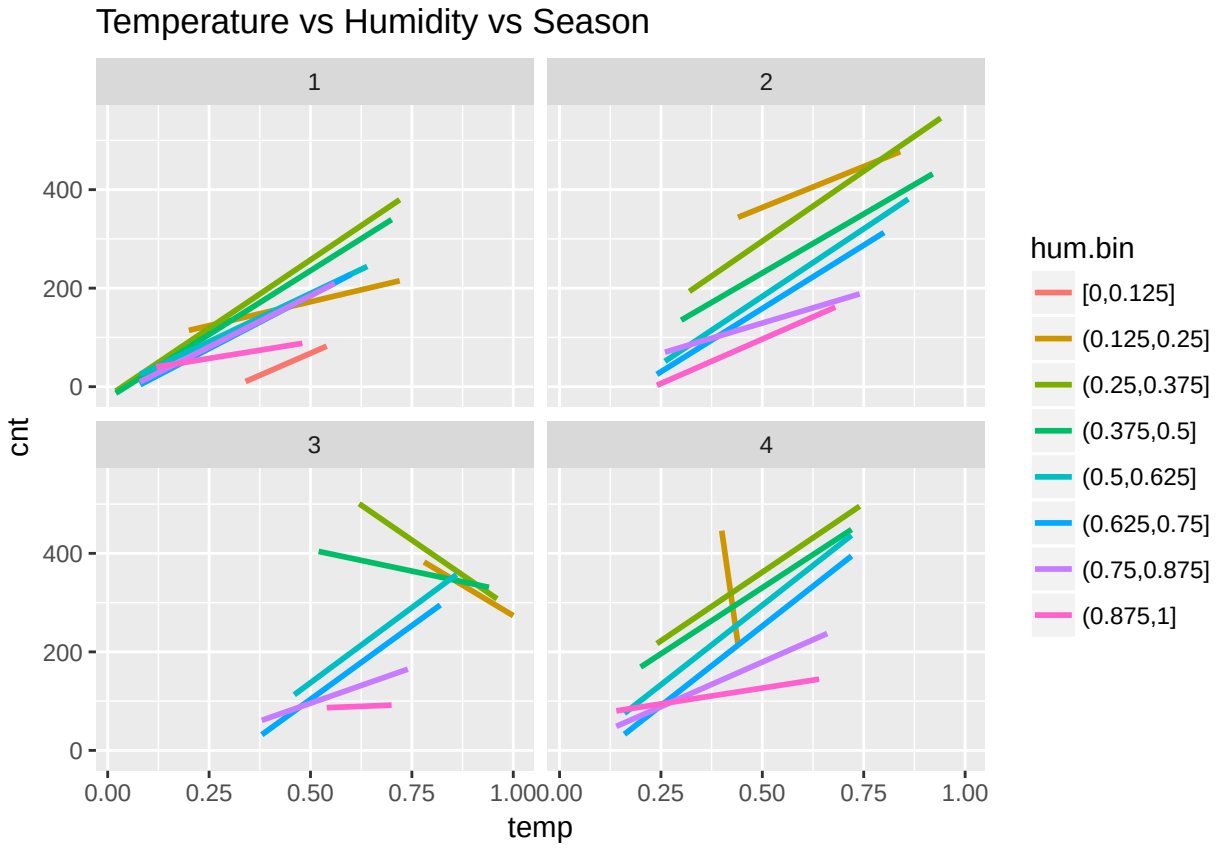
Temperature, Season and Hour

We take the various combination pairs of **partday** and **season** and construct fitted line plots with **temp**. We fix one categorical variable at one level and see if it varies while varying the other variable. We can see that there is a lot of variability in the slopes of the different lines. We will consider adding this interaction term.



Temperature, Humidity and Season

Similar to the above case, we find a lot of variation in slopes. There must be some interaction present and we will consider it in our model.



Multiple Linear Regression model with interactions

We build the model allowing for a maximum of three-way interactions with the transformed variables and a little bit of intuition and common sense and of course, our plots we had done earlier. We then use the Type II ANOVA tests based on the marginality principle to trim the regressor combination to size by eliminating insignificant interactions. We also use multiple **anova()** tests comparing a full model and a reduced model individually as the Type II ANOVA tests are not sufficient to delete some interactions, which we do not show here. We obtain a very good model fit. We limit ourselves to the three way interactions to avoid too much of complexity, and we have not included all possible three way interactions, but we have followed the marginality principle and have included all the required two-way interactions and main effects. There is some intuition behind how the three way interactions were selected.

Type II ANOVA results

Below are shown the Type II ANOVA results and variable selection in our model.

```
## Anova Table (Type II tests)
##
## Response: log(cnt)
##                          Sum Sq    Df   F value    Pr(>F)
## temp                      238.8     1 1991.1519 < 2.2e-16 ***
## hum                        12.9     1  107.7096 < 2.2e-16 ***
## sqrt(windspeed)             8.4     1   69.8702 < 2.2e-16 ***
```

```
## weathersit                    161.6     3  449.0487 < 2.2e-16 ***
## season                        413.6     3 1149.6781 < 2.2e-16 ***
## hr                          12706.4    23 4606.4596 < 2.2e-16 ***
## yr                            587.8     1 4901.1051 < 2.2e-16 ***
## weekday                        39.8     6   55.3567 < 2.2e-16 ***
## workingday                      0.9     1    7.1571 0.0074787 **
## temp:hum                        6.2     1   51.4877 7.696e-13 ***
## temp:sqrt(windspeed)            0.4     1    3.2633 0.0708744 .
## temp:season                    62.5     3  173.7371 < 2.2e-16 ***
## hum:season                     16.5     3   45.9248 < 2.2e-16 ***
## sqrt(windspeed):season          2.5     3    6.8994 0.0001228 ***
## season:hr                      20.9    69    2.5225 6.407e-11 ***
## sqrt(windspeed):hr              3.8    23    1.3624 0.1151247
## hum:hr                         13.6    23    4.9188 8.618e-14 ***
## hum:sqrt(windspeed)             0.1     1    0.7075 0.4003044
## hr:weekday                    170.1   138   10.2759 < 2.2e-16 ***
## hr:workingday                 110.3    23   39.9931 < 2.2e-16 ***
## temp:hum:season                14.4     3   40.0126 < 2.2e-16 ***
## temp:sqrt(windspeed):season     1.7     3    4.8285 0.0023237 **
## temp:hum:hr                     8.1    23    2.9497 2.749e-06 ***
## temp:season:hr                 15.6    69    1.8850 1.374e-05 ***
## hum:season:hr                  14.9    69    1.7957 6.066e-05 ***
## Residuals                    1246.1 10390
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing the fit using RMSLE

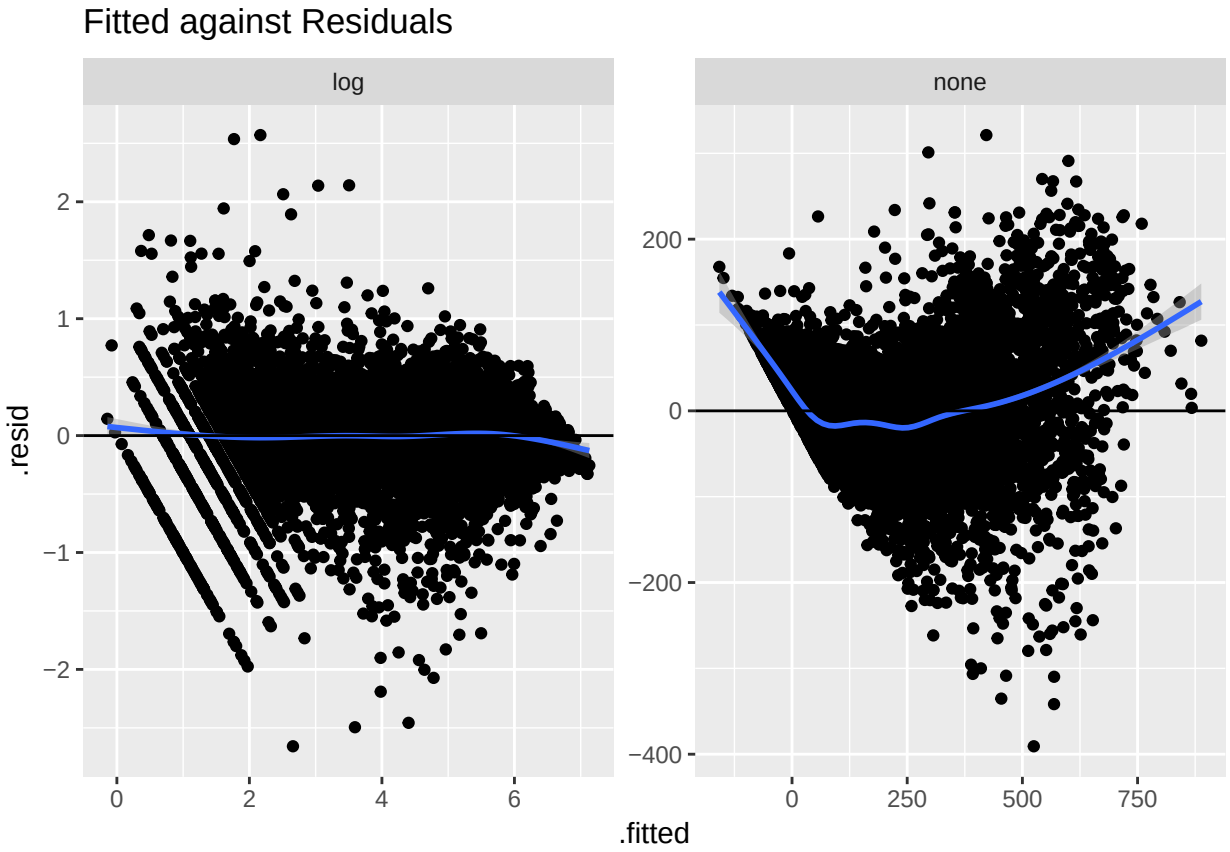
We use the Root Mean Square Log Error to check how well our model predicts on the test data. We see that the value is quite low, meaning that our model is predicting well. We also display the R-squared value of our linear model.

```
## [1] "R-squared: 0.948"
```

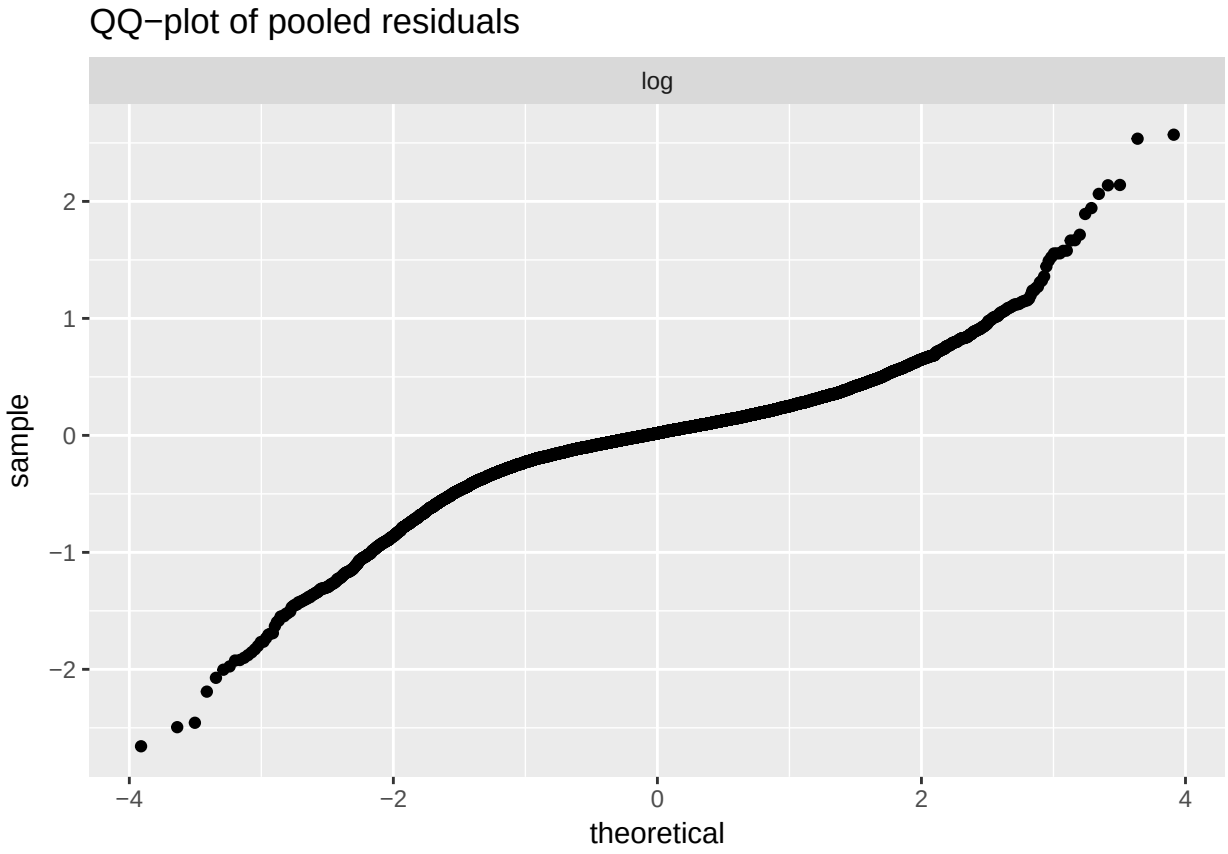
```
## [1] "RMSLE: 0.384"
```

Graphical evaluation of the model

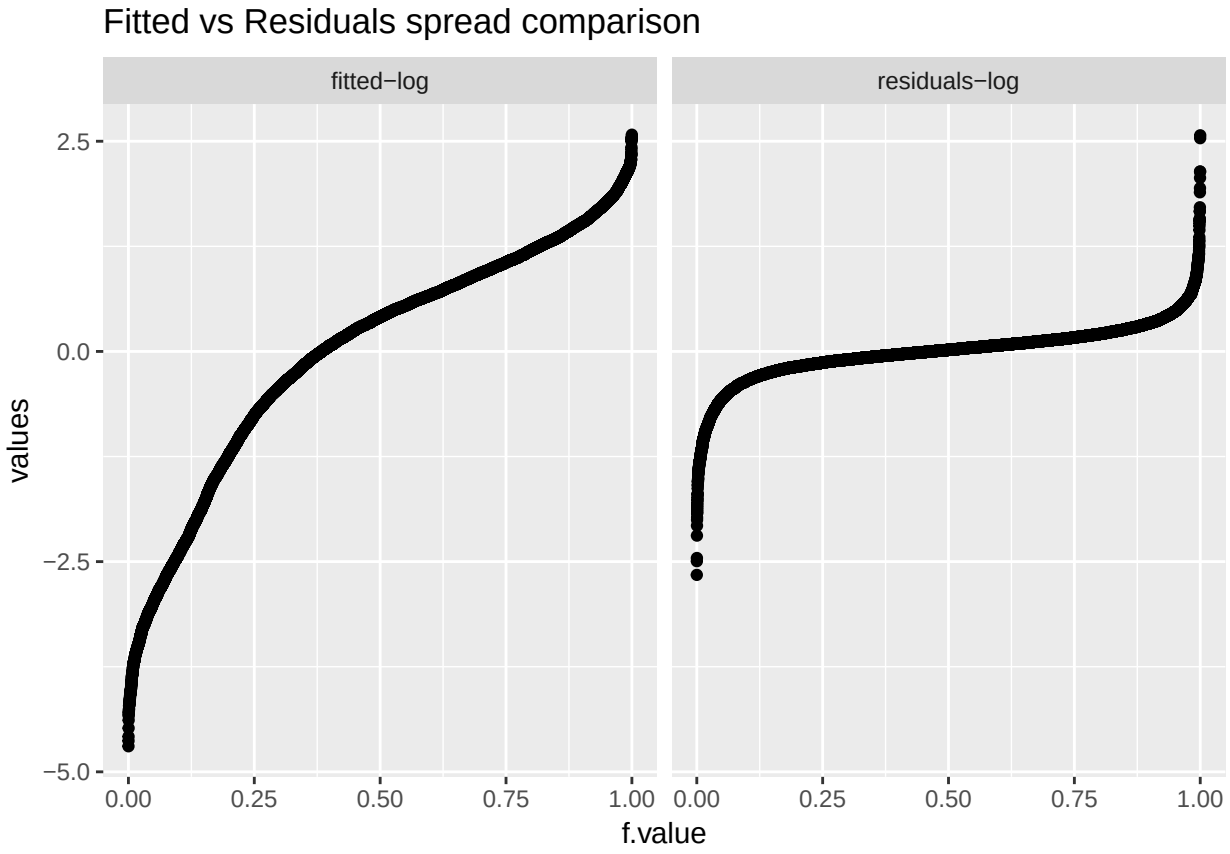
We plot the residual vs fitted plot to check for any violations of the linear regression assumptions. We draw two such plots one where we have used the log transformation on **cnt** and the other without it. We see that the loess curve in the log case engulfs a good part of the zero line, which is a good improvement after transformation. There is still a good amount of non-constant variance which is expected as the data is at a hourly level. The non-linear curve in the non-log case reflects too many model inadequacies, and the transformation significantly improves the plot.



We check for normality of residuals and we see that the residuals when pooled do not follow a normal distribution clearly seen from the QQ plots. But it can be considered as approximately normal and suits the data for prediction purposes though we would have preferred more normality as that helps with better prediction.



We plot the fitted values and residuals against a common scale to check if the model does well enough in explaining the variation in **cnt**. We see that fitted values are more spread out than the residuals. This means that our model performs very well in explaining the variance aroud the mean in the bike rented count.

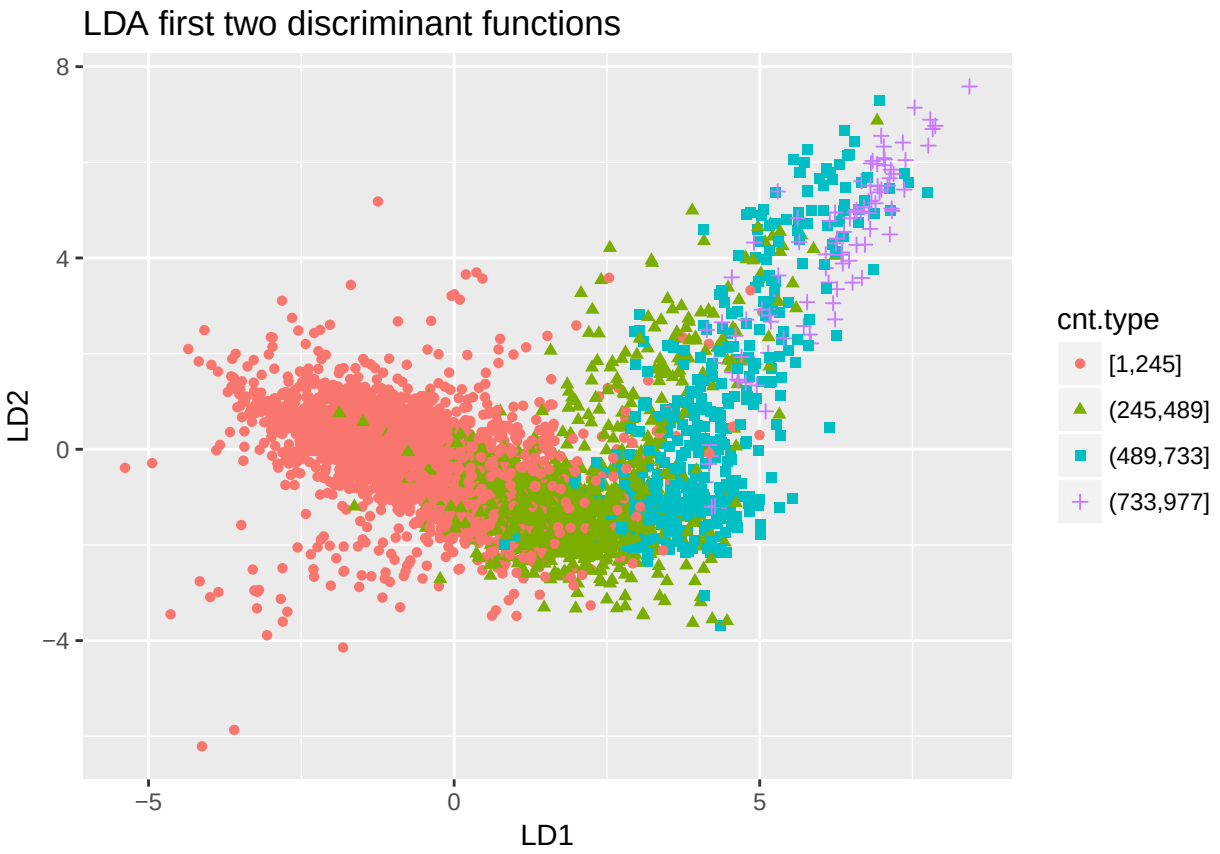


Major limitations of the linear model

Though our linear model is able to capture a lot of the variation, it is highly dependent on higher order interactions. We have to be careful in choosing the interactions as choosing too many of the higher order interactions blindly might lead to multicollinearity. We also run the risk of overfitting. Though we have not shown it here, we have run a Poisson Regression model and we faced the problem of overfitting as we got blown up predicted values using the same set of interactions.

Linear Discriminant Analysis

LDA provides a good graphical means to determine if the predictor variables are able to help distinguish between the various classes of the response varible. We bucket the variable **cnt** into four different range buckets and perform LDA to see if we achieve separation between the classes by making a plot of the two best discriminant functions against each other. We DO NOT intend to use the model to predict. Our aim is purely an investigative one, and we intend to use it to show visually that our chosen set of regressors are performing well.

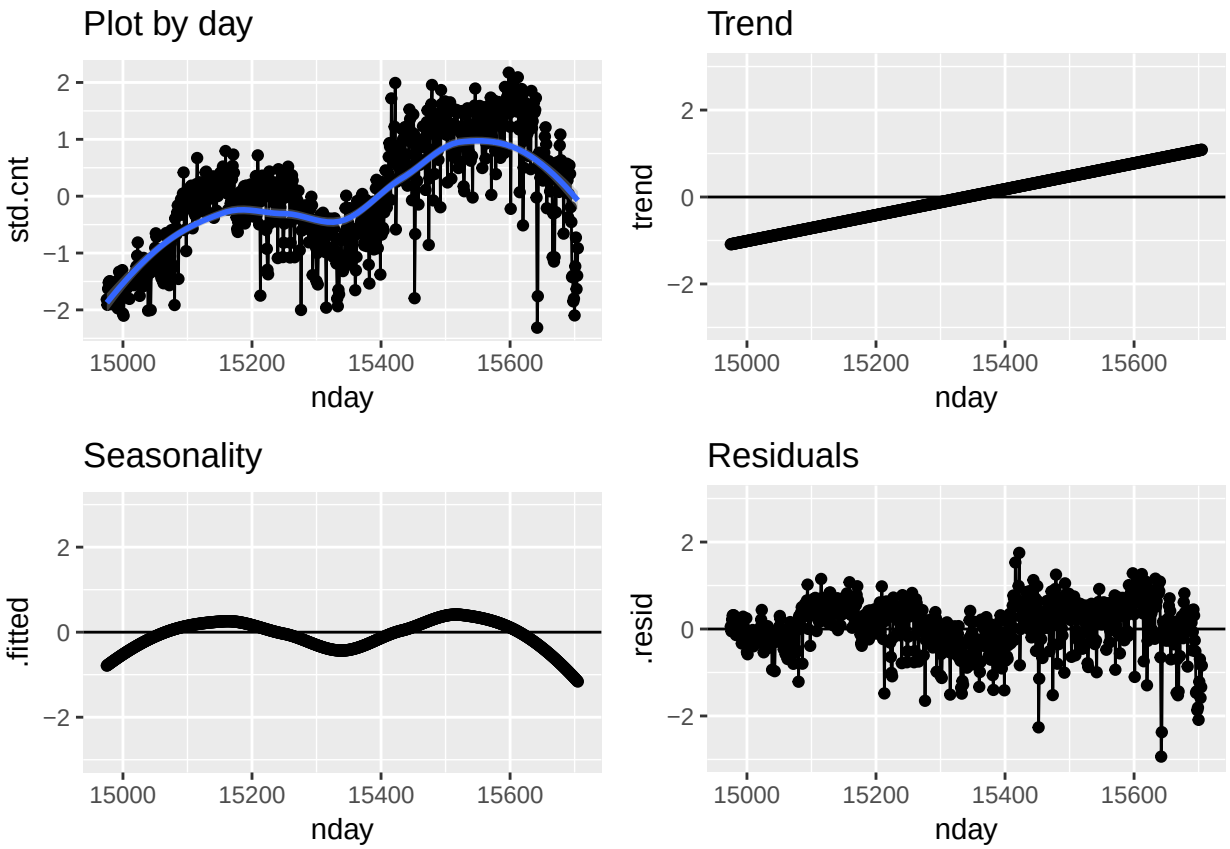


We see that we are able to achieve a decent level of separation from the plots of the best two discriminant functions. We also present the accuracy of the model just to quench your curiosity.

```
## [1] 0.8498383
```

Time components in the data

We shall try to extract the time components from the data to see to what extent a time series analysis is possible. We see from the extracted components below that it is an increasing time series and it is not difficult to figure out that there exists monthly seasonality. However the residual (noise) component of the variation is quite large with a visible pattern and looks like we may have to use some smoothing technique.



Future work

- We could try to improve our Linear Model by using Sandwich Estimators to correct for the non-constant variance.
- We could try to add polynomial weather terms to the model, though this might lead to more complexity.
- We could try to perform analysis of leverage points, influence points and outliers to further better undertand the linear model.
- ARIMA model: We could use a time series analysis to predict the count of rented bikes as our data is at a hour level.
- Decision Trees: Interactions play a huge role in our model. Decision Tree is a Machine Learning algorithm that naturally assumes interactions between the predictor variables and thus would be a good choice.
- A great problem in bike rental systems is when there is a sudden surge in the demand for bikes. We could bucket our extreme points and treat them as anomalies and try to perform Anomaly Detection to better try to understand the factors that are causing these huge bike rental demand values.

References

- STAT-S670 Course notes
- STAT-S631 Course notes