Lab Report 3: Multiple Linear Regression

MA 575 Fall 2021 - C3 Team #2

Ali Taqi, Hsin-Chang Lin, Huiru Yang, Ryan Mahoney, Yulin Li

10/25/2021

Contents

1 Introduction

In this lab report, Multiple Linear Regression (MLR) is performed on one response variable and a subset of predictors chosen from the 2011-2012 Bike Sharing dataset [1]. The dataset contains two main kinds of response variables of our concerns:

1. the count of **daily** bike rentals
2. the count of **hourly** bike rentals.

Within each kind of bike rental counts, the following 3 categories of rental counts are recorded:

1. the count of bike rentals by **casual** users
2. the count of bike rentals by **registered** users
3. the **total** count, which is the sum of casual count and registered count.

For simplicity, the **total** count of **daily** bike rentals is chosen as the response variable to be studied in this lab. The model should thus help answer the following question as mentioned in Lab Report 1:

- What are the **daily** bike rentals under different conditions? (Business owners may like to know the daily bike rentals in 2013 so that they could optimize the inventory to reduce costs, and they may also wonder whether it is worth leaving the bike-sharing system open on days with extreme weather conditions. This can be done by performing predictive modeling on the daily rental variable based on data given in 2011 and 2012.)

2 Preprocessing

2.1 Overview

Variable Interpretations (see [1])

Both hour.csv and day.csv have the following fields, except `hr` which is not available in bike-day.csv:

- `instant`: record index
- `dteday`: date
- `season`: season (1:spring, 2:summer, 3:fall, 4:winter)
- `yr`: year (0:2011, 1:2012)
- `mnth`: month (1 to 12)
- `hr`: hour (0 to 23)
- `holiday`: weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
- `weekday`: day of the week
- `workingday`: if day is neither weekend nor holiday is 1, otherwise is 0.
- `weather-sit`:
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- `temp`: Normalized temperature in Celsius. The values are divided to 41 (max)
- `atemp`: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- `hum`: Normalized humidity. The values are divided to 100 (max)
- `windspeed`: Normalized wind speed. The values are divided to 67 (max)
- `casual`: count of casual users
- `registered`: count of registered users
- `cnt`: count of total rental bikes including both casual and registered

```
# A brief look at the data structure from day.csv
head(bikedata, 3)
```

```
##   instant     dteday season yr mnth holiday weekday workingday weathersit
## 1       1 2011-01-01      1  0    1       0       6          0          2
## 2       2 2011-01-02      1  0    1       0       0          0          2
## 3       3 2011-01-03      1  0    1       0       1          1          1
##       temp    atemp      hum windspeed casual registered  cnt
## 1 0.344167 0.363625 0.805833  0.160446    331        654  985
## 2 0.363478 0.353739 0.696087  0.248539    131        670  801
## 3 0.196364 0.189405 0.437273  0.248309    120       1229 1349
```

2.2 Data Type & Value Conversion

Typically, all variables whose numerical values are not attached to actual physical meanings are treated as **categorical** variables.

```
# Boolean variables (from int to logical type)
holiday <- as.logical(bikedata$holiday)          #0 or 1
workingday <- as.logical(bikedata$workingday)    #0 or 1
```

```r
# Other categorical variables (from int to factor type)
season <- as.factor(bikedata$season)              #1 to 4
yr <- as.factor(bikedata$yr)                      #0 to 1
mnth <- as.factor(bikedata$mnth)                  #1 to 12
weekday <- as.factor(bikedata$weekday)            #0 to 6
weathersit <- as.factor(bikedata$weathersit)      #1 to 4
```

Note: Although `weathersit` (weather type) is a categorical variable, its numerical value (from 1 to 4) actually indicates a gradual change in the level of suitability for outdoor activities - the larger the number, the worse the weather condition (i.e., more fogs/rains/snows, see "Variable Interpretations" above).

Furthermore, the normalized weather condition measurements (see "Variable Interpretations" above) are also converted to their original values, so that the numerical values being used "make more sense" to us. This makes it easier for commonsense and real-life experience to be applied in later analysis.

```r
# Re-scale the normalized measurements
temp <- bikedata$temp * 41
atemp <- bikedata$atemp * 50
hum <- bikedata$hum * 100
windspeed <- bikedata$windspeed * 67
```

2.3 Visualization

2.3.1 Pairwise Relationships

Generally speaking, people's bike rental behaviors should be related to all seasonal and environmental factors, since all of them may to some degree affect people's willingness and ability to perform outdoor activities like biking. Therefore, all predictors should be taken into consideration, at least at the beginning.

Note the seasonal variables `weekday`, `workingday` and `holiday` can be viewed as roughly uncorrelated to all the environmental variables. Therefore, it would be fine to just have two separate pairs plots involving each of two uncorrelated groups of predictors, respectively (to reduce the number of variables in a single plot for better resolutions). Additionally, observe that the seasonal variables `season`, `yr` (year) and `mnth` (month) can be inferred from the variable `dteday` (day). Therefore, the two pairs plots can be made using variables grouped in the following way, to guarantee that all essential relationships can be seen in the plots:

Group 1: (Mainly) Environmental

Predictors (all environmental variables plus date):

- `dteday` (date)
- `weathersit` (weather type)
- `temp` (measured temperature)
- `atemp` (feeling temperature)
- `windspeed` (wind speed)
- `hum` (humidity)

Response:

- `cnt` (daily total count)

Group 2: Seasonal

Predictors (all seasonal variables):

- `dteday` (date)
- `season` (season)
- `yr` (year)
- `mnth` (month)
- `holiday` (holiday or not)
- `weekday` (day of the week)
- `workingday` (working day or not)

Response:

- `cnt` (daily total count)

2.3.1.1 Pairs Plot for Group 1 (Mainly Environmental)

Fig.1: Pairs Plot for Group 1

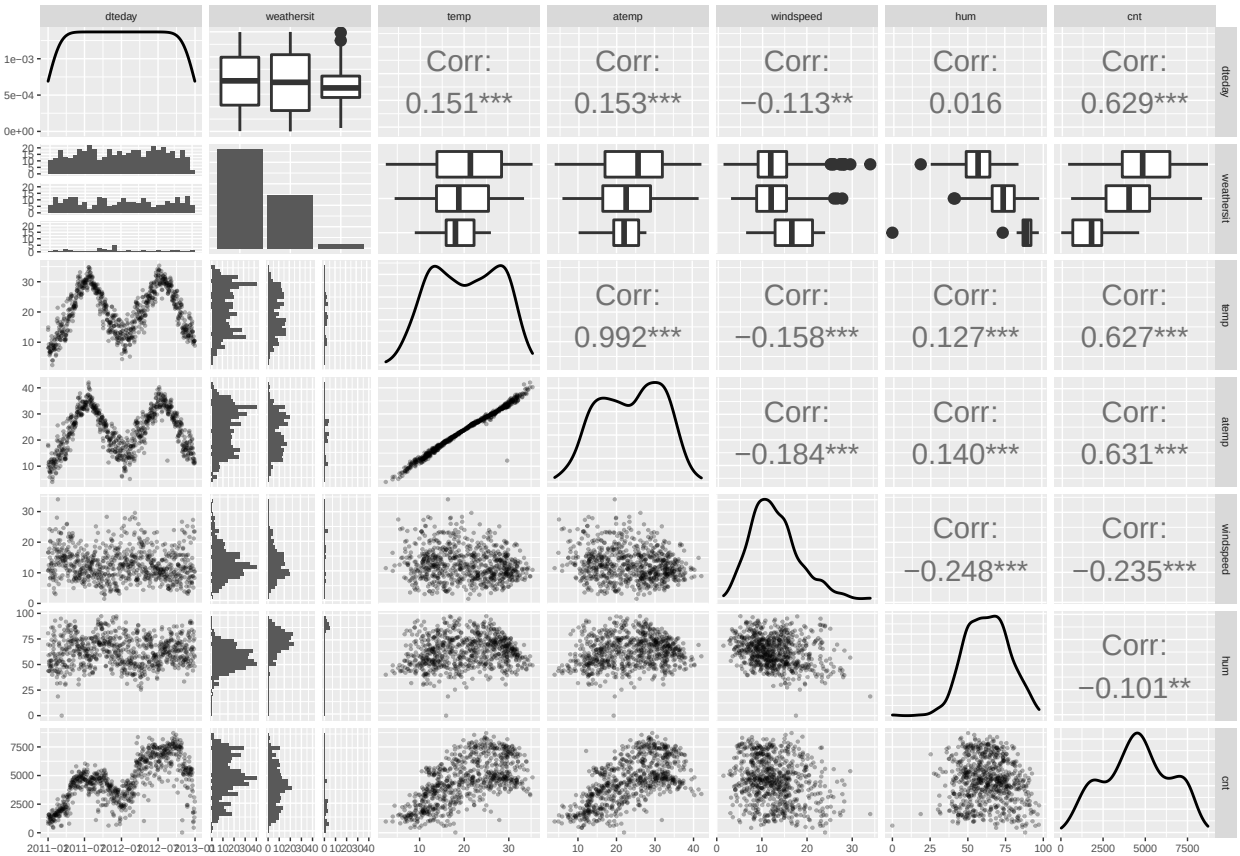


Fig.1 Column names in order: `dteday`, `weathersit`, `temp`, `atemp`, `windspeed`, `hum`, `cnt`

Observations:

Looking at the last row of the pairs plot (Fig.1), only date and temperatures seem to be significant. The corresponding correlations are also the highest (see subplot 1, 3 & 4 on the last column of Fig.1), indicating that there should be some linear relationship to be explored.

The two kinds of temperatures are highly correlated; to avoid co-linearity issues among the predictors in our model, only one of them should be kept. We choose to stick with the one having a higher correlation with rental counts, the feeling temperature, `atemp`.

Note that the significance of weather types is not immediately clear, since the distributions of rental counts under different weather types seem not to differ a lot (see subplot 2 on the last row of Fig.1), which is to our surprise. To take a closer look at the effect of weather types (`weathersit`), we choose to colored the pairs plot by it.

Fig.2: Pairs Plot for Group 1 (Colored by Weather Type)

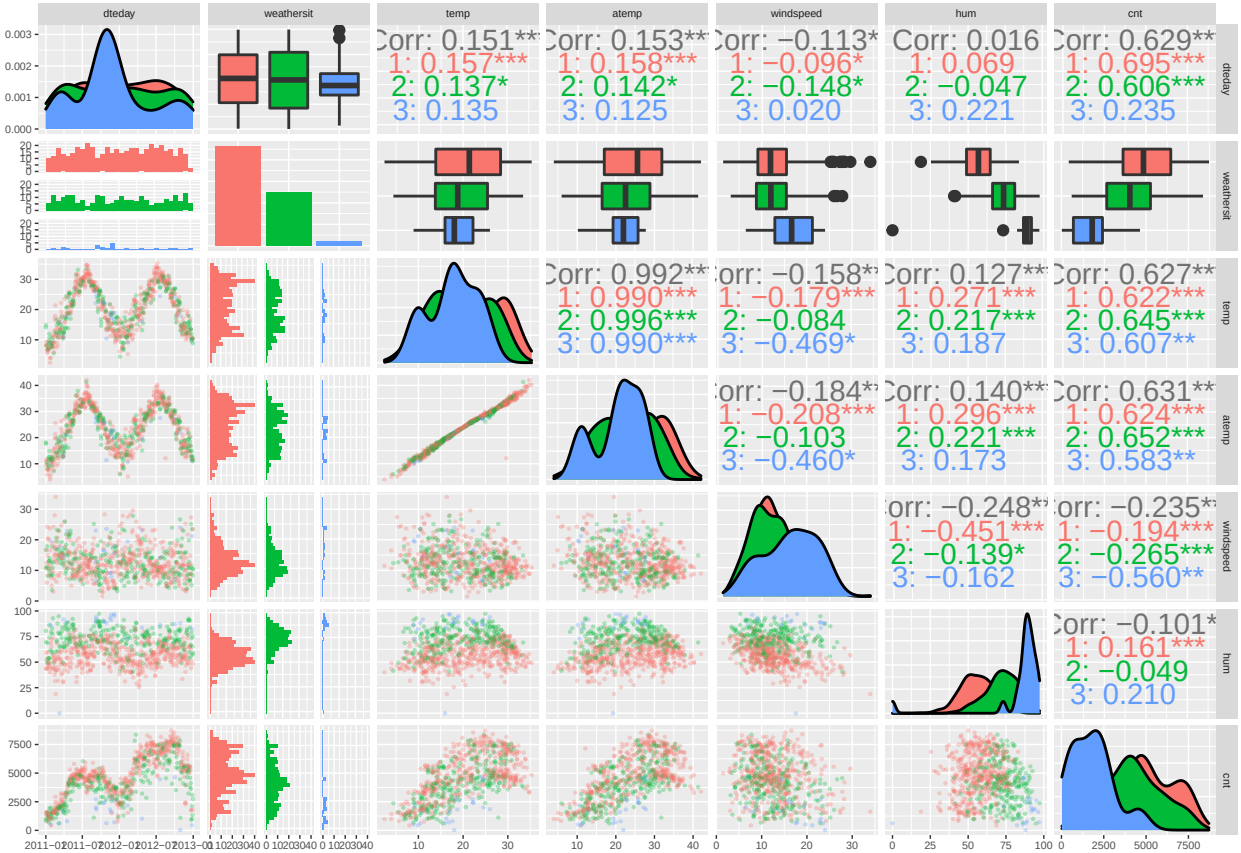


Fig.2 Column names in order: `dteday`, `weathersit`, `temp`, `atemp`, `windspeed`, `hum`, `cnt`

Observations:

From the first subplot on the last row of the above pairs plot (Fig.2), there does exist a separation of relatively higher and lower counts by weather type. This matches our commonsense - on more rainy days (weather type 2 & 3 indicated by the green and blue dots), people's biking activities do tend to less. This indicates that the weather type should be an useful predictor of bike rentals.

However, the scatter plots of rental counts against temperatures (subplot 3 & 4 on the last row of Fig.2) are still not clearly separated by weather types, which is again to our surprise. Remember that from the previous lab report [2], we have observed in these scatter plots the clustering of data points roughly into three curves at different levels of rental counts, which made us suspect the need of some categorical variable to separate the data. Now that the weather type variable does not play the role, we should attempt some other categorical variables.

After a few attempts, the desired categorical variable is found to be year, and the story starts to be clearer.

Fig.3: Pairs Plot for Group 1 (Colored by Year)

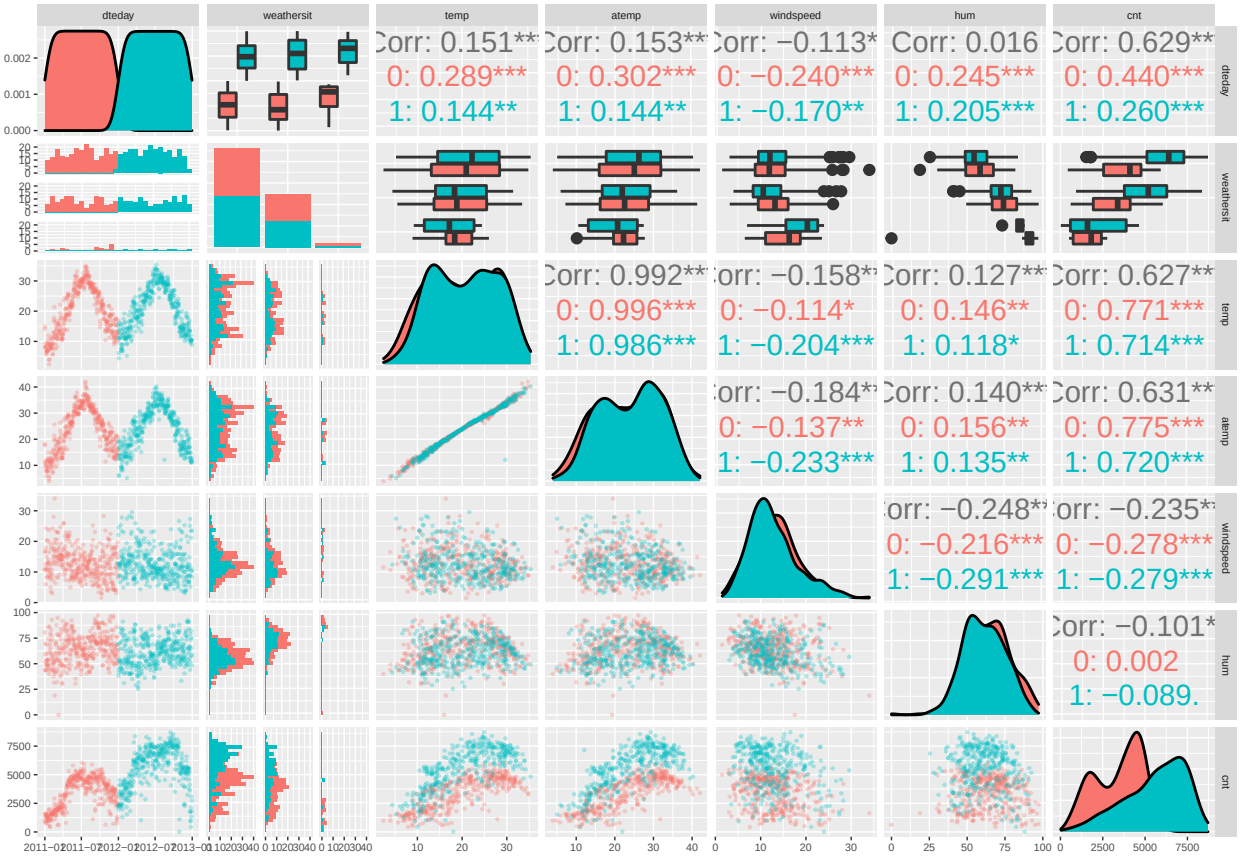


Fig.3 Column names in order: `dteday`, `weathersit`, `temp`, `atemp`, `windspeed`, `hum`, `cnt`

From the scatter-plot matrices (Fig.1-3), a time correlation is already clear. In fact, it is almost the strongest pattern, indicating the necessity of a model involving time. Before a time-series model is studied in this course, such time correlation can be utilized by noting that rental counts in the same year generally differ from counts in the other years by a similar offset, and that, similarly, rental counts in the same months also generally differ from counts in the other months by a common offset. We can then include additive terms of the categorical variables `yr` (year) and `mnth` (month) to capture the time correlation.

Furthermore, we should notice that the correlation of rental counts among different months have already been captured by another predictor, temperature. Also, the temperature variable must be highly correlated to the month variable. Then, to avoid co-linearity issues, only one of them can be kept, and in this case we should stick to temperature, because it better explains the nature of the underlying reason for that variation of bike rental counts: it is the body sensations instead of the number of months that ultimately affect people's willingness for outdoor activities. The same reasoning applies when comparing between the temperature variable and the season variable.

Conclusion:

From the above reasoning, the predictors to be kept in this group of variables should be:

- **'weathersit' (weather type)**
- **'atemp' (feeling temperature)**

The predictors `windspeed` (wind speed) and `hum` (humidity) should be dropped not only because they do not exhibit strong patterns and correlations with bike rental counts, but also because they both have a strong correlation with an existing predictor, 'weathersit' (weather type).

2.3.1.2 Pairs Plot for Group 2 (Seasonal)

Fig.4: Pairs Plot for Group 2

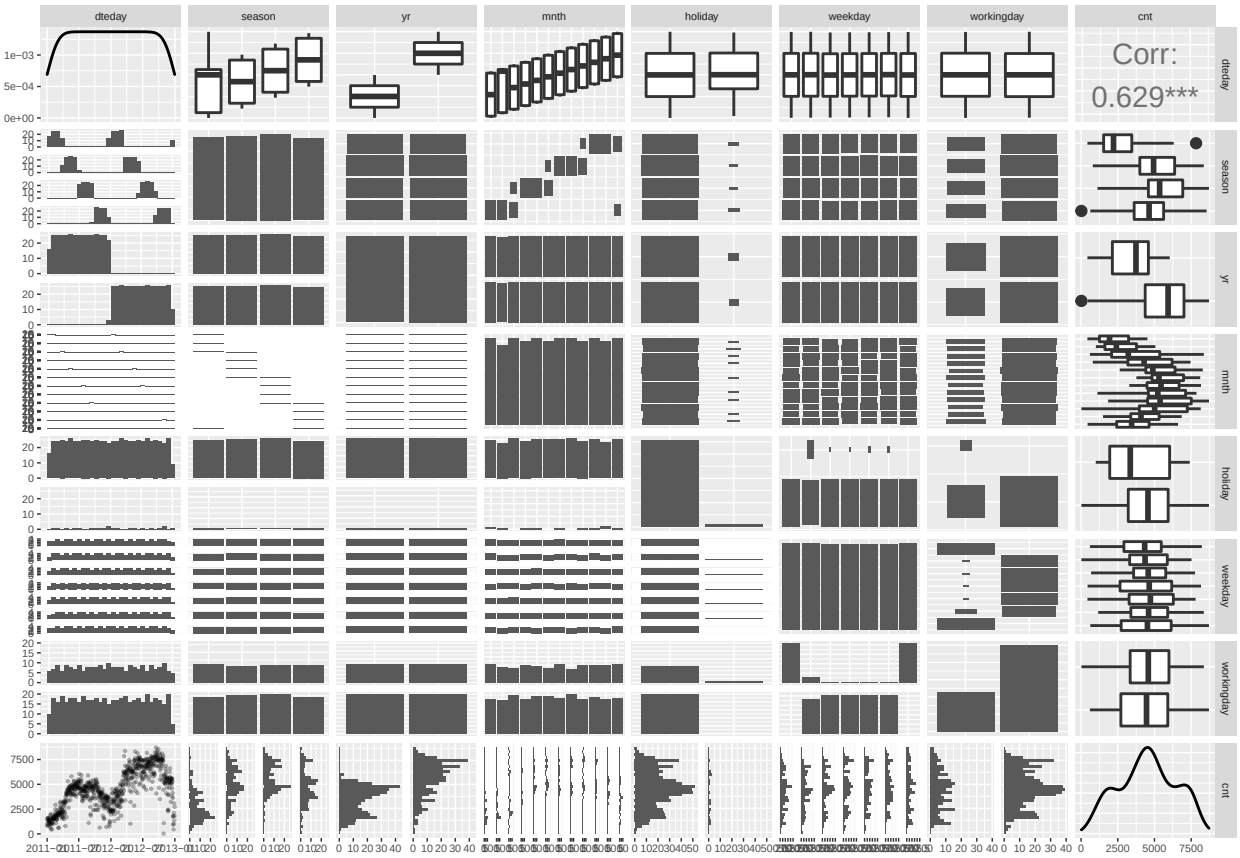


Fig.4 Column names in order: `dteday`, `season`, `yr`, `mnth`, `holiday`, `weekday`, `workingday`, `cnt`

Observations:

Most predictors in this group are categorical. To see if they are useful / significant predictors, we mainly look at how large the differences are between the average response values within different categories.

Looking at the last column of the pairs plot (Fig.4), significant predictors seem to include:

- `dteday` (date)
- `season` (season)
- `yr` (year)
- `mnth` (month)
- `holiday` (holiday or not).

Surprisingly, the weekdays and working days do not seem influential, at least for now.

Removing from the above list the variables that have strong correlations to the predictors already chosen from Group 1, we are left with the following predictors:

- **`yr` (year)**
- **`holiday` (holiday or not)**.

2.3.2 Time Variation

To further confirm the effect of the categorical variables `weekday` and `workingday`, we plot the rental counts by date and color the data points that are NOT working days (i.e., `workingday == FALSE`) (Fig.5,6), since that helps us identify the positions of weekends, which are included in non-working days, and thus we can infer from the plot where the other weekdays are. The holidays are also colored to help exclude the effect of holidays.

Observations:

Zooming in (Fig.5,6), a phenomenon is observed:

Many times, the rental count rises at the beginning of a week, reaches a maximum around the middle of the week and then falls down as the week approaches its end. This implies that the weekday variable could still be of some use.

We choose to build our MLS model first with the chosen predictors from the previous sections, and see what difference the weekday and working-day predictors can made to that model afterwards.

Fig.5: 2011 Total Count of Daily Bike Rentals by Date (with Trends)

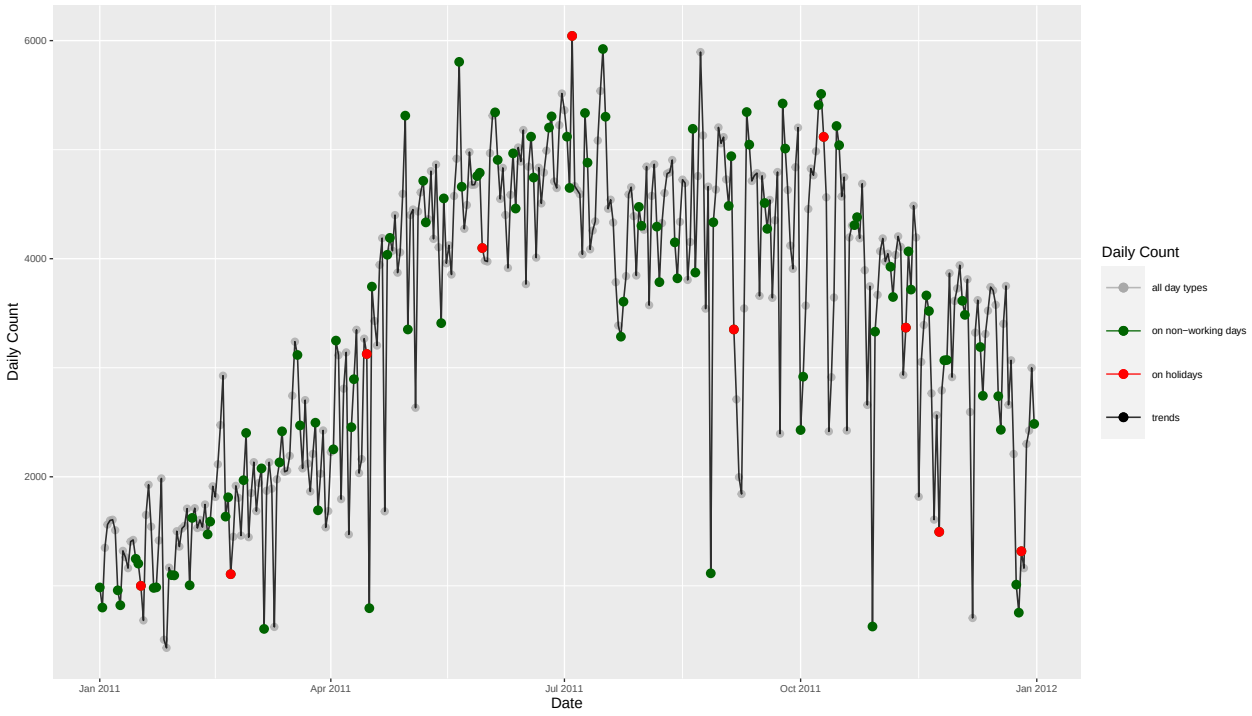
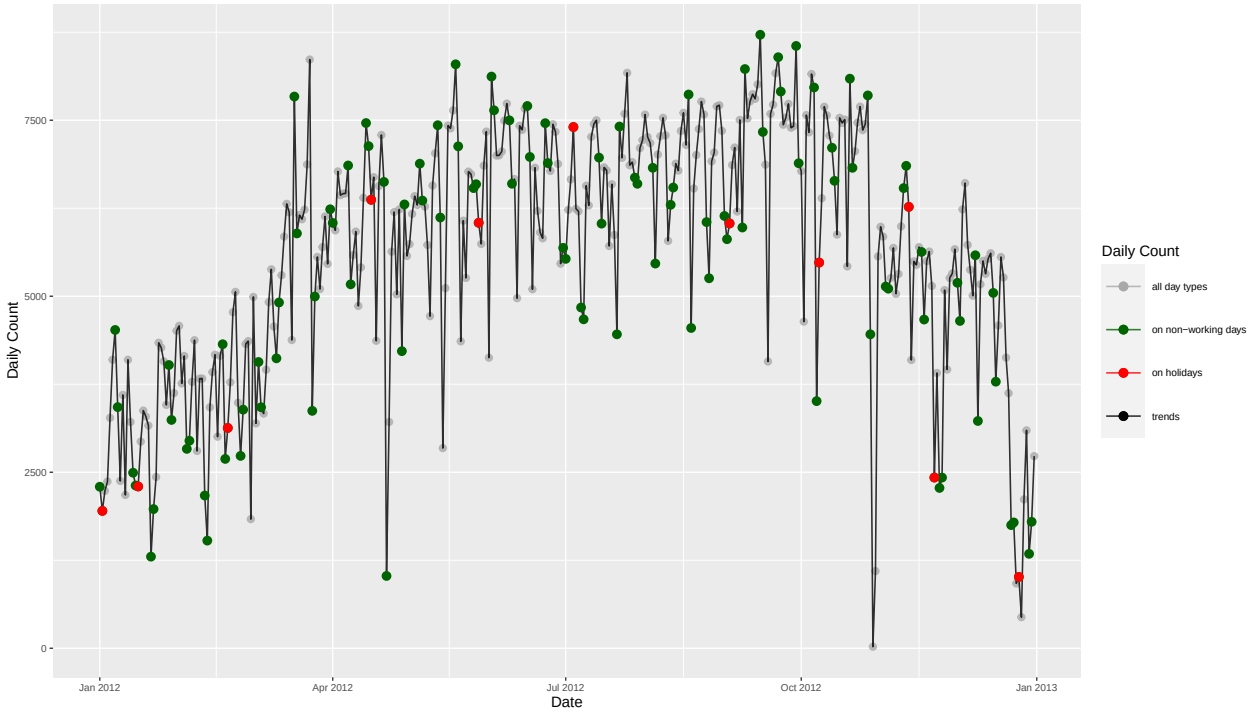


Fig.6: 2012 Total Count of Daily Bike Rentals by Date (with Trends)



3 Multiple Linear Regression (MLR) Modeling

3.1 Model Building

3.1.1 Model 1: Linear in Temperature

```
# MLR Model 1: Linear in Temperature
m.mlr.1 <- lm(cnt ~ yr + holiday + weathersit + atemp, data = bikedata)
summary(m.mlr.1)
```

```
##
## Call:
## lm(formula = cnt ~ yr + holiday + weathersit + atemp, data = bikedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4128.5  -617.1    13.0   747.3  2911.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1233.58     163.44   7.548 1.33e-13 ***
## yr           2054.20      74.65  27.517  < 2e-16 ***
## holiday      -704.05     223.29  -3.153  0.00168 **
## weathersit   -722.18      69.04 -10.461  < 2e-16 ***
## atemp        6893.97     230.79  29.871  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1007 on 726 degrees of freedom
## Multiple R-squared:  0.7312, Adjusted R-squared:  0.7297
## F-statistic: 493.7 on 4 and 726 DF,  p-value: < 2.2e-16
```

3.1.2 Model 2: Quadratic in Temperature

```
# MLR Model 2: Quadratic in Temperature
m.mlr.2 <- lm(cnt ~ yr + holiday + weathersit + atemp + I(atemp^2), data = bikedata)
summary(m.mlr.2)
```

```
##
## Call:
## lm(formula = cnt ~ yr + holiday + weathersit + atemp + I(atemp^2),
##     data = bikedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4326.5  -553.0    14.8   618.2  3254.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  -1626.76     270.62  -6.011 2.92e-09 ***
## yr            1989.29      67.83  29.328  < 2e-16 ***
## holiday       -628.92     202.39  -3.108  0.00196 **
## weathersit    -843.17      63.28 -13.325  < 2e-16 ***
## atemp        22005.64    1214.84  18.114  < 2e-16 ***
## I(atemp^2)  -16335.02    1293.59 -12.628  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 912.4 on 725 degrees of freedom
## Multiple R-squared:  0.7797, Adjusted R-squared:  0.7781
## F-statistic: 513.1 on 5 and 725 DF,  p-value: < 2.2e-16
```

3.1.3 Model 3: Cubic in Temperature

```
# MLR Model 3: Cubic in Temperature
m.mlr.3 <- lm(cnt ~ yr + holiday + weathersit + atemp + I(atemp^2) + I(atemp^3), data = bikedata)
summary(m.mlr.3)
```

```
##
## Call:
## lm(formula = cnt ~ yr + holiday + weathersit + atemp + I(atemp^2) +
##     I(atemp^3), data = bikedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4257.6  -511.1    -3.1   536.0  3320.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2201.53     519.93   4.234 2.59e-05 ***
## yr            2002.95      64.75  30.932  < 2e-16 ***
## holiday       -600.92     193.18  -3.111  0.00194 **
## weathersit    -847.83      60.39 -14.039  < 2e-16 ***
## atemp        -8593.12    3788.52  -2.268  0.02361 *
## I(atemp^2)   56203.97    8639.01   6.506 1.44e-10 ***
## I(atemp^3)  -52537.41    6192.71  -8.484  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 870.8 on 724 degrees of freedom
## Multiple R-squared:  0.7996, Adjusted R-squared:  0.7979
## F-statistic: 481.4 on 6 and 724 DF,  p-value: < 2.2e-16
```

3.1.4 MLR Model 4: Quartic in Temperature

```
# MLR Model 4: Quartic in Temperature
m.mlr.4 <- lm(cnt ~ yr + holiday + weathersit + atemp + I(atemp^2) + I(atemp^3) + I(atemp^4),
```

```
              data = bikedata)
summary(m.mlr.4)
```

```
##
## Call:
## lm(formula = cnt ~ yr + holiday + weathersit + atemp + I(atemp^2) +
##     I(atemp^3) + I(atemp^4), data = bikedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4251.3  -512.5    -2.9   534.1  3312.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2026.75     953.53   2.126  0.03388 *
## yr            2002.70      64.81  30.902  < 2e-16 ***
## holiday       -603.32     193.62  -3.116  0.00191 **
## weathersit    -847.49      60.45 -14.019  < 2e-16 ***
## atemp        -6553.98   10063.87  -0.651  0.51510
## I(atemp^2)   48284.35   37224.51   1.297  0.19501
## I(atemp^3)  -40125.41   57082.35  -0.703  0.48232
## I(atemp^4)   -6763.71   30922.25  -0.219  0.82692
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 871.4 on 723 degrees of freedom
## Multiple R-squared:  0.7996, Adjusted R-squared:  0.7977
## F-statistic: 412.1 on 7 and 723 DF,  p-value: < 2.2e-16
```

The quartic model (Model 4) ends up having all temperature terms insignificant, presumably because the influence of each individual term in the polynomial are being evened out by other terms having a similar effect (e.g., the 4th order term is able to cover some variation in the form of a 3rd-order polynomial, and the linear term is similar to the 3rd-order term in that both functions are odd).

We can verify this guess by removing one of the lower-order terms, say, the cubic term:

```
# MLR Model 4_1: Quartic in Temperature
m.mlr.4_1 <- lm(cnt ~ yr + holiday + weathersit + atemp + I(atemp^2) + I(atemp^4),
                data = bikedata)
summary(m.mlr.4_1)
```

```
##
## Call:
## lm(formula = cnt ~ yr + holiday + weathersit + atemp + I(atemp^2) +
##     I(atemp^4), data = bikedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4232.1  -513.1    -4.8   543.3  3287.4
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1433.91     444.74   3.224  0.00132 **
## yr            2001.75      64.77  30.905  < 2e-16 ***
## holiday       -611.25     193.23  -3.163  0.00162 **
## weathersit    -846.36      60.41 -14.011  < 2e-16 ***
## atemp          235.88    2823.99   0.084  0.93346
## I(atemp^2)   22330.68    4737.05   4.714 2.91e-06 ***
## I(atemp^4)  -28371.71    3355.71  -8.455  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 871.1 on 724 degrees of freedom
## Multiple R-squared:  0.7995, Adjusted R-squared:  0.7978
## F-statistic: 481.1 on 6 and 724 DF,  p-value: < 2.2e-16
```

Alternatively, removing the quadratic term also does the job (actually even better):

```
# MLR Model 4_2: Quartic in Temperature
m.mlr.4_2 <- lm(cnt ~ yr + holiday + weathersit + atemp + I(atemp^3) + I(atemp^4),
                data = bikedata)
summary(m.mlr.4_2)
```

```
##
## Call:
## lm(formula = cnt ~ yr + holiday + weathersit + atemp + I(atemp^3) +
##     I(atemp^4), data = bikedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4218.5  -515.4    -7.6   543.7  3267.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    874.66     347.03   2.520  0.01194 *
## yr            2000.52      64.82  30.864  < 2e-16 ***
## holiday       -617.91     193.39  -3.195  0.00146 **
## weathersit    -845.75      60.46 -13.987  < 2e-16 ***
## atemp         6333.01    1605.26   3.945 8.75e-05 ***
## I(atemp^3)   33314.39    7270.05   4.582 5.41e-06 ***
## I(atemp^4)  -45776.75    7184.50  -6.372 3.33e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 871.8 on 724 degrees of freedom
## Multiple R-squared:  0.7991, Adjusted R-squared:  0.7975
## F-statistic: 480.1 on 6 and 724 DF,  p-value: < 2.2e-16
```

3.1.5 MLR Model 3: Additional Models

Inspired by the result of the quartic models, we should also try removing one of the lower-order terms in the cubic model (Model 3):

```
# MLR Model 3_1: Cubic in Temperature (dropping one less-significant lower-order term)
m.mlr.3_1 <- lm(cnt ~ yr + holiday + weathersit + I(atemp^2) + I(atemp^3),
                data = bikedata)
summary(m.mlr.3_1)
```

```
##
## Call:
## lm(formula = cnt ~ yr + holiday + weathersit + I(atemp^2) + I(atemp^3),
##     data = bikedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4285.7  -515.1    19.7   550.0  3327.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1066.92     142.19   7.503 1.83e-13 ***
## yr            1995.81      64.86  30.770  < 2e-16 ***
## holiday       -603.31     193.73  -3.114  0.00192 **
## weathersit    -852.04      60.54 -14.075  < 2e-16 ***
## I(atemp^2)   36896.40    1478.50  24.955  < 2e-16 ***
## I(atemp^3)  -39165.03    1900.58 -20.607  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 873.3 on 725 degrees of freedom
## Multiple R-squared:  0.7982, Adjusted R-squared:  0.7968
## F-statistic: 573.4 on 5 and 725 DF,  p-value: < 2.2e-16
```

```
# MLR Model 3_2: Cubic in Temperature (dropping another lower-order term)
m.mlr.3_2 <- lm(cnt ~ yr + holiday + weathersit + atemp + I(atemp^3), data = bikedata)
summary(m.mlr.3_2)
```

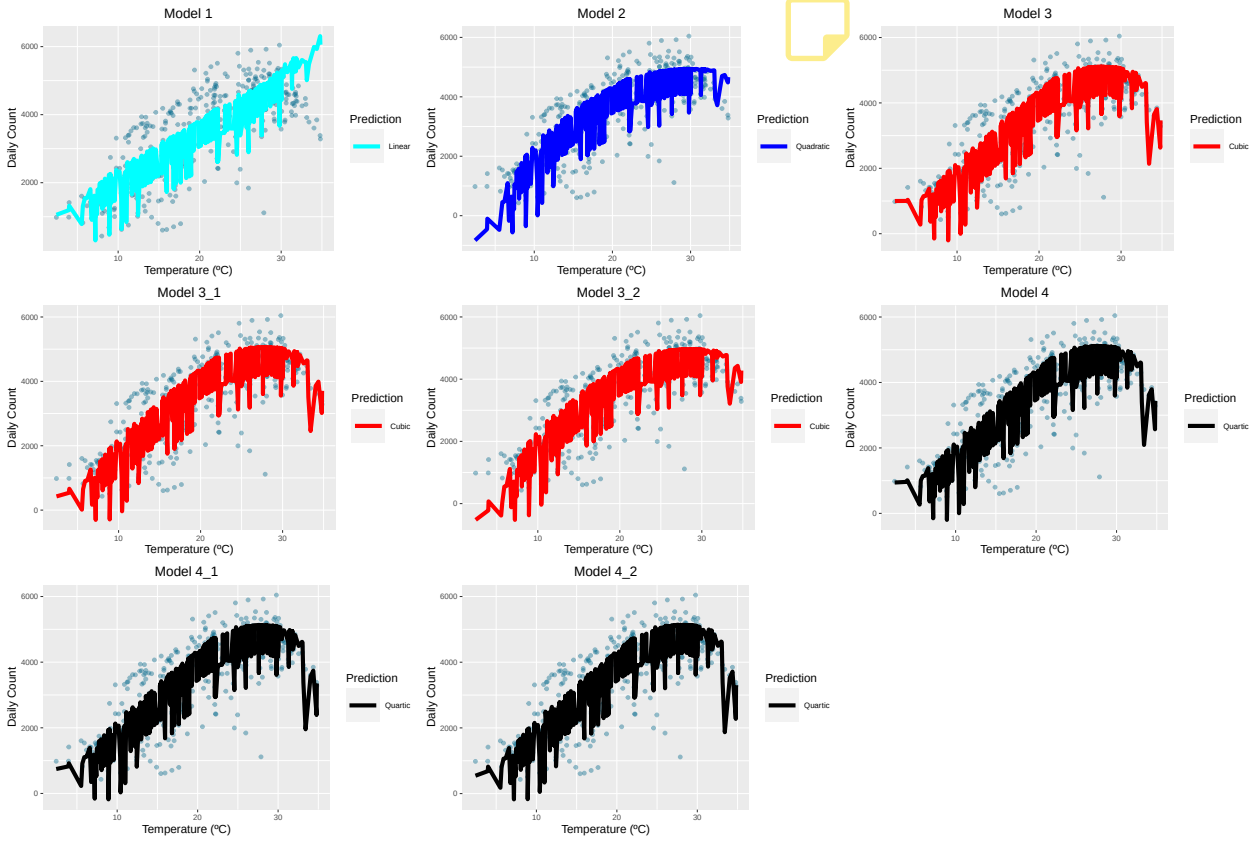
```
##
## Call:
## lm(formula = cnt ~ yr + holiday + weathersit + atemp + I(atemp^3),
##     data = bikedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4323.8  -522.7    29.3   582.3  3294.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -905.08     211.49  -4.280 2.12e-05 ***
```

```
## yr            1988.03      66.53  29.880  < 2e-16 ***
## holiday       -616.90     198.60  -3.106  0.00197 **
## weathersit    -852.80      62.08 -13.736  < 2e-16 ***
## atemp        15692.79     664.70  23.609  < 2e-16 ***
## I(atemp^3)  -12662.19     909.85 -13.917  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 895.3 on 725 degrees of freedom
## Multiple R-squared:  0.7879, Adjusted R-squared:  0.7864
## F-statistic: 538.6 on 5 and 725 DF,  p-value: < 2.2e-16
```

3.2 Model Selection

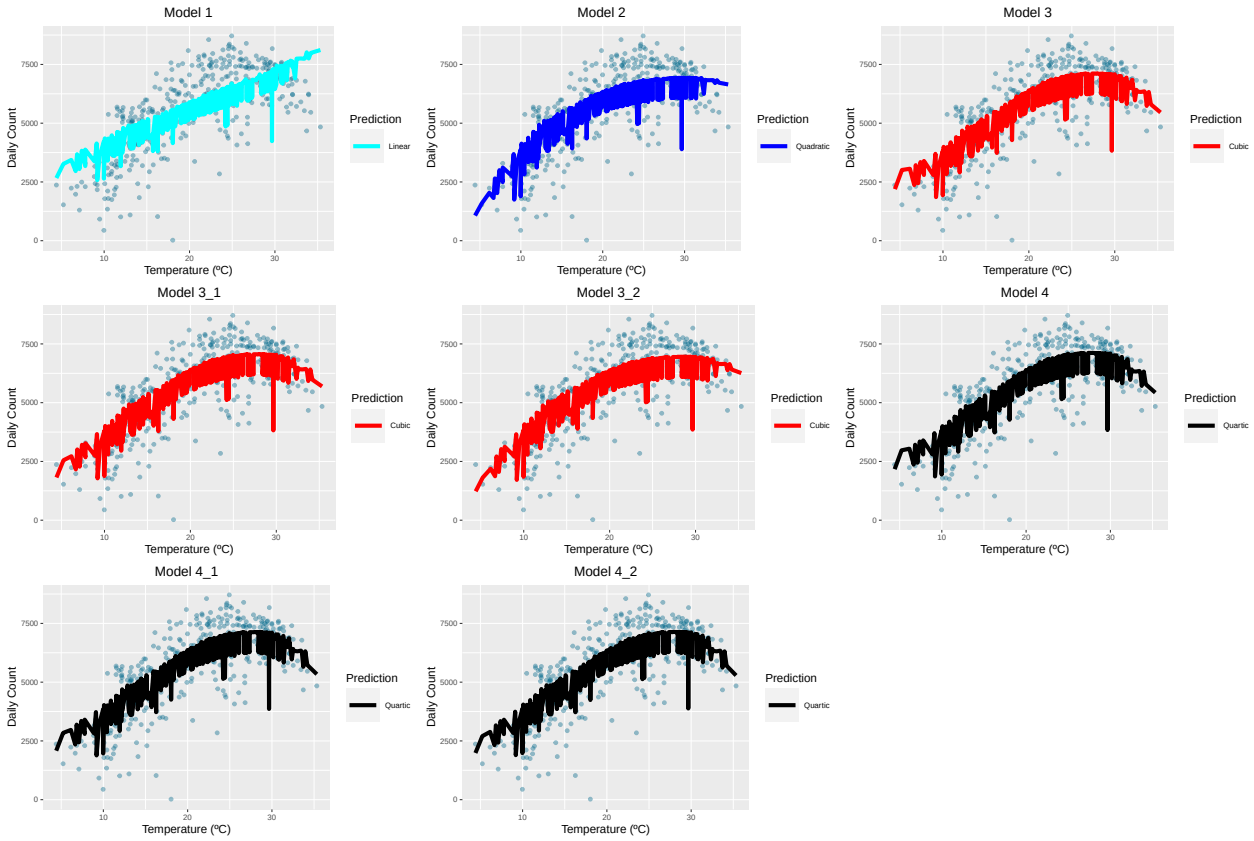
3.2.1 Fitted Values

Fig.7: 2011 Daily Bike Usage Count by Feeling Temperature, Predictions Overlayed



Observe that in 2011 (Fig.7), all models have to some degree lead to predicted counts below 0. This is a problem that should be fixed.

Fig.8: 2012 Daily Bike Usage Count by Feeling Temperature, Predictions Overlayed

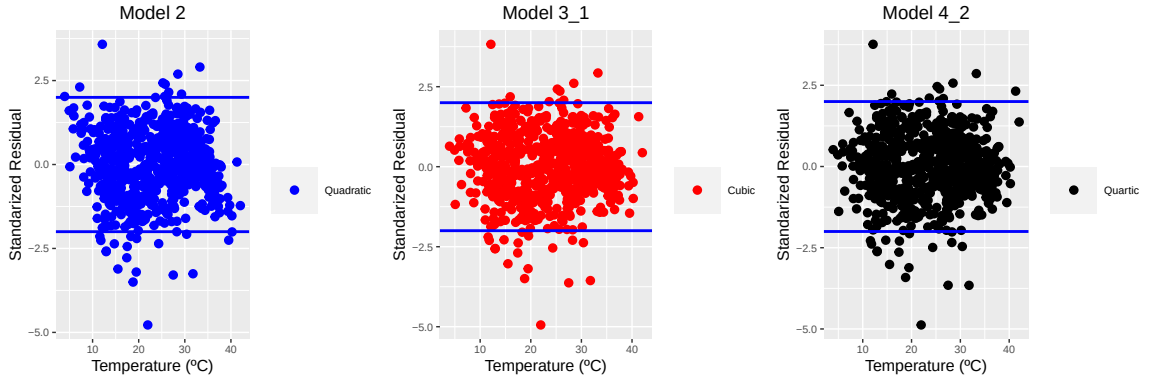


Considering both the significance of coefficients and how well the visual trends are captured (Fig.7,8), we keep Model 2, 3_1 and 4_2 for further comparisons.

3.3 Residual Diagostics

3.3.1 Residual Trends

Fig.9: Standardized Residual Plots

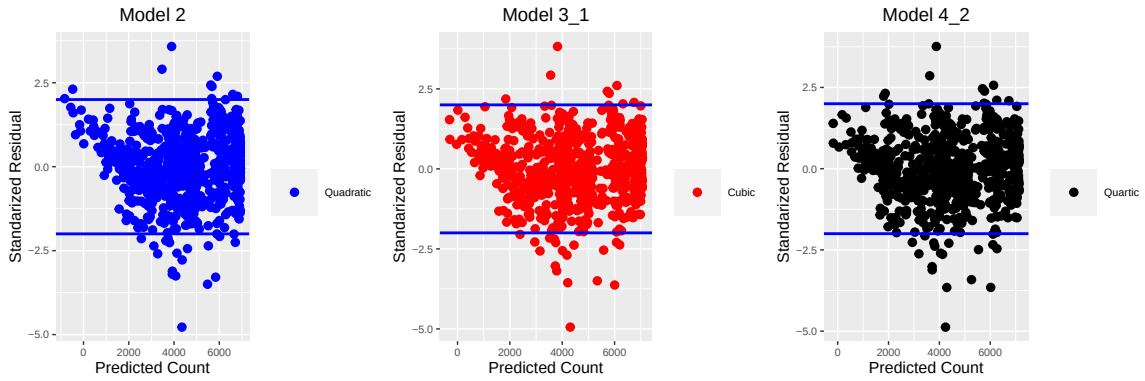


Most data points spreed evenly above and below the x-axis, which means the predicted values have captured the correct trends in response variation, which is good. Note that for the quadratic model (Model 2), there is still a slight positive bias in the lower-temperature parts and a negative bias in the higher-temperature parts, which also matches our observation in Fig.7&8. This problem is alleviated in the higher-degree models.

3.3.2 Residual Normality

3.3.2.1 Standardized Residual versus Fitted Value

Fig.10: Standardized Residual Plots

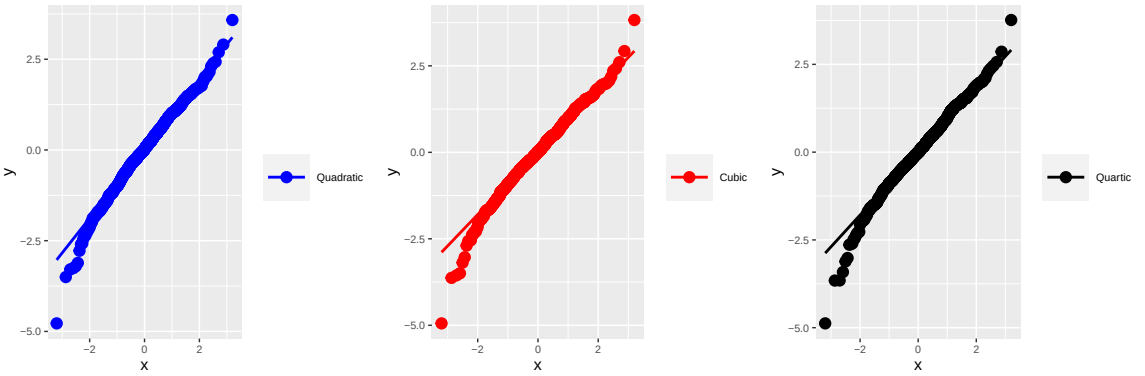


When the predicted counts are low, which corresponds to the lower-temperature parts in Fig.10, the residuals have a positive bias, i.e., the predicted values tend to be less than the true values, which matches what we have seen in Fig.7&8.

The problem that the predicted rental counts go below 0 still needs to be fixed in the later studies.

3.3.2.2 Normal Q-Q Plots

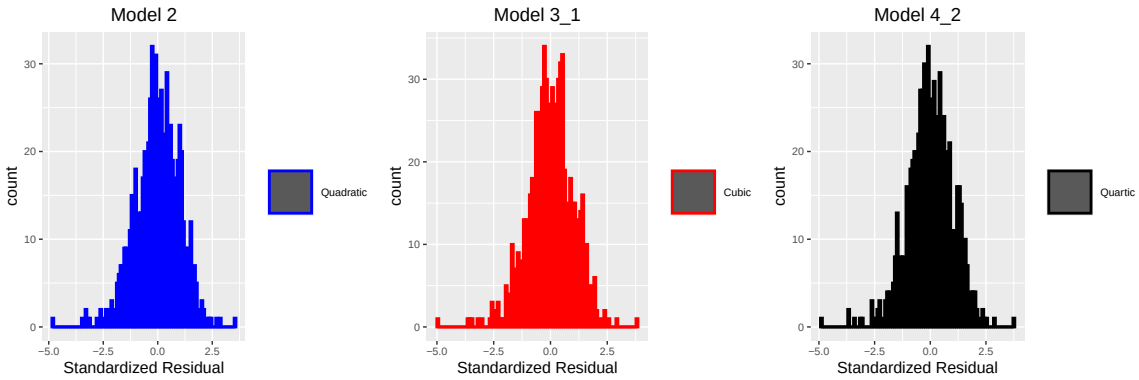
Fig.11: Normal Q-Q Plots



All of the residual distributions (Fig.11) match quite well with the standard normal distribution, which is really great, as such degree of matches can rarely be seen in general.

3.3.2.3 Residual Histograms

Fig.12: Residual Histograms



Given the nice match of distributions (Fig.11), the histograms should also match the standard normal distribution quite well, but surprisingly, the cubic model exhibits a slight bi-model issue (Fig.12). This is to be inspected in later analysis.

4 Conclusion

Given the above diagnostic results, it is still hard to decide among the 3 models.

Model 2 is good in that it involves less higher order terms, which means more stability of the model. However, it is weak in that it does not fit as well under the lowest and highest temperatures.

Model 3_1 and 4_2 are good in that they fit the best among all, have all coefficients significant, and also have nice diagnostic performance. However, the existence of higher-order terms leads to a more unstable model, and the one at a medium level of polynomial degrees, Model 3_1, happens to have the bi-modal issue in its standardized residuals.

These problems could potentially be resolved by:

1. introducing a time-series model
2. including the weekday variable
3. breaking the model into high and low temperature parts and model them separately.

Further analysis is left for the next lab report.

5 References

[1] Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.

[2] MA 575 Fall 2021 C3 Team #2, "Lab Report 2: Ordinary Least Squares".

6 Appendix: Intermediate Results

Fig.13: Pairs Plot for Group 1 (Colored by Season)

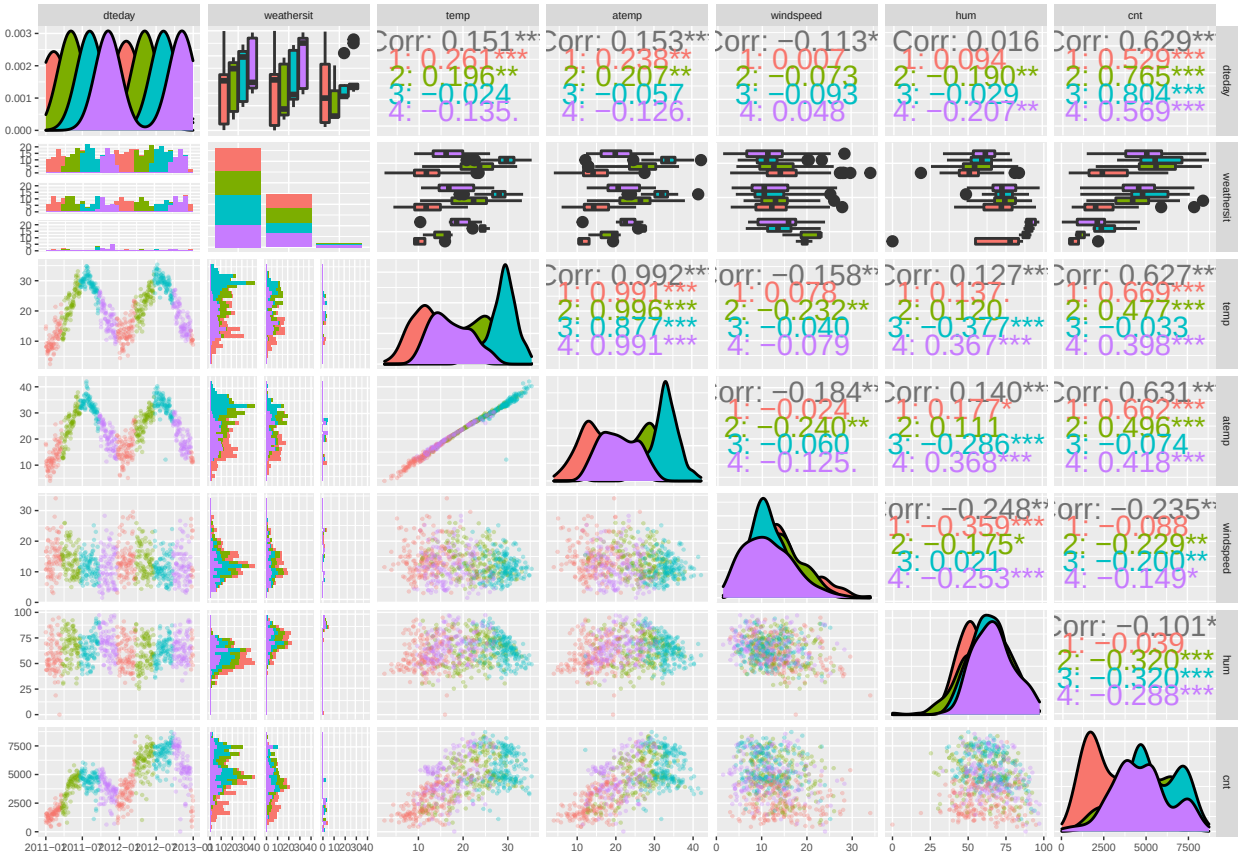


Fig.13 Column names in order: `dteday`, `weathersit`, `temp`, `atemp`, `windspeed`, `hum`, `cnt`

Fig.14: Pairs Plot for Group 1 (Colored by Month)

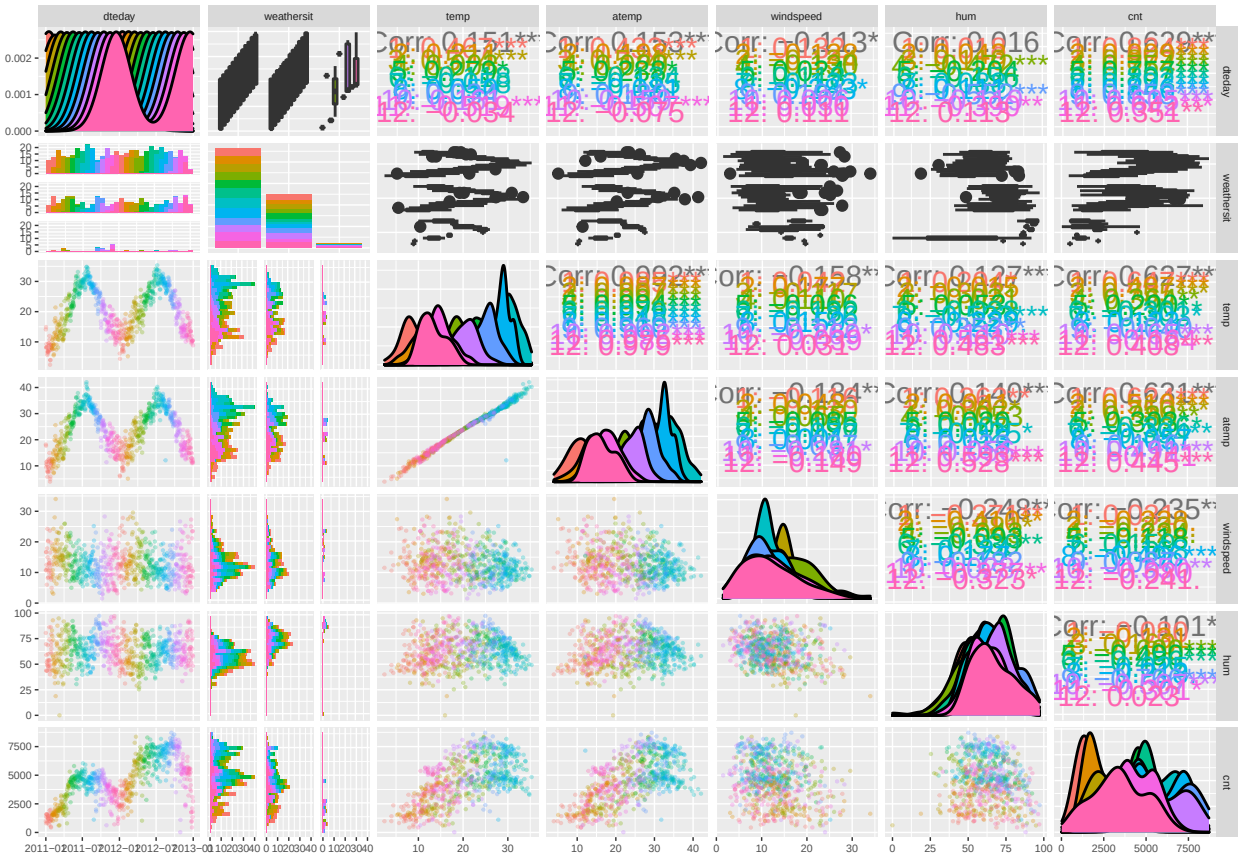


Fig.14 Column names in order: `dteday`, `weathersit`, `temp`, `atemp`, `windspeed`, `hum`, `cnt`

Fig.15: Pairs Plot for Group 1 (Colored by Weekdays)

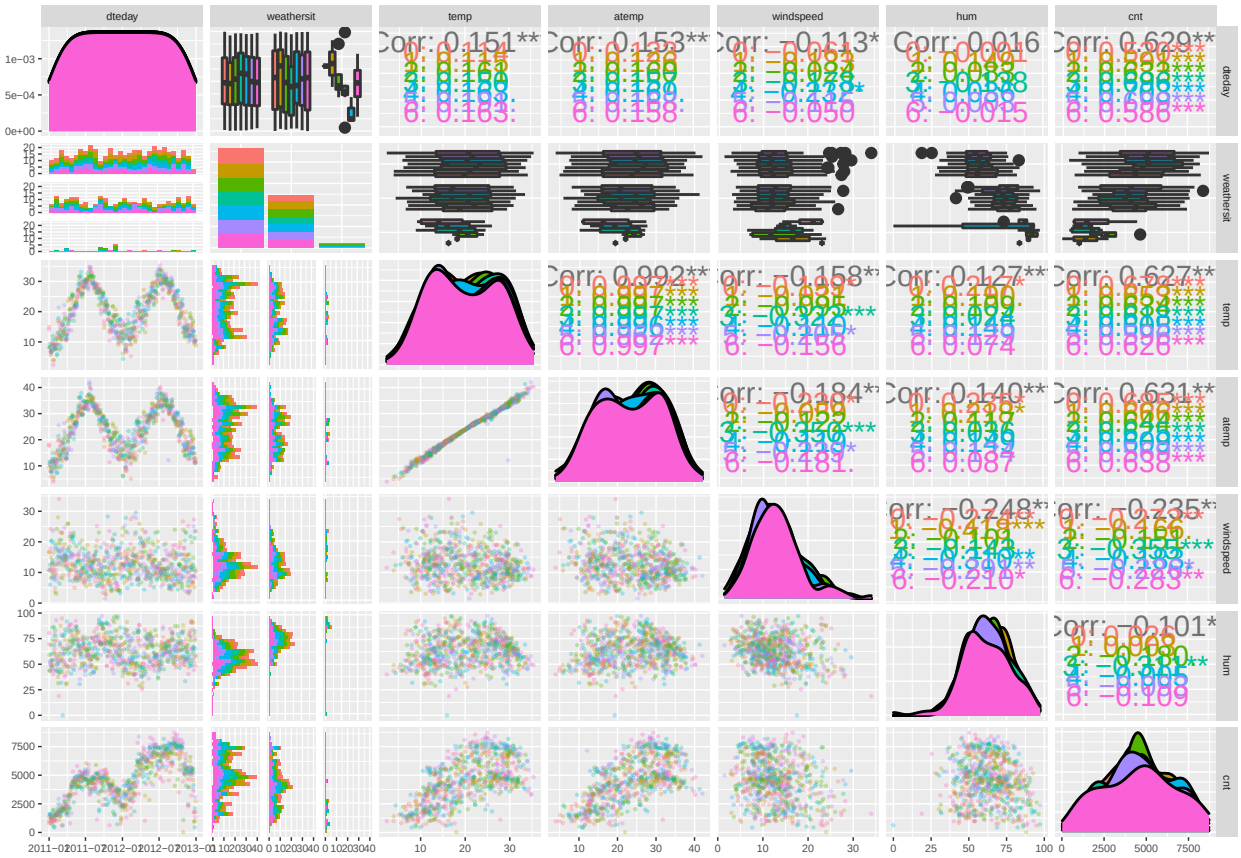


Fig.15 Column names in order: `dteday`, `weathersit`, `temp`, `atemp`, `windspeed`, `hum`, `cnt`

Fig.16: Pairs Plot for Group 1 (Colored by Working Day)

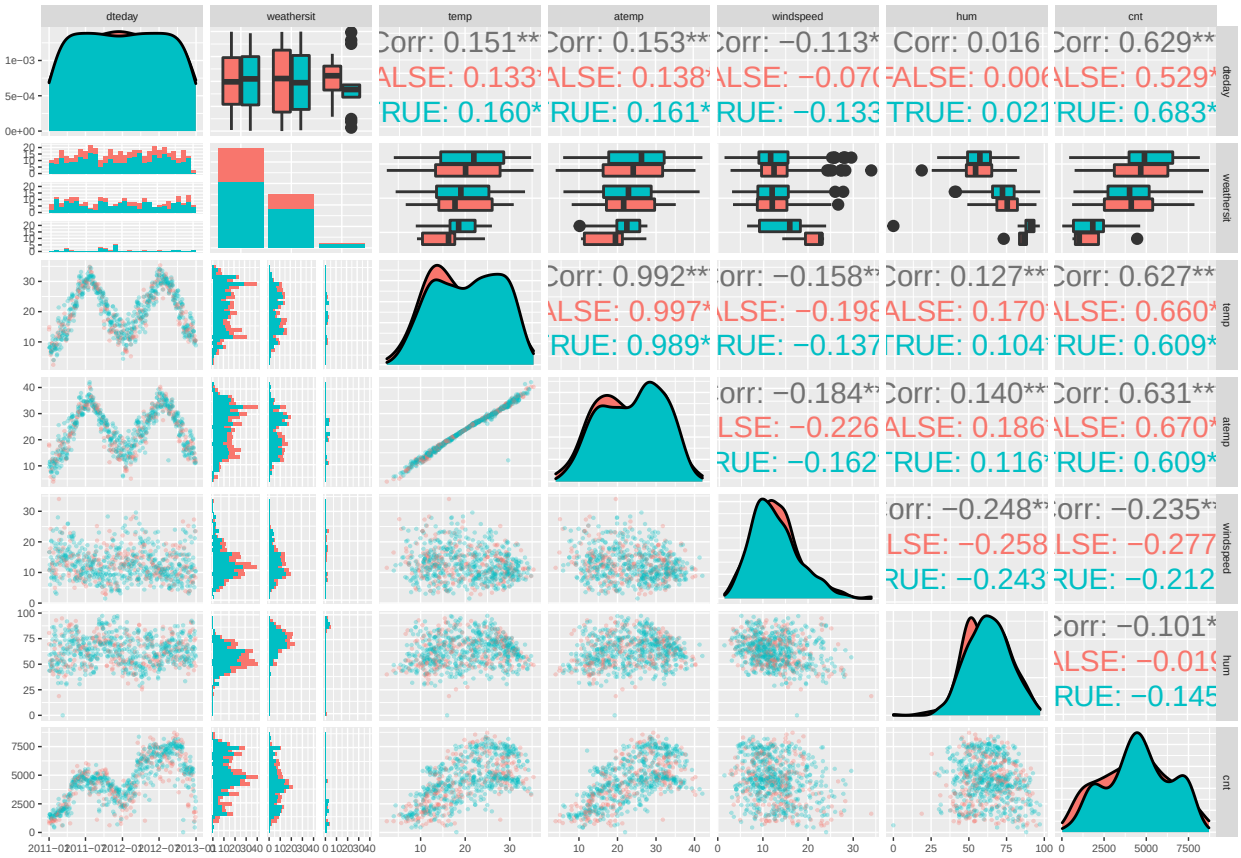


Fig.16 Column names in order: `dteday`, `weathersit`, `temp`, `atemp`, `windspeed`, `hum`, `cnt`

Fig.17: Pairs Plot for Group 1 (Colored by Holiday)

Fig.17 Column names in order: `dteday`, `weathersit`, `temp`, `atemp`, `windspeed`, `hum`, `cnt`