LAB2RMD

```
# Duomenys
# https://www.kaggle.com/datasets/brajeshmohapatra/bike-count-prediction-data-set?select=train.csv

library(tidyverse)
library(reshape2)
library(AER)
library(MASS)
library(goft)
library(lubridate)
```

```
tr <- read.csv("train.csv")
te <- read.csv("test.csv")
te$count <- te$casual + te$registered
mega <- rbind(tr, te)
mega$hour <- hour(mega$datetime)
mega$day <- day(mega$datetime)
mega$yday <- yday(mega$datetime)
mega$wday <- wday(mega$datetime)
mega <- dplyr::select(mega, -c(datetime, casual, registered))
```

```
head(mega)
```

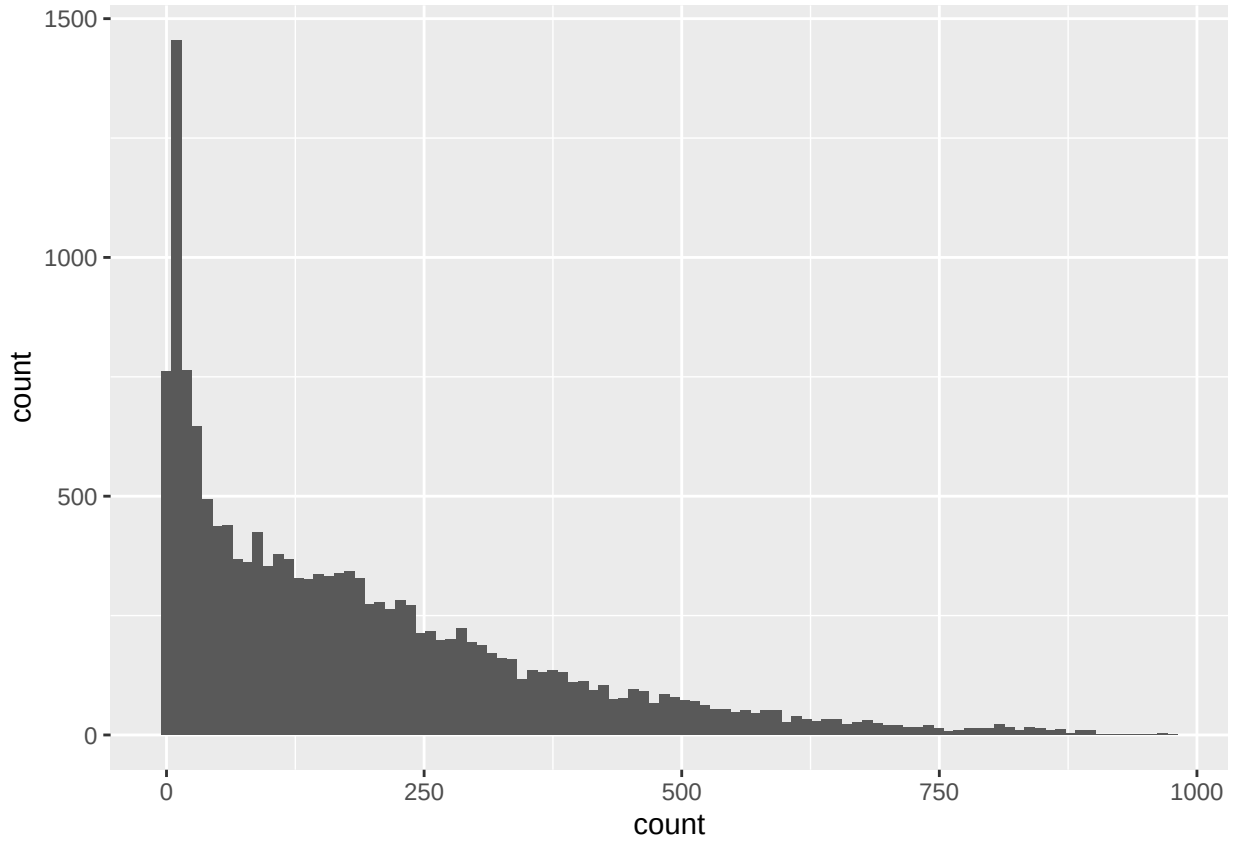
```
##   season holiday workingday weather temp  atemp humidity windspeed count hour
## 1      1       0          0       1 9.84 14.395       81    0.0000    16    0
## 2      1       0          0       1 9.02 13.635       80    0.0000    40    1
## 3      1       0          0       1 9.02 13.635       80    0.0000    32    2
## 4      1       0          0       1 9.84 14.395       75    0.0000    13    3
## 5      1       0          0       1 9.84 14.395       75    0.0000     1    4
## 6      1       0          0       2 9.84 12.880       75    6.0032     1    5
##   day yday wday
## 1   1    1    7
## 2   1    1    7
## 3   1    1    7
## 4   1    1    7
## 5   1    1    7
## 6   1    1    7
```

```
summary(mega)
```

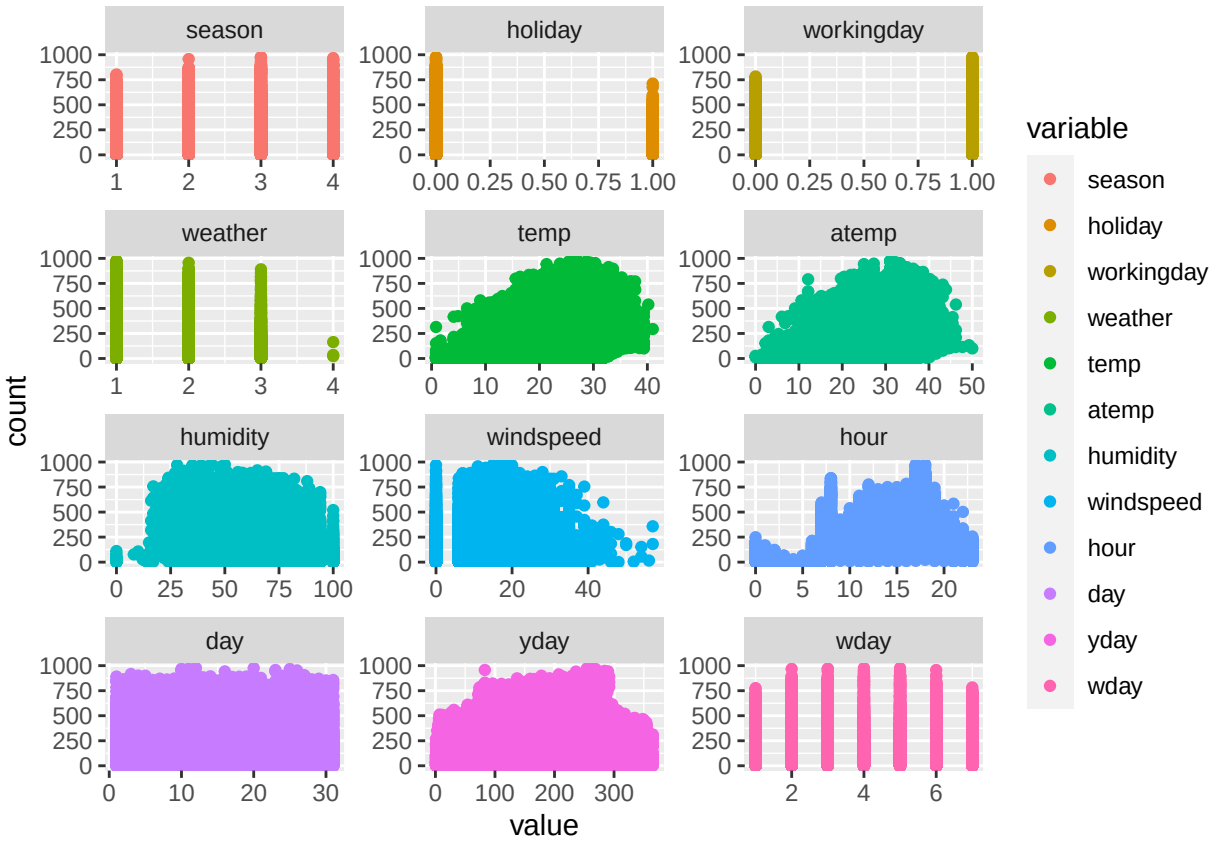
```
##      season         holiday          workingday        weather
##  Min.   :1.000   Min.   :0.00000   Min.   :0.0000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:1.000
##  Median :3.000   Median :0.00000   Median :1.0000   Median :1.000
##  Mean   :2.502   Mean   :0.02877   Mean   :0.6827   Mean   :1.425
##  3rd Qu.:3.000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:2.000
##  Max.   :4.000   Max.   :1.00000   Max.   :1.0000   Max.   :4.000
##       temp           atemp          humidity        windspeed
##  Min.   : 0.82   Min.   : 0.00   Min.   :  0.00   Min.   : 0.000
```

```
##  1st Qu.:13.94   1st Qu.:16.66   1st Qu.: 48.00   1st Qu.: 7.002
##  Median :20.50   Median :24.24   Median : 63.00   Median :12.998
##  Mean   :20.38   Mean   :23.79   Mean   : 62.72   Mean   :12.737
##  3rd Qu.:27.06   3rd Qu.:31.06   3rd Qu.: 78.00   3rd Qu.:16.998
##  Max.   :41.00   Max.   :50.00   Max.   :100.00   Max.   :56.997
##      count            hour            day             yday
##  Min.   :  1.0   Min.   : 0.00   Min.   : 1.00   Min.   :  1.0
##  1st Qu.: 40.0   1st Qu.: 6.00   1st Qu.: 8.00   1st Qu.: 93.0
##  Median :142.0   Median :12.00   Median :16.00   Median :184.0
##  Mean   :189.5   Mean   :11.55   Mean   :15.68   Mean   :183.7
##  3rd Qu.:281.0   3rd Qu.:18.00   3rd Qu.:23.00   3rd Qu.:275.0
##  Max.   :977.0   Max.   :23.00   Max.   :31.00   Max.   :366.0
##       wday
##  Min.   :1.000
##  1st Qu.:2.000
##  Median :4.000
##  Mean   :4.004
##  3rd Qu.:6.000
##  Max.   :7.000
```

```
d <- mega[0:floor(0.90 * nrow(mega)),]
ggplot(d, aes(x=count)) + geom_histogram(bins = 100)
```




```
ggplot(melt(d, "count"), aes(x = value, y = count, colour = variable)) +
  geom_point() +
  facet_wrap(~variable, scales = "free", nrow = 4)
```




```
### Poisson
m1 <- glm(count ~ ., family="poisson", data=d)
summary(m1)
```

```
##
## Call:
## glm(formula = count ~ ., family = "poisson", data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -26.565   -9.036   -3.536    3.908   43.609
##
## Coefficients:
##                Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  4.052e+00  4.231e-03  957.905  < 2e-16 ***
## season       6.378e-02  1.327e-03   48.073  < 2e-16 ***
## holiday     -1.065e-01  3.993e-03  -26.661  < 2e-16 ***
## workingday   1.171e-02  1.308e-03    8.956  < 2e-16 ***
## weather     -1.425e-02  1.109e-03  -12.858  < 2e-16 ***
## temp         8.167e-03  4.609e-04   17.719  < 2e-16 ***
## atemp        3.065e-02  4.266e-04   71.835  < 2e-16 ***
## humidity    -1.089e-02  3.718e-05 -292.993  < 2e-16 ***
## windspeed    3.753e-03  7.496e-05   50.070  < 2e-16 ***
## hour         4.573e-02  9.670e-05  472.911  < 2e-16 ***
## day         -5.165e-04  6.627e-05   -7.793 6.55e-15 ***
## yday         2.267e-04  1.443e-05   15.711  < 2e-16 ***
## wday         6.485e-03  2.962e-04   21.896  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2590437  on 15640  degrees of freedom
## Residual deviance: 1616281  on 15628  degrees of freedom
## AIC: 1715971
##
## Number of Fisher Scoring iterations: 5
```

```
cat("Deviacija padalinta is laisves laipsniu: ",m1$deviance / m1$df.residual)
```

```
## Deviacija padalinta is laisves laipsniu:  103.4221
```

```
cat("Turi buti tarp 0.7 ir 1.3, tad nebegalime naudoti puasono modelio")
```

```
## Turi buti tarp 0.7 ir 1.3, tad nebegalime naudoti puasono modelio
```

```
### Negative Binomial
m2 <- glm.nb(count ~ ., data = d)
summary(m2)
```

```
##
## Call:
## glm.nb(formula = count ~ ., data = d, init.theta = 1.17499586,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7312  -1.0025  -0.3154   0.3086   4.0209
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.5733936  0.0500612  71.381  < 2e-16 ***
## season       0.0924761  0.0149496   6.186 6.18e-10 ***
## holiday     -0.1166441  0.0472293  -2.470 0.013521 *
## workingday   0.0870829  0.0165310   5.268 1.38e-07 ***
## weather     -0.0441938  0.0130839  -3.378 0.000731 ***
## temp         0.0087633  0.0060366   1.452 0.146585
## atemp        0.0327719  0.0055487   5.906 3.50e-09 ***
## humidity    -0.0089594  0.0004701 -19.058  < 2e-16 ***
## windspeed    0.0015382  0.0009812   1.568 0.116973
## hour         0.0712705  0.0011385  62.600  < 2e-16 ***
## day         -0.0007898  0.0008524  -0.927 0.354110
## yday        -0.0003017  0.0001570  -1.922 0.054649 .
## wday         0.0085756  0.0037346   2.296 0.021661 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.175) family taken to be 1)
##
##     Null deviance: 25324  on 15640  degrees of freedom
## Residual deviance: 17695  on 15628  degrees of freedom
## AIC: 188457
##
```

```
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.1750
##           Std. Err.:  0.0122
##
##  2 x log-likelihood:  -188428.8140
```

```
cat("Deviacija padalinta is laisves laipsniu: ",m2$deviance / m2$df.residual)
```

```
## Deviacija padalinta is laisves laipsniu:  1.132281
```

Cia reikia prielaidu

```
# p1
# p2
# p3
# p4
```

```
# # # NEVEIKIA NES LIBRARY SUGADINTAS # # #
### Zero deflated
#require(foreign)
#require(ggplot2)
#require(VGAM)
#require(boot)
#ztrunc <- vglm(count ~ ., family = posnegbinomial(), data = d)
#summary(ztrunc)
```

```
### Stepwise
m2step <- stepAIC(m2, direction = "both")
```

```
## Start:  AIC=188454.8
## count ~ season + holiday + workingday + weather + temp + atemp +
##     humidity + windspeed + hour + day + yday + wday
##
##               Df    AIC
## - day          1 188454
## <none>           188455
## - windspeed    1 188455
## - temp         1 188456
## - yday         1 188456
## - wday         1 188458
## - holiday      1 188459
## - weather      1 188463
## - workingday   1 188480
## - season       1 188487
## - atemp        1 188495
## - humidity     1 188779
## - hour         1 190750
##
## Step:  AIC=188453.7
## count ~ season + holiday + workingday + weather + temp + atemp +
##     humidity + windspeed + hour + yday + wday
##
##               Df    AIC
```

```
## <none>          188454
## - windspeed   1 188454
## - temp        1 188455
## + day         1 188455
## - yday        1 188455
## - wday        1 188457
## - holiday     1 188457
## - weather     1 188462
## - workingday  1 188479
## - season      1 188486
## - atemp       1 188495
## - humidity    1 188780
## - hour        1 190748
```

```
summary(m2step)
```

```
##
## Call:
## glm.nb(formula = count ~ season + holiday + workingday + weather +
##     temp + atemp + humidity + windspeed + hour + yday + wday,
##     data = d, init.theta = 1.174943919, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7351  -1.0032  -0.3157   0.3090   4.0302
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.5621002  0.0485454  73.377  < 2e-16 ***
## season       0.0936458  0.0148871   6.290 3.17e-10 ***
## holiday     -0.1149367  0.0472175  -2.434 0.014925 *
## workingday   0.0870732  0.0165313   5.267 1.39e-07 ***
## weather     -0.0438875  0.0130808  -3.355 0.000793 ***
## temp         0.0085455  0.0060316   1.417 0.156548
## atemp        0.0329213  0.0055456   5.937 2.91e-09 ***
## humidity    -0.0089749  0.0004698 -19.102  < 2e-16 ***
## windspeed    0.0015192  0.0009810   1.549 0.121475
## hour         0.0712666  0.0011385  62.595  < 2e-16 ***
## yday        -0.0003142  0.0001561  -2.013 0.044093 *
## wday         0.0085938  0.0037346   2.301 0.021386 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.1749) family taken to be 1)
##
##     Null deviance: 25323  on 15640  degrees of freedom
## Residual deviance: 17695  on 15629  degrees of freedom
## AIC: 188456
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.1749
##           Std. Err.:  0.0122
##
```

```
##  2 x log-likelihood:  -188429.6580
```

```
cat("Deviacija padalinta is laisves laipsniu: ",m2step$deviance / m2step$df.residual)
```

```
## Deviacija padalinta is laisves laipsniu:  1.132214
```

```
# Koeficientai
est <- cbind(Estimate = coef(m2step), confint(m2step))
```

```
## Waiting for profiling to be done...
```

```
exp(est)
```

```
##                Estimate      2.5 %     97.5 %
## (Intercept) 35.2371260 31.8928539 38.9432545
## season       1.0981707  1.0650461  1.1317528
## holiday      0.8914226  0.8135570  0.9789327
## workingday   1.0909765  1.0559708  1.1269871
## weather      0.9570616  0.9317581  0.9831574
## temp         1.0085821  0.9987261  1.0189477
## atemp        1.0334692  1.0237935  1.0428419
## humidity     0.9910652  0.9901059  0.9920251
## windspeed    1.0015203  0.9995839  1.0034638
## hour         1.0738675  1.0707886  1.0769595
## yday         0.9996858  0.9993652  1.0000132
## wday         1.0086308  1.0012696  1.0160470
```

```
dopred <- function(tt, model) {
  # Index
  index <- 1:nrow(tt)
  # Ground truth
  real <- tt$count
  # Predicted
  tt <- dplyr::select(tt, -count)
  predicted <- predict(model, newdata = tt, type = "response")
  # df
  tempdf <- data.frame(real, predicted, index)

  print(mean(abs((tempdf$real-tempdf$predicted)/tempdf$real)) * 100)
  p <- ggplot(tempdf, aes(x=index)) +
      geom_line(aes(y = real), color = "#0F9D58") +
      geom_line(aes(y = predicted), color="#4285F4", linetype="twodash")
  return(p)
}
```

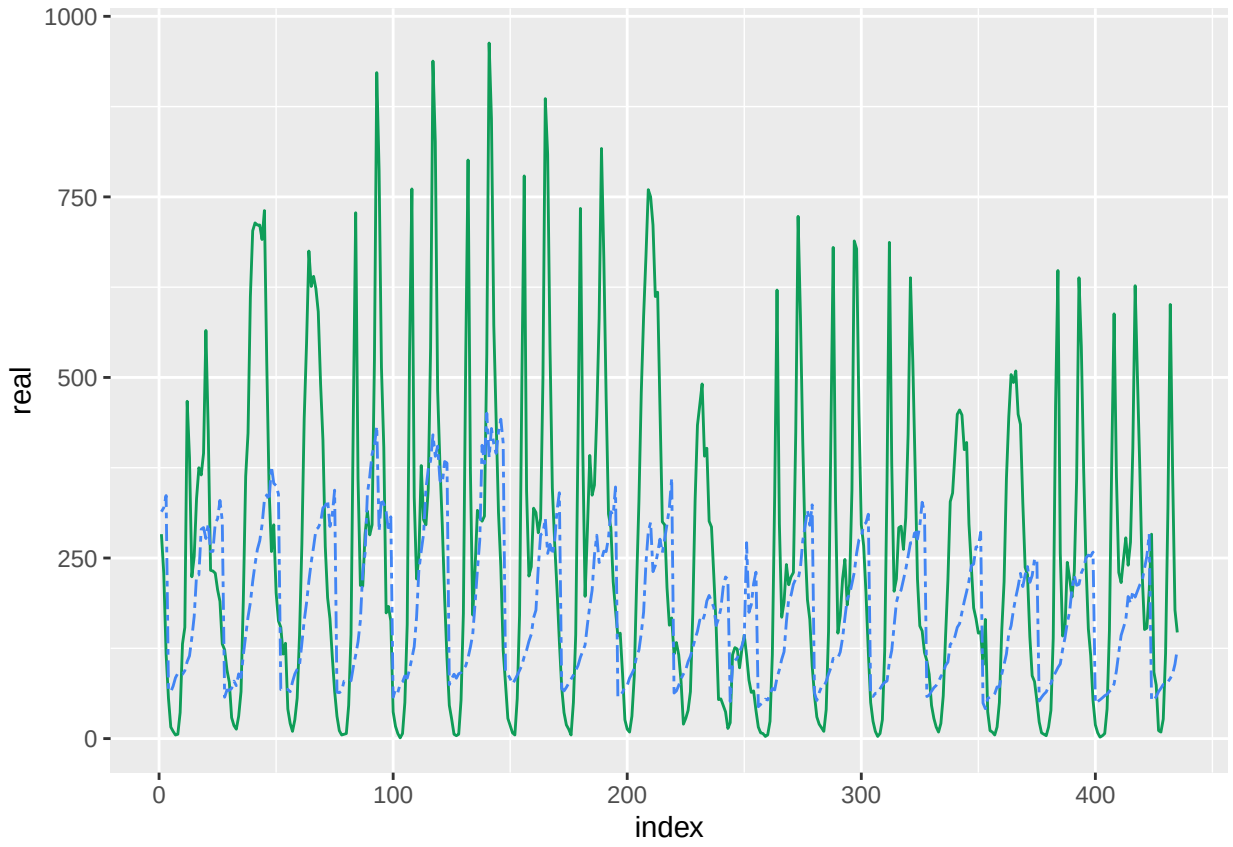
```
v <- mega[floor(0.90 * nrow(mega)):nrow(mega),]

# Split to 4 datasets
num_groups = 4
totest <- v %>%
  group_by((row_number()-1) %/% (n()/num_groups)) %>%
  nest %>% pull(data)

# Prediction for each dataset
plots <- list()
for (i in 1:num_groups) {
  plots[[i]] = dopred( data.frame(totest[i]), m2step)
```

```
}
```

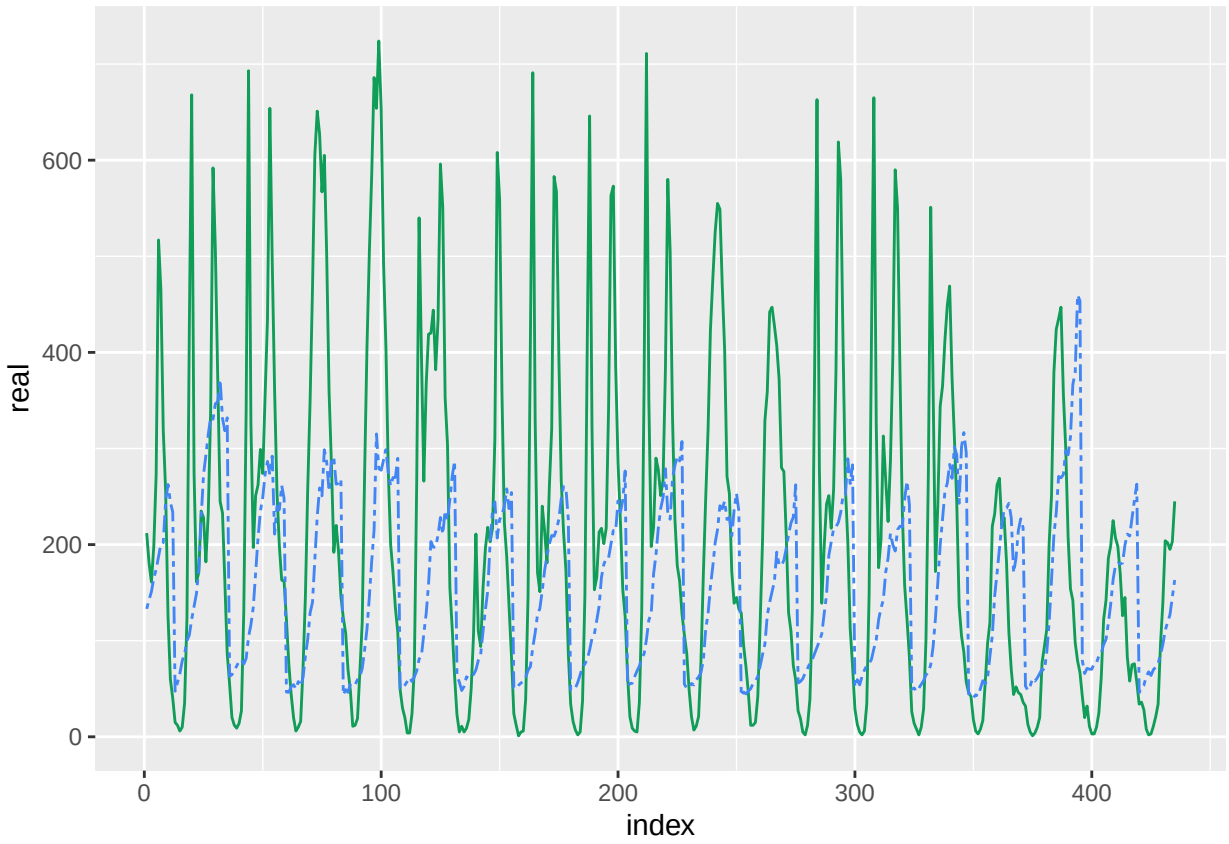
```
## [1] 181.3344
## [1] 216.1337
## [1] 193.7837
## [1] 249.3642
```

```
plots
```

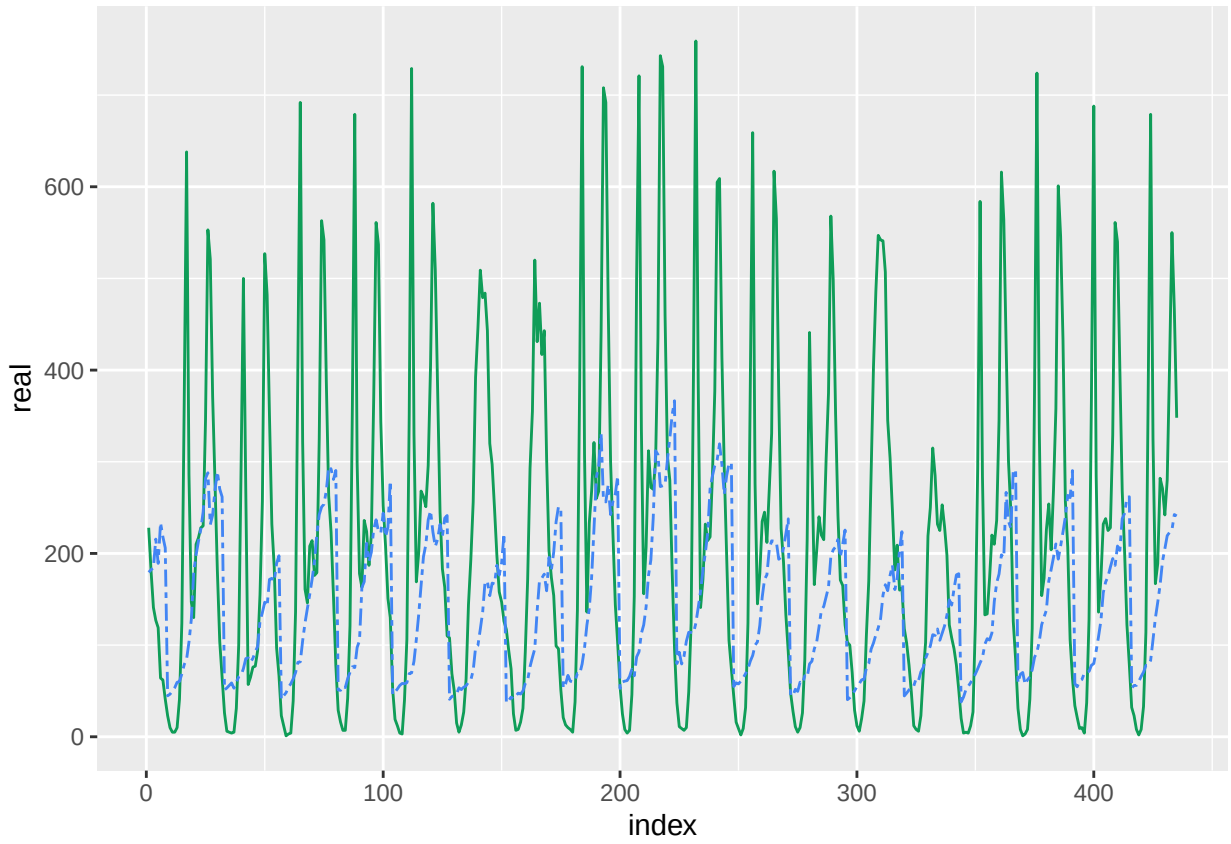
```
## [[1]]
```



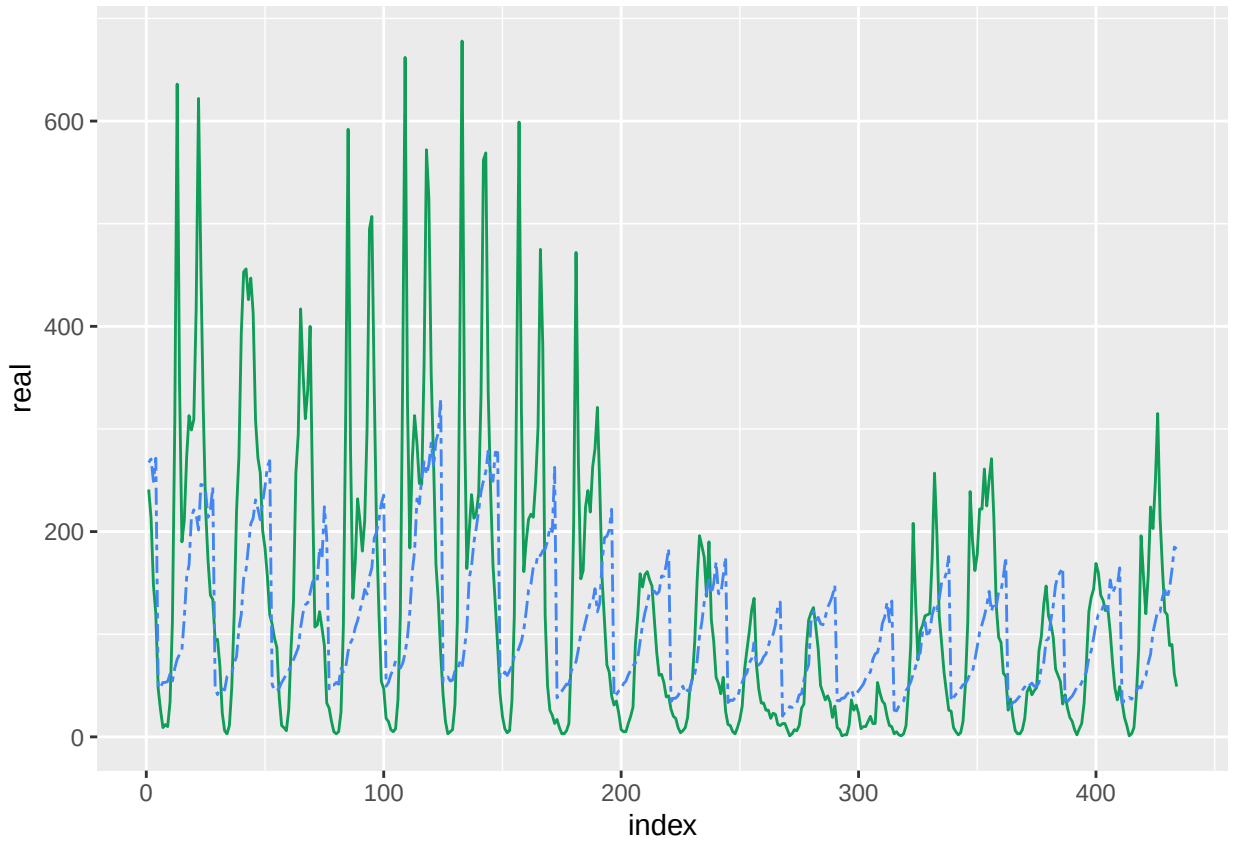

```
##
## [[2]]
```




```
##
## [[3]]
```




```
##
## [[4]]
```


600
400
200
0
real
0
100
200
300
400
index