```r
######## LOAD & CLEAN DATA ###########

# Load data
day <- read_csv("data/day.csv")
hour <- read_csv("data/hour.csv")

# make sure it is ordered correctly
hour <- hour[order(hour$dteday, hour$hr),]
day <- day[order(day$dteday),]


# Clean data
setDT(hour)
hour[, season := as.factor(ifelse(season == 1, "Spring",
                            ifelse(season == 2, "Summer",
                                   ifelse(season == 3, "Fall",
                                          ifelse(season == 4, "Winter", NA)))))]

hour[, weathersit := as.factor(ifelse(weathersit == 1, "Clear",
                              ifelse(weathersit == 2, "Misty",
                                     ifelse(weathersit == 3, "Rain",
                                            ifelse(weathersit == 4,
                                                   "Thunderstorm", NA)))))]

hour <- hour[, -c("instant")]

sumstats_day <- day
setDT(sumstats_day)
sumstats_day[, season := as.factor(ifelse(season == 1, "Spring",
                                  ifelse(season == 2, "Summer",
                                         ifelse(season == 3, "Fall",
                                                ifelse(season == 4,
                                                       "Winter", NA)))))]

sumstats_day[, weathersit := as.factor(ifelse(weathersit == 1, "Clear",
                                        ifelse(weathersit == 2, "Misty",
                                               ifelse(weathersit == 3, "Rain",
                                                      ifelse(weathersit == 4,
                                                             "Thunderstorm", NA)))))]
sumstats_day <- sumstats_day[, -c("instant")]


# dummify the data
dmy <- dummyVars(" ~ .", data = hour)
hour <- data.frame(predict(dmy, newdata = hour))

dmy <- dummyVars(" ~ .", data = sumstats_day)
sumstats_day <- data.frame(predict(dmy, newdata = sumstats_day))

# get total counts
setDT(hour)
setDT(day)
```

```r
# further cleaning
setDT(hour)
setDT(sumstats_day)
hour[, yr := ifelse(hour$yr == 0, 2011, 2012)]
sumstats_day[, yr := ifelse(sumstats_day$yr == 0, 2011, 2012)]

hour_temp <- hour[, .(mean_count = mean(cnt)), by = c("temp")]
hour_temp <- hour[, lapply(.SD, mean), by=temp]

day[, month := mnth + yr*12]

# durbin watson
dwtest(day$cnt ~ day$instant)
# partial autocorrelation
pacf(day$cnt, lag.max = nrow(day))
# auto correlation
acf(day$cnt, lag.max = nrow(day))
# mann-kendall (seasonal)
smk.test(ts_cnt)
# unit root stationarity (reject null)
summary(ur.kpss(day$cnt))
```