**Submission:**

* 1 submission per Assignment Group featuring 1 .rmd file with code and text responses through Canvas

**FPP3 text:**

“Forecasting: Principles and Practice” <https://otexts.com/fpp3/>

Datasets

**Canadian Gas**

canadian\_gas (available with library(fpp3))

**Description: type below in console**

?canadian\_gas

**Bicycle rentals data**

**Where to access:**

Teams -> Class materials -> Assignments -> bicycle.rda

**How to import:**

Save file to local directory, change working directory (setwd(“~/Documents/BANA4090…”) and

Bike <- readRDS(“bicycle.rda”)

*Be sure to convert to a tsibble before you start using the functions we’ve learned (to convert use the tsibble() function and specify the index.*

A tibble: 731 x 16

* instant: record index
* dteday : date
* season : season (1:winter, 2:spring, 3:summer, 4:fall)
* yr : year (0: 2011, 1:2012)
* mnth : month ( 1 to 12)
* holiday : weather day is holiday or not
* weekday : day of the week
* workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
* weathersit :
  + 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 derived via (t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39 (only in hourly scale)
* atemp: Normalized feeling temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-16, t\_max=+50 (only in hourly scale)
* 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 rental bikes
* registered: count of registered rental bikes
* cnt: count of total rental bikes including both casual and registered

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

Exercises

1.) Explore and Model the bicycle dataset (18 pts):

The bicycle.rda file is a tibble composed of daily total rental bikes over time with other exogenous information like whether the date was a holiday, the weather situation, temperature and windspeed. The cnt column is total bikes rented, and is the forecast variable. We will past values of total bikes rented and the attributes included in this dataset to forecast future values of total bikes rented.

1. **Import and format data (1 pts)**
   1. Import and convert bicycle.rda to a tsibble with an index of dteday (see first page for hint)
   2. Using select(), retain only these columns: (dteday, workingday, weathersit, atemp, windspeed, cnt)
   3. Create a new column called weekday\_factor which is the abbreviated day of week (hint use wday function from lubridate package and see arguments with ?wday)
   4. Create a new column called workingday\_factor which converts workingday to a factor (use factor()).
   5. Convert weathersit to factor called weathersit\_factor (use factor())
   6. Remove workingday and weathersit columns from dataset
2. **Visualize the the time series and potential predictor variables (4 pts)**
   1. Use autoplot() or ggplot’s geom\_line() to visualize the time series’ cnt column
   2. Visualize average bike rentals for weekday\_factor, workingday\_factor and weathersit\_factor to understand to understand how the day of week and generalized weather impacts average ridership
      1. Hint – depending on how you want to visualize you may need to convert the tsibble to a tibble to use the group\_by() and summarise() functions without the dteday index. If it’s a tsibble, it will always group by the columns AND the index
      2. Hint – you could create the visual with geom\_boxplot()
   3. Perform a scatter plot of cnt and the below variables. Describe the relationships.
      1. atemp
      2. windspeed
   4. Based on what you notice above, what predictor variables would you include in a time series linear model? Which of these variables would be considered dummy variables?
3. **Are total rentals per day (the cnt variable) stationary? (2 pts)**
   1. Plot a correlogram (ACF plot) of the cnt column. What does the plot tell you about the stationarity of the data? Confirm this with a KPSS unit root test.
   2. Difference the data and test stationarity with a KPSS unit root test
   3. Create new column called diff\_cnt which is difference(cnt)
4. **Fit TSLM models to the data (12 pts)**
   1. Create a new column called lag\_diff\_cnt which is lag(diff\_cnt)
   2. Create a new column called lag\_weather which is lag(weathersit\_factor)
   3. Filter out the first row of data which includes missing values for the lags just created
   4. Split the dataset into a train and test set
      1. Training set is data before 10/1/2012
      2. Training set is data on and after (>=) 10/1/2012 **and on and before (& <=) 10/30/2012**
   5. Fit the following models to the train dataset
      1. diff\_cnt = atemp + workingday\_factor + weathersit\_factor
      2. diff\_cnt = lag\_diff\_cnt + atemp + workingday\_factor + weathersit\_factor
      3. diff\_cnt = lag\_diff\_cnt+ lag\_weather + atemp + workingday\_factor + weathersit\_factor
      4. Compare AICc of the above models and report and inspect residuals of the model with the lowest AICc (glance() function)
         1. Run the model() again with the best performing formulation and use report() to view the coefficients for each of the predictor variables
         2. *For next steps see section 7.3:* Use gg\_tsresiduals() to inspect the residuals. Are the residuals autocorrelated? What does this imply about the model?
            1. Hint – if you use the features(.innov, ljung\_box) function your dof argument (degrees of freedom of the model) can be found with the glance(model) function on your model
         3. Are the residuals stationary? (use a KPSS unit root test)
         4. Plot the residuals against the predictors lag\_cnt and atemp. Do they appear to be randomly scattered?
         5. Plot the residuals against the fitted values. What do you see?
   6. Compare to benchmark methods
      1. Fit 4 models:
         1. TSLM(diff\_cnt = lag\_diff\_cnt + lag\_weather + workingday\_factor + weathersit\_factor)
            1. *note: atemp was left off intentionally*
         2. Benchmark methods (mean, naïve, drift)
      2. Plot the forecasts for the test dataset (10/1/2012-10/30/2012)
      3. Compare MAPE, MASE and MAE results on the test set. Does our linear model perform better on the test set than the benchmark methods?

FPP3 – 3.7, Exercise 3 (1pt):

Why is a Box-Cox transformation unhelpful for the canadian\_gas data?