# Project Proposal Instructions

Instructor: Arthur Small

Version of 2020-10-07

In this assignment you will develop your initial concept note into a draft of a full project proposal. Treat this assignment as a "dry run" for developing a proposal for a grant or fellowship application, or for your Ph.D. prospectus.

Your proposal should include at least the following sections and information.

**Front matter:** Descriptive title, your name, date, reference to "SYS 7030 Time Series Analysis & Forecasting, Fall 2020".

**Abstract:** A very brief summary of the project.

## Introduction

Give a narrative description of the problem you are addressing, and the methods you will use to address it. Provide context:

- What is the question you are attempting to answer?
- Why is this question important? (Who cares?)
- How will you go about attempting to answer this question?

This work addresses the question: Why do people not use probabilistic forecasts for decision-making (Council 2007)?

## The data and the data-generating process

Describe the data set you will be analyzing, and where it comes from, how it was generated and collected. Identify the source of the data. Give a narrative description of the data-generating process: this piece is critical.

Since these will be time series data: identify the frequency of the data series (e.g., hourly, monthly), and the period of record.

```
# Open connection to a remote database
# Make sure your VPN network connection is active if needed!

# if(!('RPostgreSQL' %in% installed.packages())) install.packages('RPostgreSQL')
library(RPostgreSQL)

# "my_postgres_credentials.R" contains the log-in information
source("/Users/Arthur/GitRepos/Teaching/my_postgres_db_credentials.R")

# Open connection
db_driver <- dbDriver("PostgreSQL")
```

```
db <- dbConnect(db_driver,user=user, password=password,dbname="postgres", host=host)
rm(password)

# check the connection: If function returns value TRUE, the connection is working
dbExistsTable(db, "metadata")
```

```
# library(dplyr) -- don't need this if you are loading the entire 'tidyverse' suite
library(tidyverse)

esales <- dbGetQuery(db,'SELECT * from eia_elec_sales_va_all_m') # SQL code to retrieve data from a tab
# str(esales)
esales <- as_tibble(esales) # Convert dataframe to a 'tibble' for tidyverse work
# str(esales)
```

```
# Reference: https://arrow.apache.org/docs/r/
# if(!('arrow' %in% installed.packages())) install.packages('arrow')
library(arrow)

write_feather(esales, "esales.feather")
```

```
# Close connection -- this is good practice
dbDisconnect(db)
dbUnloadDriver(db_driver)
```

## Exploratory data analysis

```
library(arrow)
```

```
##
## Attaching package: 'arrow'
```

```
## The following object is masked from 'package:utils':
##
##     timestamp
```

```
esales <- read_feather("esales.feather")
```

### Provide a brief example of the data, showing how they are structured.

```
print(esales)
```

```
## # A tibble: 233 x 4
##     value date        year month
##     <dbl> <date>     <int> <int>
##  1  8282. 2020-05-01  2020     5
##  2  7839. 2020-04-01  2020     4
##  3  8889. 2020-03-01  2020     3
##  4  9368. 2020-02-01  2020     2
##  5  9209. 2020-01-01  2020     1
##  6 10038. 2019-12-01  2019    12
##  7  9291. 2019-11-01  2019    11
##  8  8757. 2019-10-01  2019    10
```

```
##  9  9874. 2019-09-01  2019     9
## 10 10912. 2019-08-01  2019     8
## # ... with 223 more rows
```

```
# References: https://www.tidyverse.org/, https://dplyr.tidyverse.org/

esales %>%
  filter(year == 2019) %>%
  filter(value > 9000) %>%
  print()
```

```
## # A tibble: 10 x 4
##     value date        year month
##     <dbl> <date>     <int> <int>
##  1 10038. 2019-12-01  2019    12
##  2  9291. 2019-11-01  2019    11
##  3  9874. 2019-09-01  2019     9
##  4 10912. 2019-08-01  2019     8
##  5 11527. 2019-07-01  2019     7
##  6  9903. 2019-06-01  2019     6
##  7  9147. 2019-05-01  2019     5
##  8  9466. 2019-03-01  2019     3
##  9  9148. 2019-02-01  2019     2
## 10 10925. 2019-01-01  2019     1
```

```
esales %>%
  group_by(month) %>%
  summarise(mean = mean(value)) -> mean_esales_by_month
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
esales %>%
  mutate(sales_TWh = value/1000) %>%
  select(-value)
```

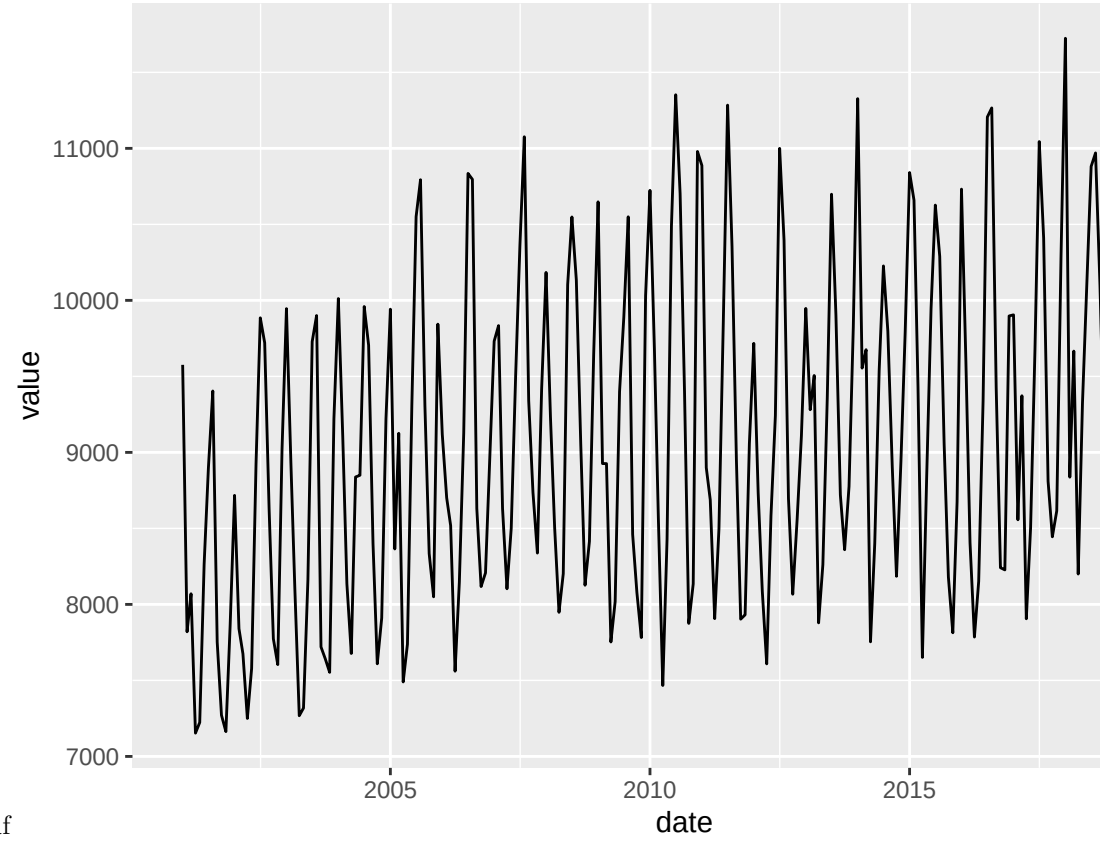
```
## # A tibble: 233 x 4
##    date        year month sales_TWh
##    <date>     <int> <int>     <dbl>
##  1 2020-05-01  2020     5      8.28
##  2 2020-04-01  2020     4      7.84
##  3 2020-03-01  2020     3      8.89
##  4 2020-02-01  2020     2      9.37
##  5 2020-01-01  2020     1      9.21
##  6 2019-12-01  2019    12     10.0
##  7 2019-11-01  2019    11      9.29
##  8 2019-10-01  2019    10      8.76
##  9 2019-09-01  2019     9      9.87
## 10 2019-08-01  2019     8     10.9
## # ... with 223 more rows
```

```
# filter(data object, condition) : syntax for filter() command
```

### Plot the time series.

```
#Reference: https://ggplot2.tidyverse.org/

ggplot(data=esales, aes(x=date,y=value)) + geom_line()
```




ggplot2 to generate a plot-1.pdf

### Perform and report the results of other exploratory data analysis.

## Statistical model

### Formal model of data-generating process

Write down an equation (or set of equations) that represent the data-generating process formally.

If applicable: describe any transformations of the data (e.g., differencing, taking logs) you need to make to get the data into a form (e.g., linear) ready for numerical analysis.

What kind of process is it? $AR(p)$? White noise with drift? Something else?

Write down an equation expressing each realization of the stochastic process $y_t$ as a function of other observed data (which could include lagged values of $y$), unobserved parameters ($\beta$), and an error term ($\varepsilon_t$). Ex:

$$y = X \cdot \beta + \varepsilon$$

Add a model of the error process. Ex: $\varepsilon \sim N(0, \sigma^2 I_T)$.

### Discussion of the statistical model

Describe how the formal statistical model captures and aligns with the narrative of the data-generating process. Flag any statistical challenges raised by the data generating process, e.g. selection bias; survivorship bias; omitted variables bias, etc.

## Plan for data analysis

Describe what information you wish to extract from the data. Do you wish to... estimate the values of the unobserved model parameters? create a tool for forecasting? estimate the exceedance probabilities for future realizations of $y_t$?

Describe your plan for getting this information. OLS regression? Some other statistical technique?

If you can: describe briefly which computational tools you will use (e.g., R), and which packages you expect to draw on.

## Submission requirements

Prepare your proposal using Markdown. (You may find it useful to generate your Markdown file from some other tool, e.g. R Markdown in R Studio.) Submit your proposal by pushing it to your repo within the course organization on Github. When your proposal is ready, notify the instructor by also creating a submission for this assignment on Collab. Please also upload a PDF version of your proposal to Collab as part of your submission.

## Comment

Depending on your prior experience, you may find this assignment challenging. Treat this assignment as an opportunity to make progress on your own research program. Make your proposal as complete as you can. But note that this assignment is merely the First Draft. You will have more opportunity to refine your work over the next two months, in consultation with the instructor, your advisor, and your classmates.

## References

Council, National Research. 2007. *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*. https://doi.org/10.17226/11699.