DATA 624, Spring 2025, Project #1

**Format:** Group effort, no interaction with others outside of your group on this assignment

**DUE:** 03/16/25 by 12 PM ET

**Submission:** Via your Group Representative to my email – scott.burk@sps.cuny.edu

**Submission:** Word Readable Document for Report (**all in one**), Excel Readable (**all in one, separate sheets**) for forecasts.

**File NAMING Convention:** Group#\_Project1\_Spring2025

**Grading Objectives:**

1. Report Quality and Readability – 55%
2. Each of the Following 3 Parts Accuracy – 15% (based on RMSE)
3. Total Possible 100%
4. Percent of Total Grade – 15%

**Report Submission – MAJOR Part of Grade, 55%**

Your group will submit a professional, easy to read report. This may include:

1. Table of Contents
2. Executive Summary
3. Interpretation of Goals – 1 per problem
   1. Data exploration
   2. Approach to take and why
   3. Walk through each part of the analysis.
   4. What you are trying to achieve, the way you are attempting to do that and why.
   5. What you are finding and how it might impact the next step.
   6. Iterate
   7. Results, interpretation of results. Predictions, etc.
   8. Summary

The consumers of this report are executives, business professionals and data scientists. You need to communicate to all audiences; therefore, you cannot just present a technical report. You should provide commentary on your approach, why you are taking this approach and your findings along the way. The report should be very easy to navigate, follow and understand. You must explain what/how/why. And submit your forecasts. Your representative will submit the materials to me in an email with a minimum of **two attachments** – A Word readable doc (Report), An Excel readable doc (my XLSX history to you with the forecasts). I will need your code in the Word document (above) so I can reproduce the results. An R Markdown file is appreciated but not required. Please include all the libraries you are using up front in the code. All code from A to Z and the code should be ***well documented*** as if you are passing it off to a data scientist.

**Part A – ATM Forecast, ATM624Data.xlsx, Accuracy – 15%**

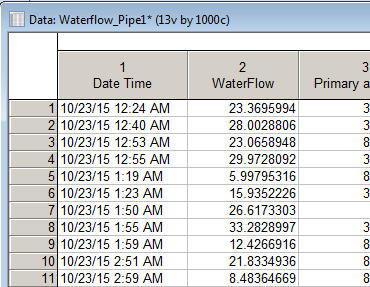
In part A, I want you to forecast how much cash is taken out of 4 different ATM machines for May 2010 (forecast this month). The data is given in a single file. The variable ‘Cash’ is provided in hundreds of dollars, other than that it is straight forward. I am being somewhat ambiguous on purpose. I am giving you data, please provide your written report on your findings, visuals, discussion and your R code all within a Word readable document, except the forecast which you will put in an Excel readable file. I must be able to cut and paste your R code and run it in R studio. Your report must be professional – most of all – readable, EASY to follow. Let me know what you are thinking, assumptions you are making! Your forecast is a simple CSV or Excel file that MATCHES the format of the data I provide – same columns.

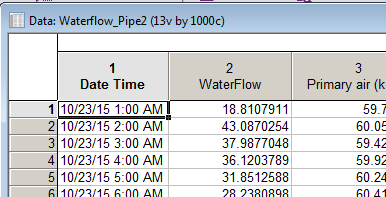
**Part B – Forecasting Power, ResidentialCustomerForecastLoad-624.xlsx, Accuracy – 15%**

Part B consists of a simple dataset of residential power usage for January 1998 until December 2013. Your assignment is to model these data and create a monthly forecast for 2014. The data is given in a single file. The variable ‘KWH’ is power consumption in Kilowatt hours, the rest is straight forward. **Add these as separate tab to your existing XLSX file above – clearly labeled.**

**Part C – Waterflow\_Pipe1.xlsx and Waterflow\_Pipe2.xlsx, Accuracy – 15%**

Part C consists of two data sets. These are simple 2 columns sets, however they have different time stamps. Your assignment is to time-base sequence the data and aggregate based on hour ***(example of what this looks like, follows***). Note for multiple recordings within an hour, take the mean. Then to test appropriate assumptions and forecast a week forward with confidence bands (80 and 95%). **Add these to your existing XLSX file above – clearly labeled.**





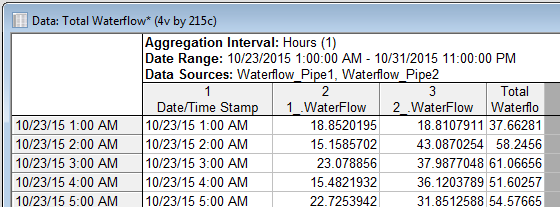


Table of Contents

1. Executive Summary
2. Interpretation of Goals
3. Problem 1
4. Data Exploration
5. Approach to Take and Why
6. Walkthrough of Analysis
7. Objective and Methodology
8. Findings and Impact on Next Steps
9. Iteration
10. Results and Interpretation (Predictions, etc.)
11. Summary
12. Problem 2
13. Data Exploration
14. Approach to Take and Why
15. Walkthrough of Analysis
16. Objective and Methodology
17. Findings and Impact on Next Steps
18. Iteration
19. Results and Interpretation (Predictions, etc.)
20. Summary
21. Problem 3
22. Data Exploration
23. Approach to Take and Why
24. Walkthrough of Analysis
25. Objective and Methodology
26. Findings and Impact on Next Steps
27. Iteration
28. Results and Interpretation (Predictions, etc.)
29. Summary
30. Summary

**Problem 3**

Problem Statement: The following assignment consists of two datasets. They are two simple 2 columns datasets, but they have different timestamps. The task is to time-base sequence the data and aggregate based on the hour. Then to test appropriate assumptions and forecast a week forward with confidence bands (80 and 95%).

**Data Exploration**

*# load libraries*

*library(readxl)*

*library(dplyr)*

*library(lubridate)*

*library(tsibble)*

*library(forecast)*

*library(writexl)*

*# reading in excel files*

*data1 <- read\_excel("Waterflow\_Pipe1.xlsx")*

*head(data1)*

*data2 <- read\_excel("Waterflow\_Pipe2.xlsx")*

*head(data2)*

**Dataset One: Waterflow\_Pipe1.xlsx** 

A screenshot of a data

AI-generated content may be incorrect.

The Waterflow\_Pipe1.xlsx dataset has 1000 records of water flow measurements taken at different times. There are two columns: DateTime and WaterFlow. DateTime column contains date and time values. WaterFlow contains water flow measurements. The DateTime column has a range from October 23, 2015, to November 1, 2015. The DateTime is represented in a standard datetime format. Each entry in this column includes both the date and time, formatted as YYYY-MM-DD HH:MM:SS.ssssss. The WaterFlow values range from approximately 1.07 to 38.91.

**Dataset Two: Waterflow\_Pipe2.xlsx**

The Waterflow\_Pipe2.xlsx dataset has 1000 records of water flow measurements taken at different times. There are two columns: DateTime and WaterFlow. The DateTime column contains date and time values. The WaterFlow column contains water flow measurements. The DateTime column has a range from October 23, 2015, to December 3, 2015. The DateTime is represented as fractional days like (example: 42300.041666666701). The WaterFlow values range from approximately from 1.88 to 78.30.

**Approach to Take and Why**

The primary difference between the two datasets is the representation of the DateTime columns. In Waterflow\_Pipe1.xlsx, the DateTime values are represented in a standard datetime format (YYYY-MM-DD HH:MM:SS.ssssss), while in Waterflow\_Pipe2.xlsx, the DateTime values are represented as fractional days since a reference date. This difference in representation affects how the data is going to be processed and interpreted. To ensure consistency and accuracy in the analysis, it is essential to standardize the DateTime formats across both datasets.

**Walkthrough of Analysis**

Objective and Methodology

Objective: The goal is to time-base sequence the data and aggregate it based on the hour, then test appropriate assumptions and forecast a week forward with confidence bands (80% and 95%).

Methodology:

1. Combine both dataframes: Merge the datasets from Waterflow\_Pipe1.xlsx and Waterflow\_Pipe2.xlsx.
2. Convert 'DateTime' column to datetime timestamp format: Standardize the DateTime representation by converting all DateTime values to a common datetime timestamp format.
3. Replace the datetime with date + hour: Simplify the DateTime values to include only the date and hour.
4. Average for each hour: Calculate the average values for each hour to create a time-based sequence.
5. Plot ACF and PACF: Plot the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) to identify an appropriate prediction model.
6. ETS vs ARIMA: Select and use the best prediction model for the data.
7. Forecast a week forward: Generate forecasts for the next week with 80% and 95% confidence bands.

Steps 1 – 4:

*# combine both dataframes*

*data <- bind\_rows(data1, data2)*

*head(data)*

*#convert 'DateTime' column to datetime timestamp format*

*data <- data %>% mutate(DateTime = as\_datetime(DateTime)) %>% mutate( date = as.Date(DateTime), hour = paste(format(DateTime, format = "%H"), ":00:00")) #a process is used to round the DateTime to the nearest hour*

*head(data)*

*# replace the datetime datetime with date + hour*

*data <- data %>% mutate(DateTime = ymd(date) + hms(hour)) %>% group\_by(DateTime) head(data)*

*#average for each hour to a specific hour*

*data <- data %>% summarise(WaterFlow = mean(WaterFlow, na.rm = TRUE)) %>% as\_tsibble(index = DateTime)*

*head(data)*

Data Aggregation Results:

A screenshot of a screen

AI-generated content may be incorrect.

Step 5:

Plot the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) function plots in order to to help to identify the appropriate model for time series analysis by showing the correlation between observations at different lags.

*# plot ACF and PACF*

*acf(data$WaterFlow)*

*pacf(data$WaterFlow)*

A graph of a line

AI-generated content may be incorrect.

ACF and PACF plots analysis:

The following ACF plot shows a strong correlation at lag 1 that seems like it might be gradually decreasing, but there is no strong evidence to prove this hypnosis. Additionally, the Augmented Dickey-Fuller Test will be performed to check for stationarity.

*# check if the combined data is stationary*

*adf\_test <- adf.test(data$WaterFlow, alternative = "stationary")*

*print(adf\_test)*

Augmented Dickey-Fuller Test:

data:  data$WaterFlow

Dickey-Fuller = -7.0271, Lag order = 9, p-value = 0.01

alternative hypothesis: stationary

The Augmented Dickey-Fuller test results suggest that the water flow data is stationary, meaning it doesn't have a trend over time.

The following PACF plot shows significant spikes at early lags but does not cut off sharply, which suggests that an ARIMA model might be a good choice.

Step 6:

Since the ACF indicates stationarity and the PACF suggests an autoregressive structure, ARIMA model is a better choice compared to ETS. Also, since data is stationary d = 0. The first ARIMA model to try is going to be ARIMA(1, 0, 1);

**ARIMA (1, 0, 1)**

*# fit ARIMA(1, 0, 1) model*

*arima\_model <- Arima(data$WaterFlow, order = c(1, 0, 1))*

*summary(arima\_model)*

ARIMA(1,0,1) with non-zero mean

Coefficients:

         ar1      ma1     mean

      0.9966  -0.9594  36.0915

s.e.  0.0031   0.0100   4.4176

sigma^2 = 213.1:  log likelihood = -4103.02

AIC=8214.04   AICc=8214.08   BIC=8233.68

Training set error measures:

                    ME     RMSE      MAE       MPE     MAPE      MASE        ACF1

Training set 0.3523111 14.57568 11.23132 -27.23585 48.26237 0.7132861 -0.02267587

*# plot residuals*

*tsdisplay(residuals(arima\_model), main = "Residuals of ARIMA(1, 0, 1) Model")*

A graph of a model

AI-generated content may be incorrect.

The residuals of the ARIMA(1,0,1) model mostly center around zero, but its variance changes over time, which can be a sign of inconsistency. The ACF plot shows most residuals within the confidence intervals, indicating no strong correlation, though a few minor spikes. The PACF plot shows no major lag dependencies, indicating no strong additional autoregressive structure. Overall, this model is not promising but might not be the best.

*# AIC value*

*AIC(arima\_model)*

[1] 8214.043

**auto.arima()**

*# auto.arima to find the best model*

*auto\_arima\_model <- auto.arima(data$WaterFlow) summary(auto\_arima\_model)*

ARIMA(0,1,1)

Coefficients:

          ma1

      -0.9629

s.e.   0.0083

sigma^2 = 212.9:  log likelihood = -4100.12

AIC=8204.24   AICc=8204.25   BIC=8214.05

Training set error measures:

                   ME     RMSE      MAE       MPE     MAPE      MASE        ACF1

Training set 0.531199 14.57593 11.14363 -26.38446 47.91859 0.7077172 -0.02311872

*# AIC - auto.arima*

*AIC(auto\_arima\_model)*

[1] 8204.239

Comparing ARIMA(1, 0, 1) AIC vs ARIMA(0, 1, 1) AIC: 8214.043 vs 8204.239

Since auto.arima  ARIMA(0, 1, 1) AIC is slightly less than ARIMA(1, 0, 1), and stationarity by was initially non-consistent,  ARIMA(0, 1, 1) can be a better option to choose for the forcasting.

Step 7:

*# forecast for the next 7 days (80% and 95% confidence lvls)*

*forecast\_011 <- forecast(auto\_arima\_model, h = 7, level = c(80, 95))*

*print(forecast\_011)*

*# save the data frame to excel file*

*forecast\_df <- data.frame( DateTime = seq(from = max(data$DateTime) + 1, by = "hour", length.out = 7), Forecast = forecast\_011$mean, 80% Lower CI = forecast\_011$lower[, 1], 80% Upper CI = forecast\_011$upper[, 1], 95% Lower CI = forecast\_011$lower[, 2], 95% Upper CI = forecast\_011$upper[, 2] )*

*write\_xlsx(forecast\_df, "ARIMA\_0-1-1\_forecast.xlsx")*

A screenshot of a spreadsheet

AI-generated content may be incorrect.

*# plot the forecast*

*autoplot(forecast\_011) + ggtitle("ARIMA(0, 1, 1) Forecast with 80% and 95% Confidence Intervals") + xlab("DateTime") + ylab("WaterFlow")*

**TBD: ANALYSIS AND CONLUSION< SUMMURY ETC**

Data Exploration: Describe the Waterflow\_Pipe1.xlsxdataset, its structure, and any initial observations.

* Approach to Take and Why: Explain the chosen methodology and rationale.
* Walkthrough of Analysis:
* Objective and Methodology: Detail what you aim to achieve and how.
* Findings and Impact on Next Steps: Discuss initial findings and their implications.
* Iteration: Describe any iterative processes or refinements.
* Results and Interpretation: Present final results, predictions, and insights.
* Summary: Summarize the key points and conclusions.