**Will my baggage make it to my final destination?**

**BAN 6025 Forecasting Project**

If you travel often, it is almost certain that you have experienced delays, cancelled flights, and other disruptions in your travel schedule. But, one of the most frustrating problems that you can experience is having the airline mishandle your baggage. When your bags are lost or delayed or sent to the wrong destination, it can really ruin your day!

The data set you will be analyzing for this project contains operational data from three airlines for the years 2004 – 2010. The data are recorded at a monthly level. The variables included are:

|  |  |
| --- | --- |
| **Variable Name** | **Variable Description** |
| Airline | American Eagle, Hawaiian, United |
| Date | Month and year |
| Month | Month |
| Year | Year |
| Baggage | The total number of passenger complaints for lost, damaged, or misrouted baggage |
| Scheduled | The total number of flights scheduled for the airline that month |
| Cancelled | The total number of flights cancelled by that airline that month |
| Enplaned | The total number of passengers who boarded a plane with the airline that month |

These data are available from the U.S. Department of Transportation, “Air Travel Consumer Report,” the Office of Aviation Enforcement and Proceedings, Aviation Consumer Protection Division. The data for baggage complaints and enplaned passengers cover domestic travel only.

Suppose that it is the end of 2010, and you have been hired by a travel agent trying to understand what to expect as she advises customers regarding their travel plans for the first six months of 2011. She would like to understand how things like the number of passengers travelling or the number flights cancelled might be related to the likelihood of having to make a baggage claim. She would also like to see a forecast of the number of baggage claims for the next six months. Further, if you determine that a factor like the number of cancelled flights seems to be related to the number of baggage claims, she would like to see a forecast for this factor as well to provide additional insight to her customers.

**Producing the Forecasts:**

As you might imagine, travel patterns (and hence related factors like baggage claims and cancelled flights) are somewhat seasonal. This suggests that your forecast will need to incorporate seasonality. This document will walk you through the process of producing a seasonal forecast for the number of baggage claims received monthly by United Airlines. The code provided includes all of the structure that you will need in order to reproduce this analysis for another airline or another factor (like the number of cancelled flights).

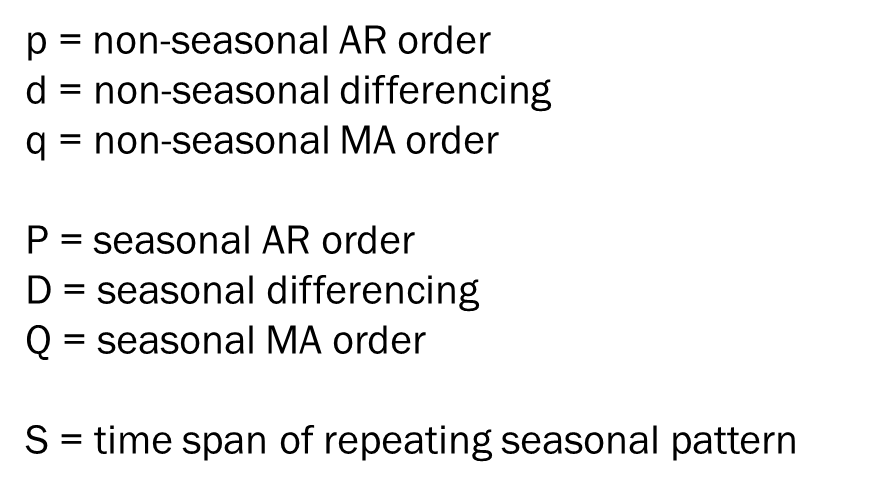
**Identifying a Seasonal ARIMA Model:**

If you recognize that your time series has seasonality (either by examining the time plot or the ACF and PACF plots), then you will need to account for that seasonality before beginning the iterative process of choosing the terms to include in the model. So, the full process goes like this:

1. Plot the time series
2. Plot the ACF and PACF.
3. Do any necessary differencing to neutralize the stationarity.
   1. If the data exhibit seasonality, and no trend, then you begin by taking a seasonal difference of lag S, where S is the length of the seasonal period. So, for example, if it takes one year (12 months) to complete the seasonal cycle, then S = 12. A seasonal difference of lag 12 is given by Yt – Yt-12
   2. If the data exhibit seasonality AND trend, then take the seasonal difference and examine the residuals. If there is still evidence of a trend, then take a first difference as well.
4. Examine the ACF and PACF of the residuals from the differenced data
   1. First look at the lower lags (1 and 2) to determine whether you believe that any nonseasonal AR or MA terms need to be included in the model. This is exactly the same process that you used for determining the values of p and q for a nonseasonal model.
   2. Next, look at the lags associated with the seasonal cycle. For example, if the seasonal cycle has S = 12 months, then look at what’s happening at lags 12, 24, 48, etc in the ACF and PACF. Are there significant values at those lags? If so, then you may need to add a seasonal AR or MA term. So, for example, if you see a significant positive value in the PACF at lag 12, you might add a seasonal AR term to the model – which means that the model would include the term Yt-12
5. Once you have decided which term to add to the model, you will fit the model and examine the residuals. You are checking to see whether the terms in the model are significant and whether the residuals are white noise.
6. You will continue to iterate by adding regular AR and MA terms as well as seasonal AR and MA terms until you achieve a “good” model – which means that all (or most) of the terms in the model are significant, the residuals are white noise, and the forecast error is generally lower than for other models being considered.

**A word about notation:**

The notation for a seasonal ARIMA models is slightly more complex than for a non-seasonal ARIMA model. The notation that we’ll use is ARIMA(p, d, q)x(P, D, Q)[S] where



So, for example, an ARIMA(2, 0, 1)x(1, 0, 1)[12] model would have 2 non-seasonal AR terms, 1 non-seasonal MA term, 1 seasonal (lag 12) AR term, and 1 seasonal (lag 12) MA term. It will have the form:

**Translating to Python:**

To fit a seasonal ARIMA model in Python, you will use the same ARIMA function that we used to fit non-seasonal ARIMA models with a few extra options (see code file for full details). Using a general notation, you will need to use the following structure:

fitmod = ARIMA(series, order=(p, d, q), seasonal\_order=(P, D, Q, S)).fit()

So, order specifies the non-seasonal terms to be included in the model and seasonal\_order specifies the seasonal terms to be included in the model. For example, to fit the model illustrated above, you would use the code

fitmod = ARIMA(series, order=(2,0,1), seasonal\_order=(1, 0, 1, 12)).fit()

To take a seasonal difference for a series with a 12-month seasonal cycle, you would use the code

fitmod = ARIMA(series, order=(0,0,0), seasonal\_order=(0, 1, 0, 12)).fit()

Again, this process is fully illustrated in the code that is provided.

So, essentially, you will need to use the same iterative process that we used for the seven series that we analyzed in class, but you will need to expand your search for a good model to include seasonal terms as well as non-seasonal terms.

**Your Analytical Task:**

**United Airlines:**

I have already created an ARIMA(p, d, q)x(P, D, Q)[S] model for baggage claims for United Airlines. You can assume that this is the final forecast model for these data. You will use this as the basis for comparison when you write your report.

1. Iteratively choose an ARIMA model for the number of cancelled flights for United Airlines and produce the forecast for the number of cancelled flights for the first 6 months of 2011.
2. Examine the correlation between the number of baggage claims and the number of cancelled flights, passengers enplaned, and flights scheduled.

**American Eagle Airlines:**

1. Iteratively choose an ARIMA model for the number of baggage claims for American Eagle and produce the forecast for the number of baggage claims expected for the first 6 months of 2011.
2. Iteratively choose an ARIMA model for the number of cancelled flights for American Eagle and produce the forecast for the number of cancelled flights for the first 6 months of 2011.
3. Examine the correlation between the number of baggage claims and the number of cancelled flights, passengers enplaned, and flights scheduled.

NOTE: You do not need to analyze the data for Hawaiian Airlines.

**Your Deliverable:**

Based on your analyses (which should include the tasks mentioned above as well as any other analysis you might choose to do for the data), you will write a short report for the travel agent that should include the following:

1. Statement of the business problem
2. Overview of the data
3. Presentation of the results for both United Airlines and American Eagle. This should include:
   1. Discussion of which factors might influence baggage claims for each airline
   2. 6-month forecast for baggage complaints for each airline
   3. 6-month forecast for cancelled flights for each airline
   4. Any other analysis that you believe is relevant to help the travel agent understand the data
4. Recommendations and conclusion – this should include a discussion of what to expect from each airline in terms of cancelled flights and baggage claims over the next six months as well as a recommendation for which airline is more likely to get your bags to their final destination without issue and what month(s) might be the best for travel in order to avoid cancelled flights and lost baggage.

Note that this report does not need to be long, but it does need to be written for a business audience. It should not include a discussion of how you arrived at your final model, but it should make some mention of the reliability or goodness of each forecast. You may include any relevant graphs and figures in line with your text. However, again, keep in mind that this report is going to a travel agent who doesn’t understand anything about forecasting methodology or Python.