**Final Project**

**Forecasting Bikeshare Usage for the City of Austin**

You may complete this project with a group of up to 5 students – or you may choose to complete the project on your own. You may choose any software package (or combination of packages) to complete the analysis. You may also need to use techniques learned outside of this class to fully address the business problem. The project is due Thursday, May 3rd at midnight.

**The problem:**

Bike shares are becoming a popular alternative means of transportation in many cities. The City of Austin has a successful bike share program. If you are not familiar with this concept, you can learn more about Austin’s program here:

<https://austinbcycle.com/>

The City of Austin has made data available for over 649,000 bike trips over 2013-2017. The data includes information on bike trip start location, stop location, duration, and type of bike share user. Information on the bike station locations is also provide.

The data have been posted in both .csv and .sas7bdat format.

**Data Dictionary for austin\_bikeshare\_trips\_fin.csv (and austin\_trips.sas7bdat)**

**bikeid:** integer id of bike

**checkout\_time:** HH:MM:SS (see start time for date stamp)

**duration\_minutes:** integer minutes of trip duration

**end\_station \_id:** integer id of end station

**end\_station\_name:** string of end station name

**month:** month of the year

**start\_station\_id:** integer id of start station

**start\_station\_name:** string of start station name (NOTE: You may need to use station name to identify the stations because station id is missing for several trips while station name is not)

**start\_time:** DDMMMYY:HH:MM:SS

**subscriber\_type:** membership type

**trip\_id:** unique trip id (interval)

**year:** year

**day:** day of the month

**date:** MM/DD/YYYY (NOTE: *This is the date that you should use for aggregating the time series. However, there are some missing values for this variable. You should be able to derive the missing values from the other date / time information that is provided if necessary for the stations that you need to analyze*)

**Data Dictionary for austin\_bikeshare\_stations.csv (and austin\_stations.sas7bdat)**

**Latitude:** geospatial latitude

**Location:** (lat, long)

**Longitude:** geospatial longitude

**Name:** string of station name

**Station\_id:** unique integer id of station

**Status:** station status (active, closed, moved, ACL-only)

The primary goal of this analysis is to forecast bike usage (e.g., number of trips initiated at the station) for the next six weeks at the top 3 active stations in terms of number of trips initiated. In order to do this, you will first need to

* Merge the two data sets
* Identify the top 3 active stations from the data
* Aggregate the data by week for time series analysis
* Use the methods learned in class to generate forecasts for the next 6 weeks for each of the top three stations

In addition, to identifying the most popular stations, the City of Austin would like to identify the bottom 3 active stations in terms of usage (e.g., number of trips initiated). They do not need to forecast usage at these stations, but they would like to see a series plot for each of those stations.

Further, the city would like to understand whether there are certain peak times for bike usage. Are there any trends? Are there any seasonal patterns?

Just looking at the series plots for the top 3 stations, are there any noticeable differences in the patterns between the stations? Are there differences in the patterns for the top 3 stations vs the bottom 3 stations?

**Your deliverable:**

You should prepare an executive summary outlining your findings for the City of Austin. The executive summary should include any relevant graphs (e.g. graphs that you reference in your discussion) and should not exceed 3 pages. Your summary should include any interesting findings that could help the City of Austin in their planning to insure that they have enough bikes to meet demand at the high usage stations.

Following the executive summary, you should include a table with the forecasted bike usage for the top 3 stations for the next 6 weeks.

Finally, you should include an appendix summarizing the process you used to generate the forecasts and including the full output from your final model for each of the top 3 stations.