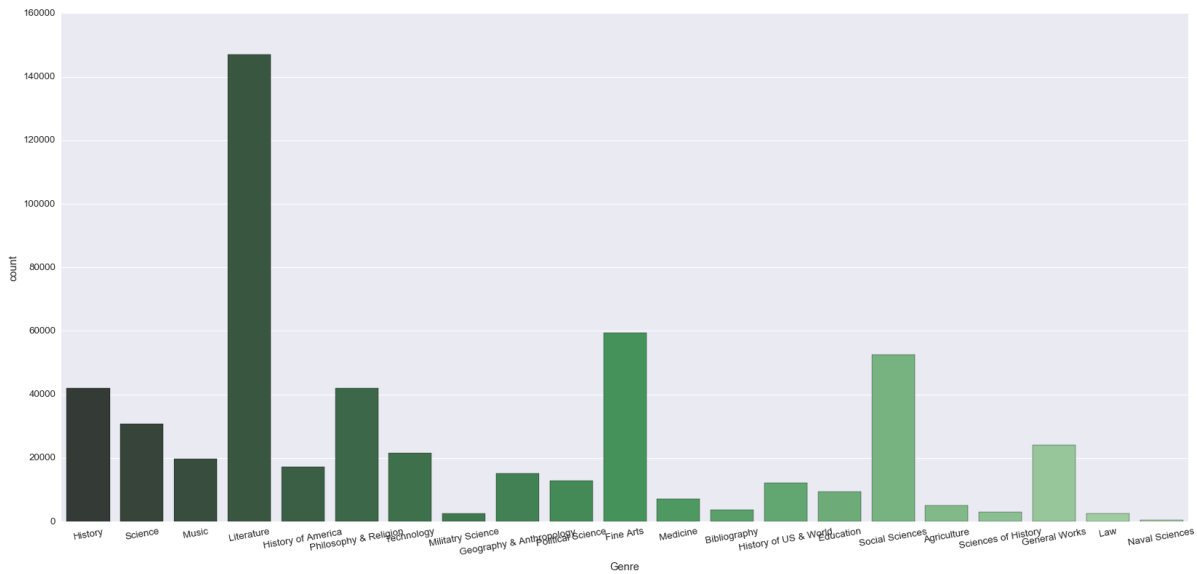
Library Demand using Time Series Modeling:

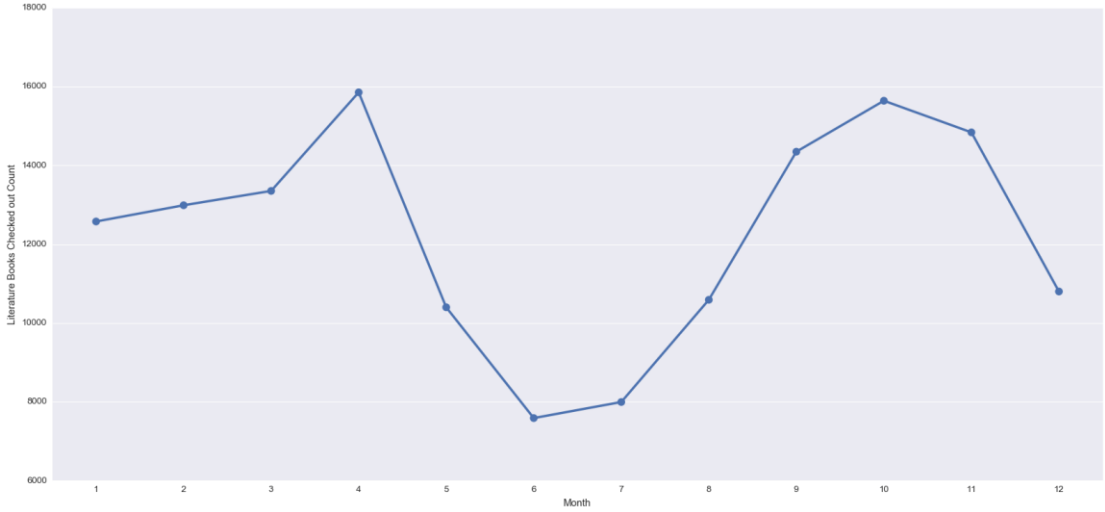
Funding to the libraries in the Universities isn't usually based on any concrete data analysis. This motivated my team and I to take up this topic and make this a data science problem. The goal of this project was to help the Library in the University of Kansas to make better decisions with the public funds received based on genre.
Since this data is confidential, I cannot expose the reader to the complete dataset and will show them as and when necessary. To begin with, I had data of every book checked out from various libraries in the University of Kansas from July 2012 to June 2016. The dataset also had book codes, which were later used to categorize them in different genres using the 'Library of Congress Classification'. To make it clear to the reader, the genres were classified as the image shown below:

```
'A' : 'General Works',
'B' : 'Philosophy & Religion',
'C' : 'Sciences of History',
'D' : 'History',
'E' : 'History of America',
'F' : 'History of US & World',
'G' : 'Geography & Anthropology',
'H' : 'Social Sciences',
'J' : 'Political Science',
'K' : 'Law',
'L' : 'Education',
'M' : 'Music',
'N' : 'Fine Arts',
'P' : 'Literature',
'Q' : 'Science',
'R' : 'Medicine',
```

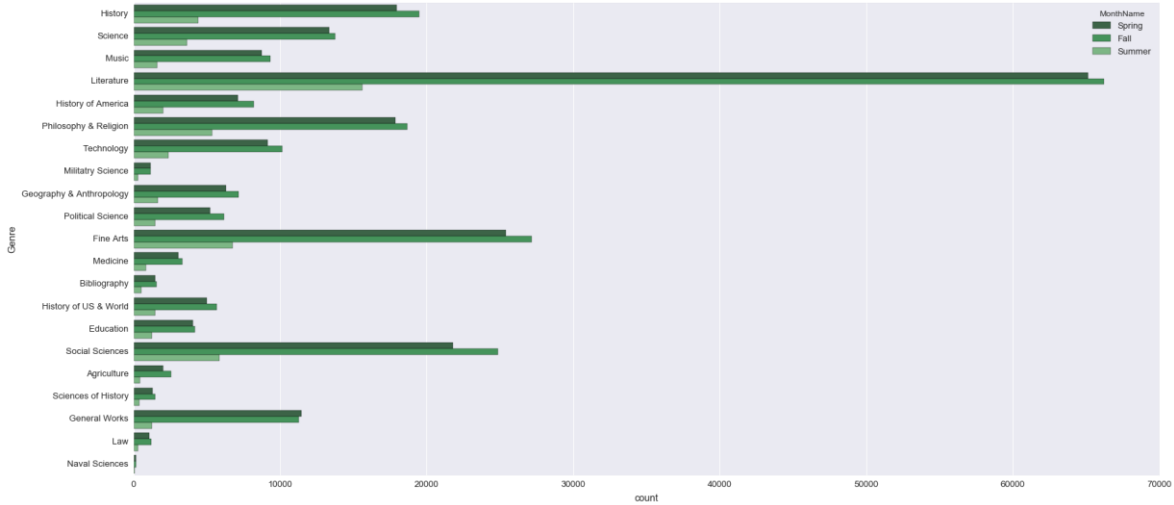
Now that we had all the genres for the books checked out from the University, I intended to do some Exploratory Data Analysis before moving to modeling. First, I wanted to know what type of books were mostly checked out from the library and to no surprise, top three most checked out genres were Social Sciences, Literature and Fine Arts, as evident from the plot below:



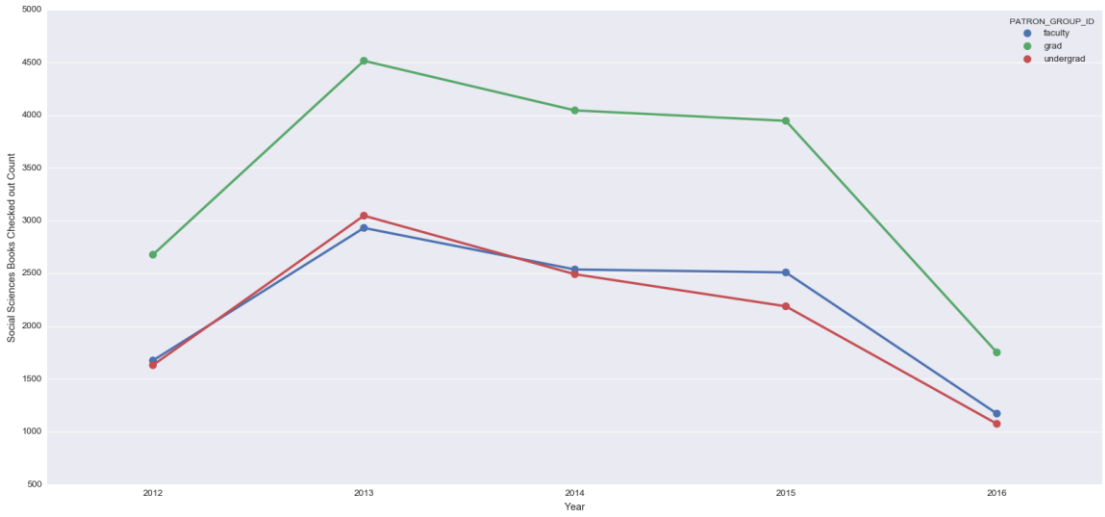
Based on the top three categories, I wanted to then see how those genres behave in each month and it was a good visual representation of what was expected. The below plot is for Literature books that were checked out in each month. As we can see, there is a dip from month 5(May) until month 7(July) which are the usual summer holiday months and there is a rise during the Spring and Fall semesters.



To even strengthen my analysis I plotted a visualization that would better help me understand the semester counts for each genre:



Another interesting observation was to see how the people in the university read. To my surprise, the graduate students whose count is very less when compared to the undergraduates, had more books checked out in some genres!



These analyses helped the library department better understand what libraries are more popular on campus amongst what students. A visualization drawback on the above plot is that it shows a sudden rise from 2012 to 2013 and then a dip from 2015 to 2016. This is because both the years have lesser data.

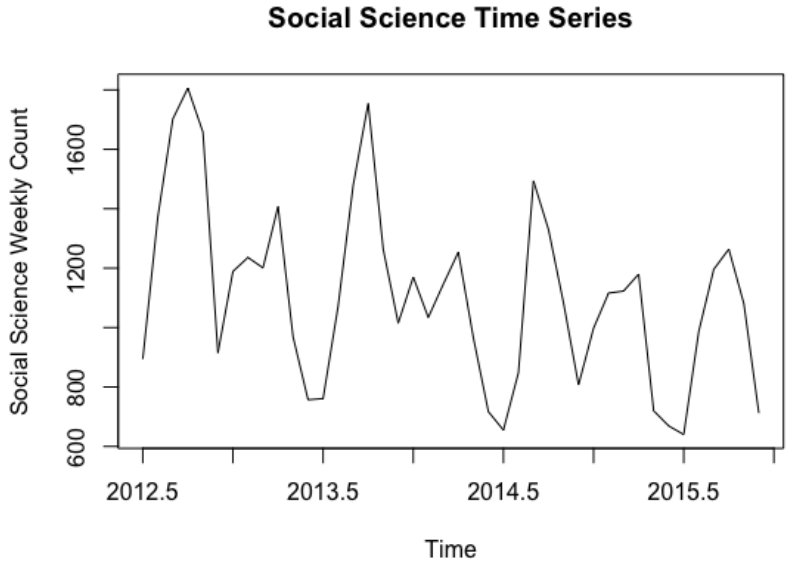
Moving to the prediction part of the problem, I wanted to build a predictive model such that it would help us determine the number of books that would be checked out in the future for a particular genre. This information can then be distributed to specific departments and it would then give rise to a data oriented funding procedure.
The essence of this problem is a Time Series model in which we have the predictor as the count of books each month for each genre. The data that we received has been cleaned, grouped and is in the form best suited for a time series model. The table shown next describes the count of Social Science books checked out each month:

Month	Year	0
7	2012	894
8	2012	1372
9	2012	1702
10	2012	1806
11	2012	1658
12	2012	915
1	2013	1189
2	2013	1236
3	2013	1201
4	2013	1407
5	2013	969
6	2013	757
7	2013	761
8	2013	1072
9	2013	1478

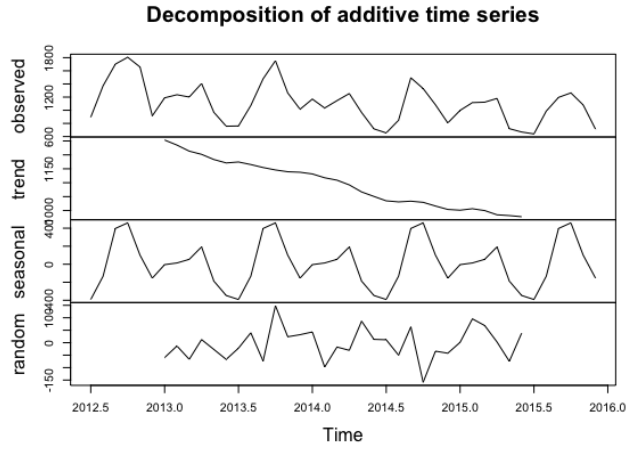
I used a couple of techniques for my time series modeling and they are listed below:

1. Holt-Winters Method:
   This is a very popular scheme to produce a smoothed Time Series where recent observations are given relatively more weight in forecasting than the older observations. Exponential smoothing is a very basic technique

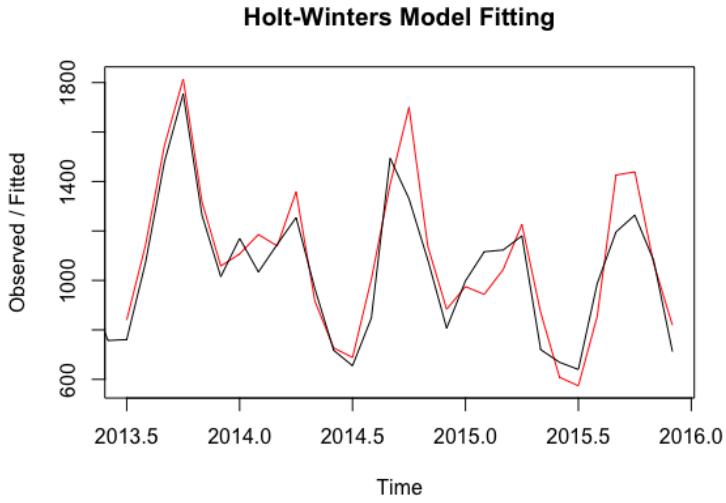
for smoothing time series data. It is an easily learned and easily applied procedure for approximately calculating or recalling some value, or for making some determination based on prior assumptions such as seasonality and trend. In the simple moving average the past observations are weighted equally, exponential window functions assign exponentially decreasing weights over time. To begin our process with Time Series, we first plot the time series:



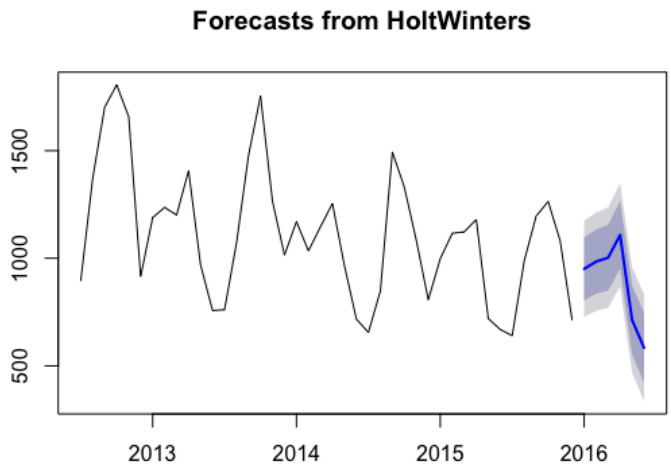
In order to proceed with our model building and prediction, we have to first cater to the stationary and seasonality in the dataset. We do this by decomposing a time series into seasonal, trend and irregular components using moving averages.
For our time series, we plot the additive-modeled series, which can explain the dataset in absolute measure. If in every month of December the number of Social Sciences books that are checked out are 200 books less than in November, the seasonality is additive in nature. It can therefore be represented by 'absolute' increase. However, if the checkouts are 10% less books in the summer months than we do in the spring months the seasonality is multiplicative in nature.



The above plot shows the original time series of the weekly data (top), the estimated trend component (second from top), the estimated seasonal component (third from top), and the estimated irregular component, which is the remainder part of the series, is represented at the bottom. We see that the estimated trend component shows decrease. The Holt-Winters method contains three variables in the equation to fit a time series model and that is why it is also called 'Triple Exponential Smoothing'. The three parameters are α, β and γ. α is the smoothing factor in the range 0 to 1 and as the value of α increases, the recent observations are given more weights and the smoothing effect is lessened. β is the trend-smoothing factor and it is the recursive application of an exponential filter twice, thus being termed "double exponential smoothing". The variable γ takes into account the seasonality quotient of the time series. Here, 'Seasonality' refers to the period of time before the behavior begins to repeat itself.
For our model, I've set β to false since the trend component does not need any smoothing. Now, since there is a seasonality component, we have to use the 'Triple Exponential Smoothing' and after plotting the model with our time series dataset, we have the below plot.



As we can see that the Holt-Winters method does a very good job at trying to find the right model to fit. With this model, we now try to predict on the test dataset, which was the holdout test set.
For our holdout set, we have the data for six months in 2016 and below we see how our model performs:



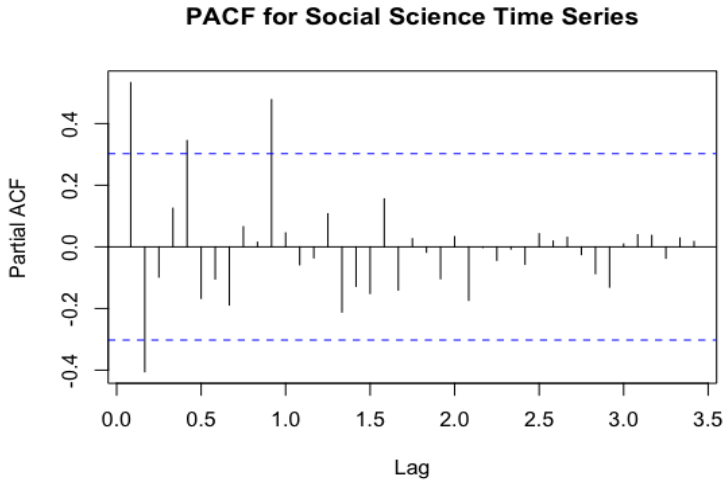

```
Smoothing parameters:
 alpha: 0.2206879
 beta : FALSE
 gamma: 0.6053309

Forecasts:
         Point Forecast    Lo 80     Hi 80    Lo 95     Hi 95
Jan 2016       950.4585 804.6985 1096.2186 727.5378 1173.3793
Feb 2016       984.5078 835.2404 1133.7751 756.2230 1212.7925
Mar 2016      1002.4539 849.7598 1155.1480 768.9284 1235.9794
Apr 2016      1108.2869 952.2412 1264.3325 869.6357 1346.9380
May 2016       713.0020 553.6753  872.3286 469.3329  956.6711
Jun 2016       583.9359 421.3944  746.4773 335.3501  832.5216
```

Reading the result, we see that α has a pretty low value, which signifies that the smoothing effect does not depend very much on the recent observations. Whereas the parameter γ signifies that the effect of seasonality is higher and recent observations have significant role in deciding the next prediction. One of the evaluation metric is RMSE and the RMSE of 114.97 is pretty good. The other evaluation metric I used is MAPE (Mean Absolute percentage error) and for the monthly Social Science count, the model's MAPE is only 1%, which tells us that our model is very accurate in making new predictions and all the predicted values lie within the 95% confidence interval.

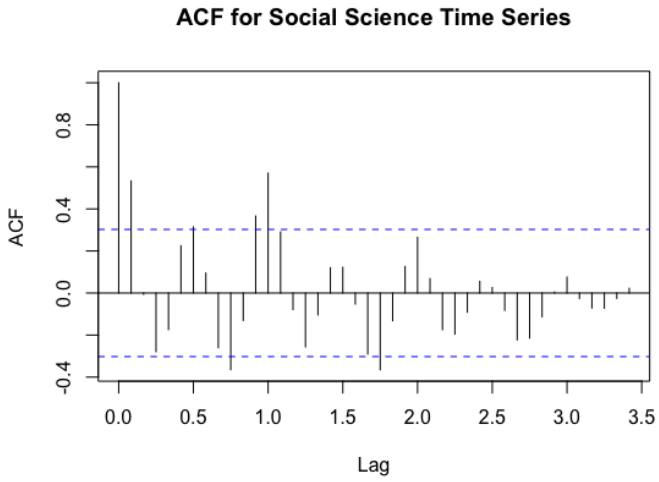
ARIMA Model:

ARIMA models, also called Box-Jenkins models, are models that may possibly include Autoregressive terms, Moving Average terms, and Differencing operations. The Autoregressive part of the model deals with how dependent/correlated an instance is with its pervious values and it is denoted by 'q'. The Autoregressive term can be estimated with the help of Partial Auto-Correlation Function (PACF) plot. Moving Average, or MA part of the model, which is denoted by 'p' is used to account for the shocks a model witnesses. This part of the model deals with sudden peaks and

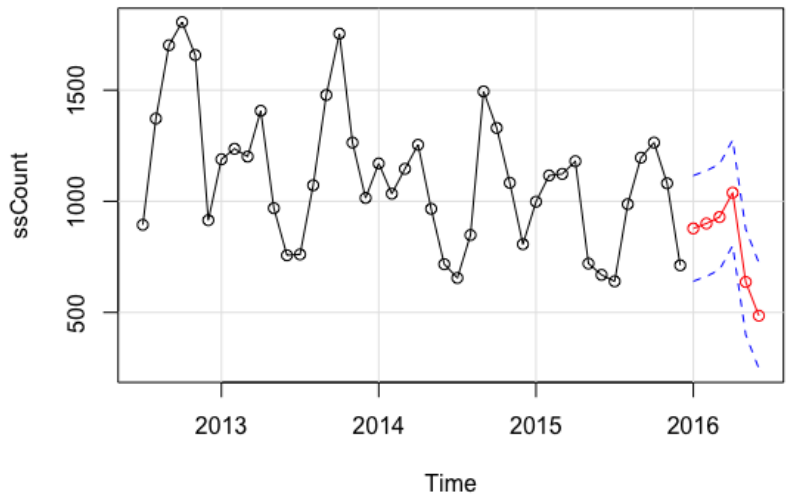
falls of the model and is estimated with an Auto-Correlation Function (ACF) plot.
For an ARIMA model, the time series should be stationary and also seasonality free. Stationary series stems from the fact that the mean and variance of the time series should be nearly constant and should not be a function of time. Seasonality refers to patterns that repeat itself over time and we clearly saw from the time series above that our dataset has seasonality patterns. To overcome these shortcomings, we first difference our time series (represented by 'd') to make it stationary and then use the Seasonal ARIMA model to tackle the problem of seasonality. While approaching the seasonal ARIMA, the parameters become P, D and Q and they are still estimated using the PACF and ACF plots.
To estimate P, D and Q we considered both ACF and PACF plots together. AR models have theoretical PACFs with non-zero values at the AR terms in the model and zero values elsewhere and the ACF plot tapers to zero in some fashion. Below is a PACF plot for Social Science genre:



As we can see that after the first couple of spikes, the plot almost tapers to zero but we do not have a definitive order for the AR part of our model.
MA models have theoretical ACFs with non-zero values at the MA terms in the model and zero values elsewhere. There is evidence from the ACF plot below about the order of our MA term, it is almost of order 1:



Since the ACF and PACF do not tail off very considerably and instead has values that stay close to the significance boundary over many lags, this indicates that series is non-stationary and differencing will be needed. In our model, the differencing needed was of order 1.
After estimating the parameters, building our model and making prediction on the hold out dataset, we get the below plot for weekly count data:



The results from ARIMA weren't better than Holt-Winters but were still very good. The RMSE of the predicted model was 179 for Social Sciences whereas the MAPE was recorded to be 17%.

Conclusion:
The models built are pretty accurate and it will definitely help the libraries make better decisions based on the untouched data they have. Time Series modeling is a beast on its own and is very different from the known statistical and machine learning methods. I had a great time learning about Time Series and hope the reader took some insight too!