Michigan Outoor Recreation Search Interest: Second Installment

Author: Dan Shaffer

==========

In the first installment, I discussed the basic topic, data, trends, and correlations for this project.

This project is to study and forecast trends for Google search interest from Michigan for 10 forms of outdoor recreation. I showed the historical trends for this data, which is sometimes trending upward or downwards and always shows various forms of seasonality. I also showed the correlation between various forms of outdoor recreation, which is mostly based on seasonality.

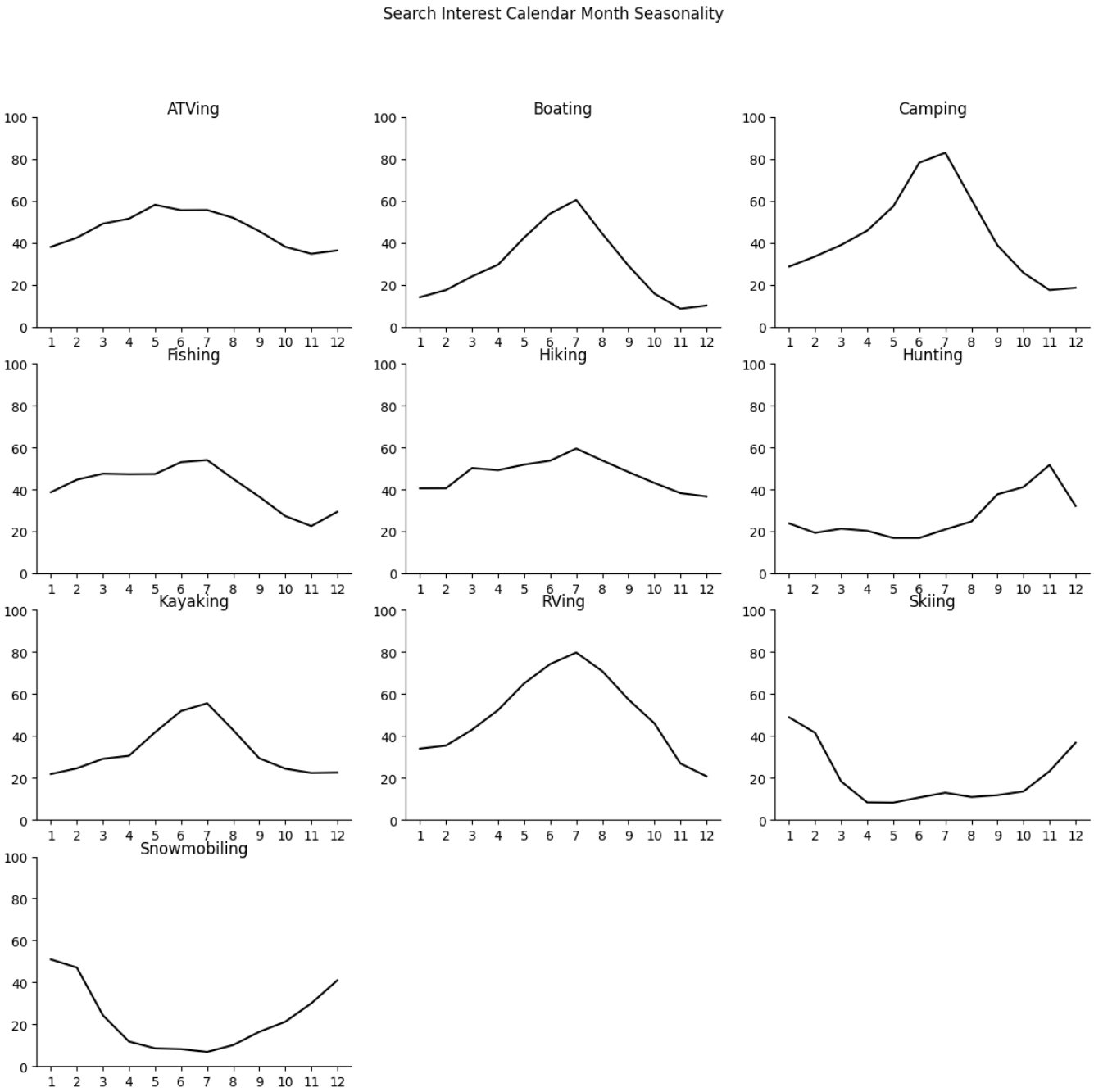
In this installment, I will perform initial statistical analysis on these data trends.

==========

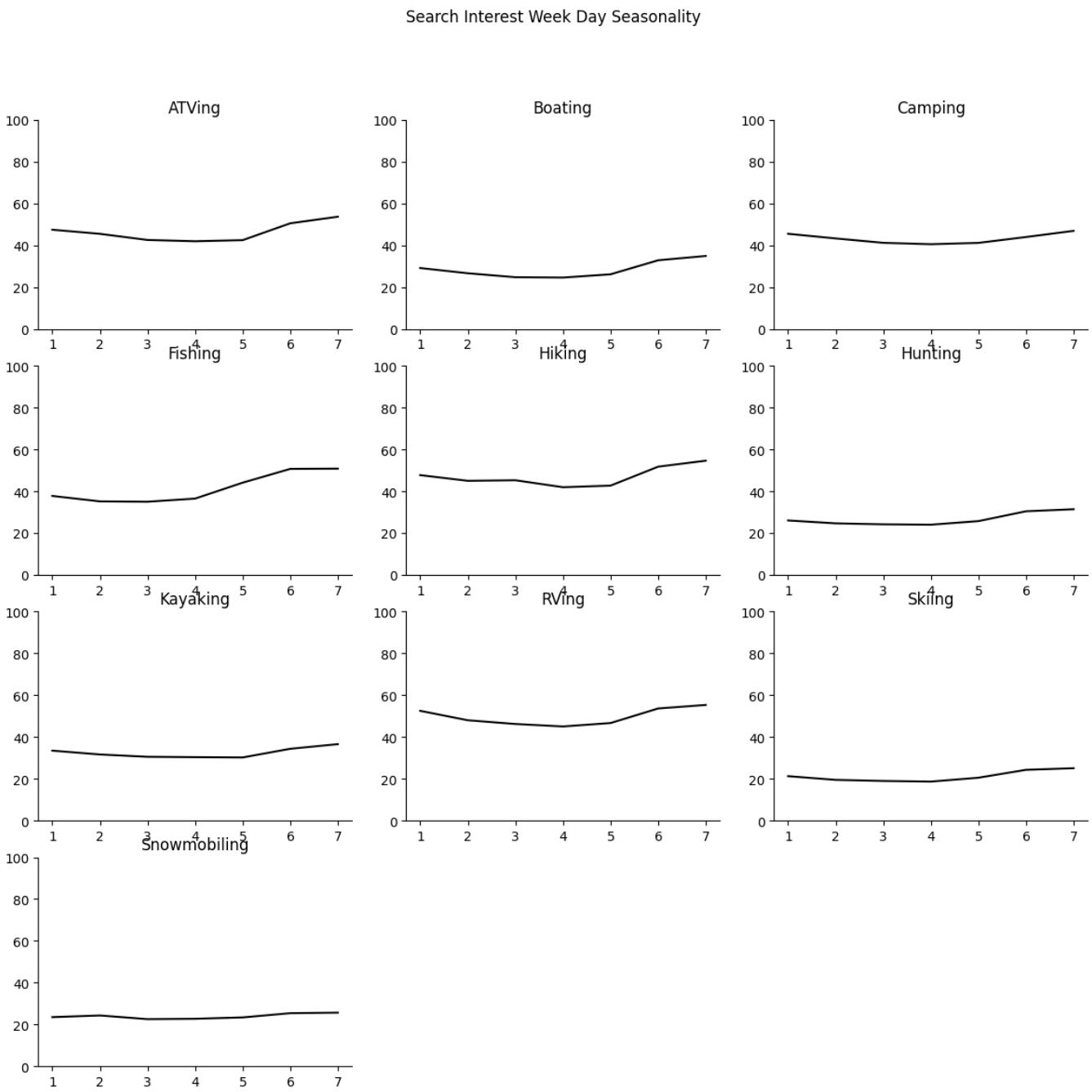
Seasonality

Given that we are working with daily data, there are two forms of seasonality that I'm concerned about. One is annual seasonality and the other is seasonality based on weekday. (Note that the technical term "seasonality" need not be based on the calendar seasons but rather any repeated periodic trend.)

Calendar year seasonality for each of these activities was discussed at length in the first installment. While the final model will likely incorporate this seasonality as a function on day or week of the year, month of the year is much easier to interpret in visualizations. Below, we see that most activities are most popular in the summer. Fishing is also popular in the winter (due to ice fishing). Hunting is most popular in the fall, but secondarily in the spring for turkey hunting. Snowmobiling and skiing are the two activities most popular in the winter.



Not discussed during the first installment was seasonality based on weekday. All the outdoor activities show highest search interest on weekends and Fridays. ATVing, boating, camping, hiking, and rving have the highest relative Sunday values (day 1). Fishing shows the highest relative search interest at the end of the week, and search interest for fishing starts growing on Thursday.



==========

Trend Stationarity

Based on the above seasonality analysis, I deseasonalize the data based on both calendar day and weekday. Next we need to know if the data is trending upwards or downwards. This could lead to spurious correlations between two variables that happen to be trending with no real relationship with each other.

The primary way to check for stationarity is the augmented dick fuller test. This test determines if we can reject the fact that the time series has a "unit root". This concept is beyond the scope of this writeup. However, it suffices to say if the p-value of the test is below 0.05 (or other appropriate significance level)

then we reject the unit root and assume the time series is stationary. In this context, this means that the time series does not have a significant trend (augmented dicky fuller is a poor test for other forms of stationarity).

Based on the ADF tests below, we can reject a unit root for half of the activities and fail to reject the other half, for a 0.05 signficance level. (At 0.1 significance level, we would reject for six of the ten activities.) This means that for about half the activities we can't reject that there is a trend. A common way to address this type of stationarity is by differencing, or looking at the change in the value from day to day. At least for now, I difference all of the data series so that they can easily be compared to each other.

```
ATVing
ADF p-value is:  0.24
Boating
ADF p-value is:  0.0
Camping
ADF p-value is:  0.51
Fishing
ADF p-value is:  0.15
Hiking
ADF p-value is:  0.0
Hunting
ADF p-value is:  0.08
Kayaking
ADF p-value is:  0.01
RVing
ADF p-value is:  0.49
Skiing
ADF p-value is:  0.0
Snowmobiling
ADF p-value is:  0.0

==========================================================================
```

Autocorrelation/Partial Autocorrelation

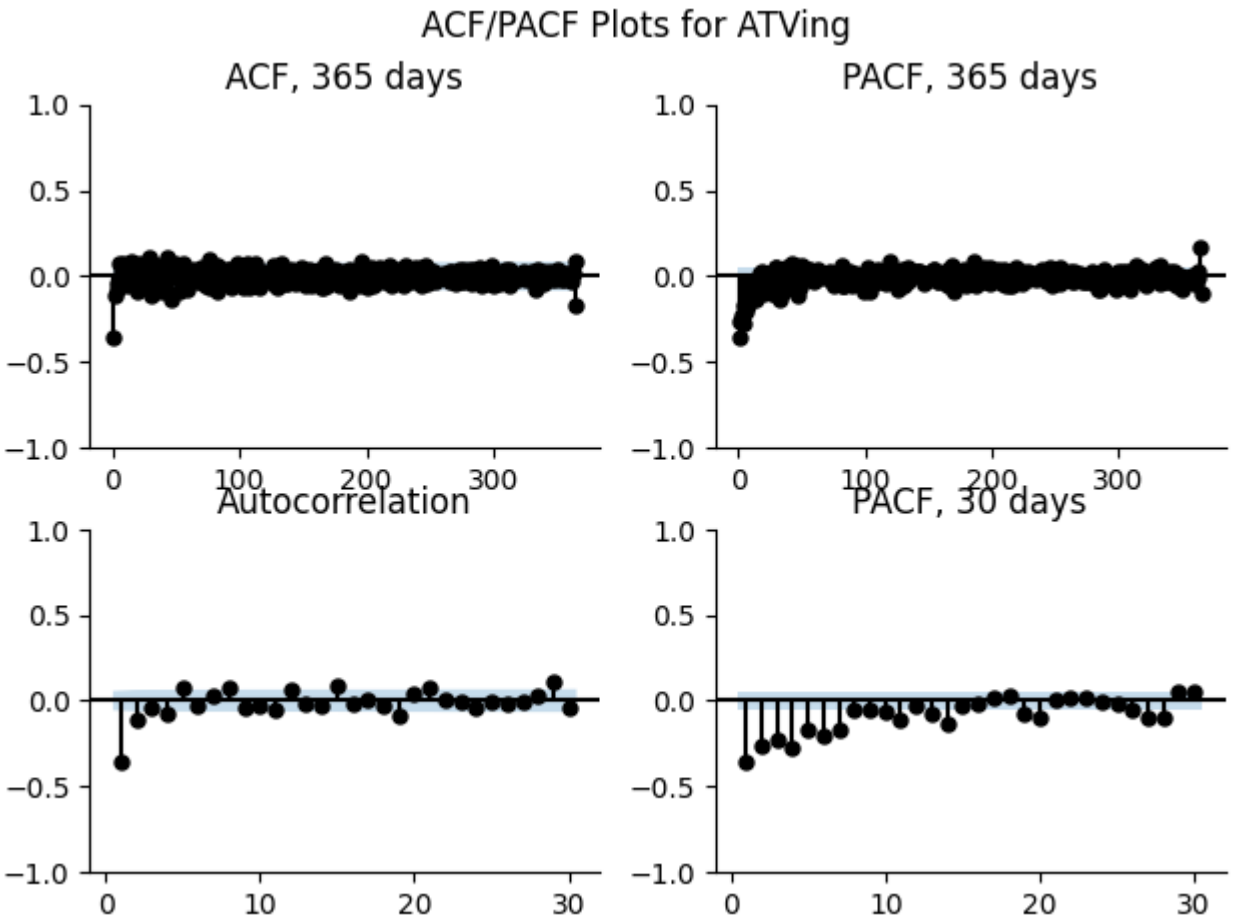
Autocorrelation is the correlation between a data series and the lagged values of the same series. In other words, the first autocorrelation is the correlation between a value and the value the day before. The seventh autocorrelation is the correlation with the value seven days before.

If values are autocorrelated with each other, they will be autocorrelated with values fairly distant in the past even if there is no direct association. In other words, even if a value were entirely unrelated to the value two days ago, if it were related to the value one day ago, the second autocorrelation would still be evident. That's because the current value is correlated with the value one day ago and the value one day ago is, in turn correlated with the value two days ago.

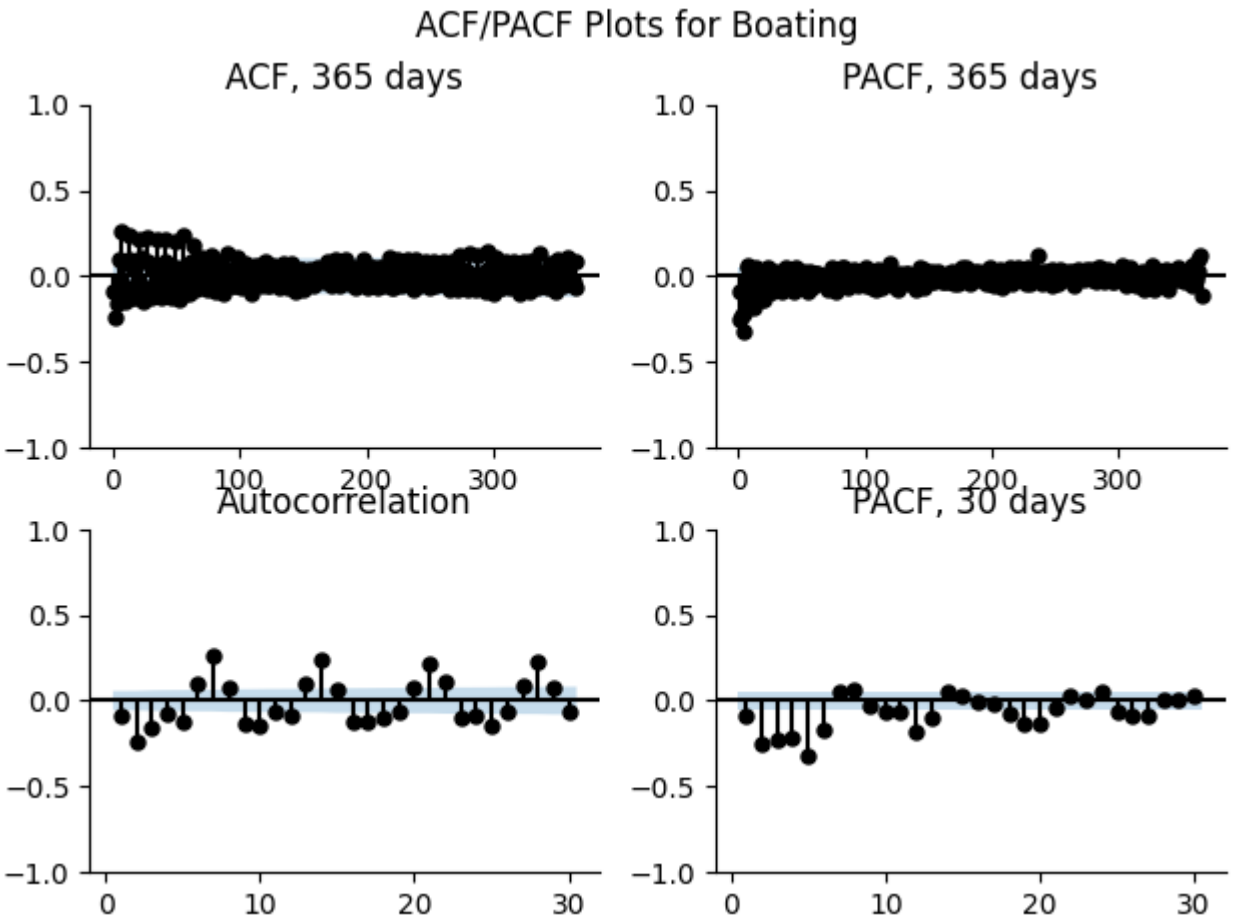
The partial autocorrelation considers the correlation after parting out the correlation from more recent lags. So the first autocorrelation is the same as the first partial autocorrelation. However, the second partial autocorrelation is only the correlation that remains after considering the first autocorrelation. The third parital autocorrelation is the correlation after considering the correlation of the first two lags.

I look at autocorrelation and partial autocorrelation for the each of the activities after deseasonalizing by calendar day, weekday, and differencing.

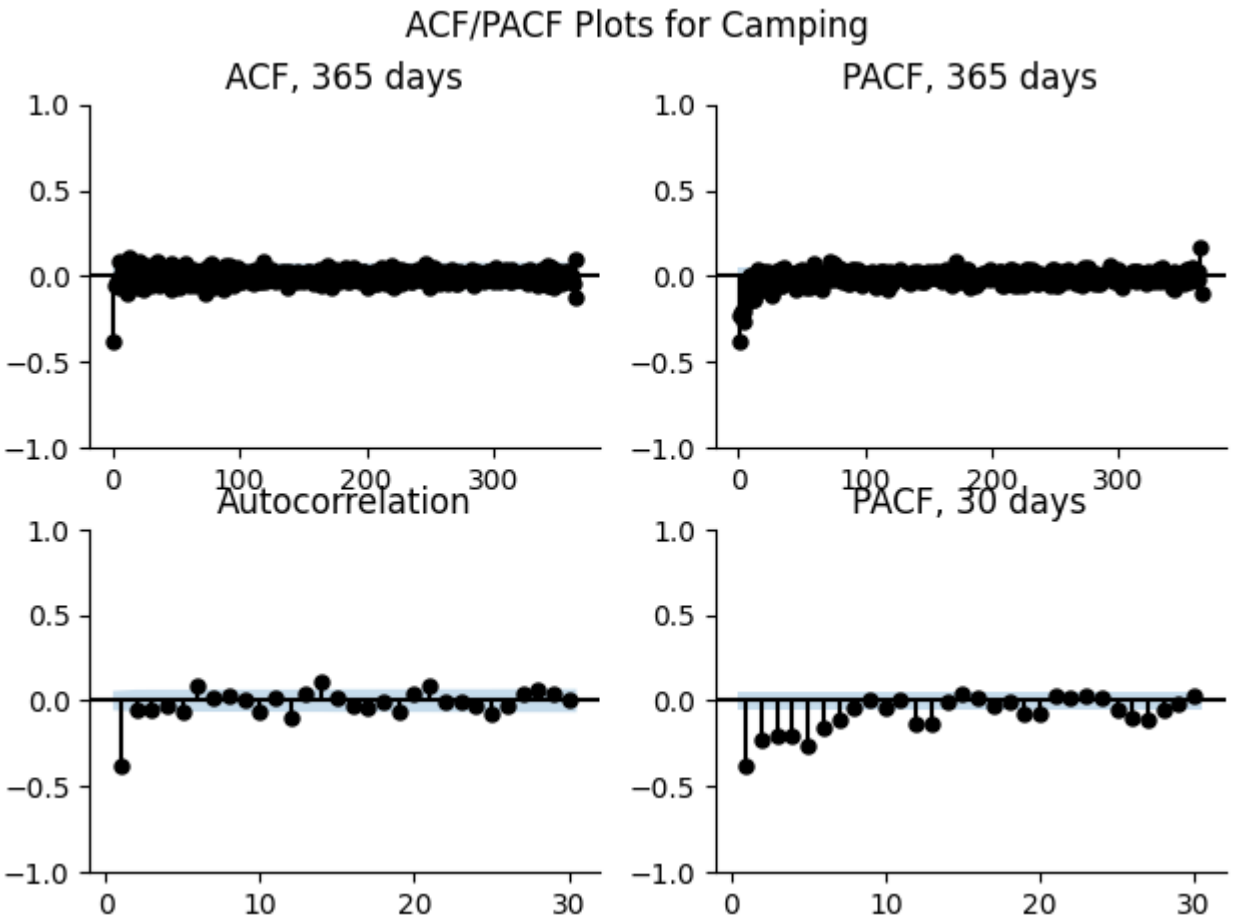
Below are the autocorrelation (ACF) and partial autocorrelation (PACF) plots for ATVing. On the top of the chart, we see the values for 365 lags. On the bottom of the chart are the values for 30 lags. We generally see that only the first few lags are outside the confidence interval for autocorrelation. For the PACF plots, there appears to be more lags outside the confidence interval, especially within the first week. There are also lags at around 14 and 21 days. This could be due to incomplete deseasonalization based on weekday. However, due to the weekday structure of the data, it also stands to reason that particularly high search interest on certain weekdays will then have a higher influence of subsequent days even after properly accounting for seasonality. In other words, it's not just the number of lags between two values that determine correlation, but rather the exact weekdays that these days occur on. In a technical sense, this means that the series is not stationary.



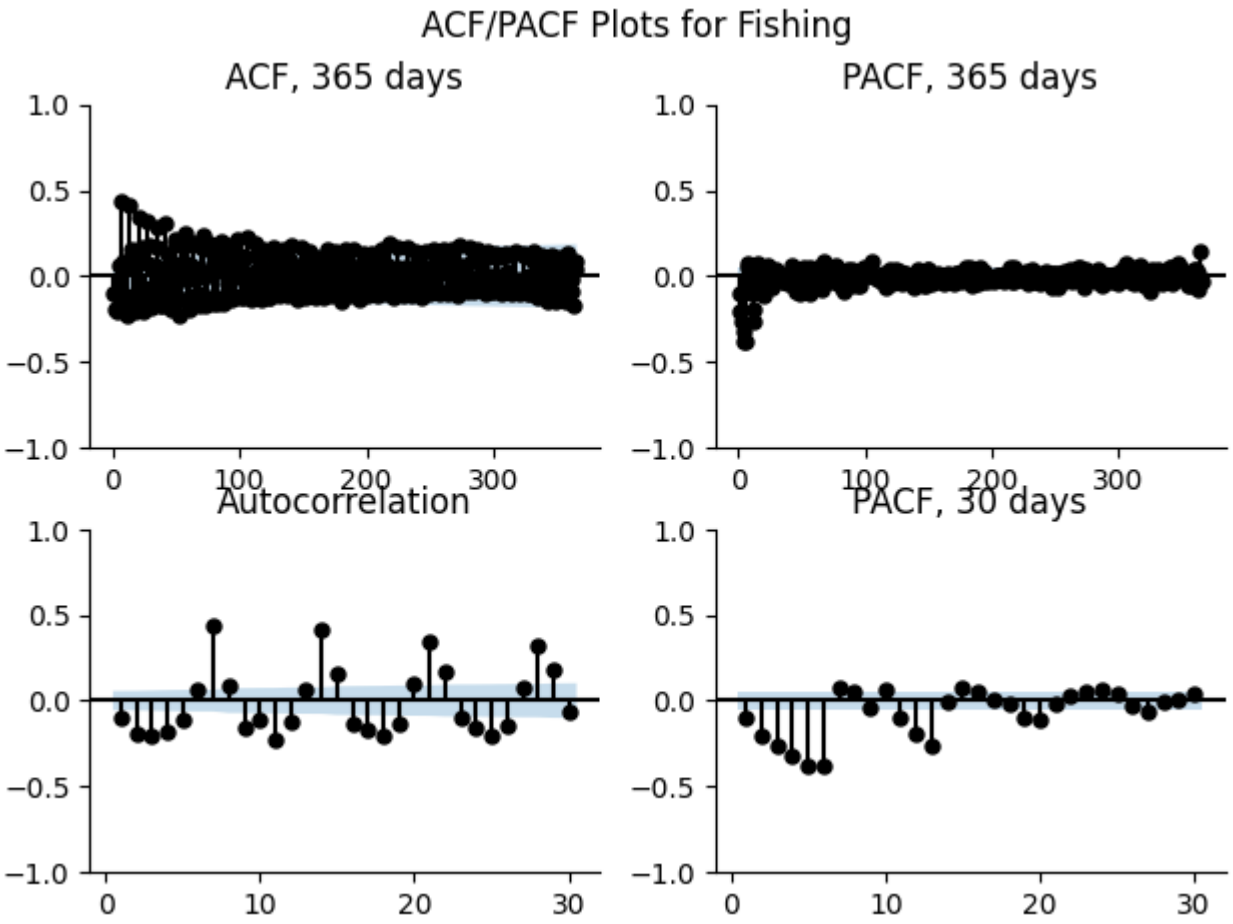
For boating, autocorrelation is much more pronounced. In the 365 day plot, we see autocorrelation persists for as much as two months. Within the weekly plots, we see a very obvious weekly pattern. For boating, autocorrelation persists more than partial autocorrelation. However, partial autocorrelation persists as well within the first week.



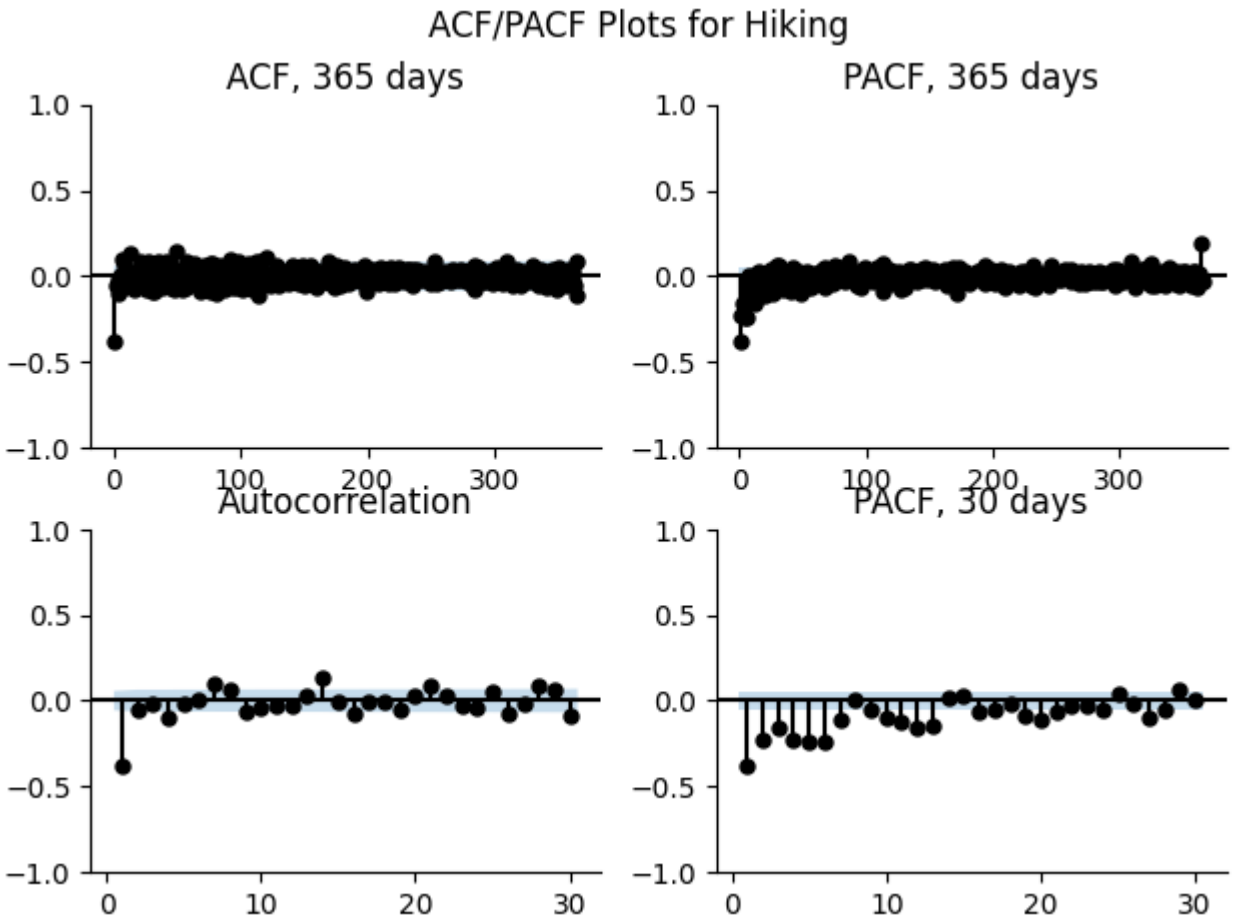
For camping, autocorrelation fades rapidly and partial autocorrelation persists for the first week and at 14 days.



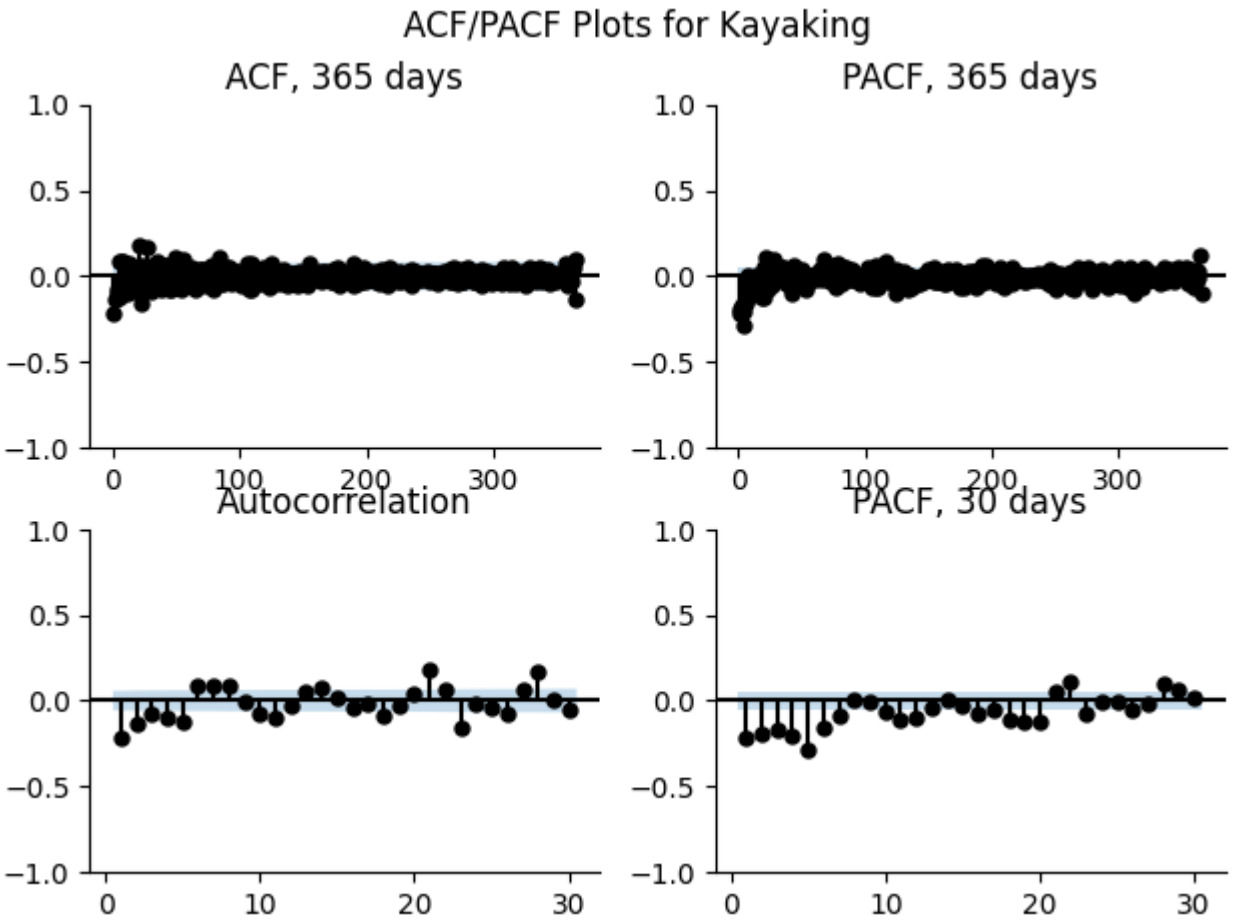
Fishing shows simlar patterns to boating with very persistent autocorrelation and partial autocorrelation that persists for the first few weeks.



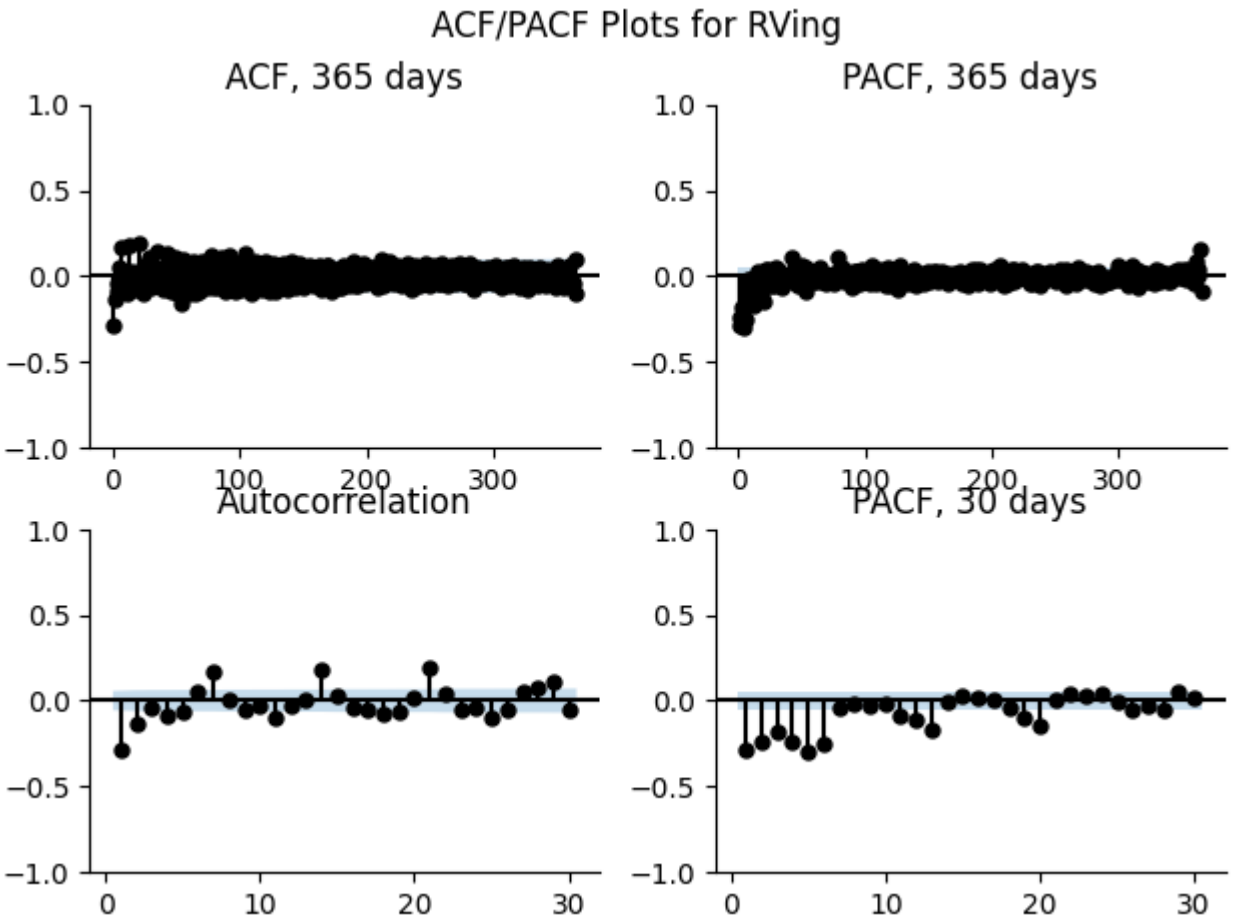
Hiking is more simlar to camping and atving with rapidly fading autocorrelation and some persistence in partial autocorrelation.



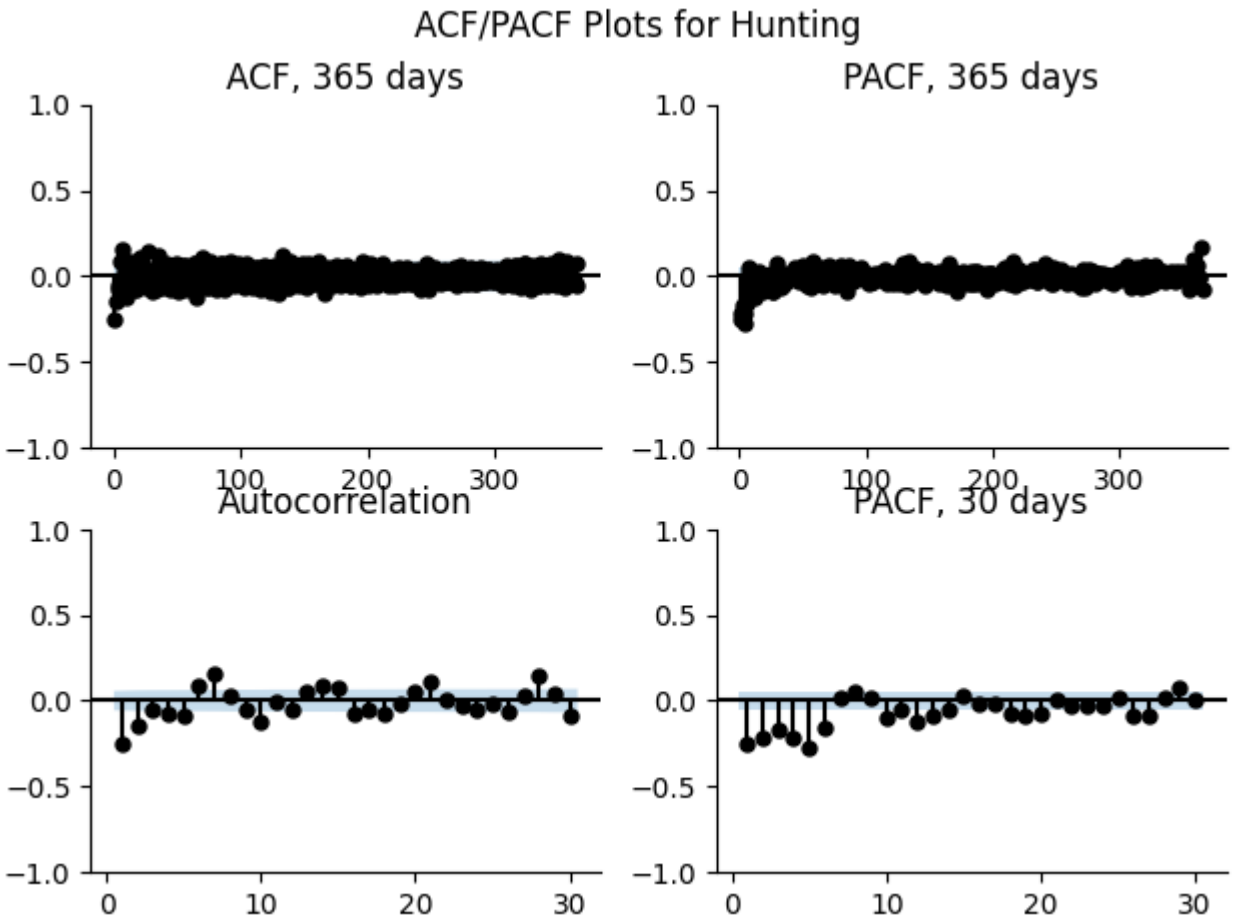
Kayaking shows less extreme autocorrelation than boating, but slightly more so than hiking. Most partial autocorrelation is during the first week.



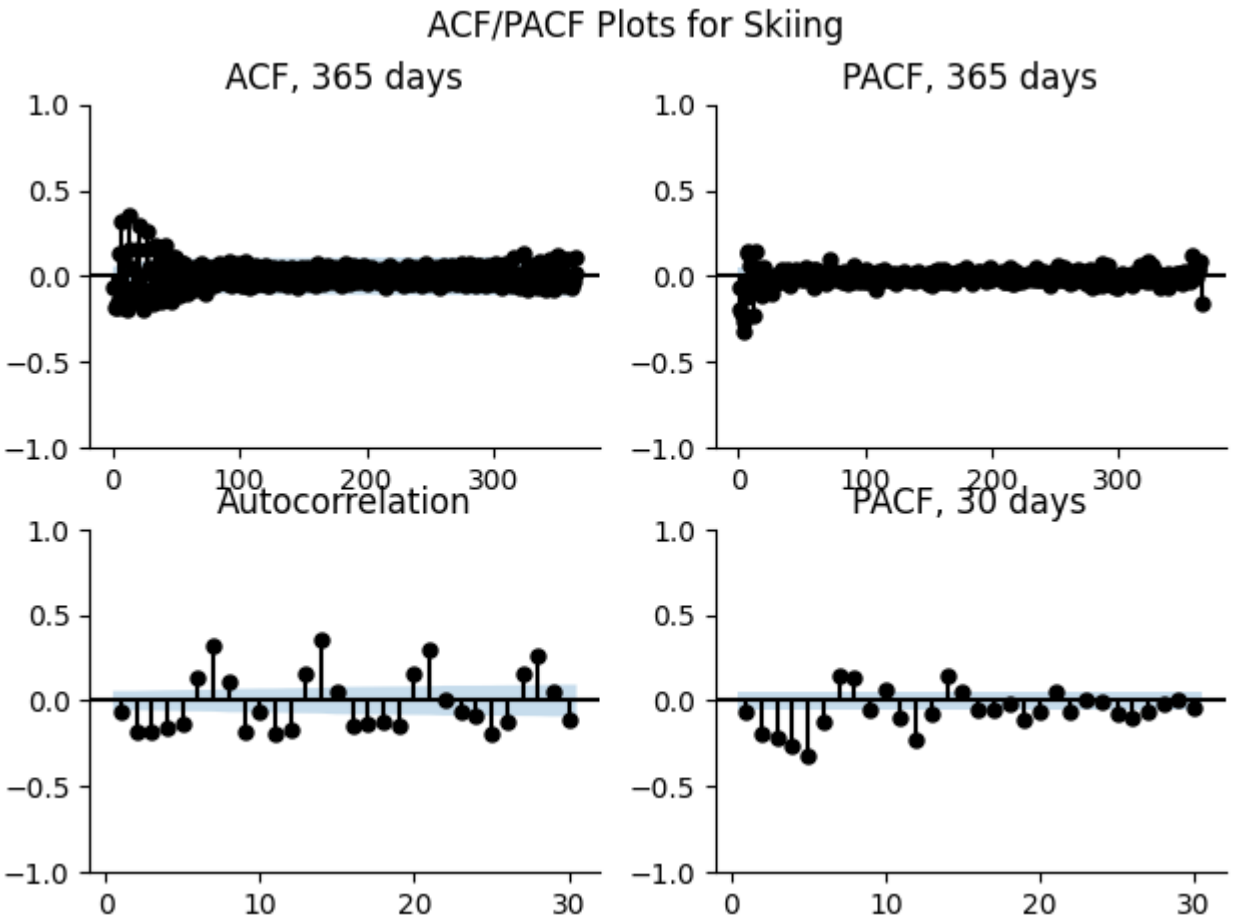
RVing show similar results to camping with rapidly declining autocorrelation and more evident partial autocorrelation during the first week and at 14 and 21 days.



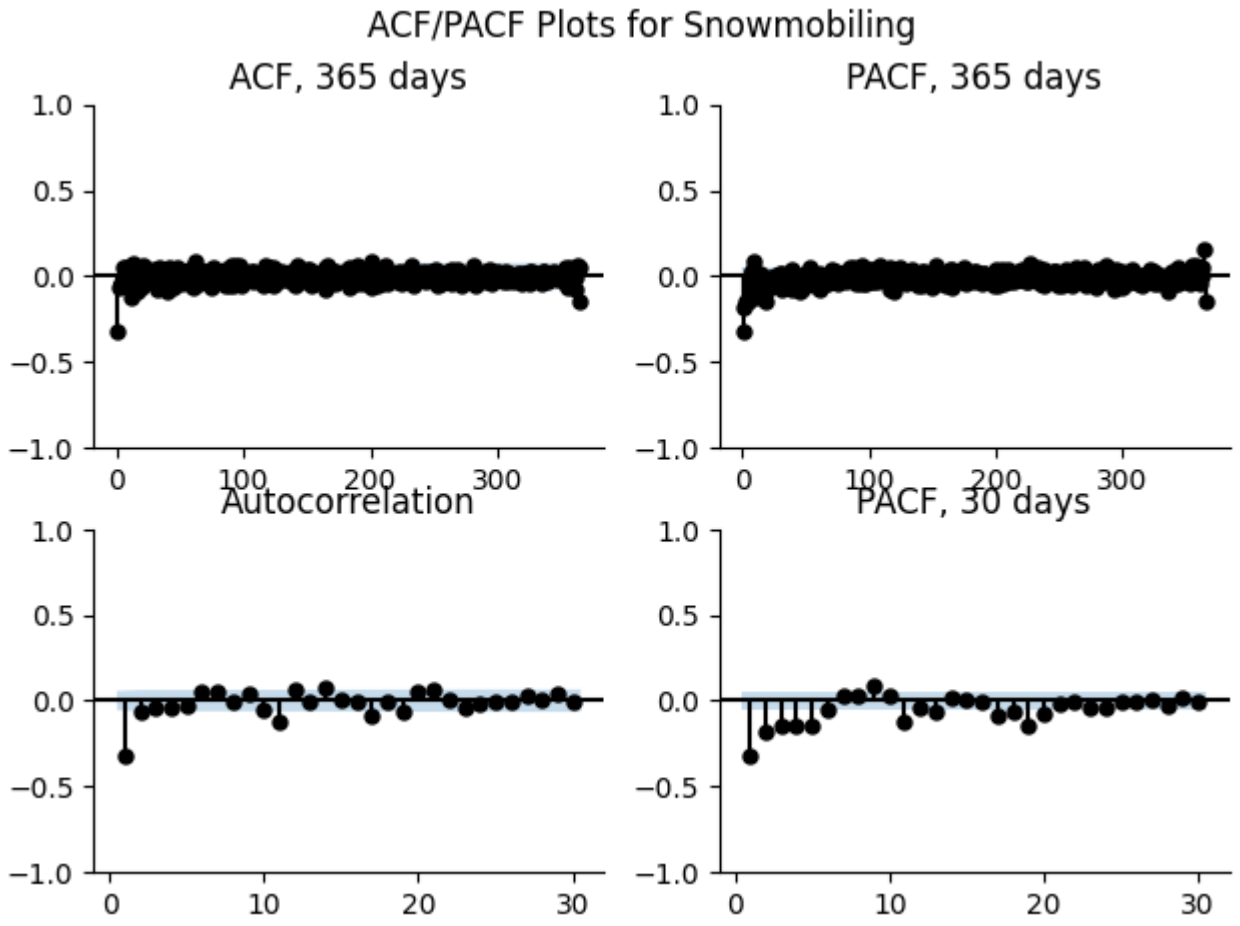
Hunting is also similar to RVing and camping with more evidence of partial autocorrelation especially during the first week.



Skiing shows the most autocorrelation but also shows partial autocorrelation during the first week and at 14 days.



Finally, snowmobiling shows limited autocorrelation but partial autocrrelation during the first week and, to some extent, 14 days.

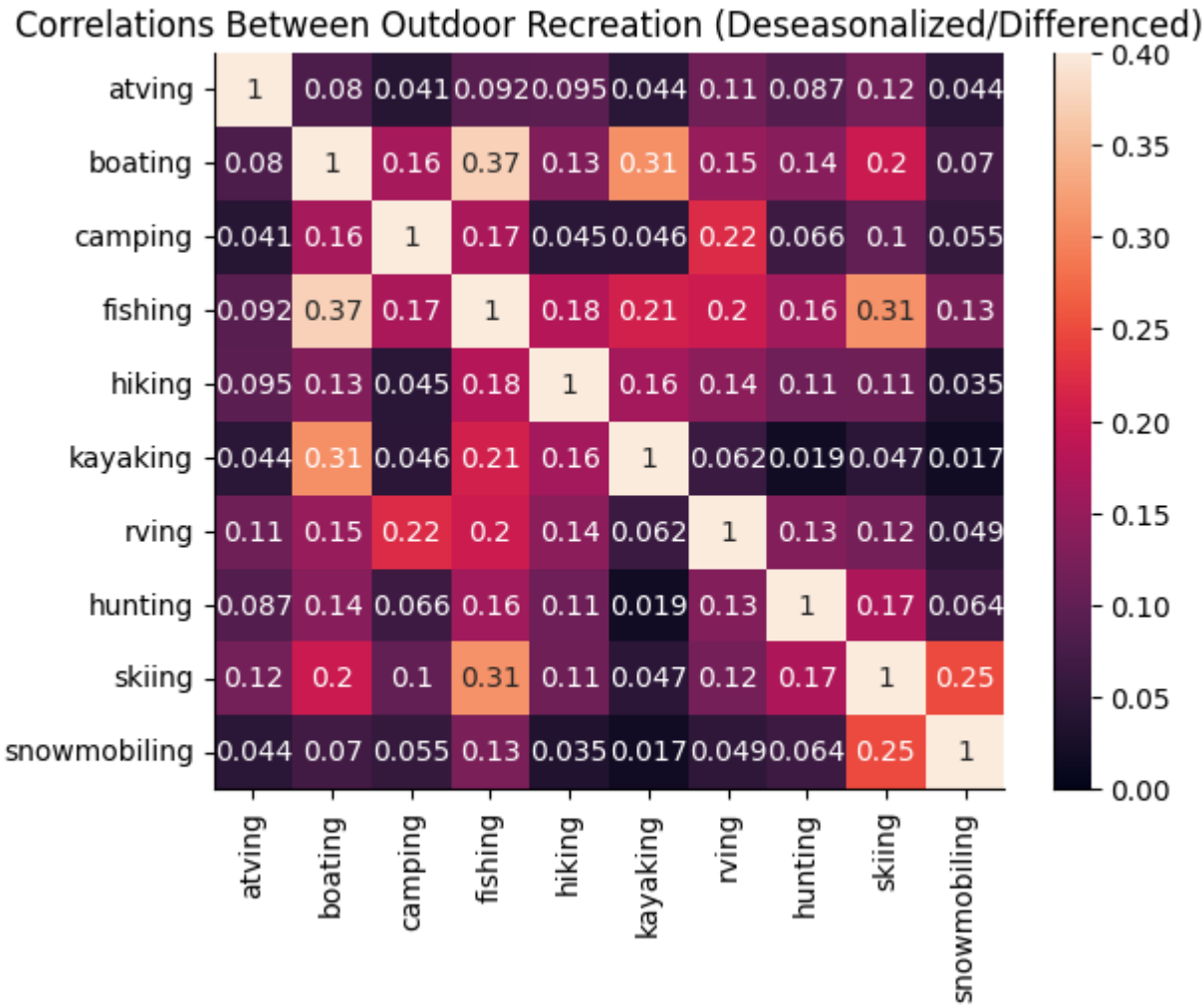


==========

Correlation between Activities

In the first installment, I presented a correlation heatmap between search interest for various activities. I explained that these correlations were largely spurious and driven by seasonality. Summer activities and winter activities were positively correlated within their group and negatively correlated between groups.

The following heatmap shows the correlation between activities after deseasonalizing and differencing the data. Correlations are now much less strong but still evident. For instance, the highest correlation is between boating and fishing, which are indeed activities frequently enjoyed together. There's also a high correlation between fishing and skiing, no doubt during the winter ice fishing season. Snowmobiling and skiing are also positively correlated. Kayaking and boating are another high value.



==========

Installment 2 Conclusions

These results exemplify a similar situation to the Northern Michigan search interest project. Specifically, this data will be difficult to fit with a traditional statistical model. While autocorrelation and partial autocorrelation plots suggest series might be fit to a fair degree using autoregressive and moving average models, these plots also show considerable evidence of nonstationarity even after differencing and deseasonalization, especially related to weekday. Also, we see the result for fishing correlation with boating and skiing. No doubt, these effects change depending on whether you are in the summer or winter. These interactions between various variables and seasonality goes well beyond simple seasonality, and matches the challenges presented by the Northern Michigan search interest project.

For Installment 3, I plan to correlate weather variables with outdoor recreation activity search interest.