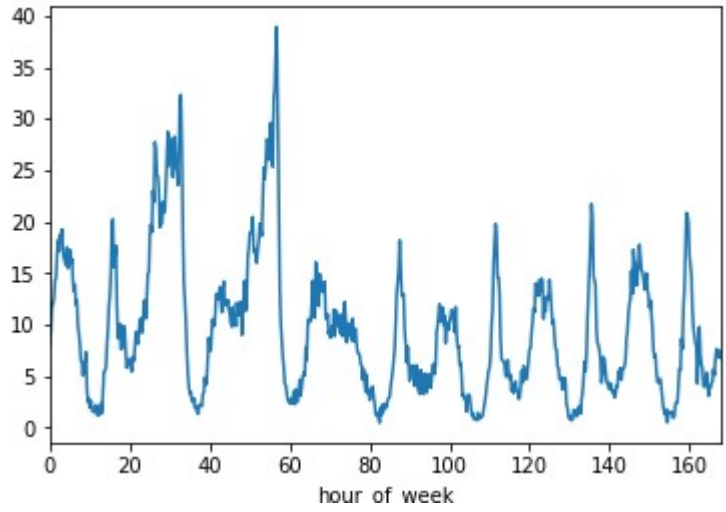
Ultimate Exploratory Data Analysis

- There is a problem with the data. The timestamps begin with 1970-01-01, which is an implausible date, since there was no Internet and no rideshare industry at the time. That date is commonly used as a base date, so it's likely that there has been an error in format conversion at some point (perhaps in saving data to CSV from an Excel spreadsheet), in which some other base date was replaced with 1970-01-01.

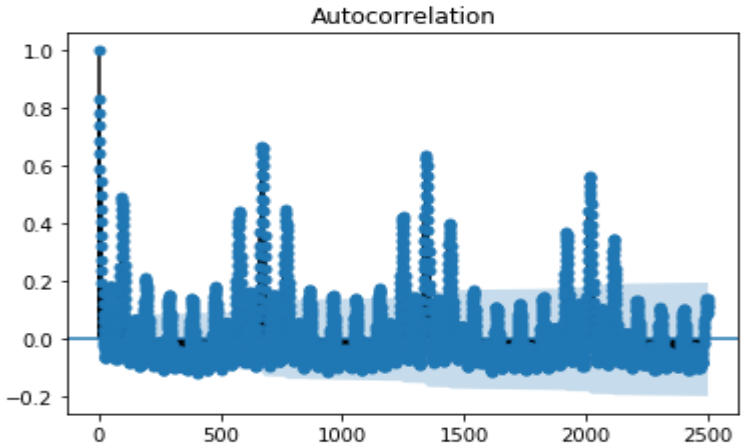
- Because of the data problem discussed above, it appears that the days of the week are also wrong. The problem is apparent in the following chart, which shows an aggregation (by mean) across weeks for the 15-minute interval observations:



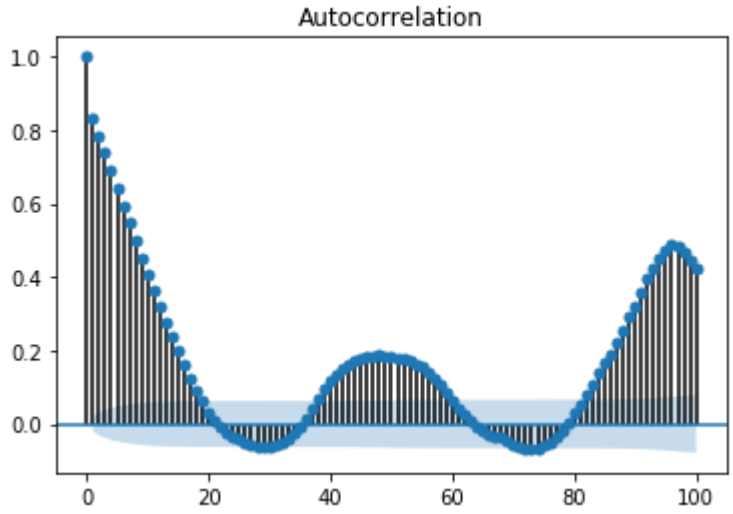
The zero on the X-axis above represents the first point in the data series (aggregated with the analogous points in each subsequent week). If we take the data at face value, that point is 8PM on Wednesday. But if we accept that interpretation, we see an implausible pattern: Thursday and Friday nights (and Friday and Saturday during the daytime) are unusual, while Saturday nights seem more like ordinary weekday nights, and Sunday during the day seems like an ordinary weekday.

- In the light of the problem discussed above, I will make the assumption that the data actually begin on Thursday and not Wednesday. I will also assume that the times of day are correct, but since we know the days are wrong, I would caution against accepting my analysis at face value before verifying the times of day.

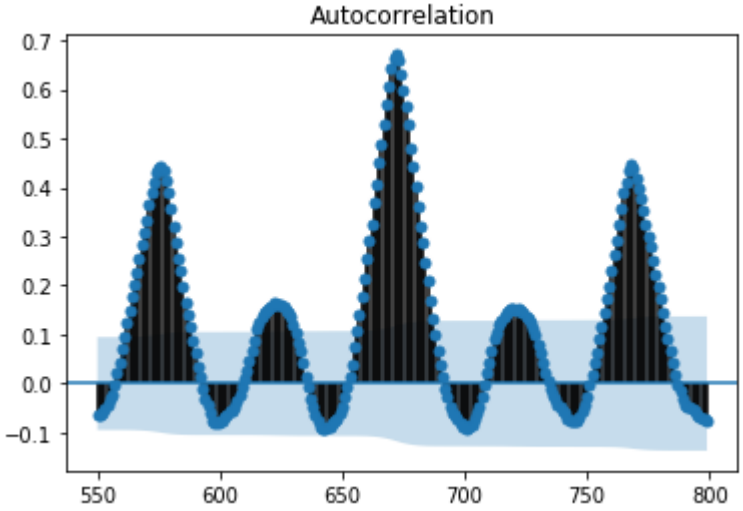
- I plot the autocorrelation function to give an idea of the general cyclical properties of the data:



Clearly there are cyclical patterns, as indicated by the strong peaks in the autocorrelation function. In particular, if we look at the shorter lags, we see a 12-hour (48 lags) and 24-hour (96 lags) cycle:



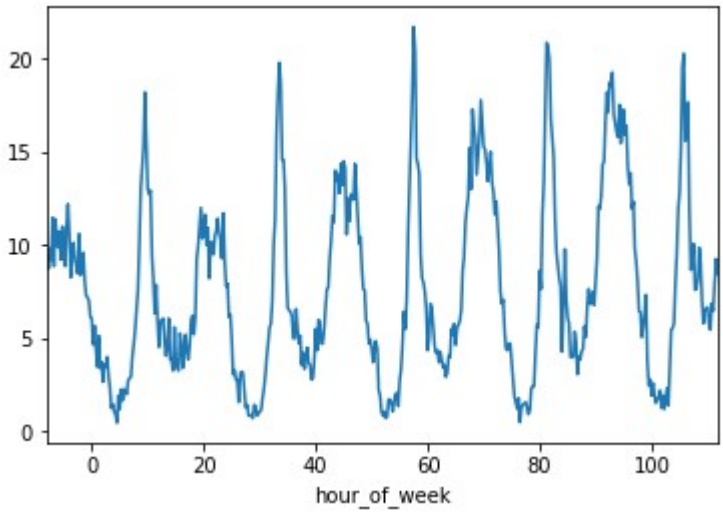
Moving the focus outward, we see a weekly cycle (672 lags), as well as 6-day and 8-day cycles:



The 6-day and 8-day cycles are likely explained by the presence of weekends. In other words, after a Sunday, we see a repeat of weekend-like activity 6 days later, on the subsequent Saturday; and after a Saturday, we also see similar weekend-like activity 8 days later, on the subsequent Sunday.

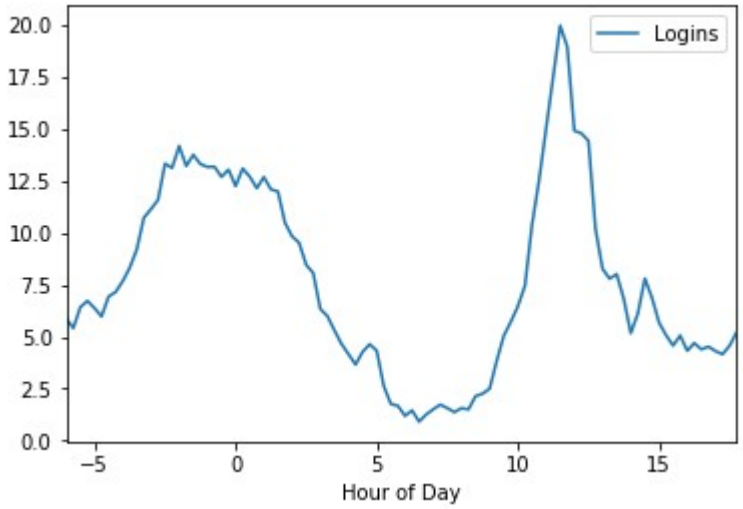
There is no evidence of a monthly or hourly cycle, but if these exist, they might be hard to detect, as there are only four months of data and only four observations per hour.

- Given my assumption about days of the week, we can divide weeks into weekdays and weekends. There are several possibilities for how to do this, but I find the data consistent with my impression that weekends begin on Friday evening (maybe around 6PM) and end on Sunday evening (maybe around 6PM, or maybe later, but using the same cutoff point makes the analysis more straightforward). Thus we can define weekdays as the period from Sunday at 6PM to Friday at 6PM. The following chart shows aggregate (mean) weekday activity over the sample:



Zero on the X-axis represents midnight Sunday. What we see here is largely a regular pattern, where each weekday is like the others, although there is also an upward trend in the nightly peaks over the course of the week from Monday night to Thursday night. During the nightly peak periods on Sunday and Monday, there are about 10 logins per interval (40 per hour), and by Thursday night, there are about 17 per interval (68 per hour).

We can also aggregate across days of the week, to get a clearer picture of the daily pattern:



There is a plateau at night and a sharp peak during the early middle of the work day.

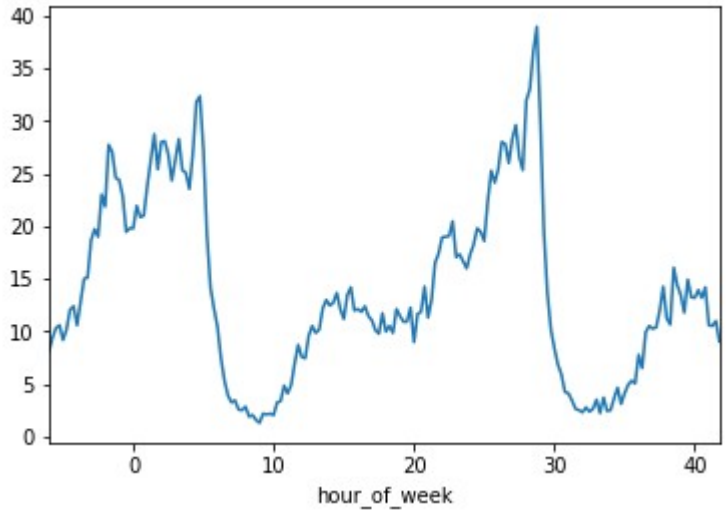
A closer look at the run-up to the nightly plateau shows that the run-up begins around 8PM, and the plateau is reached at around 9:30

Hour of Day	
-4.00	7.671429
-3.75	8.327619
-3.50	9.210476
-3.25	10.729524
-3.00	11.140000
-2.75	11.582857
-2.50	13.336190
-2.25	13.131429
-2.00	14.188571

A closer look at the daytime behavior shows that the run-up begins at the beginning of the work day, the sharp peak occurs around 11:30, and activity has largely died down again by 2PM.

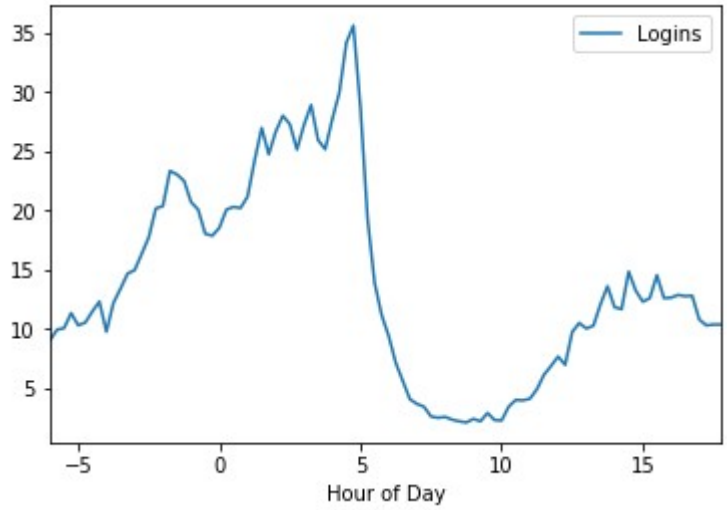
Hour of Day	
9.00	2.495238
9.25	3.834286
9.50	5.040952
9.75	5.714286
10.00	6.465714
10.25	7.449524
10.50	10.378095
10.75	12.530476
11.00	15.027619
11.25	17.507619
11.50	19.988571
11.75	18.960952
12.00	14.920000
12.25	14.796190
12.50	14.434286
12.75	10.194286
13.00	8.273333
13.25	7.792381
13.50	8.015238
13.75	6.834286
14.00	5.168571

We can also look at the typical weekend activity:



Here zero on the X-axis represents midnight Friday. We see a pattern essentially repeated twice, although the early nighttime peak is higher on Friday night than Saturday, while the late nighttime peak is higher on Saturday night (or Sunday morning, depending on how you think of it).

We can aggregate the two days of the weekend as follows:



In contrast to the weekday behavior, the main peak is at night, or, more precisely speaking, in the early morning. The peak occurs very late, which indicates to me that these data are *not* from a city where the bars close at 2AM.

The run-up in activity during the evening occurs between 8PM and 10PM:

Hour of Day	
-4.00	9.800000
-3.75	12.233333
-3.50	13.433333
-3.25	14.700000
-3.00	15.000000
-2.75	16.366667
-2.50	17.766667
-2.25	20.200000
-2.00	20.400000

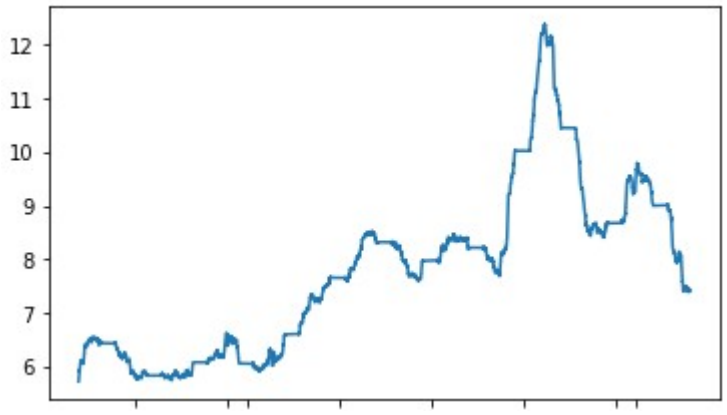
The weekend nightly peak occurs at around 4:30 AM, followed by a sharp drop-off over the next two hours:

Hour of Day	
4.00	27.733333
4.25	29.966667
4.50	34.200000
4.75	35.633333
5.00	28.666667
5.25	19.366667
5.50	13.933333
5.75	11.200000
6.00	9.433333
6.25	7.166667
6.50	5.633333
6.75	4.100000
7.00	3.700000

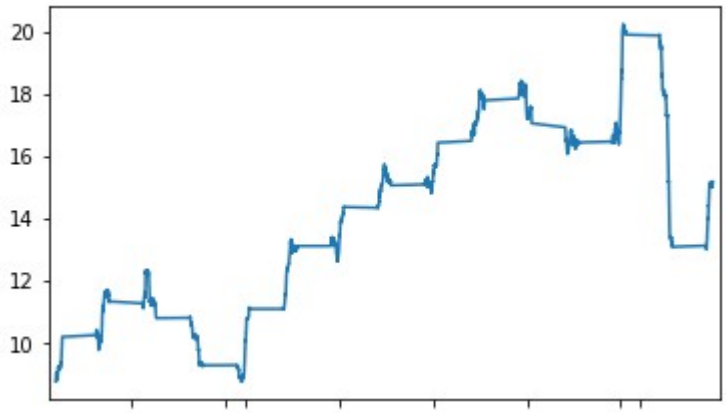
The autocorrelation function suggested a 12-hour cycle, but it's perhaps more accurate to say that there are two different 24-hour cycles in opposite phase. On both weekdays and weekends, the nighttime peak and the daytime peak have a different character.

- Beyond cyclical characteristics, we will want to know if there is a trend in the data. Or rather, since we know that weekends and weekdays behave differently, we want to know if there is a trend in either the weekday data or the weekend data or both.

This is what a rolling average for weekdays looks like over the course of the data:

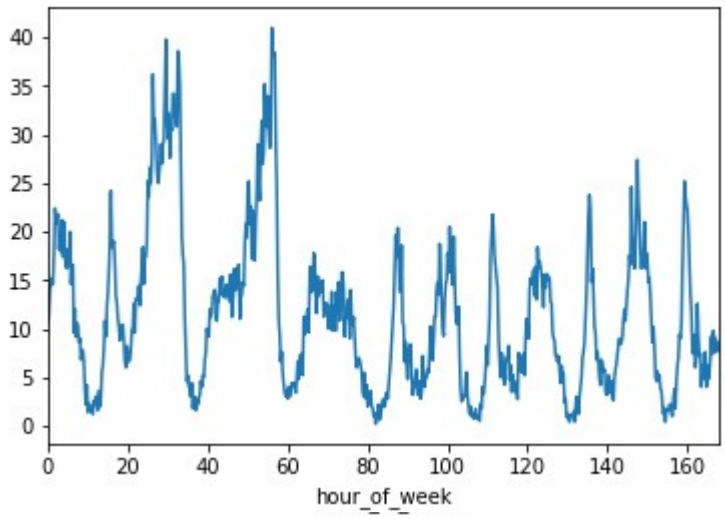


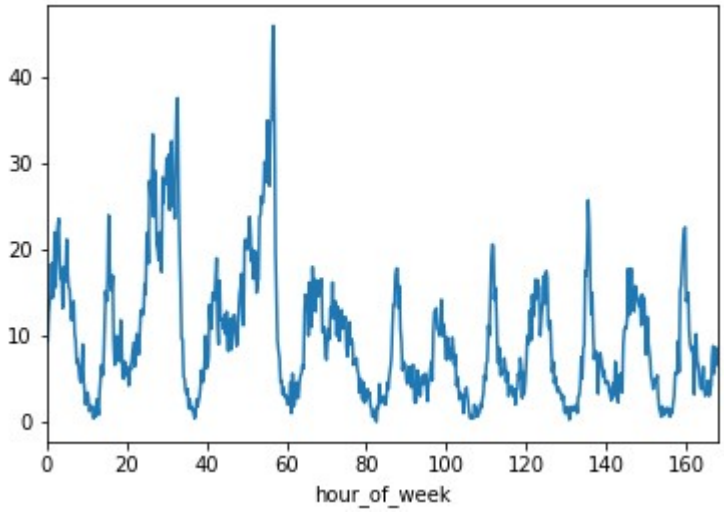
This is what it looks like for weekends:

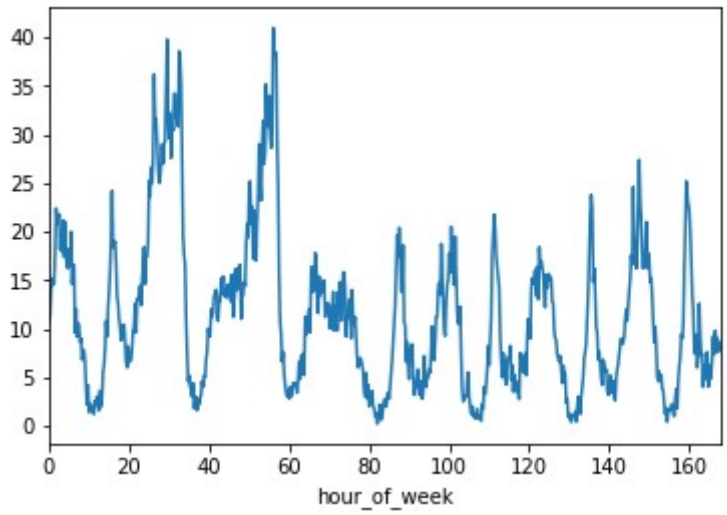


In both cases, we see a period of low and relatively stable activity, followed by a period of moderate and rising activity, followed by a period of high and volatile activity.

We can divide the data into thirds and look at whether the weekly pattern seems to be different in each period. (See next page, and note that zero on the X-axis here refers to the hour at which the original data begin, so 8PM on Thursday.)







The pattern looks pretty much the same in all three periods.

Note: Code for this analysis is contained in eda.ipynb