Analysis of Chicago Train Ridership Data using Power Bi

Introduction

The Chicago train dataset was used by Taylor & Francis in their book to specify predictors that can be used to build an efficient model which can be used to predict short-time ridership volume. This would help forecast the demand for cars sent to service passengers coming out of stations. In this report, we will focus on analysing the ridership data using Power Bi. We will:

- manipulate and prepare data for reporting.
- model, visualize and explore data.
- build and share a dashboard with others

Daily ridership data were obtained for 126 stations between January 1, 2001, and November 30, 2016. Ridership is measured by the number of entries into a station across all turnstiles every day over the above mentioned period.

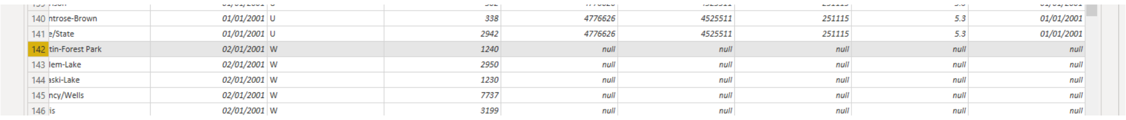
We will check if there are factors such as gasoline prices and the employment rate that may affect ridership. Would more people give up using their cars and use public transportation instead as the gas price rises? The same question might arise when the unemployment rate decreases.

To answer these questions, average weekly gasoline prices were obtained for the Chicago region (U.S. Energy Information Administration) from 2001 through 2016. For the same period, monthly unemployment rates were pulled from the United States Census Bureau.

Data preparation

The granularity of the date column in the tables of **Chicago_gas_prices** and **Chicago_unemplyment** does not match that in the **entries.** Having transformed the data, the queries of Chicago_gas_prices and Chicago_unemployment are merged with the date column selected as a matching column.

If we look closer at the table of the unemployment rates, we realize that the rate is computed on the first of each month. Merging the Chicago_unemplyment query with entries query will populate the unemployment_rate column with null values over the period from the second day to the last day of each month as shown in the table below.



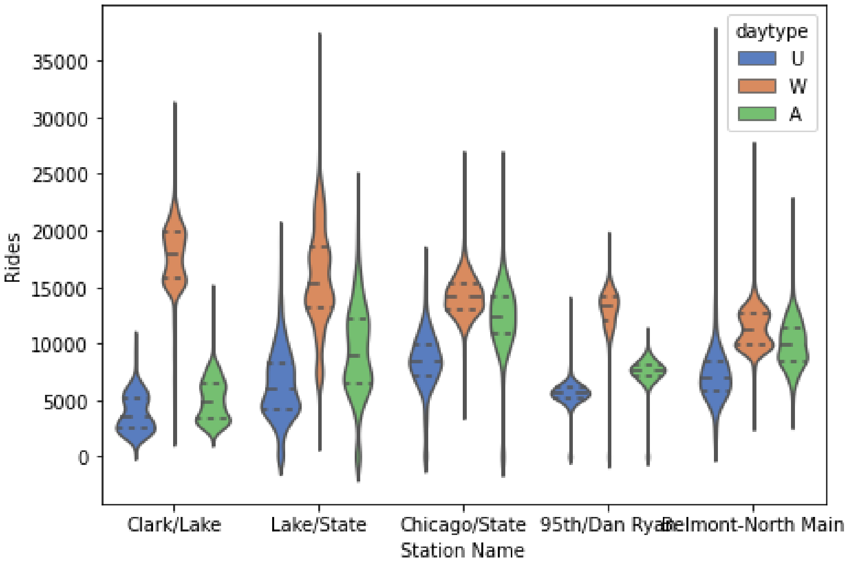
140	ntrose-Brown	01/01/2001	U	338	4776626	4525511	251115	5.3	01/01/2001
141	e/State	01/01/2001	U	2942	4776626	4525511	251115	5.3	01/01/2001
142	tin-Forest Park	02/01/2001	W	1240	null	null	null	null	null
143	lem-Lake	02/01/2001	W	2950	null	null	null	null	null
144	aski-Lake	02/01/2001	W	1230	null	null	null	null	null
145	ncy/Wells	02/01/2001	W	7737	null	null	null	null	null
146	is	02/01/2001	W	3199	null	null	null	null	null

Merged entries and unemployment_rate queries

Filling the unemployment_rate column over the period from 2nd to the last day of each month with the value on the first day would make two benefits: firstly, the queries of entries and unemployment rate will have the same date column which would make it suitable for reporting later. Secondly, using the average value of the unemployment rate over the whole month would remain fixed.
The same method is used to merge the queries of entries and Chicago_gas_prices. Now we can start visualizing our data.

Ridership distribution

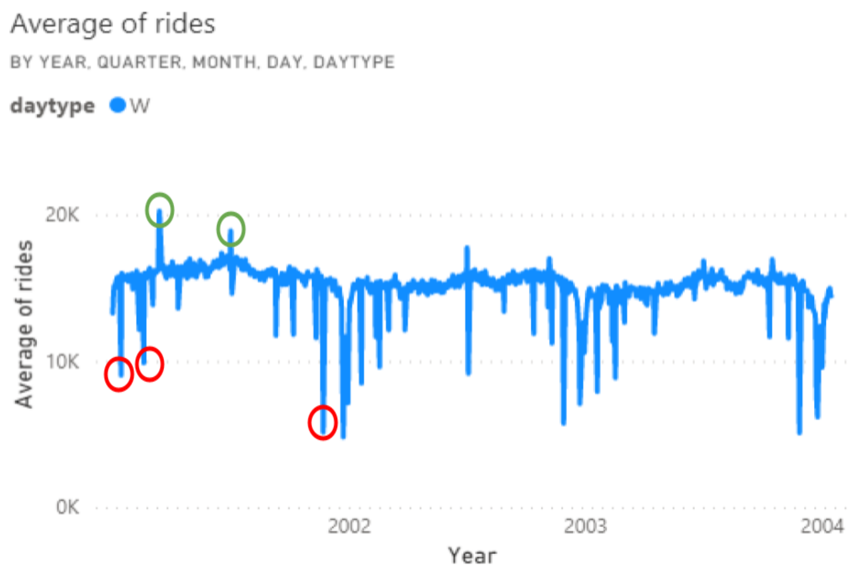
The ridership distribution across the five busiest stations is produced using a violin plot as shown in the figure below.



The violin distribution of ridership for the five busiest stations over weekdays and weekends.

In this figure, the ridership distribution is ordered from the largest median (left) to the smallest median (right). It is worth noting that busier stations have greater variability (wider distribution). For all stations, the ridership is stronger on weekdays than at weekends.

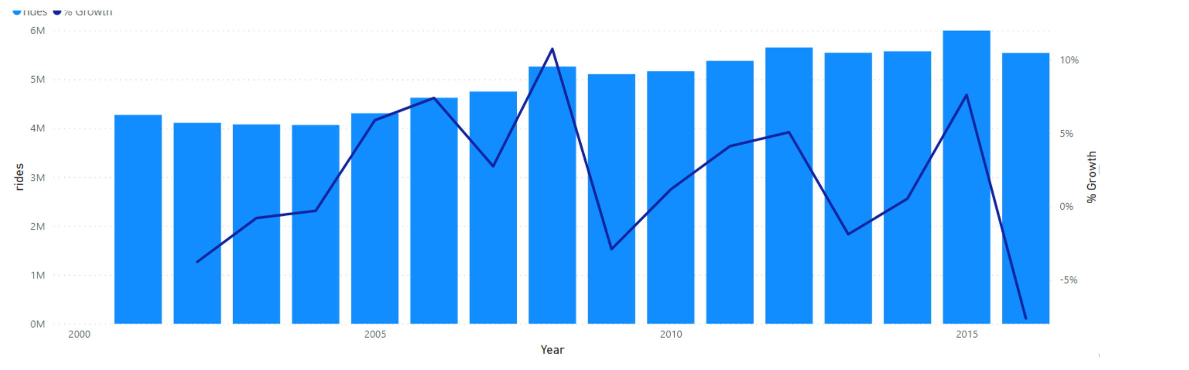
In this report, we will focus on the busiest station, namely Clark/Lake. Although there is a high number of entries into the station on some weekdays, the station suffers from low ridership on some weekdays. To reveal those days with very high and low ridership, the weekday ridership for the Clark/Lake station is plotted against day as follows:



The ridership on weekdays at the Clark/Lake station over the period between 2001 and 2004. Red and green circles indicate low and high ridership, respectively.

Ridership becomes low on weekdays celebrated by Americans such as Black Friday following Thanksgiving Day, Martin Luther King Day which is observed every year on the third Monday of January, and the President’s Day on the third Monday of each year.

It is also interesting to notice the spikes indicating high ridership on some special weekdays such as July 3, the day before Independence Day which is a federal holiday.

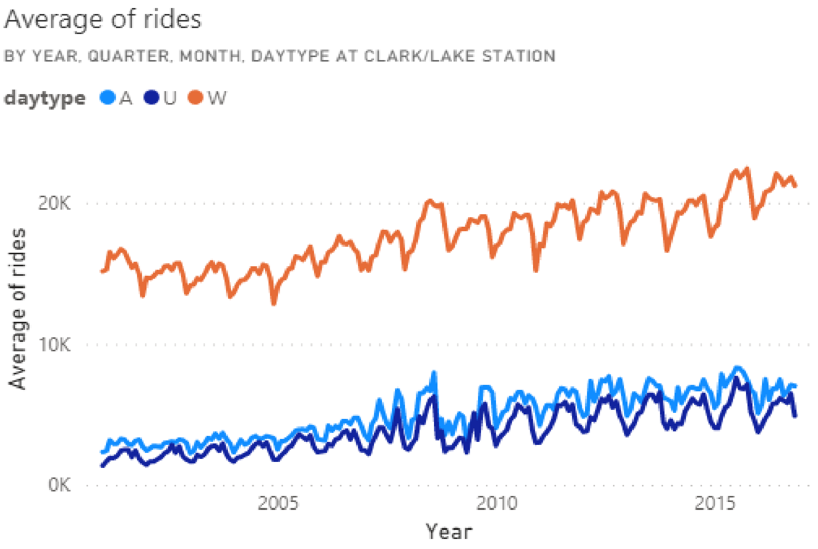


Annual ridership growth versus year

Notice that the growth fell remarkably in 2009, 2013 and 2016

Monthly average ridership

The ridership is averaged over each month and plotted as shown in the figure below. The repetitive patterns show annual trends for both weekday and weekend ridership.

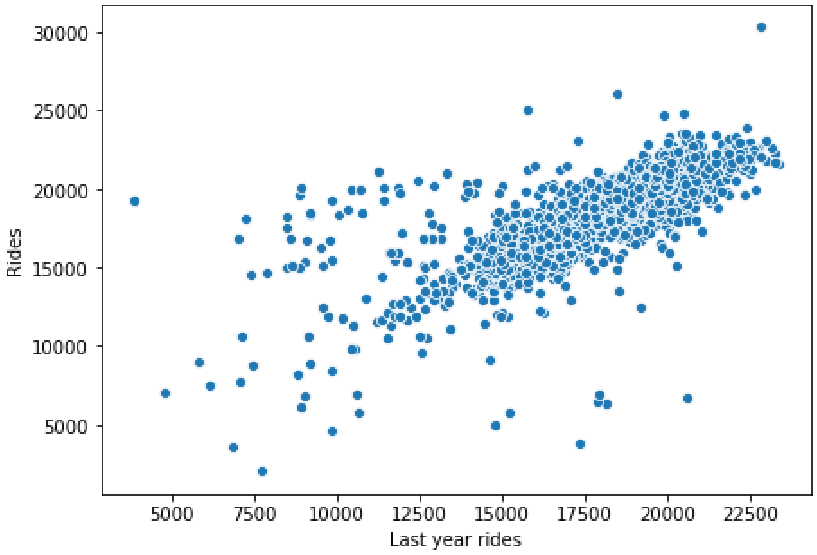


Monthly average ridership for the Clark/Lake station on weekdays (W), Saturdays (A) and Sundays and holidays (U)

The local minima of the time series (W) showed the average ridership on weekdays in December between 2001 and 2016. Most of the local minima of the time series (U and A) correspond to January and February. In May 2009, the Clark/Lake station received the lowest mean number of riders on weekends and public holidays.

Year-based lag ridership

The relationship between the year lag weekday ridership at the Clark/Lake station and the current-day ridership at the same station is plotted as shown in the figure below.



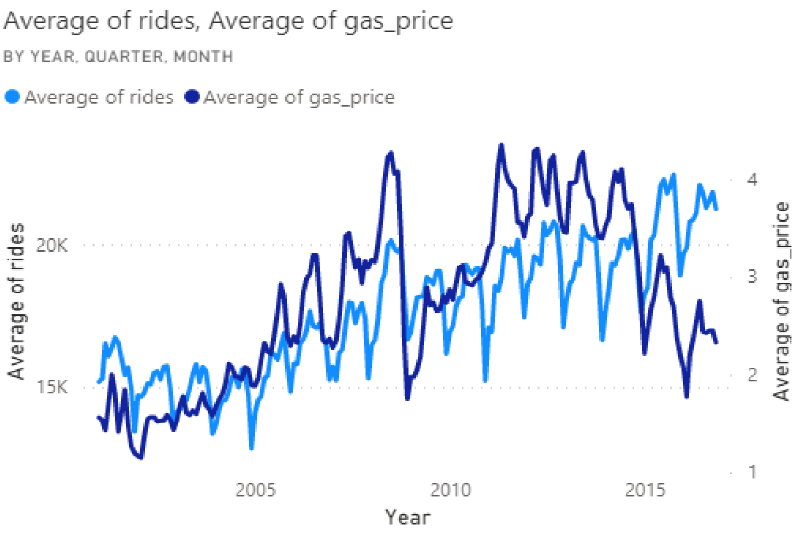
A scatter plot of the year lag ridership at the Clark/Lake station and the current-day ridership at the same station.

The figure shows a linear relationship between the two predictors which can be used to build a model to predict long-term future ridership. For example, forecasting could be used to extend or reduce the train network in response to population change.

Some samples lie far off from the overall pattern since some celebrated holidays are floating and so will be the following days. For example, Black Friday is the fourth Friday of November.

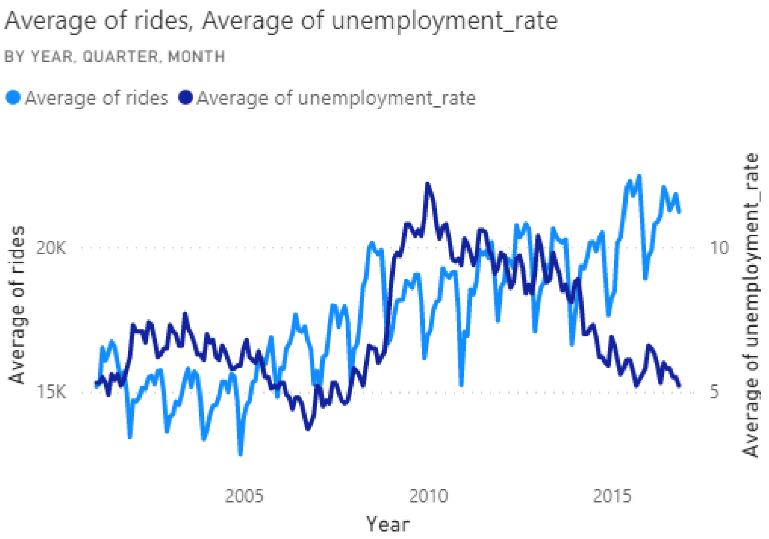
The effect of gas price

let’s see if a relationship can be established between gas prices and ridership. The monthly average ridership and the monthly average gas price are calculated and plotted as shown in the figure below.



Monthly average ridership versus monthly average gas price

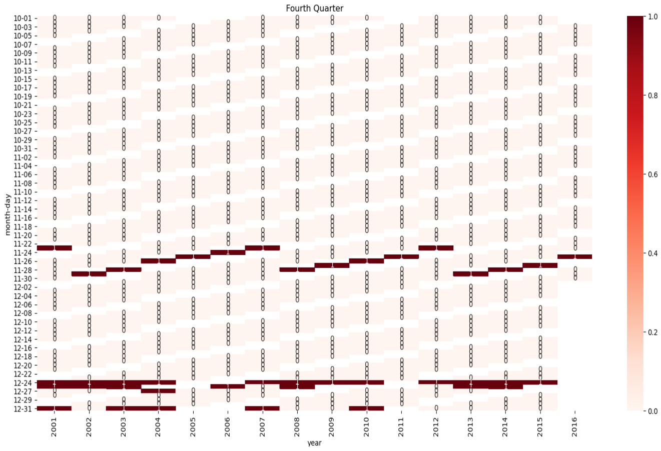
There is a positive association between these two quantities, which probably caused variation in ridership.   The correlation coefficient is calculated and is 0.8.



Monthly average ridership versus monthly average unemployment rate

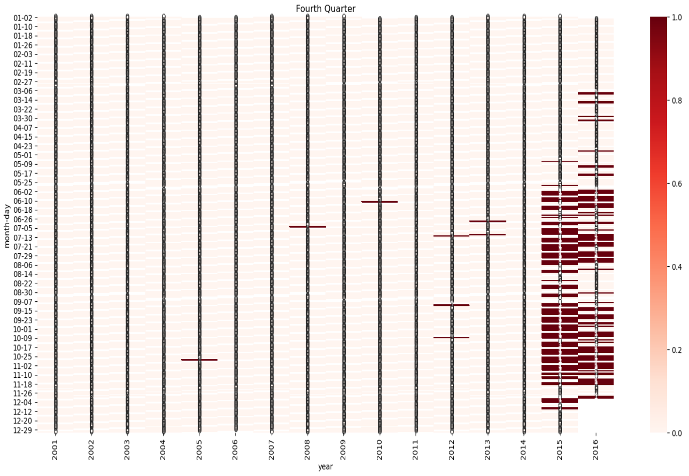
Anomalies of ridership

An indicator of weekday ridership which is less than 10,000 rides is created and a heatmap is built with the x-axis representing the year and the y-axis representing the month-day. The map brings out annual patterns that can be observed in the fourth quarter of each year. It is interesting to note that the staircases that appear in the middle of the heatmap indicate low ridership occurring on Black Friday which is a floating day following Thanksgiving Day. Also, in late December the Clark/Lake station experienced a low number of riders.



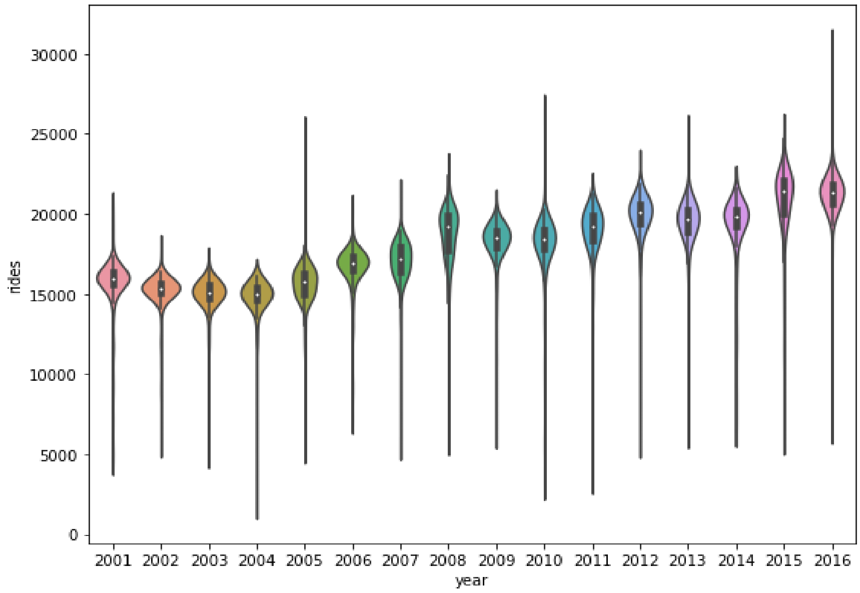
Heatmap of weekday ridership <10000 at the Clark/Lake

Another heatmap is built for weekday ridership with a value greater than the 95th percentile. The weekdays that have high ridership lie in 2015 and 2016.

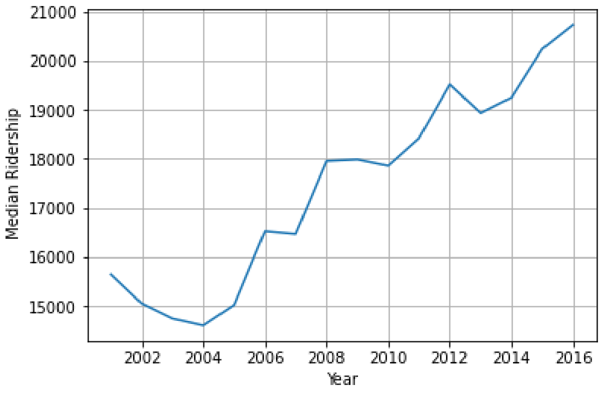


Heatmap of ridership from 2001 through 2016 for weekdays with a value greater than the 95th percentile of the ridership at the Clark/Lake station.

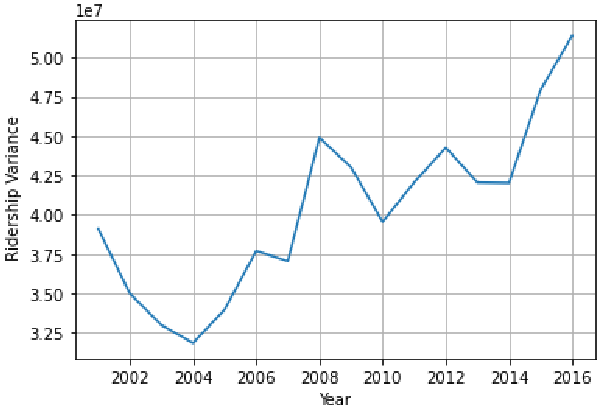
The distributions across the years of study for weekday ridership are provided. Variability in ridership increases as median ridership increases, as shown in the figures below. The values of median and variance are unusually low in 2004 whilst unusually high in 2015 and 2016.



Distribution of weekday ridership at the Clark/Lake from 2001 through 2016



Median ridership at the Clark/Lake station



Median ridership at the Clark/Lake station

Reference
http://www.feat.engineering/index.html