Tool Exploration – Revolution R Enterprise

Business Application Description

On a recent flight I took from New York City to San Francisco, my departure was delayed by 30 minutes. However, it worked to my advantage. I was running 15 minutes late and the pilot was able to make up for lost time. We landed 5 minutes ahead of schedule. So the big question I had was, “Do delayed flights typically fly faster? Or was I just really lucky?”

We can try and answer this question by analyzing a dataset containing flight arrival and departure details for all commercial flights within the USA in 2007.

Data

The dataset used for this study is sourced from FAA (Federal Aviation Administration). It is a csv file with about 7.5 million observations and 29 variables. It is downloadable from <http://stat-computing.org/dataexpo/2009/the-data.html> . I have used the data of the year 2007 for the sake of convenience of analysis.

Tool Description

The data used in this study has 7.5 million observations. R doesn’t have the capability to deal with this huge volume of data as it is a memory-bound language. Even for modern computers with 64-bit address spaces and huge amounts of RAM, dealing with data sets that are tens of gigabytes and hundreds of millions of rows (or larger) can present a significant challenge. Revolution Analytics has addressed these capacity, performance and scalability challenges with its “Big Data” initiative to extend the reach of R into the realm of production data analysis with terabyte-class data sets.

Environment Set-up

1. Install MRO-3.2.2-for-RRE-8.0.1-Windows
2. Install Revolution R Enterprise (Note that Revolution R can only be installed on Windows)
3. Install the package RevoScaleR

RevoScaleR Features

The RevoScaleR package provides a mechanism for scaling the R language to handle very large data sets. There are three major components to this package:

• A new file format especially designed for large files

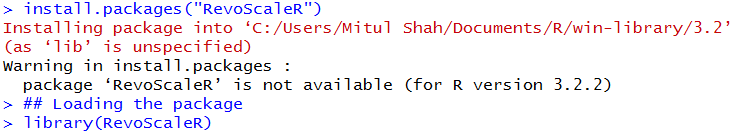
• External memory implementations of the statistical algorithms most commonly used with large data sets

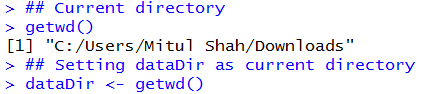
• An extensible programming framework that allows R and later C++ programmers to write their own external memory algorithms that can take advantage of Revolution R Enterprise’s new Big Data capabilities.

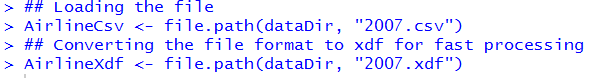
RevoScaleR XDF File Format

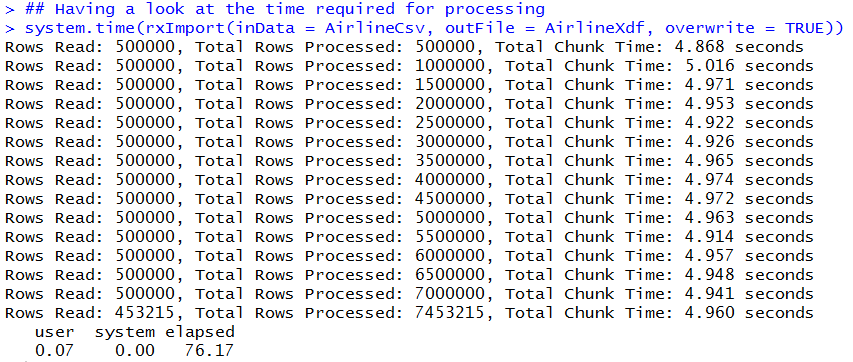
RevoScaleR provides a new data file type with extension .xdf that has been optimized for “data chunking”, accessing parts of an Xdf file for independent processing. Xdf files store data in a binary format. Methods for accessing these files may use either horizontal (rows) or vertical (columns) block partitioning. The file format provides very fast access to a specified set of rows for a specified set of columns. New rows and columns can be added to the file without re-writing the entire file. RevoScaleR also provides a new R class, RxDataSource, that has been designed to support the use of external memory algorithms with .xdf files

R code with results



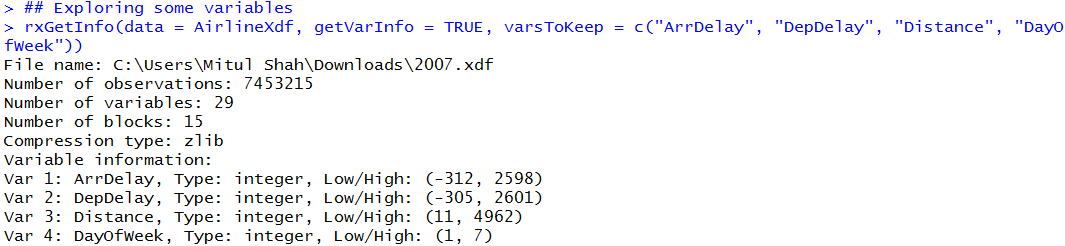




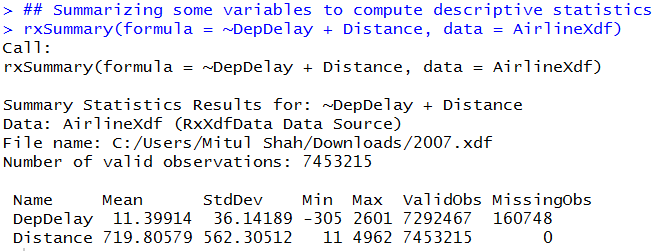


Thus, we see that 500000 rows were read and processed, then another 500000 rows were read and processed and so on. It took about 76 secs to read and process 7.5 million rows.



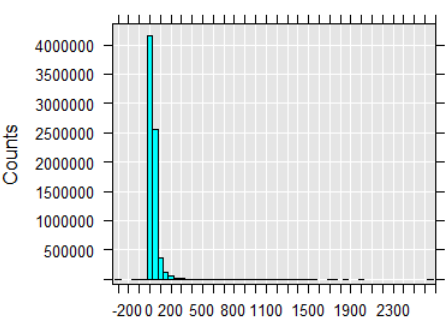


Here, we see that some flights left as early as 5 hours before the scheduled time. While some flights might leave some minutes earlier than scheduled, leaving 5 hours before the scheduled time is extremely unlikely. This might be a data quality issue.

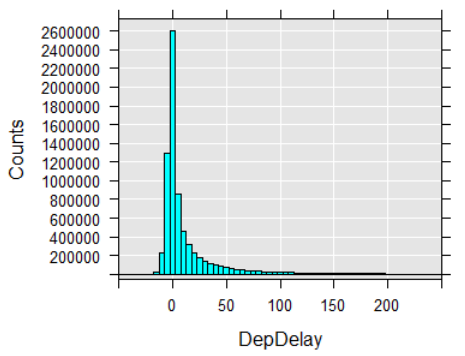


There are some differences between the summary function in Revolution R as compared to that of open-source R. First, we specify the names of the variables we want to summarize. Second, the summary function here does not provide with the quantiles. Instead, it provides with the standard deviation information.



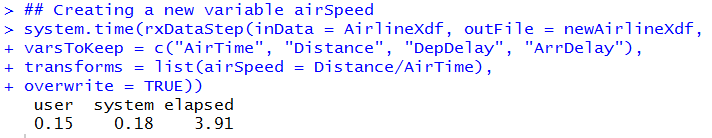




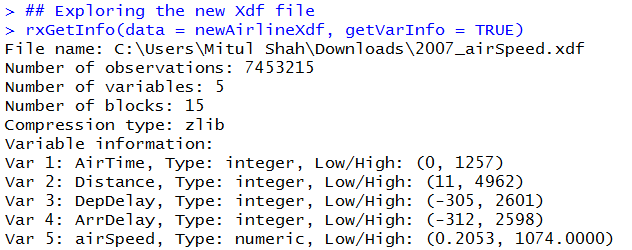


Till now, we have explored the data. Now, let’s try to get back to the question. “Do delayed flights fly faster?” In order to answer this question, we need Air Speed. However, there is no such variable in our dataset. However, we can create this variable airspeed from the existing variables.

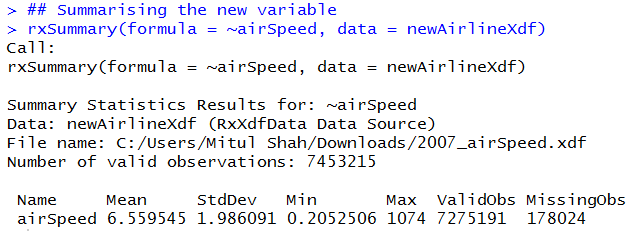




Note that it took just about 4 secs to create this new variable. This is quite a good speed up as it took about 76 secs to read and process the file with the same number of rows.

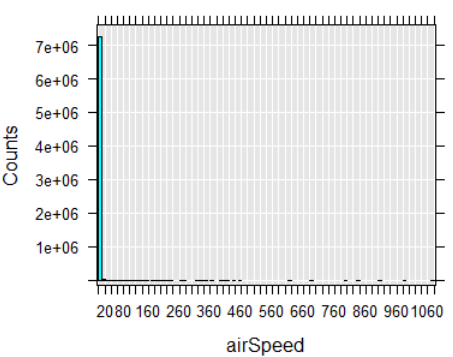


Now, this new file has airspeed variable in it.

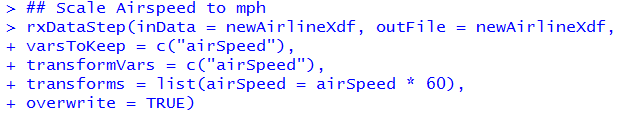


From this descriptive statistic about airspeed, we see that this distribution is quite skewed. The maximum value is about 500 standard deviations away from the mean. We should probably check that whether this is a single outlier or the distribution extends the high in a more continuous fashion.

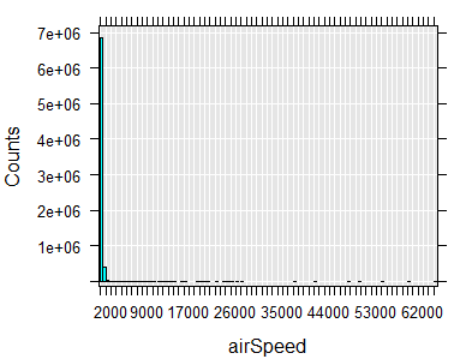




The values of airSpeed seems to be strange. When we created this variable, the variable ArrTime was in minutes. So our airSpeed has the unit miles/minute. Let’s convert this variable to miles/hour.



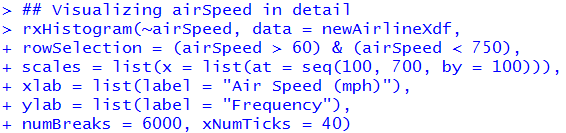


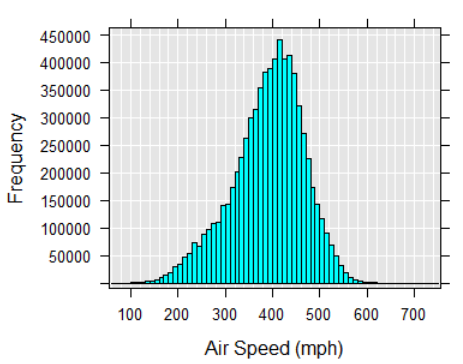


Now, the shape of our histogram looks the same but the values of the airspeed seems to be changed.

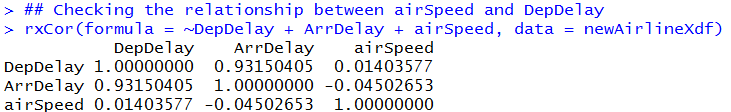
Most flight speeds are between 60 miles/hour to 750 miles/hour. So let’s remove the speeds which are not in this range.

Let’s look at this interesting part of the data.

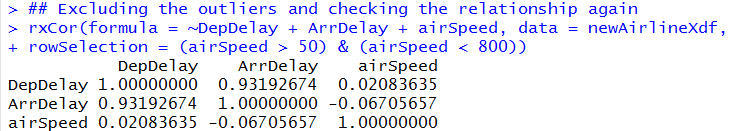




Now, we can create a linear regression model to check whether the delayed flights fly faster. But first, we might want to check whether there is any co-relation between these variables.

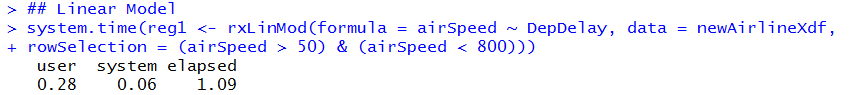


Let’s cut down the airSpeeds and just look at the co-relations for flights having airSpeeds between 50 miles/hour and 800 miles/hour.

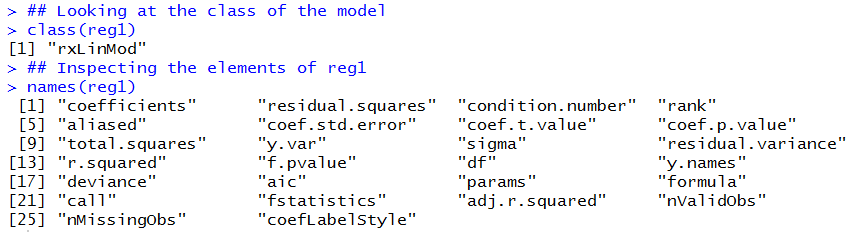


Now, the co-relation gets a little stronger between DepDelay and airSpeed.

Let’s create a linear model now on this subset of the data.

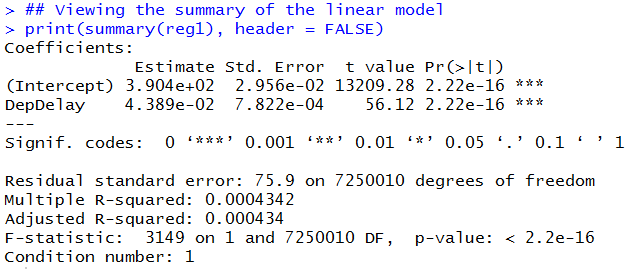


Let’s inspect the class and the elements of this model.



Many elements which are found in standard lm model are missing here. For example, there is no residuals element here. This is an explicit design decision. As rxLinMod is designed to work with extremely large datasets, quantities such as residuals and fitted values are not included in the returned object. However, these values can be easily obtained once the model is fit.

Let’s look at the summary of the model now.

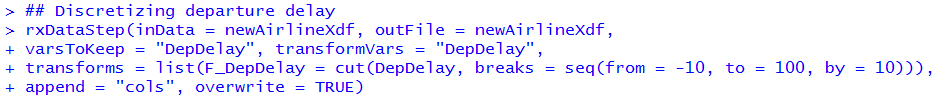


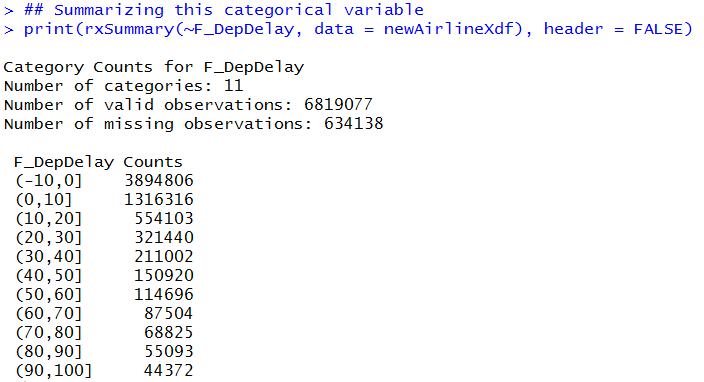
This analysis indicates that, when the flight is late to leave, the plane does actually flies faster. This co-efficient is statistically significant. But it is worth noting that the co-efficient is miniscule. For every minute of delay, the plane flies 400th of a mile faster. Thus, my experience of departing 30 minutes late and arriving 5 minutes earlier appears to be pretty far from the normal. Maybe I was pretty lucky and the weather was working in my favor.

It is also possible that this model does not capture the relationship between the Departure delay and airSpeed. Or maybe there is a non-linear relationship between the two that is being masked by this linear regression model. For example, if there is a case that if the flights have small amount of time delay, pilots feel like they can make up for that time by speeding up. But, if the delay is substantial, the pilots tend to give up and do not try to make up for any lost time.

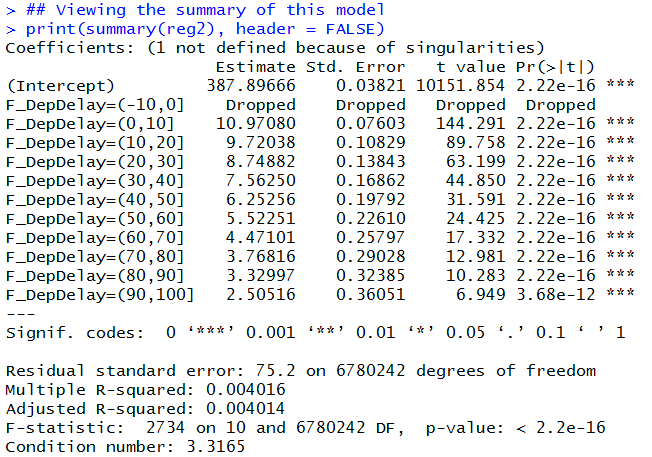
We can examine this hypothesis pretty easily by discretizing departure delay.











We can see that we now have different co-efficient for different ranges of departure delay. The intercept corresponds to the average air speed for flights that had departure delays of less than or equal to 0 by virtue of the fact that that level was dropped. We also see that the flights that have a departure delay between 0 to 10 minutes are predicted to fly at about 11 miles/hour faster than the flights that depart on time. Another 10 minutes of delay would be a little bit slower and so on. The co-efficients get smaller and smaller as the departure delay increases. But again, these values are still miniscule.

The experience of departing 30 minutes late and arriving 5 minutes early still appears to be pretty far from the normal. I was lucky that there were another variables influencing my flight speed. The pilot did mention that the weather was working in our favor.