Nova Scotia Government Employee Absentee Analysis

Chris Selig

August 20th, 2017

## Summary

The Nova Scotia government, like many governments now a days, are allowing people to access some of their data from their open data portals. For this analysis, I wanted to access the [Employee Absenteeism](%22data.novascotia.ca/Public-Service/Nova-Scotia-Government-Employee-Absenteeism/3kpf-veux%22) data. I didn't have any particular expectations of what I would find, but I did hope to forecast how many hours employees would be absent on any given day. The code for this project can be found in my github page.

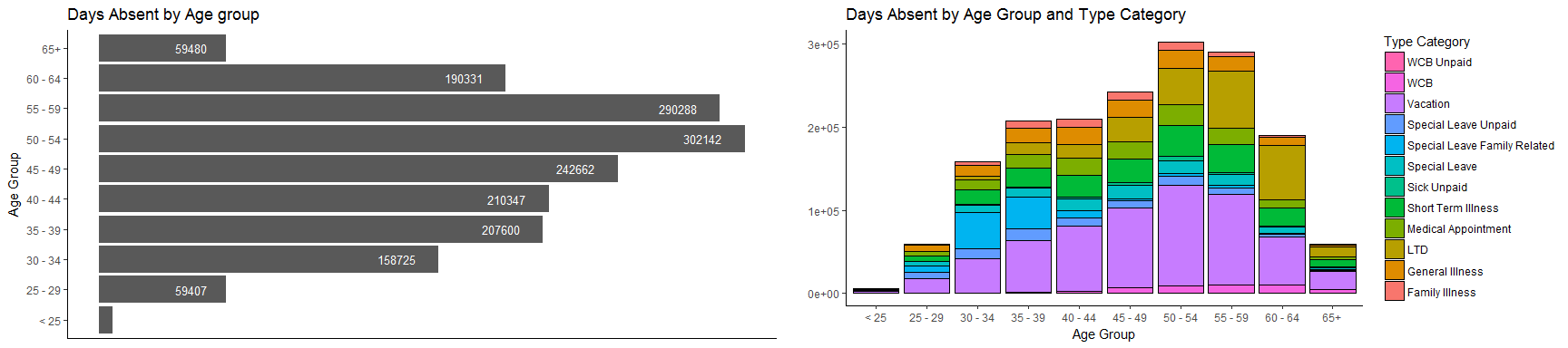
For an extra little bit of fun, I have included a simple app that allows you to forecast how many hours a day the government employees will not work.

## Exploratory Analysis

The absentee data has seven variables: 1. absence\_date: The date of absence for the employee 2. absence\_hours: The amount of hours the employee was absent 3. absence\_type: The reason for the absence 4. absence\_type\_category: The absence reasons are grouped into standard categories 5. age\_cohort\_on\_absence\_date: The age range of the employee, the day they were absent 6. employee\_type: Whether the employee was union, or non-union 7. gender: The gender of the employee

Data in this set runs from January 2014, to March 2017.

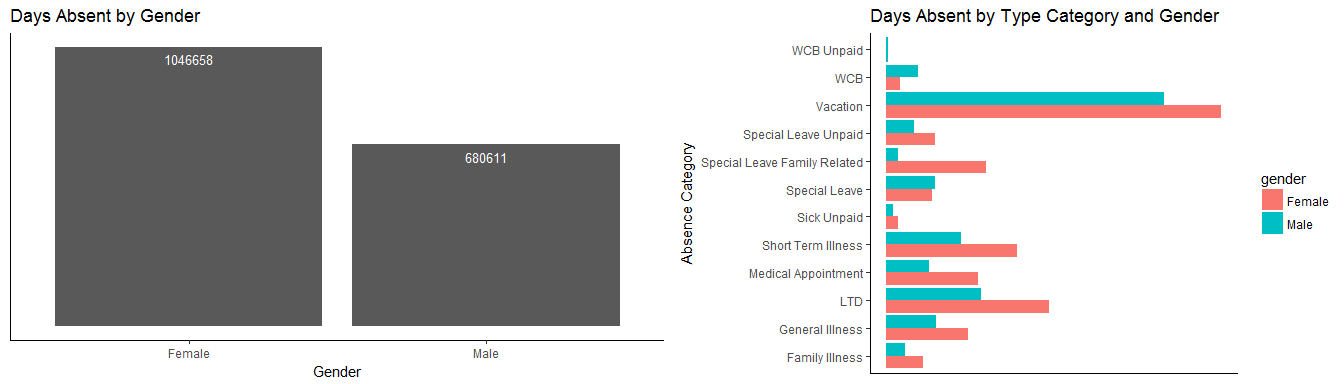
Now that we know what types of data are in the dataset, let's take a look at it to glean some insights. Taking a look at the days absent by age group.



The age group 50-54 has the most absences out of any other age group, with 55-59 as a pretty close second. Without knowing the actual number of people in each age group we should be careful not to infer too much from the findings. Personally, what I find most interesting is that the 65+ and 25-29 age group have approximately the same amount of days absent. With a typical retirement age of 65, I was expecting the 65+ age group to be pretty much non-existent, like the under 25 age group is.

On the right, you can see the absences based on the absence type category. There are some expected observations here. For example, as people get older there are more days absent due to vacations. That makes sense because normally, the longer service time with an employer, the more vacation time you get. Long term disability (LTD) and short term illness both increase with age, which also makes sense.

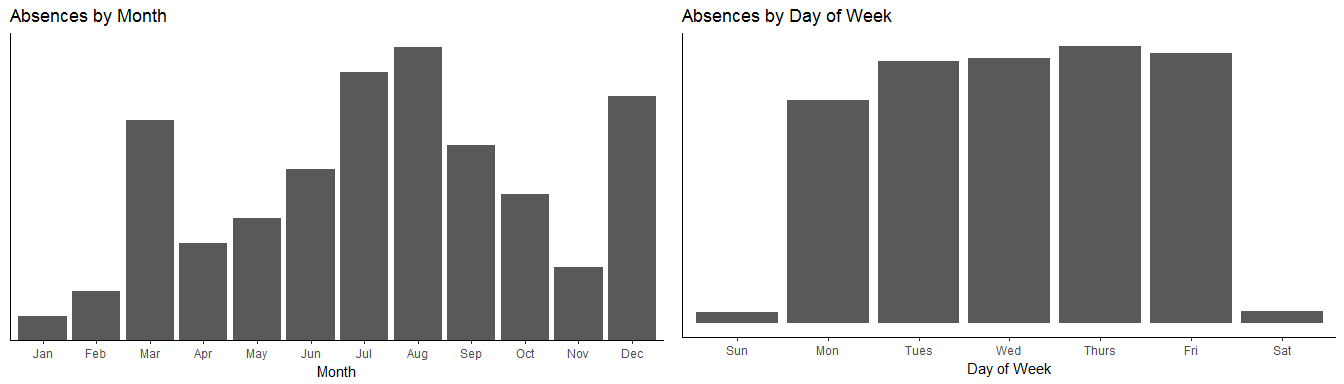
Next, we will take a look at the absensences based on gender.



On the left, in the image above, females have 42.4% more days absent then men. This makes me wonder what the ratio of women to men is in the dataset. Based off this chart alone, I would suspect there are more females working for the government than men.

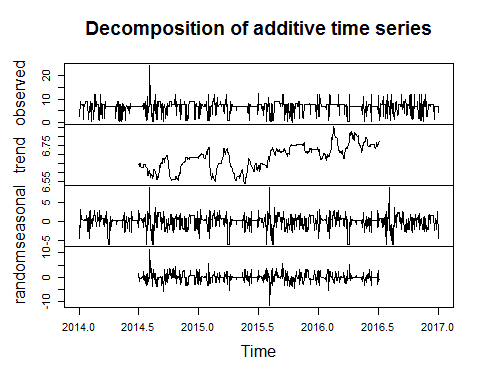
On the right, you can see another breakdown by gender, but this time, I have added the absence type category. There are some interesting observations from this graph. First, more men have absences based on Workers Compensation (WCB). This could be because of traditional gender roles with more men in manual labor type roles than women. Second, for many other type categories (for example, long term disability (LTD), family illness, etc) it appears that females have roughly 40% more absences than men, leading credibility to the fact that maybe there are more females than men in this dataset. However, what I found most interesting is that the vacation type category does not have a significant difference between the two genders. Does this mean that females are not taking as many vacation days? Men have more vacation days than women? Maybe there are not significantly more females than men in the dataset? Without more information, it is hard to tell what the case may be.

Now, we will look at the absences over time. We will look by month as well as by day of the week. Typically, certain months are known to have more absences. March, is normally the spring/March break time in Nova Scotia so people will take a vacation with their kids. July and August are typical months for vacation because school is out and the weather is nice. Below, this is the trend that we see, with the only other abnormal month being December. This also isn't surprising because it is the Christmas holidays and people will take time off to spend with family.



On the right, are the absences by the day of the week. I was really expecting to see both Monday and Friday as the days with the most absences because they strattle weekends. Instead, Monday has the least amount of absences (ignorming weekends because most government employees do not usually work weekends). Thursday has the most. It would be interesting to see if that is statistically significant or not, but that is not the goal of this exercise so I will leave it up to the reader. One theory why Tuesday through Thursday have many absences may be because people take those days off to run errands.

For the final section of the exploratory analysis, we will look at some of the underlying information of the time series itself. Decomposing the time series into it's trend, seasonal and random components, it appears that there isn't a trend component, but there might be component of seasonality.



Let's test for stationarity. We will use the [augmented Dickey-Fuller test](%22https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test%22), which is a unit root test for stationarity.

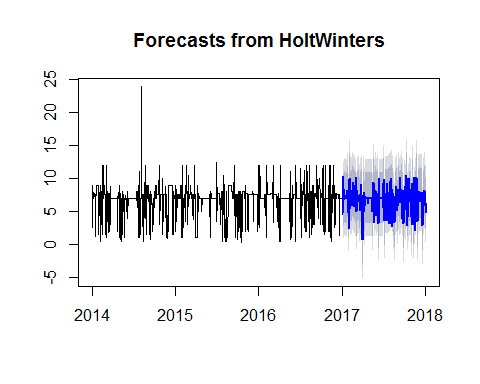
##   
## Augmented Dickey-Fuller Test  
##   
## data: timeseries  
## Dickey-Fuller = -10.478, Lag order = 10, p-value = 0.01  
## alternative hypothesis: stationary

After running the test, we get a p-value of 0.01, which is less than 0.05, so our time series is stationary. This is a good result because stationarity is required for most time series models. Stationarity just means that with the time series, its statistical properties such as mean, variance, and autocorrelation are constant over time.

## Time Series Modeling

Next, I wanted to forecast the hours absent on any given day. I chose hours because you can only be absent for partial days (to go to appointments, etc). I tried two different models, ARIMA and Holt Winters. The rest of this analysis will focus on the Holt Winters model, but if you want to see the results of the ARIMA model, feel free to check out my github page.

Holt Winters is an expontential smoothing method developed by C.C. Holt and his student, Peter Winters. This method is popular because it can produce short term forecasts on data that contains trend and a seasonal pattern, is simple, has low data storage requirements, and adapts to changes in trends and seasonal patterns as they occur.

Without further ado, after creating the Holt Winters model, I have forecasted the next 365 days. 

The blue line is the point forecast for each day. The darker grey area is the 80 percentile area, where 80% of the actual values should appear in this range. The light grey is the 95%.

## Conclusion

If you have enjoyed reading this post, feel free to review my other posts on my [website](%22www.chrisselig.com%22). Thank you for reading.

## License Information

This document [contains information licensed under the Open Government Licence - Nova Scotia.](%22https://novascotia.ca/opendata/licence.asp%22) Dataset and further information can be found by following this [link](%22data.novascotia.ca/Public-Service/Nova-Scotia-Government-Employee-Absenteeism/3kpf-veux%22). Information was accessed August 19th, 2017.