Data Machine Team June 14, 2020

DataFest 2020: Data Analysis Write-up

As the COVID-19 pandemic remains to be highly contagious, different cities around the globe began to be in lockdown since the beginning of 2020. Citizens should have generally stayed at home most of the time except for grocery and necessary outings, which infers they stopped taking part in various outdoor activities. Under this unusual time, people are forced to engage in more indoor activities. Our team wishes to explore this societal impact on Canadians with the aid of Google Search data by asking the following question: How does the lockdown change people's interest in indoor and outdoor activities? We looked at Google Trends data to see if that was the case.

Our analysis compares the date ranges from two periods, before and after the national lockdown date of March 17th, and the COVID-19 infection rates in the country. We used the *gtrendsR* package to extract the Google Engine search data of different activities that stand for indoor and outdoor activity. For example, we used 'hair cut' to indicate people's intention to style their hair at home, and 'hair salon' to indicate people's intention for going to a physical hair salon. We observed a trend of search volume for hair salons being higher from February to late March and from late May to June 2020. Starting from there, we hypothesized that in general, the intent for outdoor activities is higher than indoor activities if not due to COVID-19.

To further explain the trends, we fit a linear regression model to compare both periods to explore if there is a significant difference in the search volumes of indoor and outdoor activities, as well as to see if there was enough evidence to indicate a linear relationship between the search popularities and infection rates. We calculated the mean and variance for the number of confirmed cases in Canada and compared them with the ones generated from Google search volume of indoor and outdoor activities to evaluate if there is any correlation. Looking at the linear correlation coefficient between them, we determined the strength and direction of the linear relationship. As a matter of fact, the sign of r indicated the direction of the linear relationship between the x variable (number of cases) and the y variable (number of searches on indoor and outdoor activities).

In our main analysis, the independent variable is the simple linear regression of daily new confirmed cases, while the dependent variables are indicators of search hits on indoor activities and outdoor activities. We set up the Null hypothesis as no linear relationship between y and x, and the alternative hypothesis as there is a linear relationship between y and x. From this, we analyzed the linear relationship between indoor activities and number of cases before the lockdown firstly, we saw that at a significance level of 5%, since p-value < 0.05, we reject the null hypothesis which means that there is a linear relationship between the number of cases and the number of search terms on indoor activities. With an R^2 =0.6938, this 69% indicates the linear regression model explains most of the variation in the dependent variable (number of searches on indoor/outdoor activities) around its mean. Meanwhile, there is a linear relationship between the number of cases and the number of search terms on outdoor activities because the p-value is also less than 0.05, but R^2 lower which is 0.59. Additionally, before the lockdown, there is a stronger indication of a positive linear relationship between the number of cases and search of indoor activities. Compared with estimates computed from the data before the lockdown, the indication of a linear relationship for both indoor and outdoor activities is relatively weaker (r_indoor=0.37; r_outdoor =-0.33) after the lockdown. However, average search hits for both indoor and outdoor activities are higher after the lockdown.

To conclude, the number of searches surged during the lockdown, as expected, the average hit for outdoor activities in both periods are higher than indoor activity. From our data, we suggest a positive linear relationship between the number of confirmed cases and search of indoor activities, and surprisingly, the number of confirmed cases of COVOID-19 is negatively correlated to outdoor search hits.