Final-Project-Statistical-Modelling-with-Python

By : Brigitte Sullivan Submitted on: Friday, September 1st 2023 Lighthouse Labs Data Science Program

Project/Goals

Project description and goals:

- Perform Statistical Modelling to determine the relationship between the prevalance of "green spaces" / outdoor spaces and city bike stations in Montréal. Specifically, the relationship between number of city bike slots and the top 10 most common outdoor spaces within 1km of city bike stations.
- chose Montréal because of its ample outdoor spaces and "picnic"-culture
- There was no explicit mention that free bikes is the required independent variable in compass content and felt that number of slots represented overall supply/demand of bikes better than a point in time number of free bikes.

Null hypothesis:

there is no relationship between the number bike slots and the number of outdoor POI's

Alternative hypothesis:

there is a relationship between number of bike slots and number of outdoor POI's

Process

The steps taken to complete this project were:

1. Extract Data from city bikes and foursquare (Part 1 & 2)

- used api get request to gather the data.
- gathered locations for all city bikes stations in montreal, used these station locations to use as the center of the radius search for the foursquare request
- focused on bringing in as much data from city bikes as possible to have the most options better likelihood of building a model that was good at predicting the target variable (this created a challenge discussed later)
- Decided to focus on first gathering data from foursquare and if time, return to extract yelp data.
- Chose to focus foursquare data on any POI's that had the parent category of "Landmarks & Outdoors" (e.g., anything with a categoryid in the 16000's). Removed additional categories as a data cleaning step (seemed simpler to do using a dataframe filter than in the API call)

2. Join Data Sets & and initial exploration (Part 3)

- joined city bikes data and foursquare data based on station location. City bikes data had one location per station. However the foursquare data could have many POI's for one location (meaning it's possible to have multiple outdoor spaces within 1km of a bike station in Montreal).

- performed QA /data validation to ensure the join of two data sets performed as expected. QA done by taking a sample station and examining what the output should be then confirming that the joined data set contained this output.
- EDA showed no obvious relationship between any of the possible target variables and possible independent variables

3. Statistical Modelling

Original approach: I first attempted simple linear regression on all possible combinations (8) of independent variables:

- outdoor_space_num
- num_parks

with target variable:

- free_bikes
- empty_slots
- slots
- ebikes

The highest adjusted r-square value was 0.0055 of all the 8 models I created. Given I had the time, I decided to revise my approach in the hopes that I would find a 'better' model before going back to get yelp data.

Revised approach: Decided to restructure the data to have the number of outdoor space by type as seperate columns so that there are more options for independent variables, and narrowed the independent variable to number of slots.

The steps taken in the revised approach were:

- Restructure Data
    - pivot the category name column so that each outdoor space category had a column with a numeric value
- Address NaN values
    - pivoting the category name column created many NaN values that needed to be resolved.
- Perform multivariate linear regression with backward selection using number of slots as dependent variable and each outdoor space category as independent variables.
    - Target Variable : # slots
    - Independent Variables:
        - Park
        - Playground
        - Monument
        - Farm
        - Garden
        - Dog Park
        - Campground
        - Hiking Trail
        - Landmarks and Outdoors

    - Historic and Protected Site

Results

Here are the results of the statistical modelling performed:

- The adjusted R-quare value in the final model after performing backward selection was 0.192.
- Meaning that about 19.2% of the variance in the dependent variable can be explained by the independent variables. This is generally not considered high adjusted r-square value, meaning the correlation is not strong.
- In model output A (the model previous to the final model "B"), Park is technically not statistiscally significant with a pvalue of 0.062 (but very close to 0.05 threshold). When Garden is removed from Model B, the pvalue for Park drops and Park becomes a statistically significant.
- **Historic and Protected sites** is the outdoor space category that makes the largest contribution to the model with the highest coefficient value, followed by **monuments**. Both categories have a P-value of 0.000 indicating high significance.
- Conclusion: There is a relationship between the number of bike slots and the number of these outdoor spaces within 1km of the station (in order of descending significance):
    - Historic and Protected Sites
    - Monuments
    - Farms
    - Parks
- although my original thought that parks would be the strongest indicator the model found that there are other outdoor spaces that better predict the number of slots than Parks (like Historic Sites, Monuments, Farms)

Challenges

The challenges I encountered in this project were:

- determining the dependent and independent variables too late. I admitedly wanted to pick "the right variables" that would return the "best" model without any information meant I delayed an important decision for perhaps too long. Not having the question fully formed prior to data extraction meant having too many options for dependent variables and to few options for independent varibales.
- not finding any "good" models during first iteration of modelling and needing to restructure data.
- The first iteration, the two posible independent variables outdoor space num and num parks violated the independence assumption since the num_parks variable was a direct subset of the outdoor space num. This was another reason for why I needed to revise the my approach.

Future Goals

(what would you do if you had more time?)

- Explore the Yelp API data to see if the results are any different
- Analyse Landmarks & Outdoors as an entire category
- Perform multivariate analysis on all outdoor categories
- change the dependent variable to number of free bikes to see if the model outputs improve