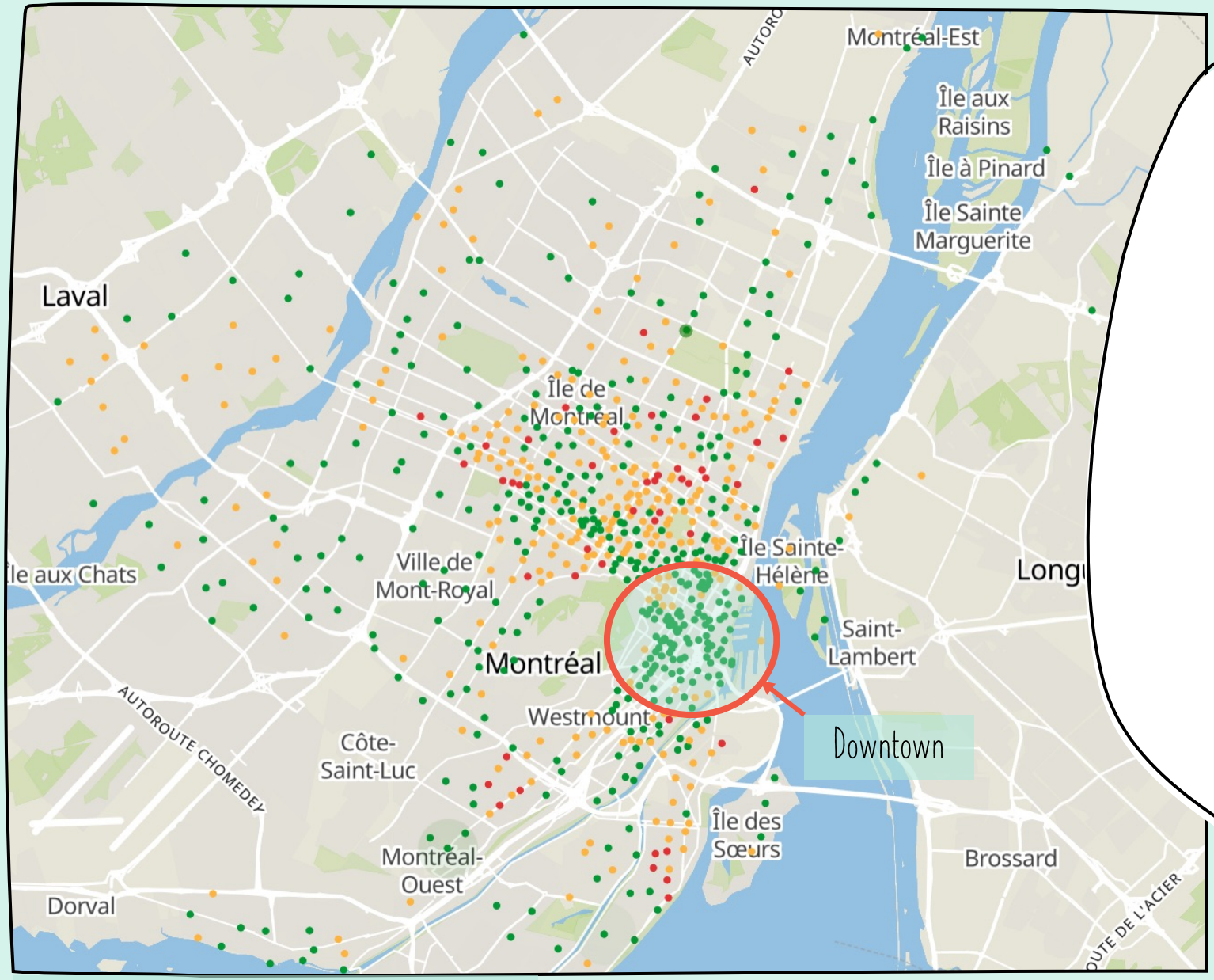
Statistical Modelling with Python

Presented by
Stefan Stefanovic
Jun 5th 2023

PROJECT FLOW STRUCTURE

1. Explored the structure of the APIs and choose Montréal as the city
2. Gathered data from CityBik, Foursquare and Yelp
3. Joined and cleaned gathered data
4. Exploratory Data Analysis
5. Regression Model Development
6. Model Evaluation and Interpretation



Screenshot taken on Monday, June 5th 2023 at 01:26 PM

BIXI, MONTRÉAL, QC

- 744 bike stations
- 10,000 bikes
- 1,534,698 trips in July 2022

REGRESSION MODEL DEVELOPMENT

Predicting AVAILABLE BIKES using:

- Station Number,
- Latitude and Longitude,
- Average rating of parks,
- Total ratings count,
- Number of Parks (1,000m)

REGRESSION MODELS

Model 1 - used data from Yelp and Foursquare

Model 2 - used data from Foursquare

Model 3 - used data from Yelp

	coef	std err	t	P>|t|	[0.025	0.975]
const	990.5679	1567.721	0.632	0.528	-2102.792	4083.928
station_no	0.0009	0.002	0.385	0.7	-0.004	0.005
latitude	-23.7312	14.85	-1.598	0.112	-53.032	5.569
longitude	-1.3271	17.231	-0.077	0.939	-35.327	32.673
avg_rating	1.0747	1.53	0.702	0.483	-1.944	4.094
total_review_count	0.0089	0.007	1.272	0.205	-0.005	0.023
POI	0.3028	0.18	1.683	0.094	-0.052	0.658
avg_rating_fsq	-0.895	0.701	-1.277	0.203	-2.278	0.488
total_ratings_count_fsq	0.0058	0.002	2.357	0.02	0.001	0.011
POI_fsq	-0.5337	0.461	-1.158	0.248	-1.443	0.375

MODEL 1 -DATA FROM YELP AND FOURSQUARE

R-squared: 0.092

Adj. R-squared: 0.047

	coef	std err	t	P>|t|	[0.025	0.975]
const	1850.6452	822.86	2.249	0.025	229.484	3471.806
station_no	0.0018	0.002	0.996	0.32	-0.002	0.005
latitude	-11.7354	10.733	-1.093	0.275	-32.882	9.411
longitude	17.737	8.417	2.107	0.036	1.155	34.319
avg_rating_fsq	-0.5472	0.582	-0.94	0.348	-1.694	0.599
total_ratings_count_fsq	0.0048	0.002	2.248	0.025	0.001	0.009
POI_fsq	-0.3829	0.387	-0.988	0.324	-1.146	0.38

MODEL 2 -DATA FROM FOURSQUARE

R-squared: 0.049

Adj. R-squared: 0.025

	coef	std err	t	P>|t|	[0.025	0.975]
const	3024.037	826.461	3.659	0	1400.799	4647.275
station_no	-0.0007	0.001	-0.521	0.602	-0.003	0.002
latitude	-39.3574	8.585	-4.585	0	-56.218	-22.497
longitude	16.7528	9.306	1.8	0.072	-1.525	35.031
avg_rating	1.4531	0.7	2.075	0.038	0.077	2.829
total_review_count	0.0106	0.004	2.538	0.011	0.002	0.019
POI	0.2515	0.099	2.546	0.011	0.057	0.446

MODEL 3 -DATA FROM YELP

R-squared: 0.091

Adj. R-squared: 0.082

BIGGEST CHALLENGES

Limited Data Availability

Data for bike stations was collected only in one instance

IF GIVEN MORE TIME

Conduct Deeper Analysis: Explore additional variables and gathering more comprehensive data for bike stations. This includes extending the data collection period to cover longer durations, capturing different times of the day to account for variations in bike usage patterns.

Perform External Validation: Validate the regression model's findings by comparing them with external data sources.

Fine-tune the Model: Refine the regression model by incorporating additional variables or considering different modeling techniques to improve its predictive power and accuracy.

THANK YOU FOR YOUR TIME AND ATTENTION

Stefan Stefanovic
Data Research Analyst