CMPT353: Airbnb in Vancouver

I. Introduction

- When travelling to Vancouver, where should a tourist book an Airbnb.
    - Results will be based on most nearby amenities
- Where are Airbnbs with the most nearby amenities concentrated in Vancouver
- How are amenities and reviews, price and the number of people an Airbnb accommodates related

II. Data cleaning

We performed data cleaning on the provided osm data of amenities in Metro Vancouver and Airbnb listings active as of September 17, 2019, for the city of Vancouver. For both dataframes, all rows with NaN value are removed, to ensure data analysis can proceed smoothly.

Since the osm data frame contained data for all of Metro Vancouver, we filtered out amenities outside of the boundary of Vancouver, to obtain a more accurate and meaningful results based on distance. We then added weights to each of the amenity, based on some predetermined criteria in *weights.csv*, in preparation for analysis.

III. Data Analysis and Results

osm.py

- For each amenity in amenities-vancouver.json, we gave them a weight between 0-5, based on their usefulness to tourists (as assigned in *weights.csv*)

airbnb.py

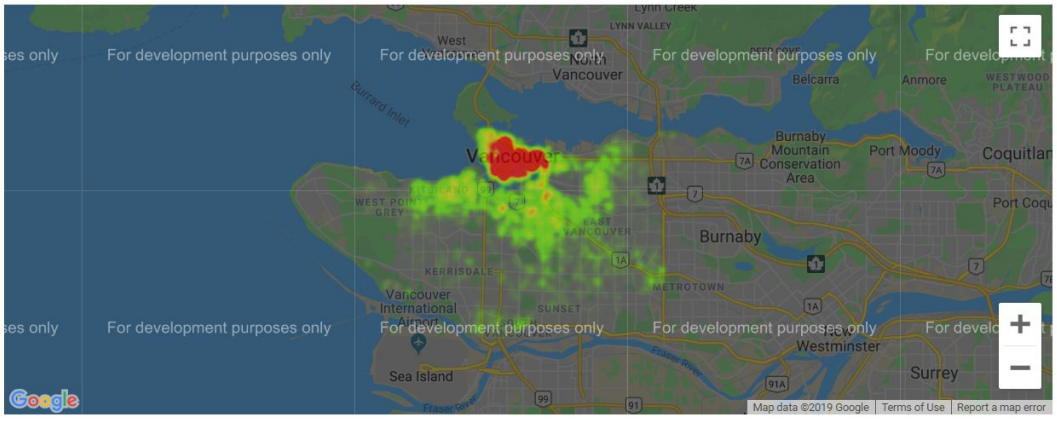
- Calculate distance between one airbnb and every amenity using Haversine formula
- All distances less than 1km are counted toward *count_amenities* and *weight_score*

project.ipynb

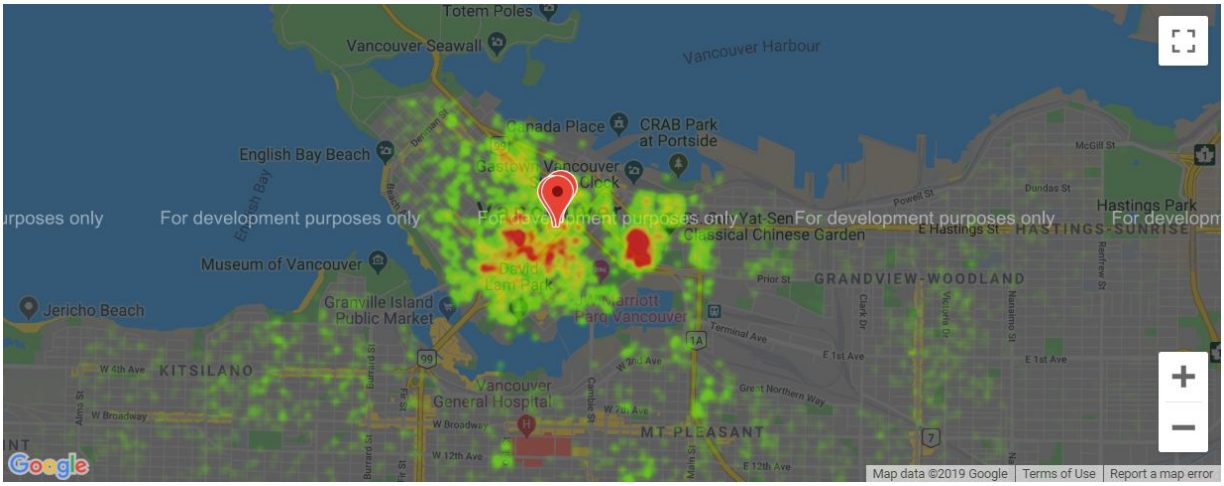
- With the count_amenities and weight_score now added, we look at visualization of these parameters using Jupyter gmaps plugin
- We generated a heat map with weights equal to the number of amenities

Heatmap of Airbnb locations weighted by number of amenities

```
fig1 = gmaps.figure()
fig1.add_layer(gmaps.heatmap_layer(locations, weights=count, max_intensity=30, point_radius=7.0))
fig1
```



Zoom level 2 (Downtown Vancouver and vicinity)



- When we compared the heatmaps where weight was equal to the number of amenities, vs the weight score we calculated, we noticed they only differed slightly. Because of this, we decided to only display the heatmap where the weight is the number of amenities.
- The markers on the zoomed in heatmap indicate the top 3 Airbnb's that have the most nearby amenities.
- The heatmap shows a hotspot, indicating:
    - There are a lot of Airbnb's located in this area; and/or
    - There are Airbnb's that have a lot of nearby amenities

stats.py

- To further analyze the relationship between the number/score of amenities and the other variables of an airbnb, we performed the following statistical analysis:
- Scatterplot
    i. Review rating vs. number of amenities

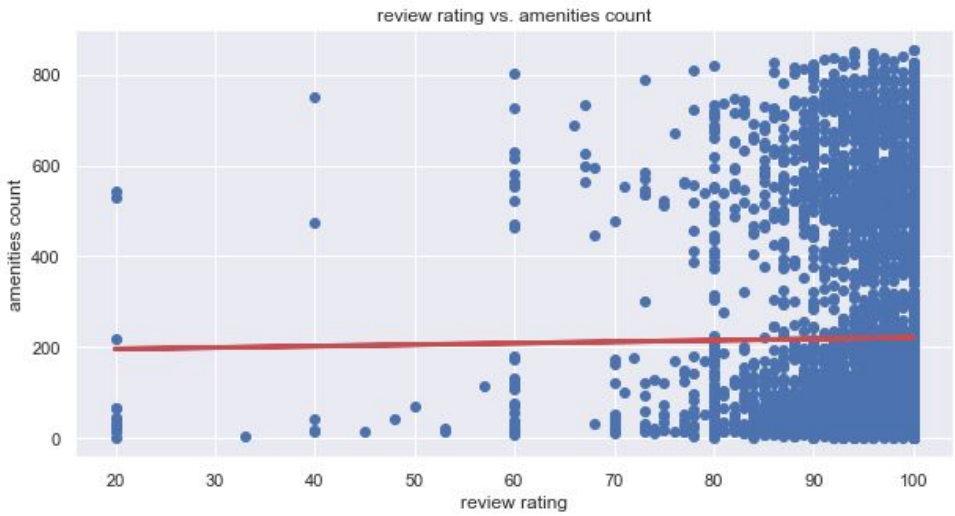


Figure 1- review rating (/10) vs. count of amenities. The gap at 370 amenities indicates a possible separation between Downtown Vancouver and the rest of Vancouver. The r^2 value is 9.65e-05 and the p-value is 0.471, this means there is no significant linear relationship between these two variables. This is likely due to the rating bias, see limitation.

ii. Price vs. number of amenities

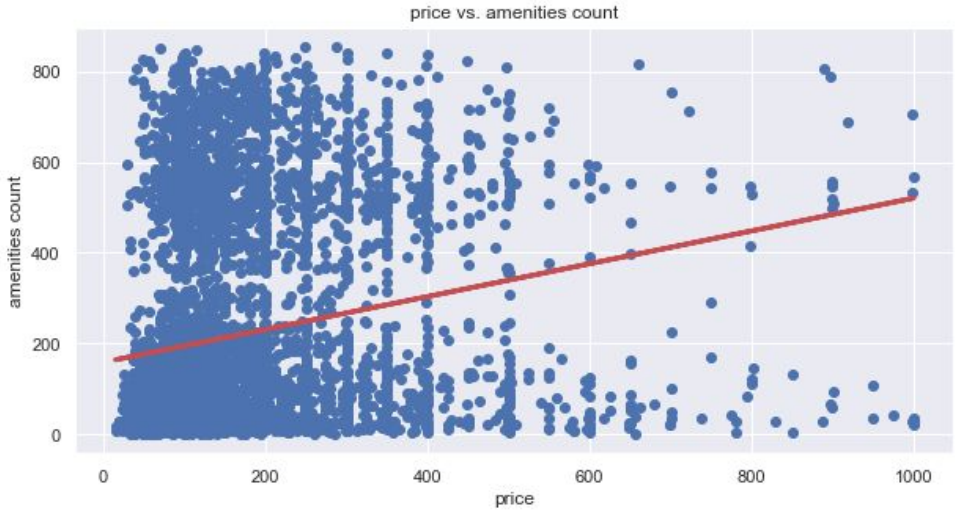


Figure 2- price vs. count of amenities. Like in figure 1, the gap at 370 amenities indicates a possible separation between Downtown Vancouver and the rest of Vancouver. The r^2 value is 0.0392 and the p-value is 2.55e-48, this means there is a statistically significant linear relationship between price and count of amenities. The scatterplot suggests that there is ~4% increase in price due to the number of amenities. Based on our findings, we further analyzed this relationship by looking at the OLS summary to obtain a more significant result.

- OLS summary of price vs. [accommodates, amenities]

OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	0.394
Model:	OLS	Adj. R-squared (uncentered):	0.394
Method:	Least Squares	F-statistic:	1736.
Date:	Mon, 02 Dec 2019	Prob (F-statistic):	0.00
Time:	11:28:13	Log-Likelihood:	-37081.
No. Observations:	5340	AIC:	7.417e+04
Df Residuals:	5338	BIC:	7.418e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x	0.8359	0.034	24.877	0.000	0.770	0.902
z	6.8236	1.721	3.966	0.000	3.450	10.197

Omnibus:	284.454	Durbin-Watson:	1.849
Prob(Omnibus):	0.000	Jarque-Bera (JB):	342.664
Skew:	0.555	Prob(JB):	3.90e-75
Kurtosis:	3.554	Cond. No.	106.

- Key statistical value:
  - **R-squared value**: 0.394, **Prob (F-statistic)** : 0.00

The R-squared value 0.394 indicates there is a relationship between price vs [# of amenities, accommodates], where accommodates is how many people the airbnb can accommodate. Since the F-statistic is 0, we can reject the null hypothesis that there is no relationship between price and [amenities, accommodates], and that indicates a statistically significant relationship as given by the R-squared value. Given a price, we can

explain 39.4% of it because of the relationship between price and # of amenities, and how many people the AirBnB can accommodate.

IV. Limitations

- The AirBnB data was limited to Vancouver, whereas the OSM data, contained amenities for all of Greater Vancouver.
- Rating Bias: People are likely to only leave ratings when they highly favored or disliked the airbnb - oftentimes not meaningful (people usually leave 1/5 or 5/5) - Source: Greg
- If we had more time
    - We could have been able to use the larger OSM data set, and the AirBnB data for all cities available, taking a Big Data approach to our question.
    - Additionally, the results from the OLS summary were significant, and could have been a project itself. If we had been given more time, we would have spent more time analyzing these results.
- Downtown Eastside and the gap that appears in the heatmap might be outlier

V. Project Overview (Accomplish statements for all group members)

Manjot:

- Cleaned OSM data to prepare for analysis
- Implemented Haversine formula to calculate the distance between each airbnb to all amenities
- Implemented OLS and reported analyzed results

Melody:

- Cleaned Airbnb data for data analysis
- Using Jupyter Gmaps, generated heatmaps for visualization of airbnb amenities count and score
- Graphed results of linear regression for visualization of statistical analysis

Overview:

- Using Jupyter Gmaps, generated heatmaps for visualization of Airbnbs, weighted by nearby amenities, found that Downtown Vancouver is the best place to rent an Airbnb
- Determined that the price of an Airbnb is significantly related to its nearby amenities, and the number of people it accommodates, using OLS to analyze this relationship