NYC Green Taxi Trips 2015

CURRENCYFAIR DATA CHALLENGE

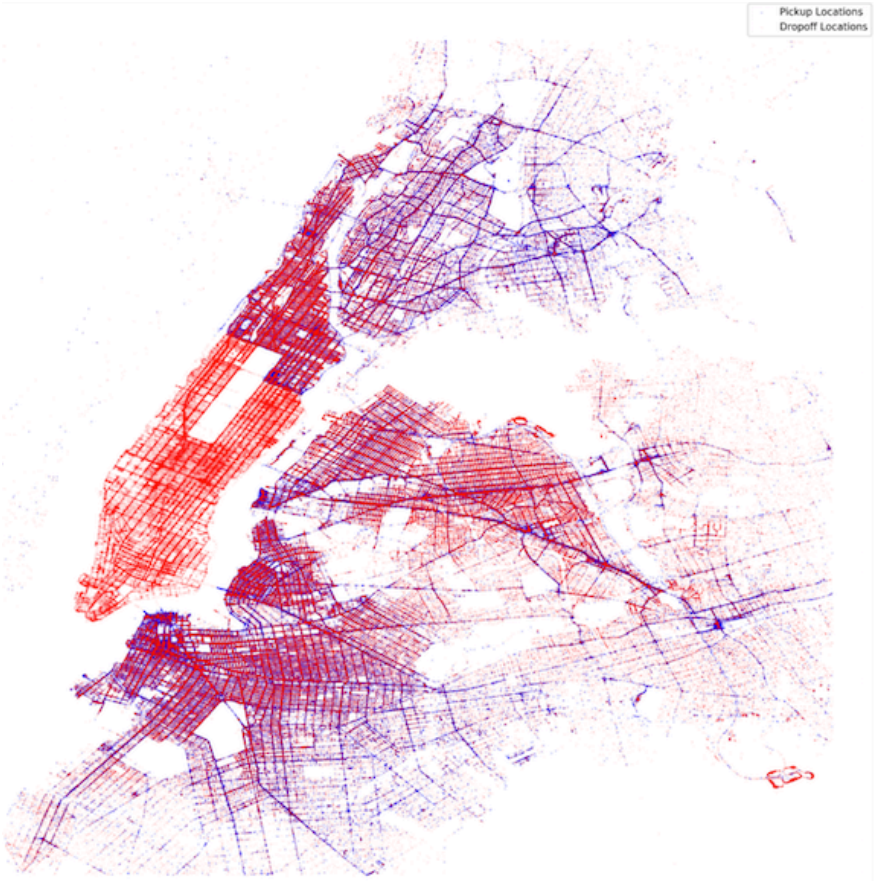
ANDY MCSWEENEY

Challenge

- The objective is to find a few recommendations that you would give to a green taxi driver in NYC.
- This exercise will help demonstrate some of the following analytics skills:
  - Data acquisition & wrangling
  - Data visualisation
  - Predictive modelling

Idea

- Originally planned to try and predict fares/tips
- But this had already been completed many times
- Imagined myself as being a consultant for a NYC Green Taxi Driver who enjoys his current neighbourhoods working hours but would like a change in scenery



Scenario

- Tom is a NYC Green Taxi Driver
- Tom current lives in Westchester Village/Unionport (BX59) but want's a change of scenery
- He likes his current work schedule and would like to move to a neighbourhood with similar working hours and if possible, a more profitable neighbourhood
- Question: Can I build a unsupervised machine learning model to identify similar neighbourhoods to suggest to Tom?

Solution – Data Pre-processing

- Reduced the dataset to focus on August instead of the whole year
  - Over 19 million lines of data, reduced to a more manageable 1.5 million
- Mapped latitude and longitude values for pick-up locations and drop-off locations to their respective NYC neighbourhood
- Converted pick-up time and drop-off time to date-time datatype and captured the pick-up hour as it's own specific feature

Solution – Model Preparation

- Group pickup locations together and split dataset into number of pickups per hour

pu_nbrhood	total	hour_1	hour_2	hour_3	hour_4	hour_5	hour_6	hour_7	hour_8	hour_9	...	hour_15	hour_16
BK09	30284	511.0	296.0	137.0	107.0	163.0	336.0	842.0	1666.0	1658.0	...	1620.0	1966.0
BK17	4907	151.0	96.0	52.0	69.0	32.0	49.0	143.0	237.0	234.0	...	225.0	292.0
BK19	429	17.0	6.0	5.0	2.0	8.0	1.0	4.0	12.0	24.0	...	38.0	38.0
BK21	2530	114.0	68.0	32.0	15.0	15.0	15.0	56.0	85.0	69.0	...	124.0	142.0
BK23	1118	85.0	61.0	25.0	14.0	2.0	3.0	4.0	6.0	8.0	...	44.0	48.0

Solution – K-Means Model

- Use these columns as input arrays for Sci-Kit Learn's K-Means clustering algorithm

```
kmeans100 = KMeans(n_clusters=12, random_state=0, n_init=100).fit(new_final_df)
```

```
np.unique(kmeans100.labels_, return_counts=True)
```

```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11], dtype=int32),
 array([ 3,  3,  3,  1,  1,  1, 74,  1,  1,  2,  1, 83]))
```

- Final used parameters
  - 12 clusters
  - Algorithm ran 100 times with different seed centroids

Results

- Westchester Village/Unionport
  - 2100 pick-ups in August 2015
  - Mean distance travelled was 3.17 miles
  - Mean total fare was 13.64 dollars

- Using these as minimum criteria, the cluster Westchester Village/Unionport was in, reduced from an initial 82 suggested neighbourhoods to a final 9

Results

pu_nbrhood	pass_count	distance	fare	tip	total	payment_type	trip_type	count	clusters
Belmont	1.347596	2.903305	12.312433	0.642372	13.982074	1.735126	1.140587	2454	11
East Tremont	1.236364	3.042269	12.494624	0.516521	14.023760	1.680165	1.157025	2420	11
Spuyten Duyvil/Kingsbridge	1.210696	2.905067	12.162656	0.737782	14.037446	1.717842	1.088520	2169	11
Van Nest/Morris Park	1.278280	3.073842	12.187434	0.541388	13.724373	1.745211	1.198771	2767	11
Hamilton Heights	1.229236	2.921072	11.827800	1.023102	14.078917	1.587225	1.031938	23984	11
Manhattanville	1.249103	2.867386	11.696914	0.992775	13.914921	1.580956	1.032622	11710	11
Morningside Heights	1.314280	2.758586	11.739215	1.276069	14.342717	1.490644	1.007886	46920	11
Upper East Side South	1.356322	2.662498	11.676133	1.479648	14.388796	1.382691	1.003719	2958	11
Rego Park	1.525585	3.127697	13.036200	0.710555	15.050502	1.739967	1.005769	11960	11


None
Original Neighbourhood
Suggested Neighbourhoods
All Other Neighbourhoods

Recommendation

- Based on the final 9 suggested neighbourhood, my recommendations would be as follows:

- If Tom wanted to move to a neighbourhood far from his current neighbourhood, I'd suggest moving to Rego Park in Queens.

- If higher tips is what is more important to him however, I'd suggest he move to either Morningside Heights or the Upper East Side South