Data science and machine learning presentation

- Presented by: By Aly Elhadad, Finnian Logan-Riley and Arran Logan-Riley.
- Date: 2/05/2022
- University: Hochschule Furtwangen
- Course: Data Science and Machine Learning in Business

Question definitions

1. Higher average income less than 30km or more than 50km compared to total revenue.
2. Find out if more taxi rides are starting in the north or south
   - Starting in the north of New York vs the South of New York?
3. Find out if and how the tip depends on the trip distance.
   - Which trip distance has the most and least tips?
4. Our findings
   1. What areas have the highest and lowest tips?

Take home messages

1. Earnings with range constraints of taxis
   - 30km range is the optimum range for the taxies and >30km accounts for 0.009% of trips and increasing the usable range to 50km will not be worth the investment
- 2. Percentage of rides starting in the north vs the south (taking Bushwick, Brooklyn as the geographic center)
  - 95.8% of the rides started in the north of NYC
  - 4.2% of the rides started in the south of NYC
- 3. Tips
  - $0.25/mile including no tips.
  - $0.50/mile only when tips are given.
- 4. Higher tips
  - Lower Manhattan to 9th Ave and Midtown East extending into Rockefeller Center.

Problem

- Increased competition
  - Recent agreement between Uber and New York taxi companies
  - Allows uber users to use New York taxis from the app
  - Direct competition with Uber
  - (Hu and Rosenthal, 2022)
- COVID-19
  - Less customers
  - Increased competition

•Hu, W. and Rosenthal, B. (2022). Call for an Uber, Get a Yellow Taxi. The New York Times. [online] 24 Mar. Available at: https://www.nytimes.com/2022/03/24/business/uber-new-york-taxis.html [Accessed 1 May 2022].

Approach to findings

1. Data exploration
2. Data sampling code
    - Code used to sample large datasets
3. Use of excel, PyCharm, python and python libraries
4. Data filtering
5. Scatter plot with density
    - Gaussian density colouring
6. Average calculations
7. Correlation lines
    - Linear regression and polynomial fit
8. Data scaling
    - Scaling data between 1 and 0
9. Location based results

- Code
    - Available at https://github.com/FinnianHBLR/data-science-ml

HOCHSCHULE FURTWANGEN UNIVERSITY | HFU

Overview

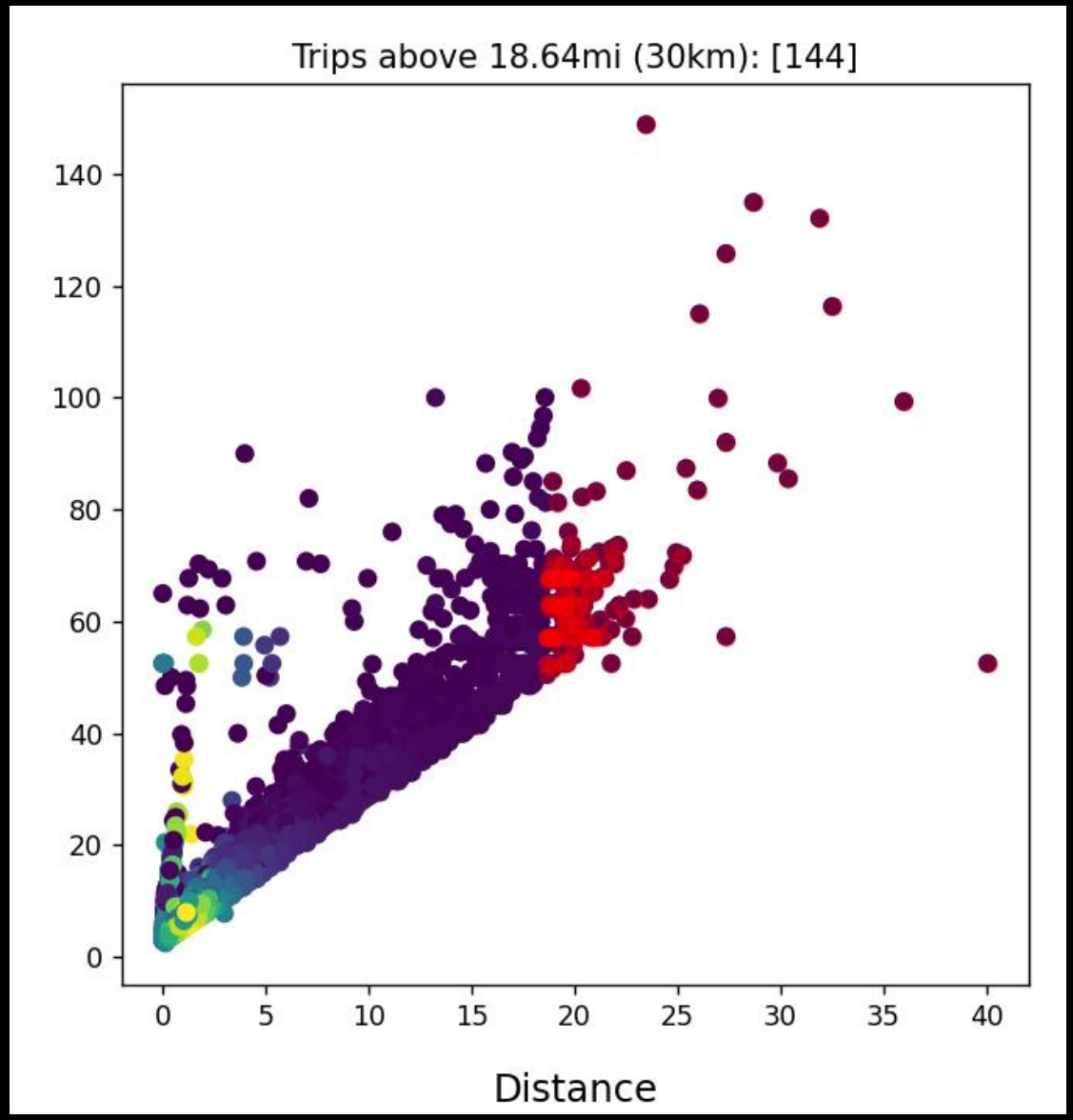
1. Constrained taxi range
   - The results of this show that having a range above 30km does not significantly increase the profitability of a taxi service.
   - The methods that were used was a scatterplot with gaussian density according to different ranges of a taxi service.
2. North or South starting locations
   - The result of this shows that most (95.8%) of the rides occur in the north while only 4.8% occur in the south
   - The methods used are a scatterplot using the longitude and latitude data and matplotlib
3. Tips and distance
   - Results show For every 1 mile there is a tip of 0.25¢
   - Used methods are scatterplots with linear regression, Polynomial fit and density
   - Data was processed without and with $0 tips
4. Tip locations
   - Method used is Geographical map Scatterplot with data normalization
   - Two main areas of interest
     Lower Manhattan and Midtown East

Earnings with range constraints of taxis

1. How much do taxis earn per trip if they can only drive 30km or 50km per trip, compared to taxis that have no limit?
    1. 0.001% of trips are above 30km in New York City
    2. No increase in revenue if the range of taxi is greater than 30km
        1. <30km range = $14.053 average revenue (0.99% of trips)
        2. >30km range = $68.02 average revenue (0.009% of trips)
        3. <50km range = $100 average revenue (0.0002% trips)
        4. Small reputation risk
    3. Averages for all trips
        1. Cost: $14.74
        2. Distance: 3.4km

Range Constraints Density Chart

1.Observed 15000 entries
2.Removed faulty data
3.Split data into relevant distances (<30,>30, <50...etc)
4.Plotted data onto a scatter plot with gaussian density
5.Calculated averages for different distance ranges



Red dots are above 18.64 miles

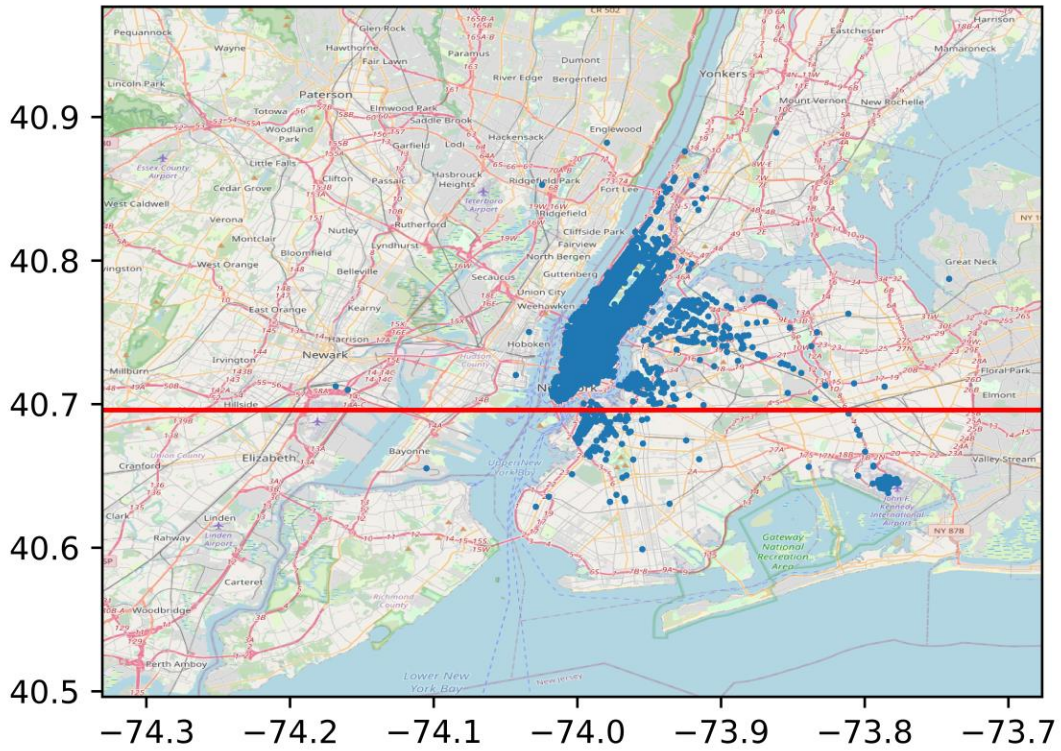
North or South starting locations

Methods:

- Filtering the data to remove irrelevant data.
- Researching the map of Newyork city to identify the geographic center.
- Dividing the data to figure out what percentage of the rides started in the north and which in the south.
- Scattering the data on the map of NYC to visualize the data.

Observations:

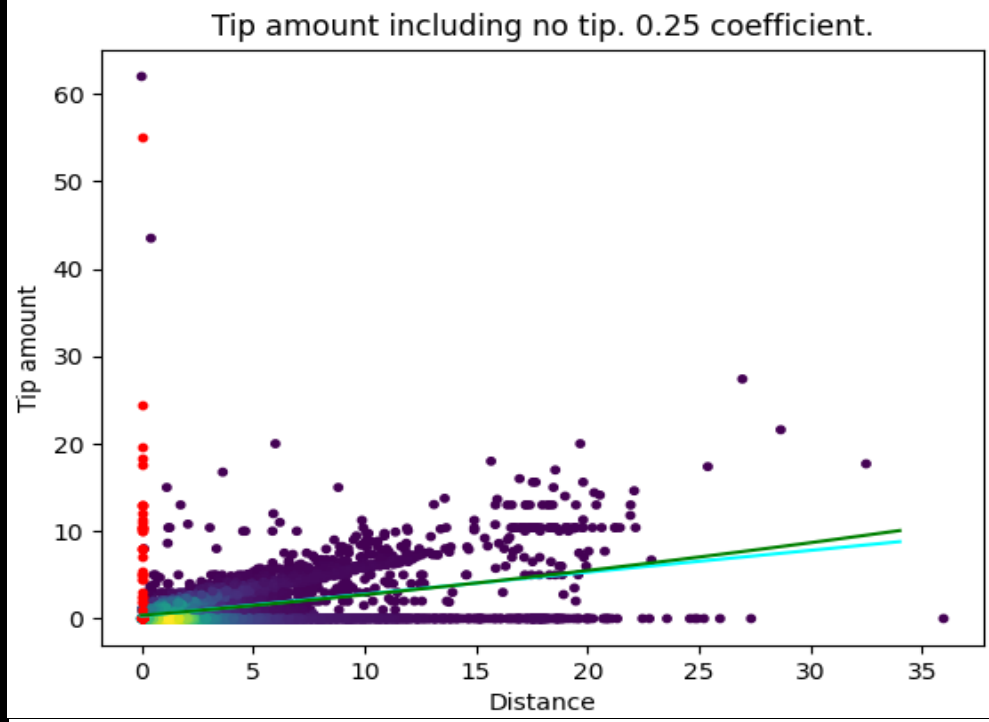
- **Bushwick, Brooklyn** is identified as the center of NYC.
- **95.8%** of rides started in the north.
- **4.2%** of the rides started in the south.

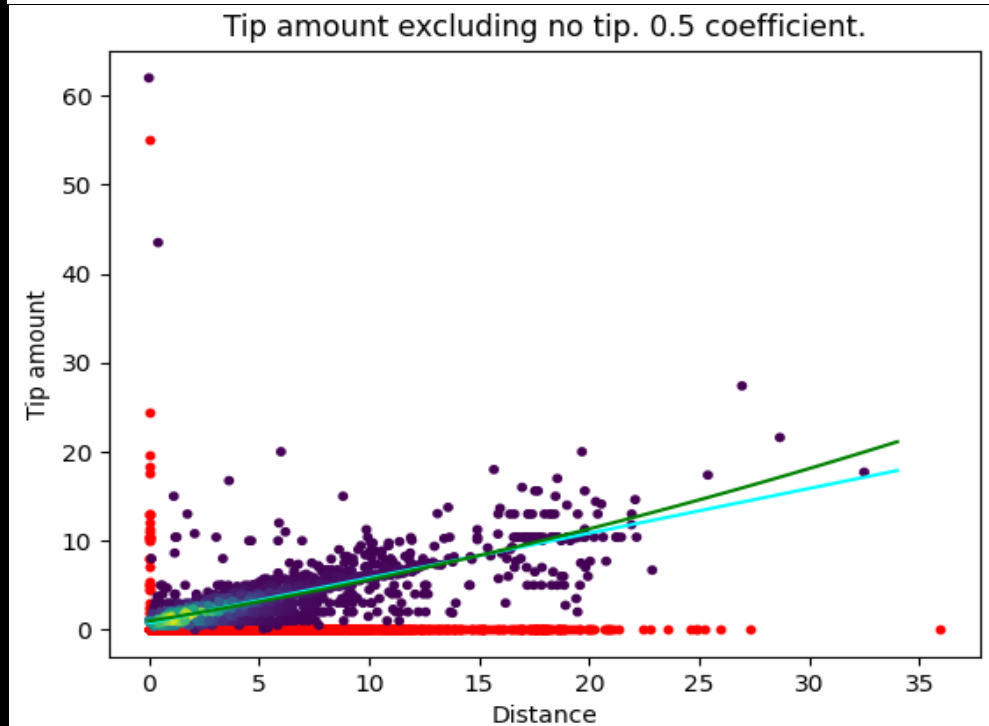


HOCHSCHULE FURTWANGEN UNIVERSITY | HFU

Tips and distance

- Scatterplot
  - Density
    - Kernel-density estimate using Gaussian kernels
  - Correlation lines
    - Linear regression
    - Polynomial fit
  - Exclusions
    - First plot - when distance is 0
    - Second plot – when distance and tip is 0
- Observations
  - Tip amount including $0 tips is $0.25/mile
    - For every 1 mile there is a tip of $0.25
  - Tip amount not including $0 is $0.50/mile
    - For every 1 mile there is a tip of $0.50
  - There is a slight increase with the polynomial towards 30 miles, but this is not very significant
  - The largest density is in the 1–3-mile range
- Code
  - Using Pandas, NumPy, Matplotlib and SciPy

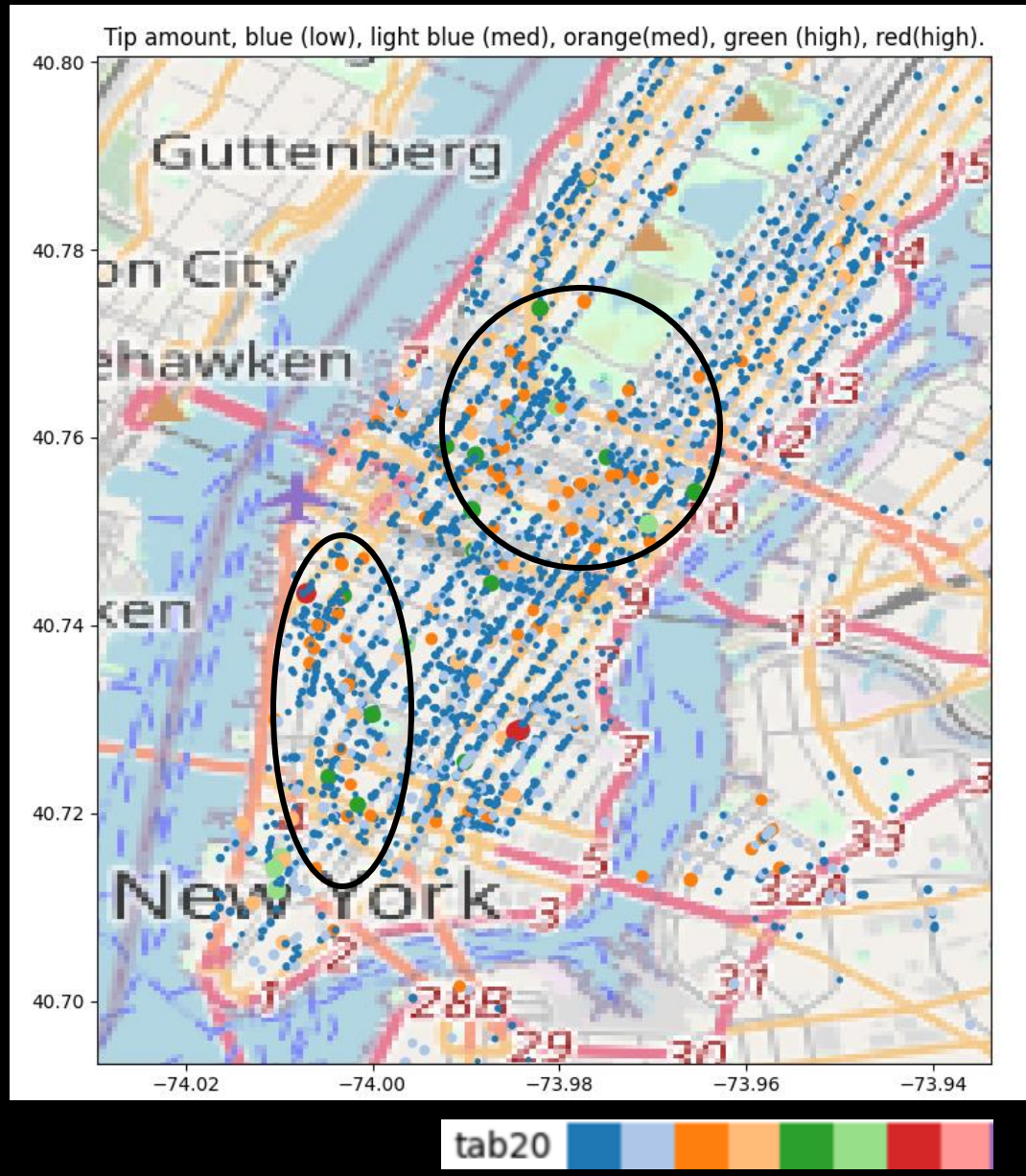




HOCHSCHULE FURTWANGEN UNIVERSITY | HFU

Tip locations

- Scatter plot
  - Larger plots are significantly larger tips.
  - Tab20 colour map to clearly show high tips.
- Observations
  - Midtown East extending into Rockefeller Center, 9th Ave to Lower Manhattan.
  - Lower tip areas Upper East, Lower East side and Upper West Side.
  - Correspond with major roads into NYC
- Data normalisation (min max scaling)
  - Tips started off as $5-$20 value and needed to be ranged between 0 (lowest) and 1 (highest). E.g., 0.5 * 255 = 127.5.
  - However, outliers exist
- Code
  - Using Pandas, NumPy, Matplotlib and Sklearn



HOCHSCHULE FURTWANGEN UNIVERSITY | HFU

Reflection

- What value can you derive from our insights?
  - By decreasing the amount charged in areas where tips are higher, the cost to drivers can be offset by the increased tips and you can charge less than the competition.
  - Focusing on increasing performance in the southern part of NYC.
  - By emphasizing tips, you can increase tip amount up to a maximum of 25%.
  - If restricted higher range vehicle could be available on request or mid trip vehicle swap
  - Longer range trips are very profitable but rarely happen
- Reasoning for methods?
  - Standard data processing: Scatterplots, normalization, linear regression
- What could be improved/ Further research
  - 3D scatterplot of New York for tip amount with K-Means Clustering
  - Standardization
  - AI models
  - New data especially for COVID-19 and recent competition metrics
  - Further tip data processing with a street-by-street analysis

Agenda overview and questions

1. Research
   - Increased competition and COVID-19 led to the need of optimizing the business
2. Range constraints
   - Increasing the range of taxis beyond 30km will not affect the profits of the company however over millions of trips the difference will start to build up, but may it will not be worth the extra investment
3. North or South starting locations.
   - Percentage of the rides starting from the north plotted vs the south 95.2% North vs 4.8% South
4. Tips
   - $0.25/mile including no tips.
   - $0.50/mile only when tips are given.
5. Higher tips
   - Lower Manhattan to 9th Ave, Midtown East extending into Rockefeller Center
6. Reflections.
   - Plans to improve performance based on acquired insights such as implementing AI models and further detailed analysis

Questions?

- Code
  - Available at https://github.com/FinnianHBLR/data-science-ml