CSP 571 Project Proposal & Outline

Analysis of Chicago dataset

Jean-Charles Louis	**Raphaël Cohen**	**Pranav Behari Lal**
jlouis@hawk.iit.edu	rcohen6@hawk.iit.edu	plal2@hawk.iit.edu
A20431556	A20432518	A20417025

1. Project Proposal

1.1. Objective

Our main objective is to analyse the taxi data for the city of Chicago for the time period of 2013 - 2018 (*look at actual dates from the dataset)*. Using the correlation between taxi usage with weather, CTA and crime data, we want understand why are people taking taxis and what can we recommend to increase usage of public transports in the Chicago area.

1.2. Deliverables

- ➔ **Analyse the decline of the taxis usage**
  - How big is the decline?
  - Which area are more affected, link to crime?
  - Can we find a trend in the decline?
- ➔ **Can we correlate trips with external data?**
  - Weather
  - CTA incidents / distance from CTA bus and train
  - Events in Chicago
- ➔ **Making a case for TaxiPool**
  - Can we cluster users with similar pickup + dropoff + time?
- ➔ **Dynamic pricing**

1.3. Methodology

Assumptions

- Not all trips are reported but the City believes that most are.
- We assume that the % of trip recorded didn't change with time.

Data Preparation

Our dataset is ~43Gb with 113 million entries and 23 features. To bring this dataset into a usable format, we already have:

- Removed unnecessary or redundant data (`trip_id`…)
- Stored data more efficiently (remove redundant $ signs, encode time as timestamps…). Making these changes has reduced the size of our dataset to ~18 GB.
- Split the data into multiple smaller files. We will look at partitioning data based on the trip date.
- Add weather data to each trip: `was_raining` and `temperature` (very high, high, medium…)
- Some location data is missing for privacy reasons, we will need to handle this appropriately.

2. Project Outline

2.1. Related work

Resources using the same main dataset

- 2016 Chicago Cabs Analysis, *Yiming Wu* (2016)
  https://nycdatascience.com/blog/r/2016-chicago-cabs-analysis
- Chicago's Public Taxi Data, *Todd Schneider* (2017)
  http://toddwschneider.com/posts/chicago-taxi-data

Other resources

- Chicago: A Uber Case Study, *Uber* (2015)
  https://uber-static.s3.amazonaws.com/web-fresh/legal/Uber_Chicago_CaseStudy.pdf
- Case Study: New York City Taxis, *Regulatory Reform Team - Ash Center at Harvard Kennedy School* (2014)
  https://datasmart.ash.harvard.edu/news/article/case-study-new-york-city-taxis-596
- Research on Optimization of Vehicle Routing Problem for Ride-sharing Taxi, *Y Lin, W Li, F Qiu, H Xu* (2012)
  https://www.sciencedirect.com/science/article/pii/S1877042812010038

2.2. All data sources

Main dataset

- The main dataset is “Taxi Trips” from City of Chicago Data Portal.
  https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew/

- ❏ `unique_key` : Unique identifier for the trip
- ❏ `taxi_id` : A unique identifier for the taxi
- ❏ `trip_start_timestamp` : When the trip started, rounded to the nearest 15 minutes
- ❏ `trip_end_timestamp` : When the trip ended, rounded to the nearest 15 minutes
- ❏ `trip_seconds` : Time of the trip in seconds
- ❏ `trip_miles` : Distance of the trip in miles
- ❏ `pickup_census_tract` : The Census Tract where the trip began
- ❏ `dropoff_census_tractT` : The Census Tract where the trip ended
- ❏ `pickup_community_area` The Community Area where the trip began
- ❏ `dropoff_community_area` The Community Area where the trip ended
- ❏ `fare`, `tips`, `tolls`, `extras`
- ❏ `trip_total` : Total cost of the trip, the total of the fare, tips, tolls, and extras
- ❏ `payment_type` Type of payment for the trip
- ❏ `company` The taxi company
- ❏ `pickup_latitude`, `pickup_longitude`, `pickup_location`
- ❏ `dropoff_latitude`, `dropoff_longitude`, `dropoff_location`

Other datasets

- Chicago weather
- CTA data
- Crime dataset (crime areas)

2.3. Tools

- We plan to use the R language for this project.
- We also plan to use the *maps* library from CRAN and other mapping tools.
- For the TaxiPool part, we plan to use a clustering algorithm.
- We’re looking into using PostgreSQL because of the size of the dataset and to manage localisation data more easily.
- We’re using GitHub to share code.