Differences between Taxi and Uber

Realtime & Big Data Analytics Final Project

Speaker: CHEN-YUAN HO

Big Data Analytics Symposium - Summer 2017

Analytics Project:
Differences between Taxi and Uber

Team:
Chen-Yuan Ho

Abstract:
- ✓ Goal: know differences between Uber and Taxi
- ✓ Tool: NYU HPC Dumbo, Hive
- ✓ Analytics: weather/month/hour vs. rides

Motivation

Who are the users of this analytic?

✓ Taxi

✓ Uber

Who will benefit from this analytic?

✓ Taxi

✓ Uber

✓ Startup companies who are interested in sharing economy

Why is this analytic important?

✓ Uber has posing great impact on shared transportation.

✓ Locate the competitive edges of both Uber and taxi, by identifying intrinsic factors of them.
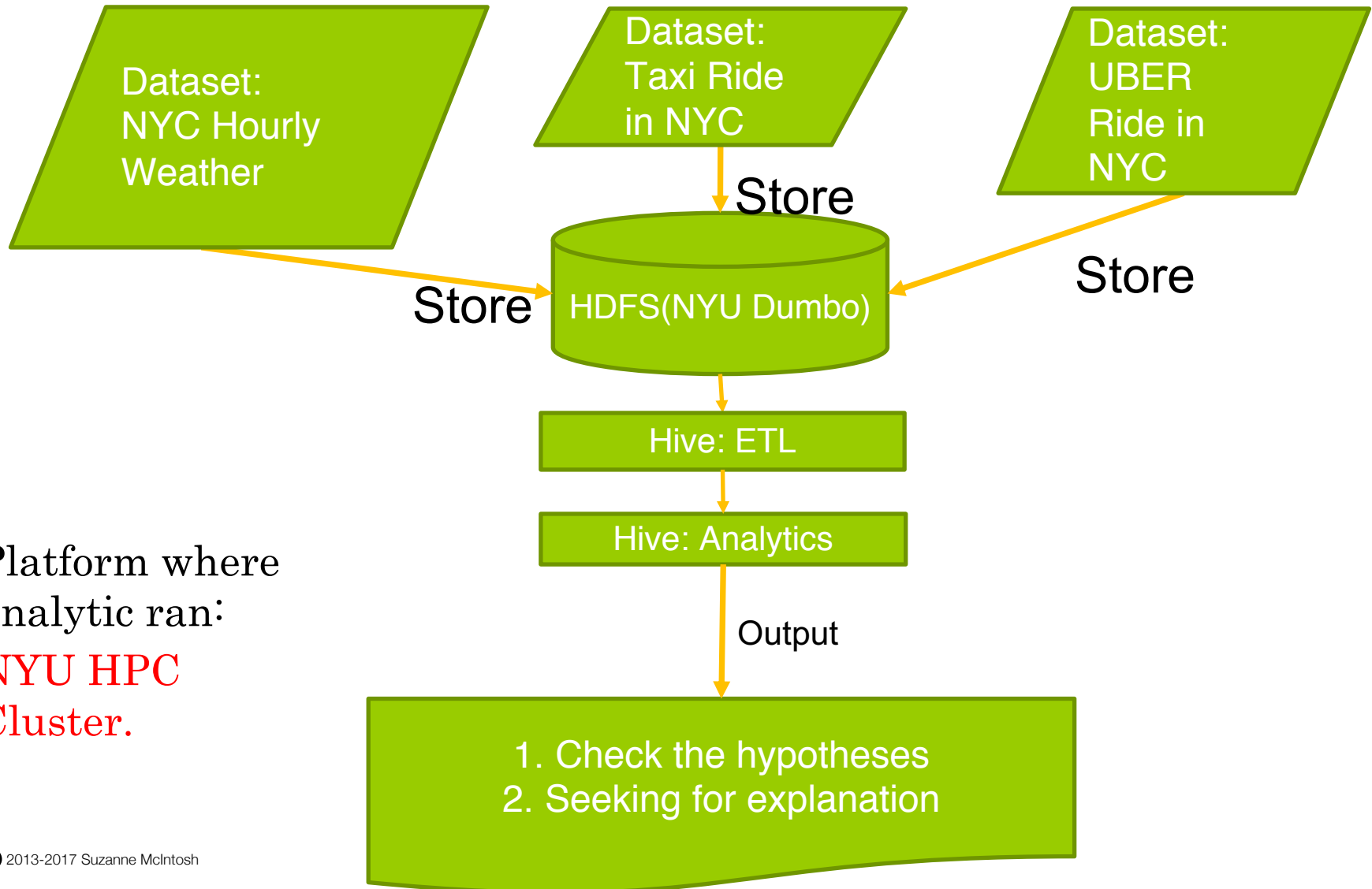
Goodness

What steps were taken to assess the 'goodness' of the analytic?

1. Understood the shared transportation industry before collecting the datasets.

2. Source datasets are from official website.

3. Check the analytic results with relative researches.

Hypotheses

◆Both occurred the most in rainy days.

◆Both occurred the most in hot and cold temperature.

◆Both occurred the most in rush hours and late night.

◆No significant differences between the distribution of both Uber and taxi rides on monthly basis.

Design Diagram

Dataset:
NYC Hourly
Weather

Dataset:
Taxi Ride
in NYC

Dataset:
UBER
Ride in
NYC

Store

Store

Store

HDFS(NYU Dumbo)

Hive: ETL

Hive: Analytics

Output

Platform where
analytic ran:

NYU HPC
Cluster.

1. Check the hypotheses
2. Seeking for explanation

Data Sources I

Name:
NYC TLC Trip Data 2014 – Yellow Taxi

Description:
Collected from NYC Open Data. Including pickup & drop-up date-time, and fare, distance, location of pickup & drop-up from April to September in 2014

Size of data:
12.9 GB

Data Sources II

Name:
Uber Pickups in NYC 2014, from FiveThirtyEight

Description:
Collected from Fivethirtyeight.com. The dataset includes pickup date-time and location from April to September in 2014.

Size of data: 192 MB

Data Sources III

Name:
Historical Weather Dataset in NYC of 2014

Description:
Parsed from UndergroundWeather.com. Including temperature and weather condition on hourly basis for everyday from April to September, 2014.

Size of data:
176KB
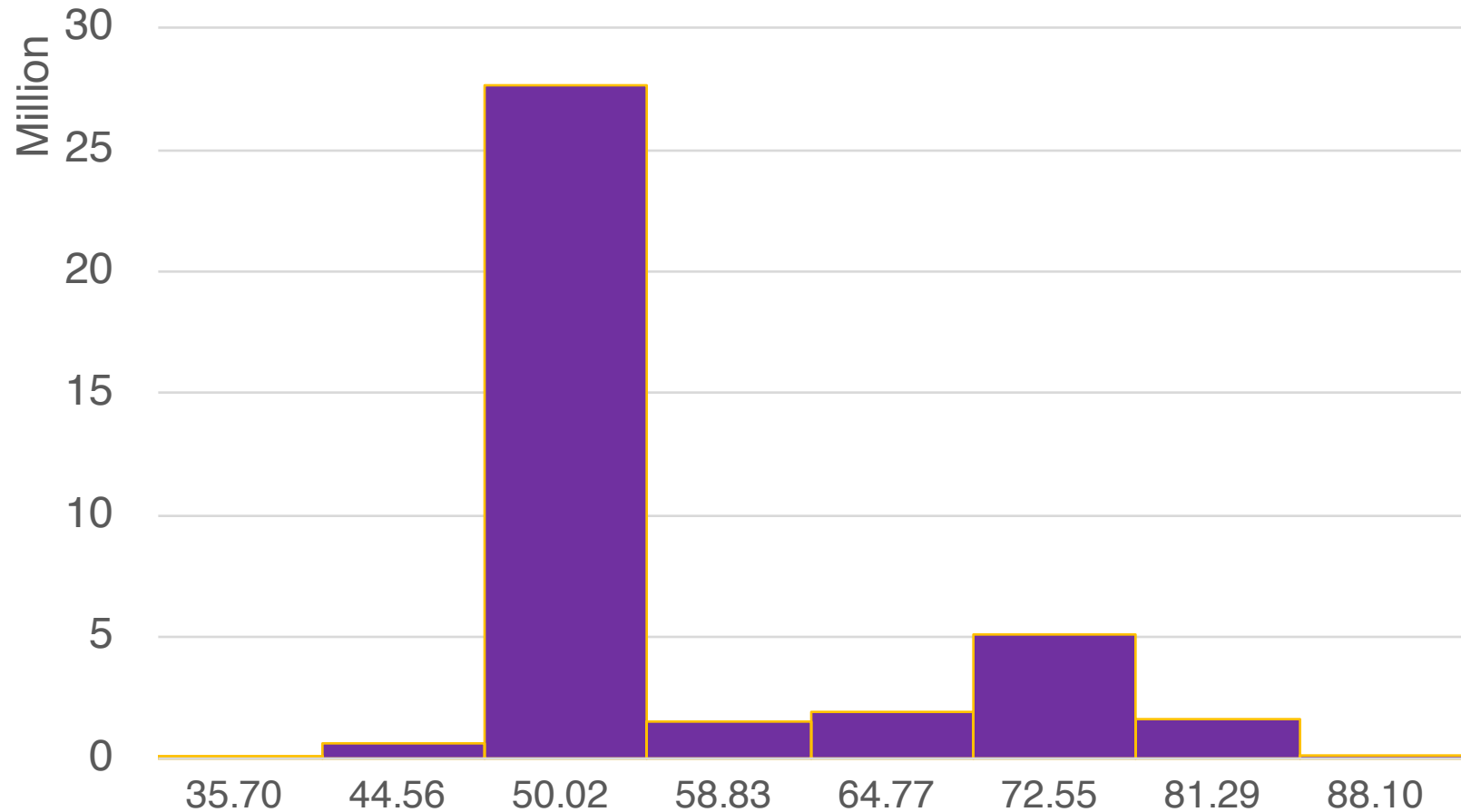
Analysis: Check the Hypothesis I

UBER vs Weather

weather_condition	count
Heavy Rain	18453802
Rain	9438232
Clear	4492004
Overcast	2481666
Mostly Cloudy	1232854
Light Rain	947512
Partly Cloudy	679642
Scattered Clouds	591626
Haze	238254
Unknown	105230
Light Snow	9952
Fog	5222

TAXI vs Weather

weather_condition	count
Clear	86077294
Overcast	43968102
Mostly Cloudy	21161340
Light Rain	14853332
Partly Cloudy	11629956
Scattered Clouds	10434858
Rain	5491188
Heavy Rain	4316672
Haze	4059788
Unknown	1719996
Light Snow	434262
Fog	102052

UBER vs. Temperature

TAXI vs. Temperature

Analysis: Check the Hypothesis III
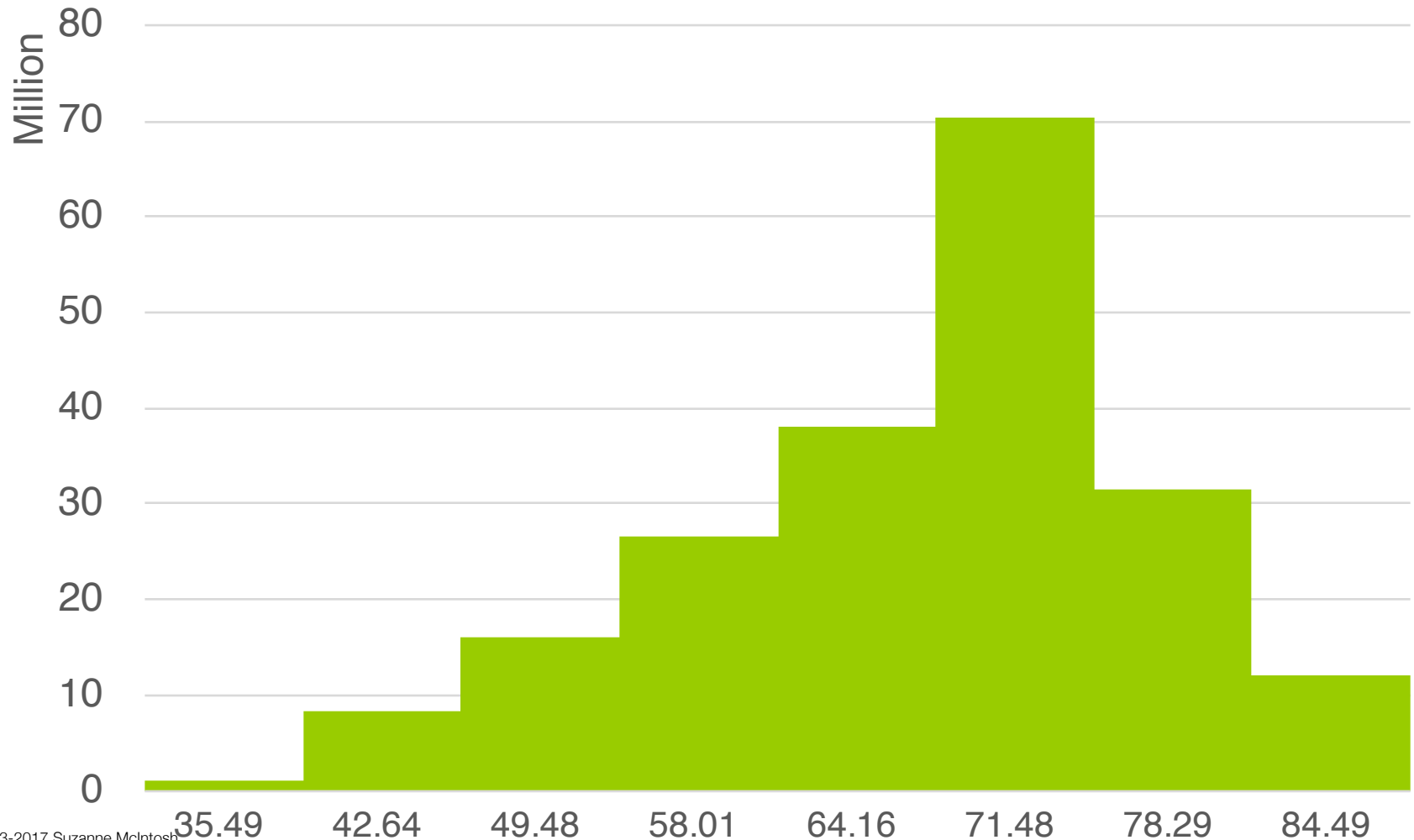
HOUR vs. Rides

Analysis: Check the Hypothesis IV

Rides vs. Month

mean	std_dev
13729181.17	835,031.291

mean	std_dev
1295522	1,707,556.186

UBER — TAXI
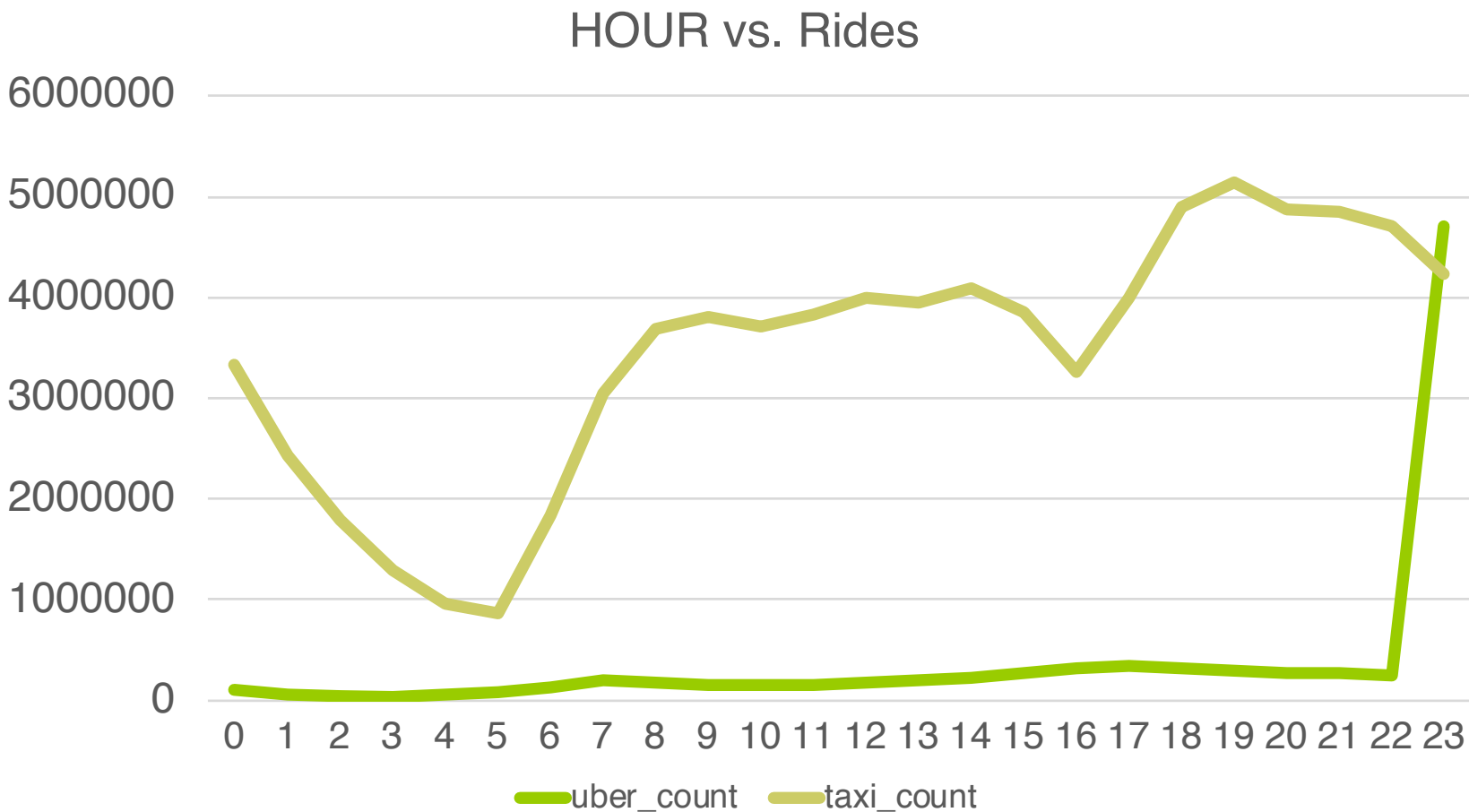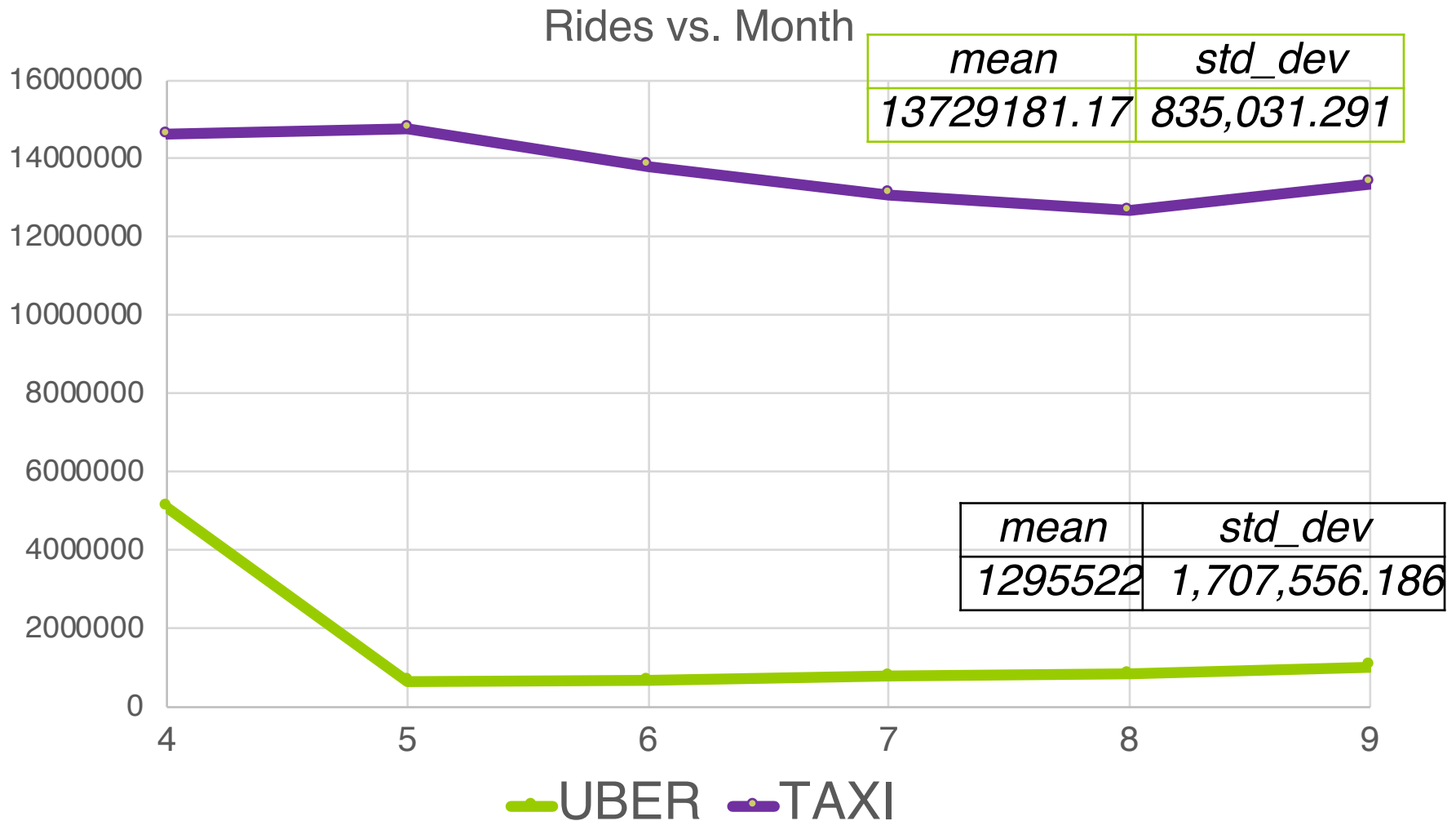
Difference between Taxi and Uber

Obstacles

1. \<Obstacle1\> Limitation of the Scope
    ➢ The content of Uber Dataset limits the scope of the topic.

2. \<Obstacle 2\> Factor Identifying
    ➢ Cannot collect all possible factors in the beginning
    ➢ Lack of thorough domain knowledge

Summary

1. Most of hypotheses in the beginning are challenged.
2. There are some intrinsic factors behind the pattern.
3. Identified patterns can be a reference of strategy modification for both Uber and taxi companies.
4. Unknown factors behind the pattern: business opportunities.

Acknowledgement

◆ Suzanne McIntosh, Courant Institute of Mathematical Sciences, New York University.

◆ Santhosh Konda, High Performance Computing Center, New York University.

Difference between Taxi and Uber

References

[1] https://data.cityofnewyork.us/view/gn7m-em8n

[2] https://github.com/fivethirtyeight/uber-tlc-foil-response/tree/master/uber-trip-dataList the work you referenced

[3] https://www.wunderground.com/history/airport/KJFK/2014/1/2/DailyHistory.html

[4] Hian Chye Koh, Gerald Tan. Data Mining Applications in Healthcare. Nanyang Business School, Nanyang Technological University, Singapore. April, 2005.

[5] Chris Rygielski, Jyun-Cheng Wang, David C. Yen. Data Mining Techniques for Customer Relationship Management. Department of DSC & MIS, Miami University, Oxford, OH, USA. Department of Information Management, National Chung-Cheng University, Taiwan. November, 2002>

Thank you!