Taxi Hotspot Prediction

Lars Fikkers (108012050)
Pascal Roose (108012051)

3 January 2020
Big Data Mining and Applications (Term Project)

Table of contents

- Project approach
- Environment setup
- Data Analysis
- Model preparation
- Results
- Conclusions

Project approach

- Responsibility
  - Pascal: Data analysis and preparation
  - Lars: Predicting through AI
- Planning
  - Defining requirements
  - Analysing data
  - Discussing approach
  - Execute plan
  - Prepare presentation

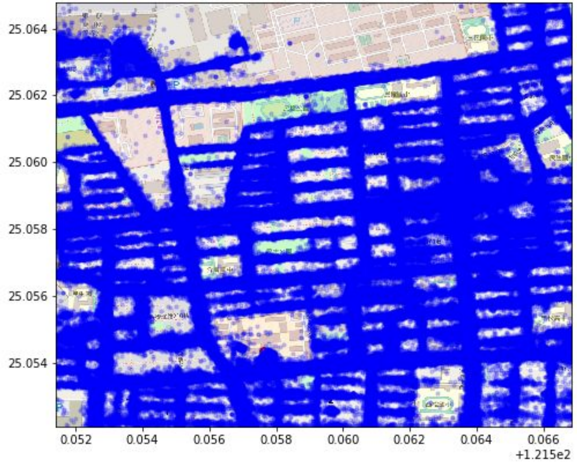
Environment Setup

- Hosted on DigitalOcean
  - Upgrade the system specifications in minutes
  - Anywhere between 4 - 64 GB RAM
  - Anywhere between 2 - 32 vCPUs used
- Ubuntu 18.04 LTS
- Spark with Python (PySpark)
- Jupyter Notebook for quick analysis

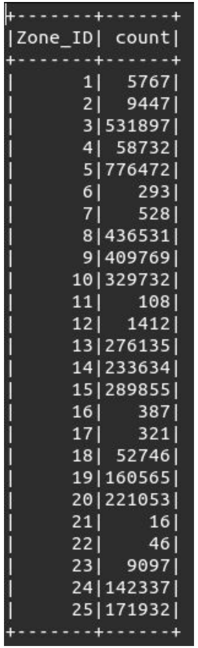
Data analysis - Location

Do we see any obvious taxi hotspots? No

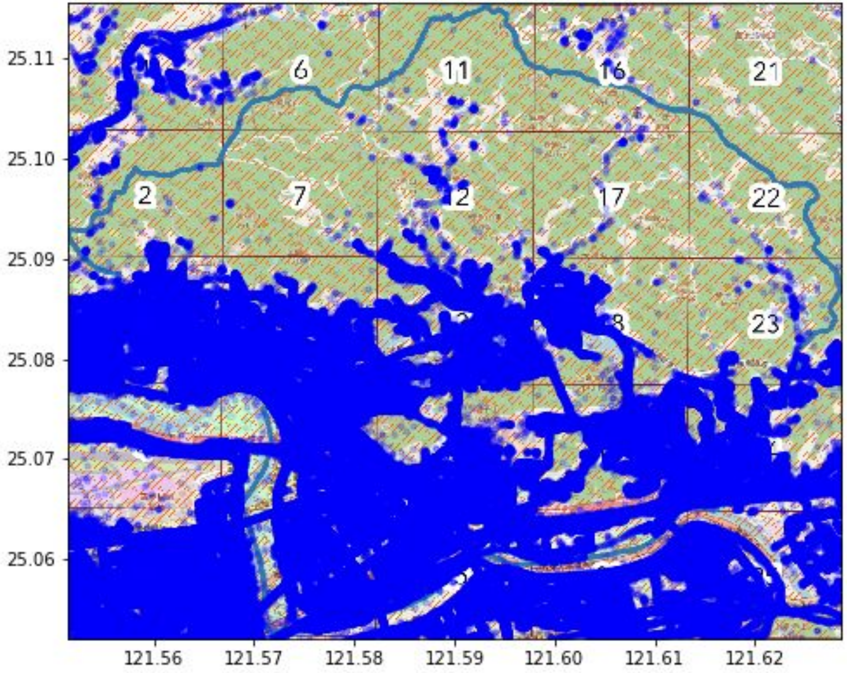
Does it matter? No



Plotting Taxi GPS pickup points - Zone 5



```
+-------+------+
|Zone_ID| count|
+-------+------+
|      1|  5767|
|      2|  9447|
|      3|531897|
|      4| 58732|
|      5|776472|
|      6|   293|
|      7|   528|
|      8|436531|
|      9|409769|
|     10|329732|
|     11|   108|
|     12|  1412|
|     13|276135|
|     14|233634|
|     15|289855|
|     16|   387|
|     17|   321|
|     18| 52746|
|     19|160565|
|     20|221053|
|     21|    16|
|     22|    46|
|     23|  9097|
|     24|142337|
|     25|171932|
+-------+------+
```

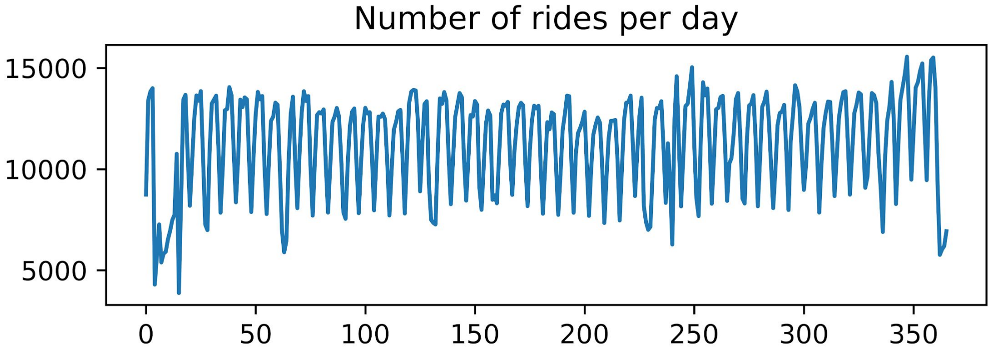


Plotting Taxi GPS pickup points

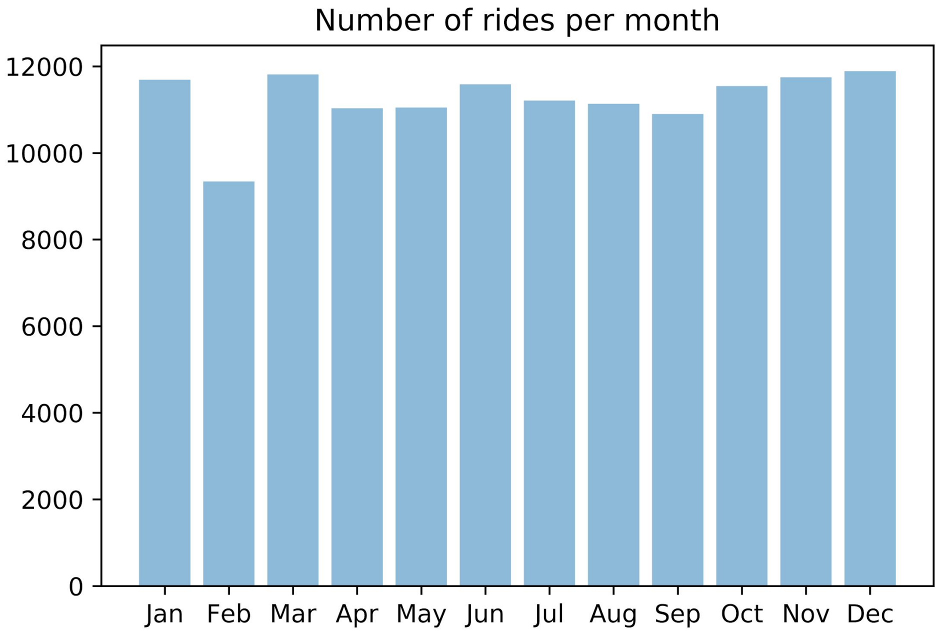
Data analysis - Per day of the year

What do we see?

- No notable growth or decline
- Wave pattern, weeks?
- Big drop in the beginning



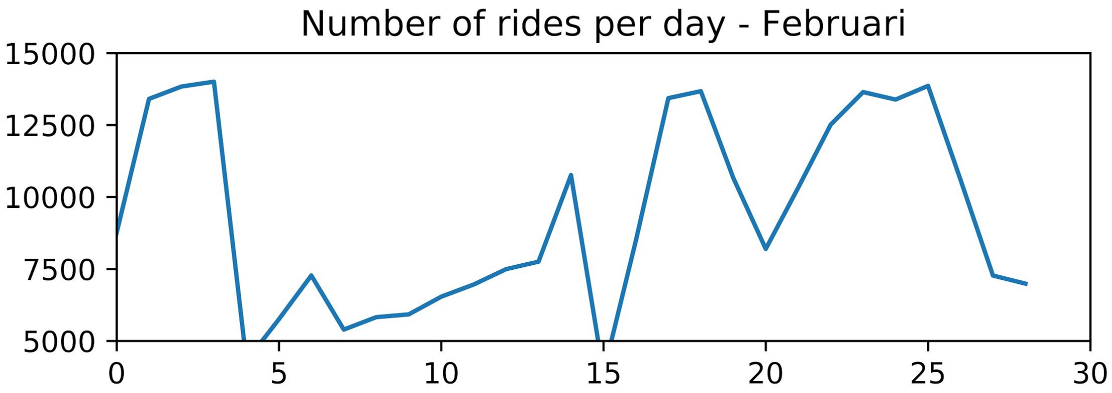
Data analysis - Month of the year

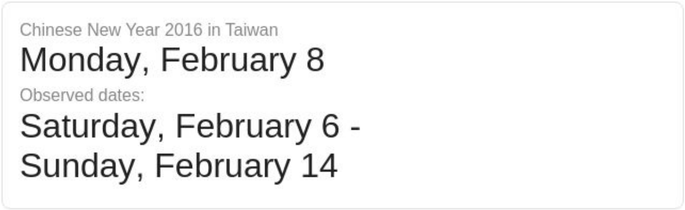


Data analysis - February

Big drop between 4th and 15th of february 2016

This is around Chinese New Year

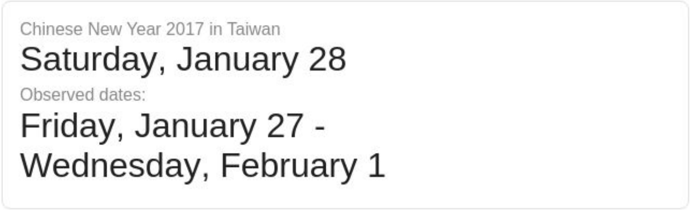



Chinese New Year 2016 in Taiwan

Monday, February 8

Observed dates:

Saturday, February 6 -
Sunday, February 14

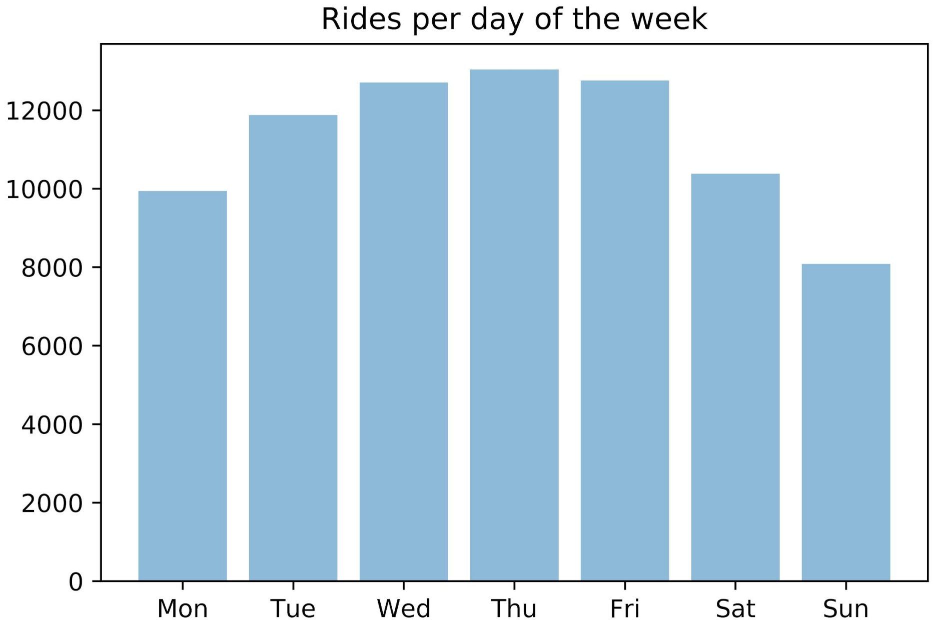

Chinese New Year 2017 in Taiwan

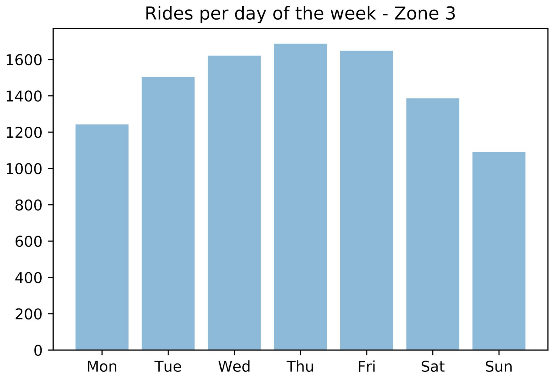
Saturday, January 28

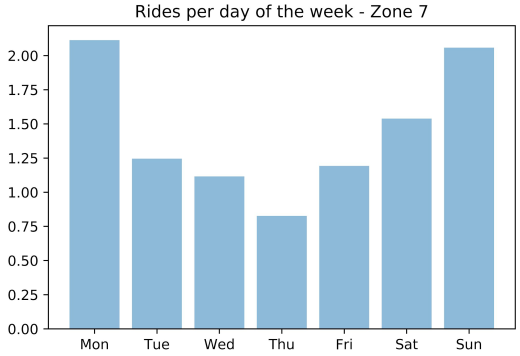
Observed dates:

Friday, January 27 -
Wednesday, February 1

Data analysis - Day of the week

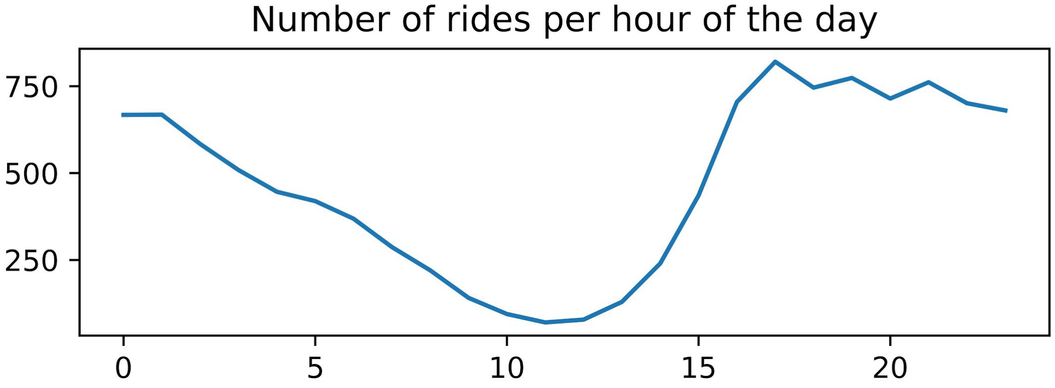


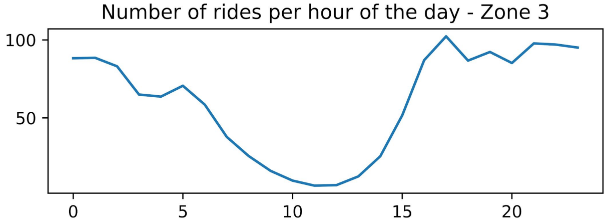


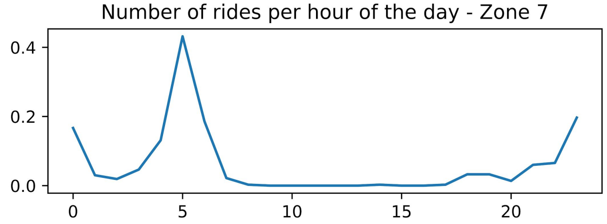


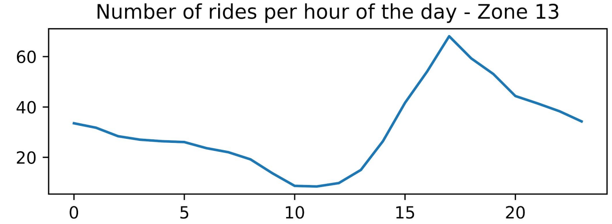
Data analysis - Hour of the day

- Not a less rides during work hours
- Pretty stable after 17h (5PM)
- Slow decline at night till the early in the morning

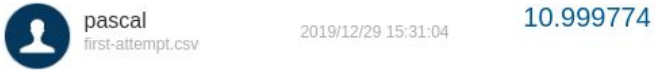




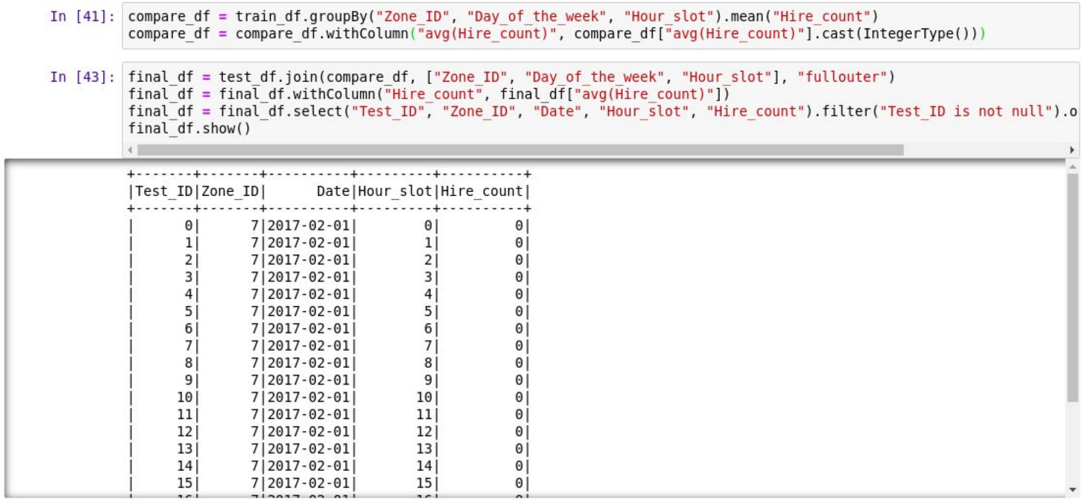




Data analysis - Combined Test Upload



pascal
first-attempt.csv
2019/12/29 15:31:04
10.999774



```
In [41]: compare_df = train_df.groupBy("Zone_ID", "Day_of_the_week", "Hour_slot").mean("Hire_count")
         compare_df = compare_df.withColumn("avg(Hire_count)", compare_df["avg(Hire_count)"].cast(IntegerType()))
```

```
In [43]: final_df = test_df.join(compare_df, ["Zone_ID", "Day_of_the_week", "Hour_slot"], "fullouter")
         final_df = final_df.withColumn("Hire_count", final_df["avg(Hire_count)"])
         final_df = final_df.select("Test_ID", "Zone_ID", "Date", "Hour_slot", "Hire_count").filter("Test_ID is not null").o
         final_df.show()
```

```
+-------+-------+----------+---------+----------+
|Test_ID|Zone_ID|      Date|Hour_slot|Hire_count|
+-------+-------+----------+---------+----------+
|      0|      7|2017-02-01|        0|         0|
|      1|      7|2017-02-01|        1|         0|
|      2|      7|2017-02-01|        2|         0|
|      3|      7|2017-02-01|        3|         0|
|      4|      7|2017-02-01|        4|         0|
|      5|      7|2017-02-01|        5|         0|
|      6|      7|2017-02-01|        6|         0|
|      7|      7|2017-02-01|        7|         0|
|      8|      7|2017-02-01|        8|         0|
|      9|      7|2017-02-01|        9|         0|
|     10|      7|2017-02-01|       10|         0|
|     11|      7|2017-02-01|       11|         0|
|     12|      7|2017-02-01|       12|         0|
|     13|      7|2017-02-01|       13|         0|
|     14|      7|2017-02-01|       14|         0|
|     15|      7|2017-02-01|       15|         0|
```

Model preparation

- 2 types considered
  - SVM
  - Decision tree
- Decision tree
- Classifier
- Regressor
- Random forest

Results

No. of trees	Max depth	RMSE
101	5	21.925095
75	20	14.316507
200	12	14.597748
200	7	17.763558
100	15	14.421029
100	10	14.958575

Conclusions

- AI
  - Depth improves prediction
  - Number of trees around 100
- Possible improvements
  - More features to classify
  - Other AI model