Car Park Availability Analysis
&
Predictive Modelling

Presented by: Group - DS - 07
Chow ,Michael
Lee Yat Shun, Jasper
Mohammad Shehzaad
Hindy Yuen Shing Yan

Workflow

1. Problem Introduction
2. Datasets
3. Exploratory Data Analysis
4. Data Preprocessing
    a. Data Cleaning
    b. Dimensionality Reduction
    c. Feature Engineering
5. Machine Learning Models
6. Model Evaluation
    a. Evaluation Metrics
7. Future Development
    a. Enhancing Predictions Accuracy
    b. Practical Usages
    c. Hyperparameters Tuning

Problem Background

Objective: Car Park Availability Analysis and Predictive Modelling across all the HK Districts

Problem: Regression (Supervised Learning)

User input:

1. Full date
2. Time (Rounded off to the nearest 15 min interval)
3. District

Output (Prediction):

1. List of car park with the corresponding vacancy prediction for enquired district and time

Dataset Information

Dataset: Parking vacancy data (01/05/2021 - 31/05/2021)

Source: https://data.gov.hk/en-data/dataset/hk-td-tis_5-real-time-parking-vacancy-data

Number of Carparks that fulfill criteria: 148

Number of Districts: 18

Extraction Methodology: JSON, requests, Pandas libraries

Dataset Information (External Data)

Dataset: population2020

Source:

http://www.censtatd.gov.hk/en/web_table.html

Dataset Dimension: (18, 8)

Extraction Methodology:

Customized and downloaded directly

	STAT_VAR	STAT_PRES	CCYY	DC	Sex	Age	OBS_VALUE	SD_VALUE
0	PP	Raw_per_n	2020	A	NaN	NaN	236000	NaN
1	PP	Raw_per_n	2020	B	NaN	NaN	173300	NaN
2	PP	Raw_per_n	2020	C	NaN	NaN	537900	NaN
3	PP	Raw_per_n	2020	D	NaN	NaN	260800	NaN
4	PP	Raw_per_n	2020	E	NaN	NaN	323000	NaN

Dataset Information (External Data)

Dataset: population_sex

Source:

http://www.censtatd.gov.hk/en/web_table.html

Dataset Dimension: (252, 8)

Extraction Methodology:

Customized and downloaded directly

	STAT_VAR	STAT_PRES	CCYY	DC	Sex	Age	OBS_VALUE	SD_VALUE
0	PP	Raw_per_n	2020	A	M	0 - 14	11500	NaN
1	PP	Raw_per_n	2020	A	M	15 - 24	10200	NaN
2	PP	Raw_per_n	2020	A	M	25 - 34	15100	NaN
3	PP	Raw_per_n	2020	A	M	35 - 44	14000	NaN
4	PP	Raw_per_n	2020	A	M	45 - 54	14000	NaN

Dataset Information (External Data)

Dataset: area

Source :

http://www.censtatd.gov.hk/en/web_table.html

Dataset Dimension: (18, 2)

Extraction Methodology:

Customized and downloaded directly

	District	Area (km2)
0	Central and Western	12.44
1	Eastern	18.56
2	Southern	38.85
3	Wan Chai	9.83
4	Sham Shui Po	9.35

Dataset Information (External Data)

Dataset: district_borders (visualization only)

Source: https://www.had.gov.hk/psi/hong-kong-administrative-boundaries/hksar_18_district_boundary.json

Dataset Dimension: (18, 5)

Extraction Methodology:

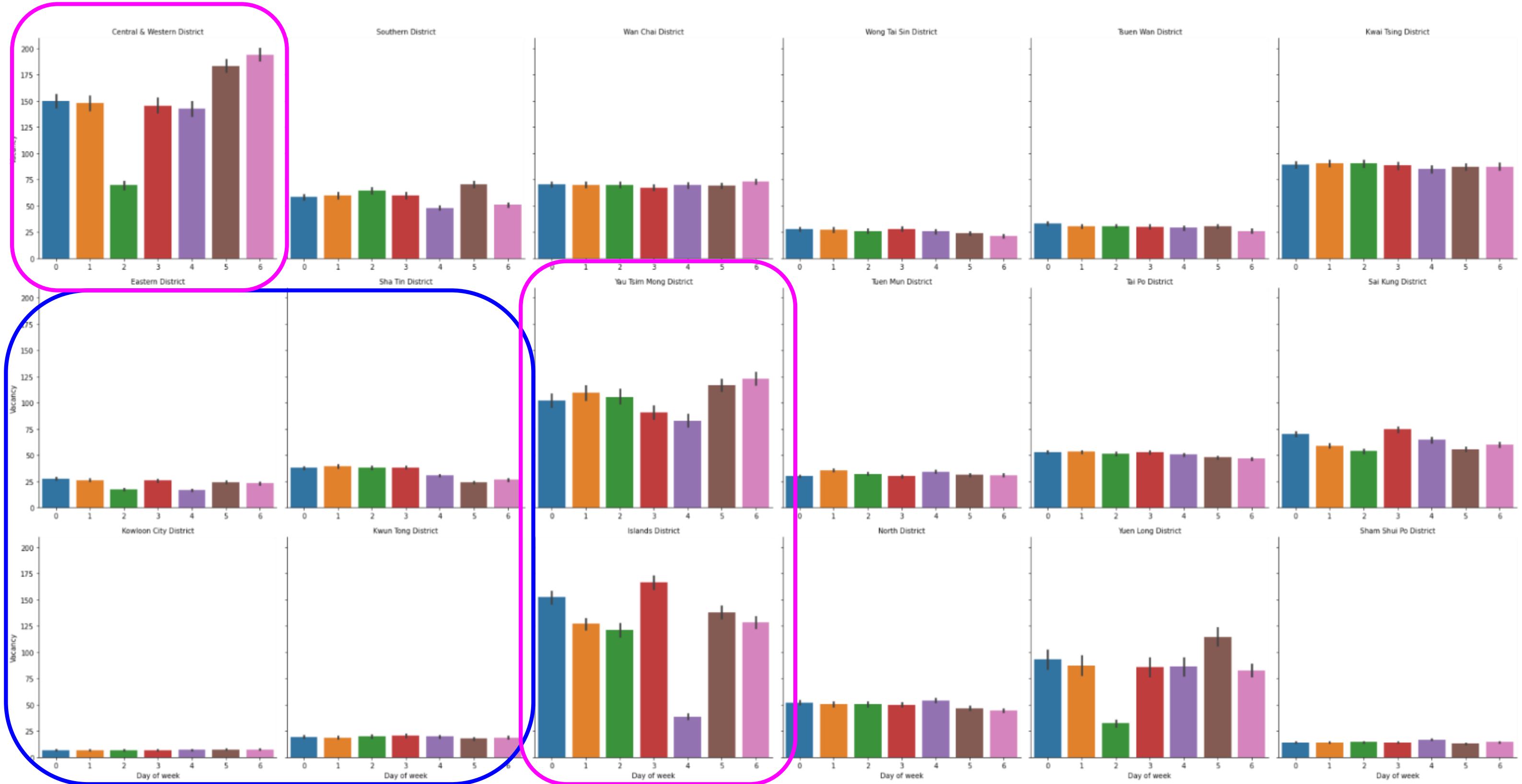
JSON, requests,

Pandas libraries,

geopandas

	地區號碼	District	地區	Administrative District Boundary of Hong Kong	geometry
0	A	Central & Western	中西區	Administrative District Boundary of Hong Kong	POLYGON ((114.14951 22.29748, 114.10987 22.297...
1	B	Wan Chai	灣仔	Administrative District Boundary of Hong Kong	POLYGON ((114.19771 22.25790, 114.19778 22.258...
2	C	Eastern	東區	Administrative District Boundary of Hong Kong	POLYGON ((114.21939 22.29568, 114.21087 22.300...
3	D	Southern	南區	Administrative District Boundary of Hong Kong	POLYGON ((114.16338 22.25882, 114.16305 22.258...
4	E	Yau Tsim Mong	油尖旺	Administrative District Boundary of Hong Kong	POLYGON ((114.17448 22.32665, 114.17417 22.326...

Exploratory Data Analysis



Figure

mean vacancies - day per 18 districts.

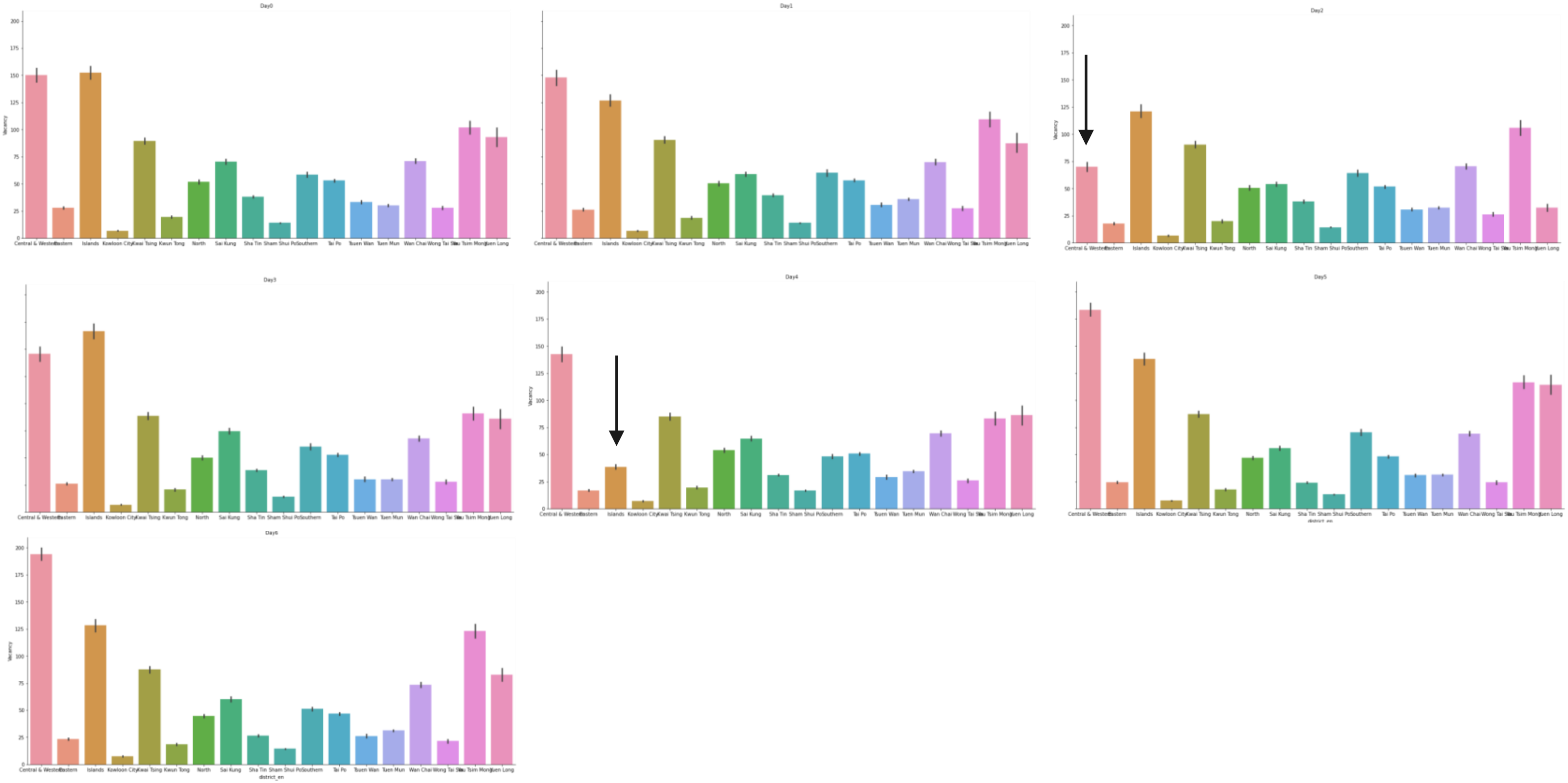
Findings:

Some district have significantly less vacancies which is likely due to the number of parking spots in those districts

On graph

Pink: High Vacancy
Blue: Low Vacancy

Exploratory Data Analysis



Figure

mean vacancies - district per day

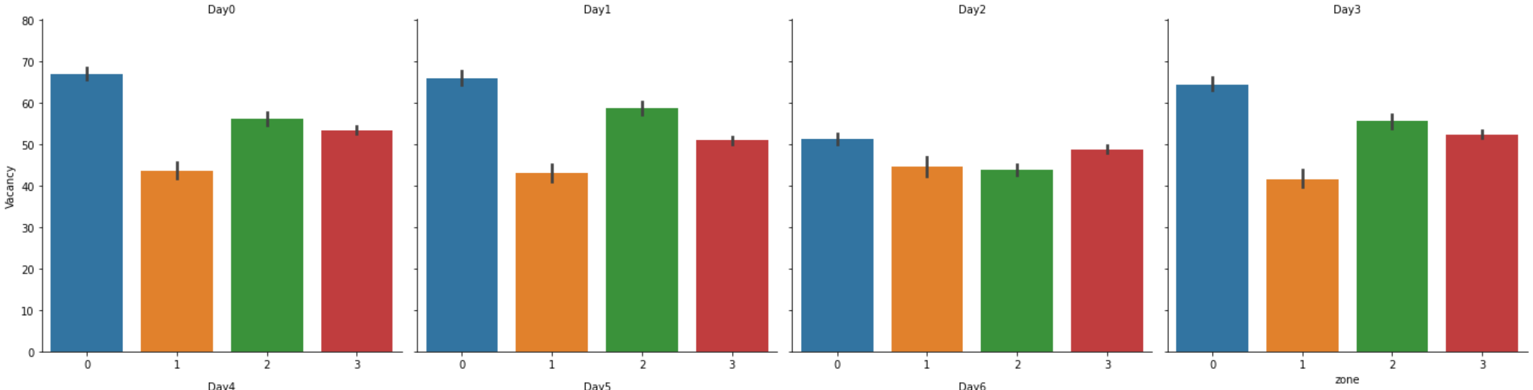
Findings:

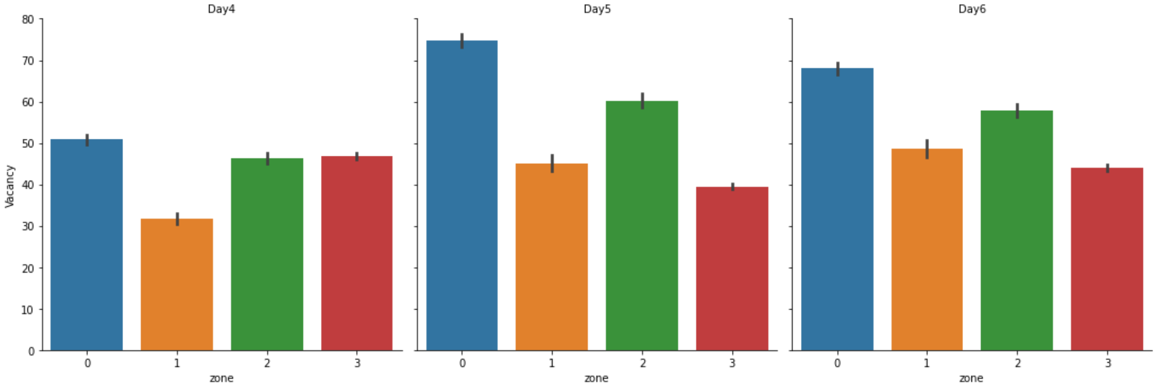
Parking vacancies does not vary with day for almost all district

On graph

Arrow:
Sudden drop in vacancy

Exploratory Data Analysis





Zone 0 (HKI)	Central & Western, Southern, Wan Chai, Eastern
Zone 1 (KLN)	Sham Shui Po, Wong Tai Sin, Kowloon City, Kwun Tong, Yau Tsim Mong
Zone 2 (NTW)	Islands,Tsuen Wan, Kwai Tsing, Yuen Long, Tuen Mun
Zone 3 (NTE)	Tai Po, Sai Kung, North, Sha Tin

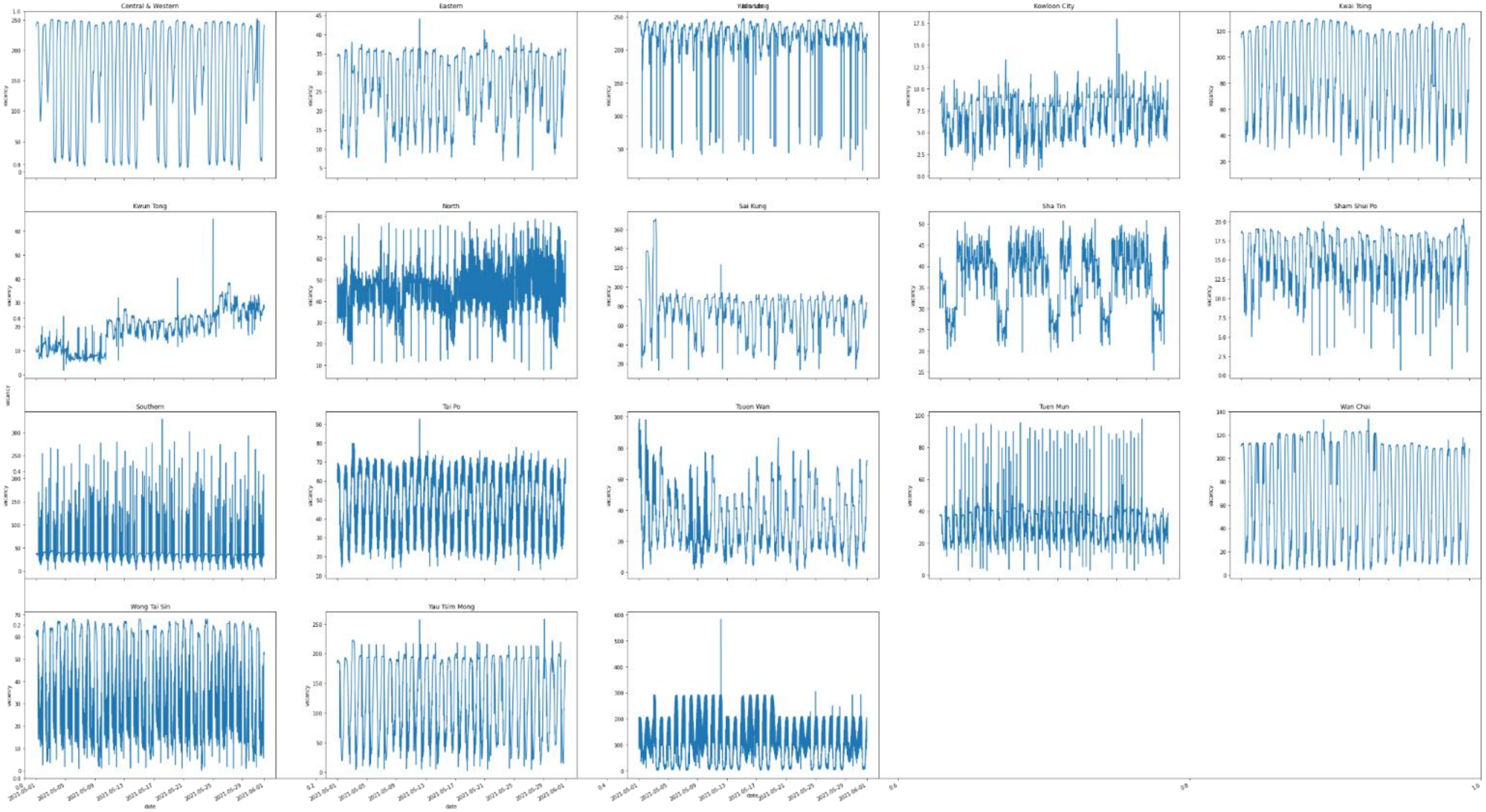
<u>Figure</u>

mean vacancies - zone per day

<u>Findings:</u>

Parking vacancies does not vary much with day

Exploratory Data Analysis



Figure

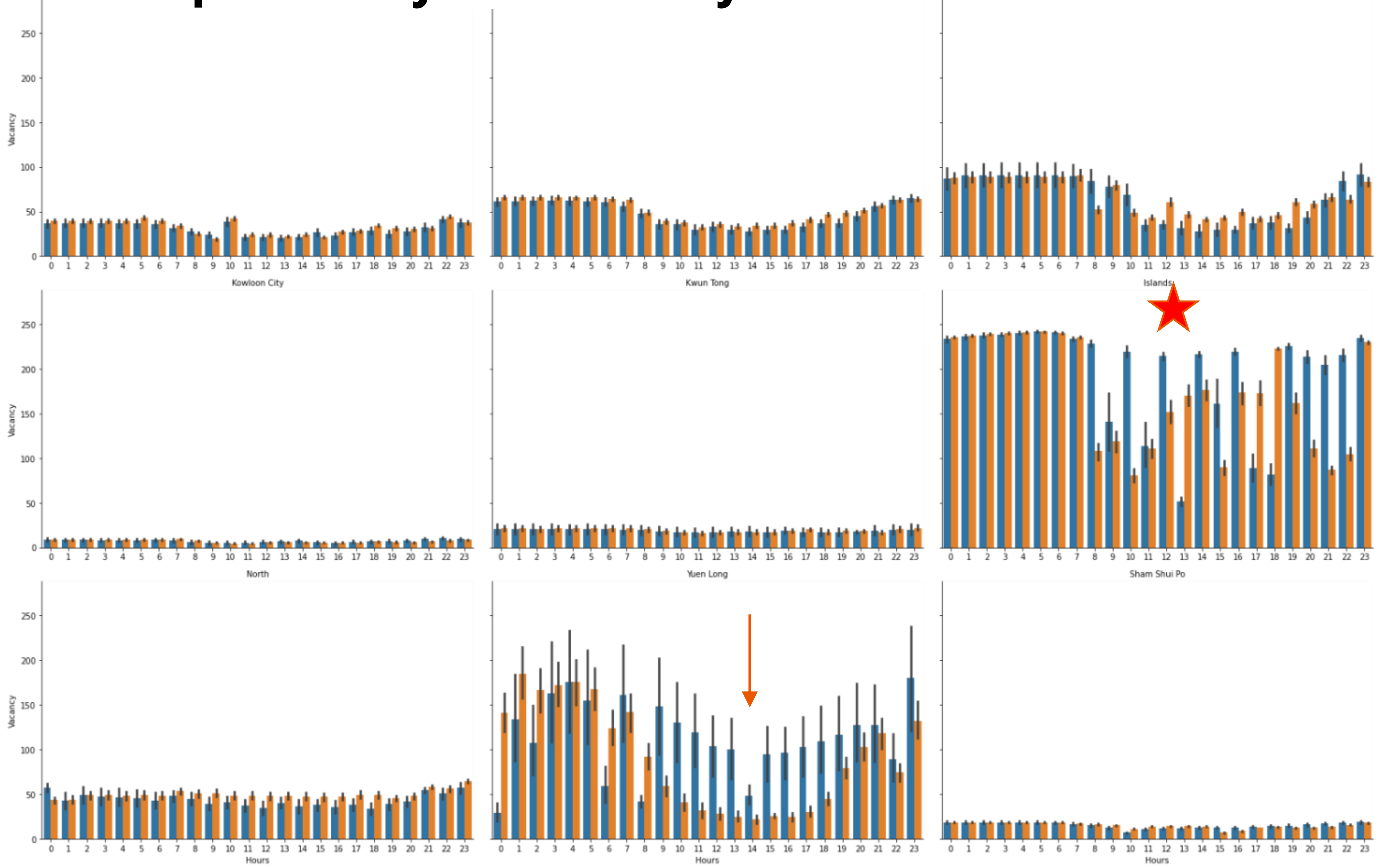
mean vacancies - date per district

Findings:

Parking vacancies mostly vary in a periodic but stable manner

This suggests a small correlation between vacancy and date

Exploratory Data Analysis



Figure

mean vacancies - hours per district

Findings:

Parking vacancies drop usually during the day, lowest around noon.

Parking vacancies doesn't vary much for districts that have a low average vacancy

Irregular Pattern in parking vacancies for Southern, Islands

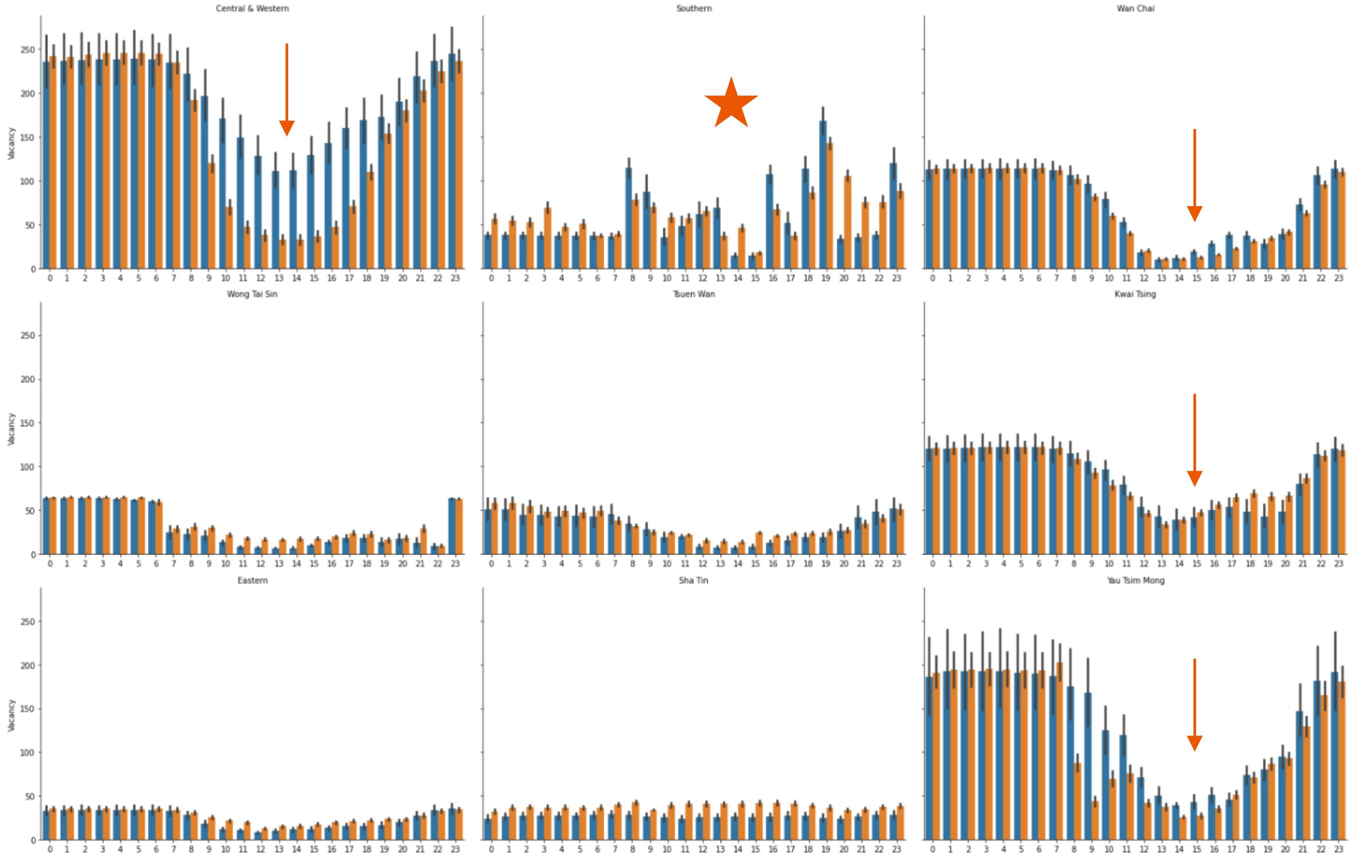
On Graph:

Arrow:
Significant drop in vacancy

Star:
Irregular Pattern

Weekdays
Weekends

Exploratory Data Analysis



Figure

mean vacancies - hours per district

Findings:

Parking vacancies drop usually during the day, lowest around noon.

Parking vacancies doesn't vary much for districts that have a low average vacancy

Irregular Pattern in parking vacancies for Southern, Islands

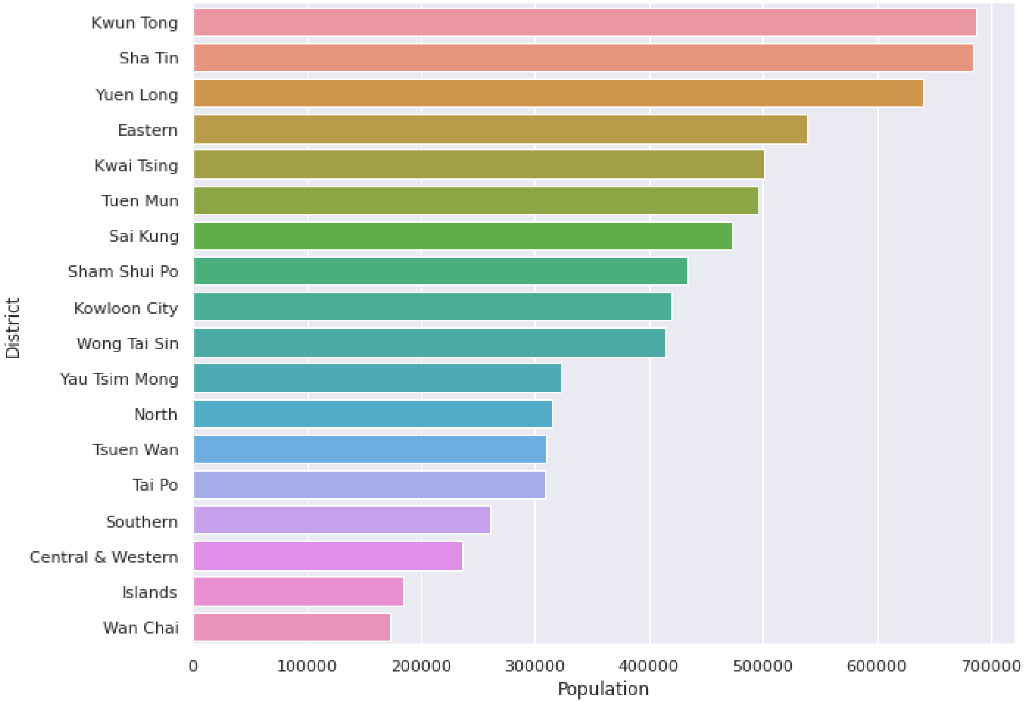
On Graph:

Arrow:
Significant drop in vacancy

Star:
Irregular Pattern

Weekdays
Weekends

Exploratory Data Analysis



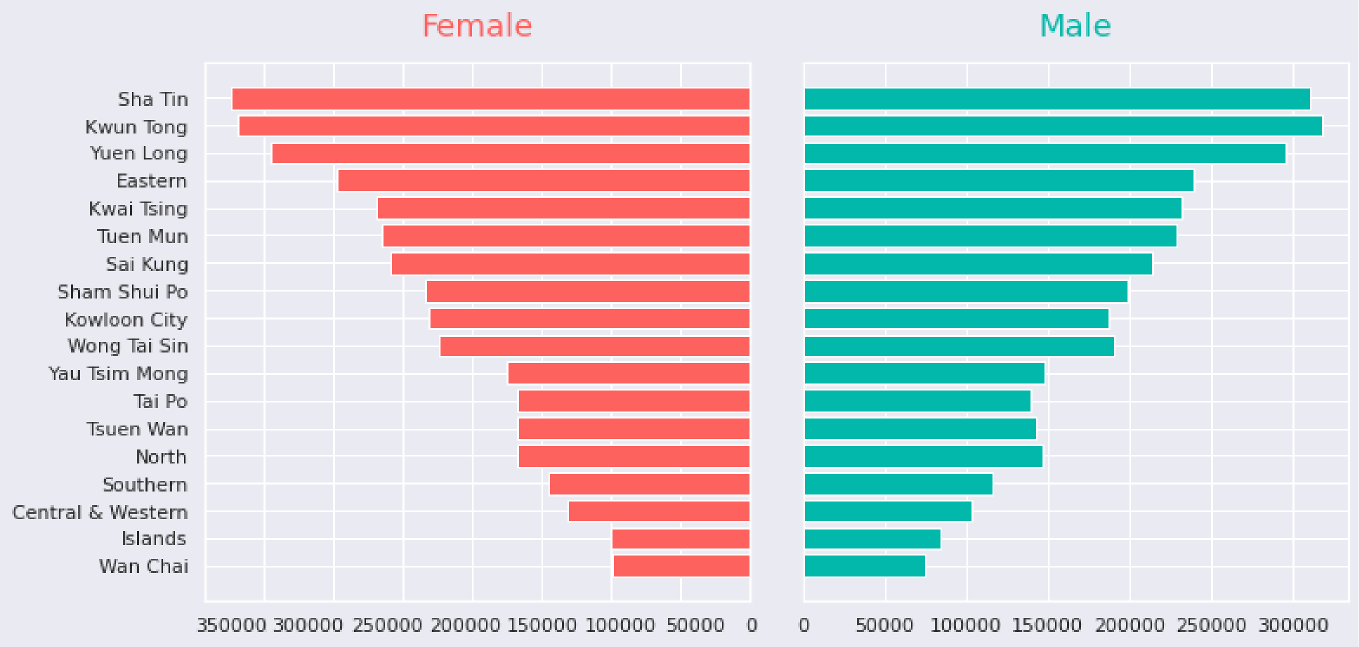
Figure

Population of each district

Findings:

Kwun Tong, Sha Tin, & Yuen Long are the top 3 most populated districts

Exploratory Data Analysis



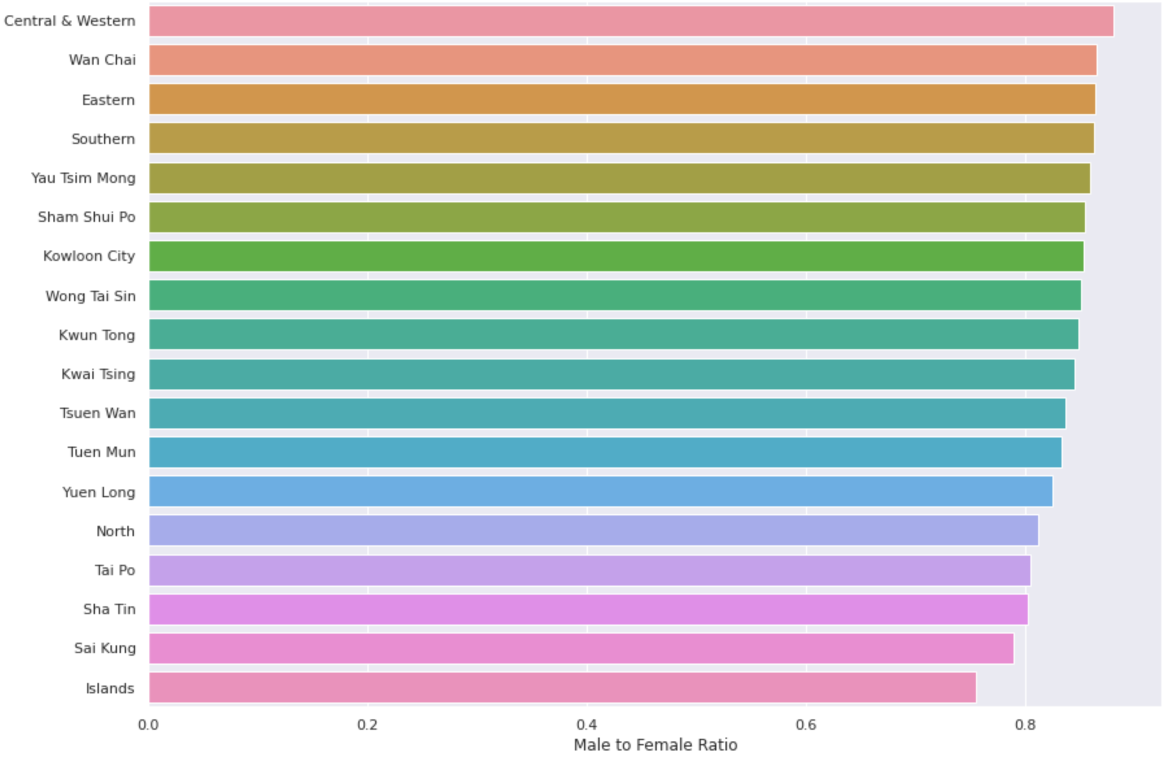
Figure

Population and gender distribution of each district

Findings:

Kwun Tong, Sha Tin, & Yuen Long are the top 3 most populated districts

Exploratory Data Analysis



Figure

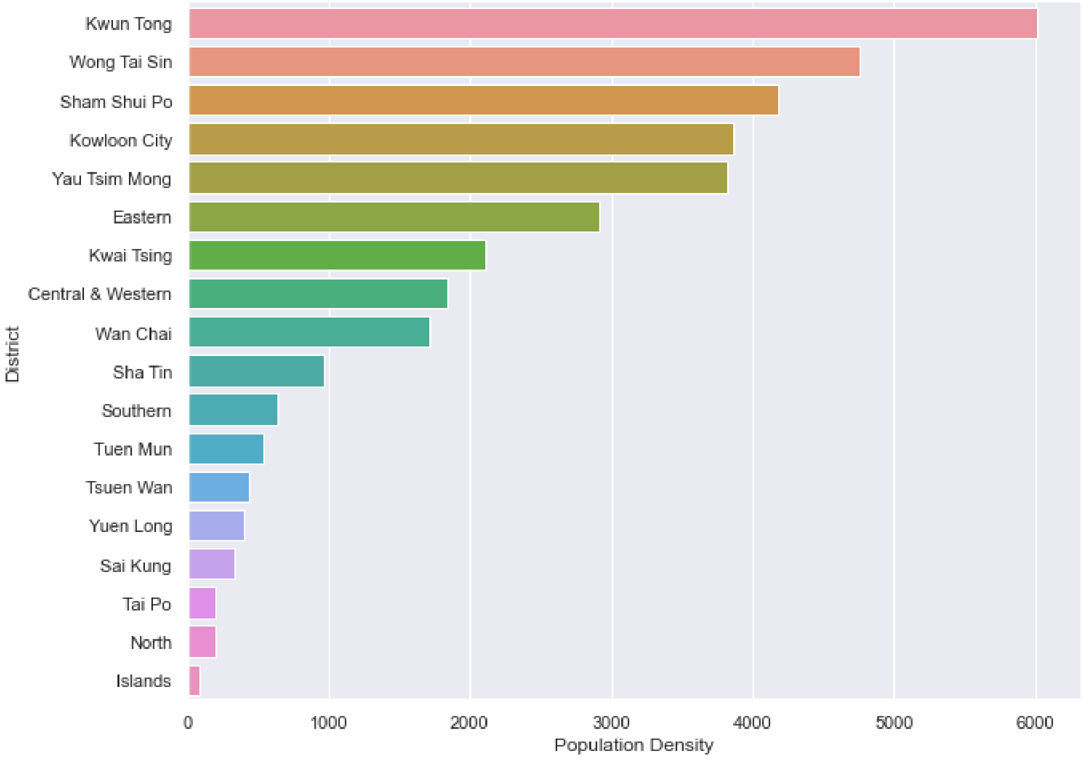
Male-to-Female Ratio

Findings:

The differences of Male-to-Female Ratio between each district are noticeable but small

Nothing really insightful

Exploratory Data Analysis



<u>Figure</u>

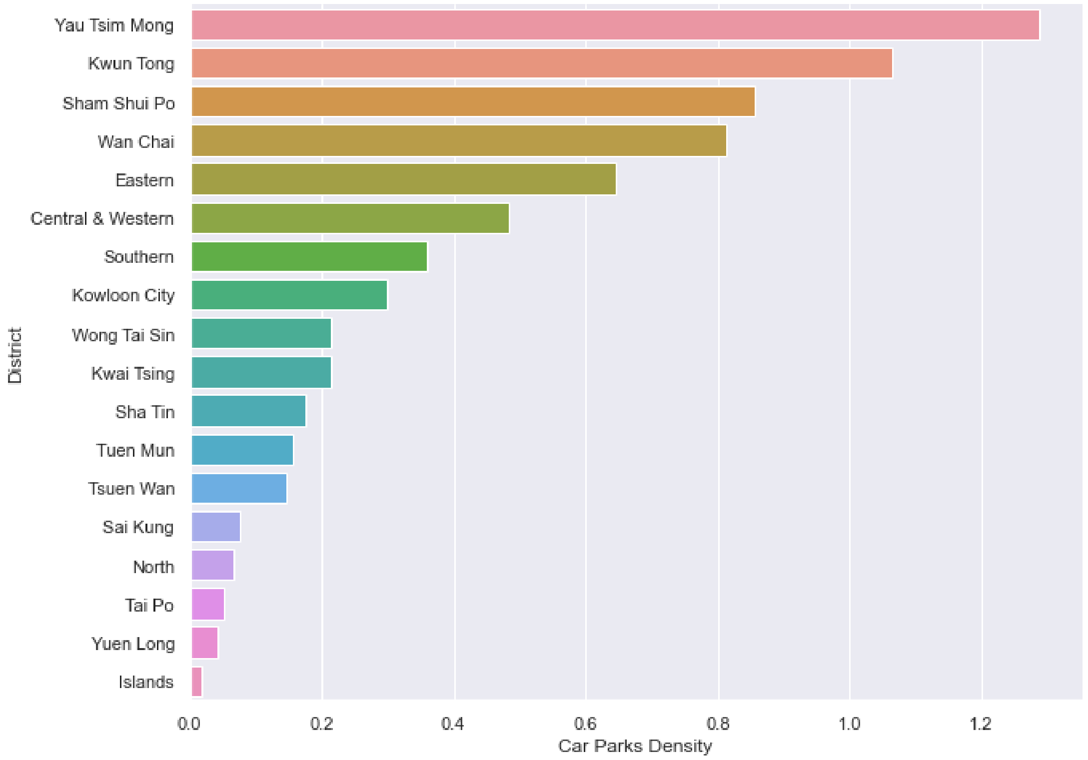
Population Density

<u>Findings:</u>

The differences of Population Density between each district are significant.

Which might cause a impact on performance of our model as a potential feature.

Exploratory Data Analysis



Figure

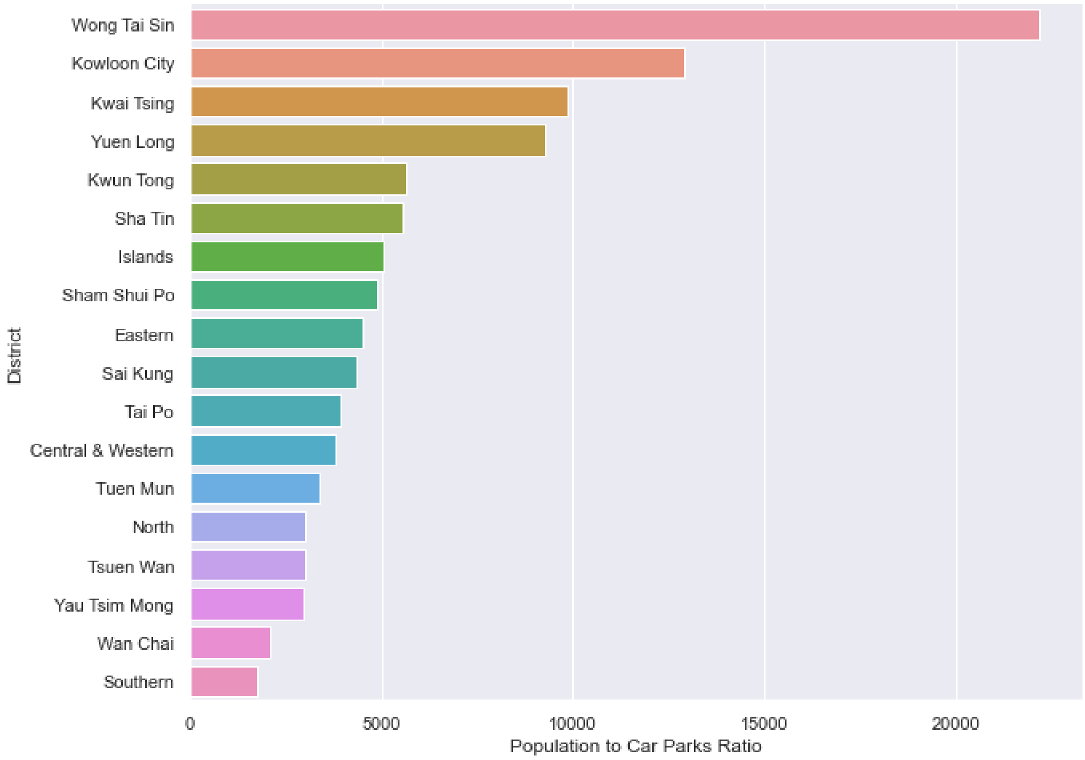
Car Parks Density

Findings:

Same as Population Density, the differences of Car Parks Density between each district are significant.

Which might also cause a impact on prediction performance of our model as a potential feature.

Exploratory Data Analysis



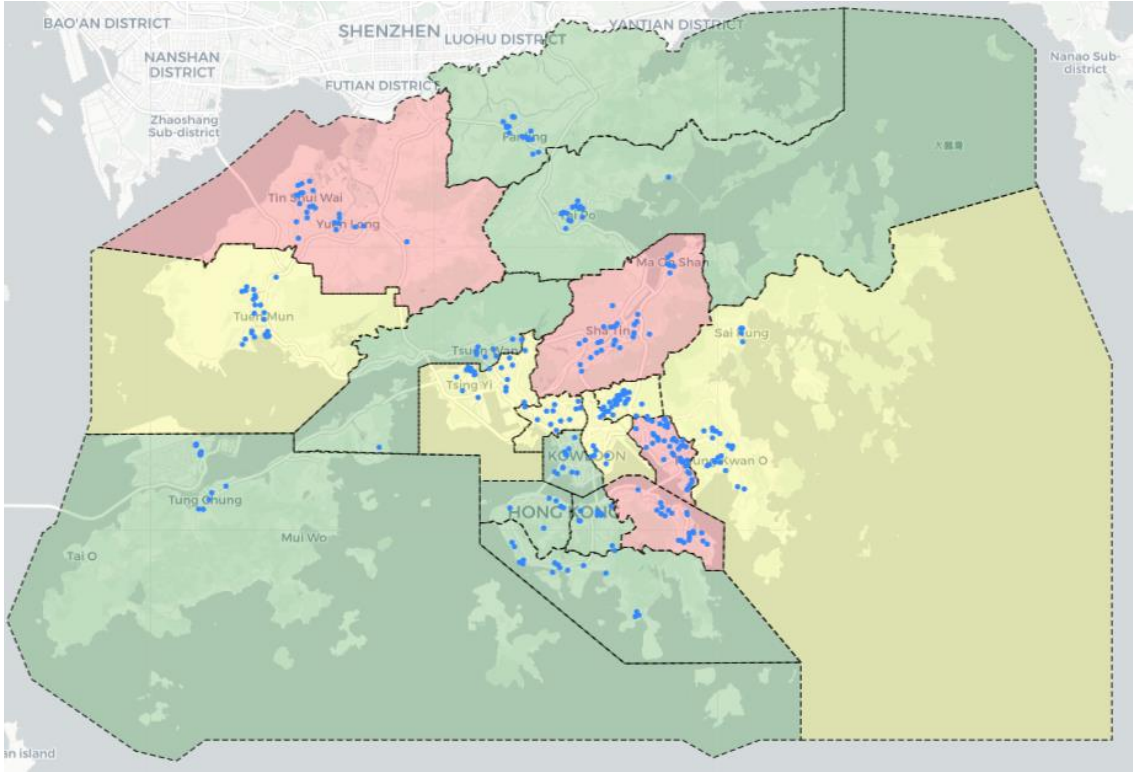
Figure

Population-to-Car Parks Ratio

Findings:

Wong Tai Sin has the highest Population-to-Car Parks Ratio. Kowloon City, Kwai Tsing and Yuen Long roughly shares similar Population-to-Car Parks Ratio

Exploratory Data Analysis



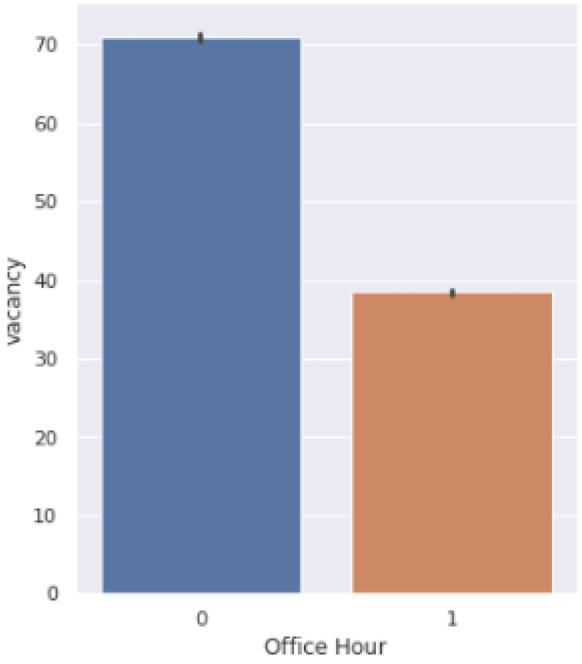
Figure

Car parks and population distribution overlaid

Findings:

Population Density is **positively correlated** with number of car parks, rather than the area of the districts.

Exploratory Data Analysis



Figure

vacancy in office hour (9:00 - 18:00)

Findings:

The mean vacancy in non-office hour is around 70 and in office hour is almost 40, greater by a factor of ~2

This feature might be useful for our model to predict.

Exploratory Data Analysis

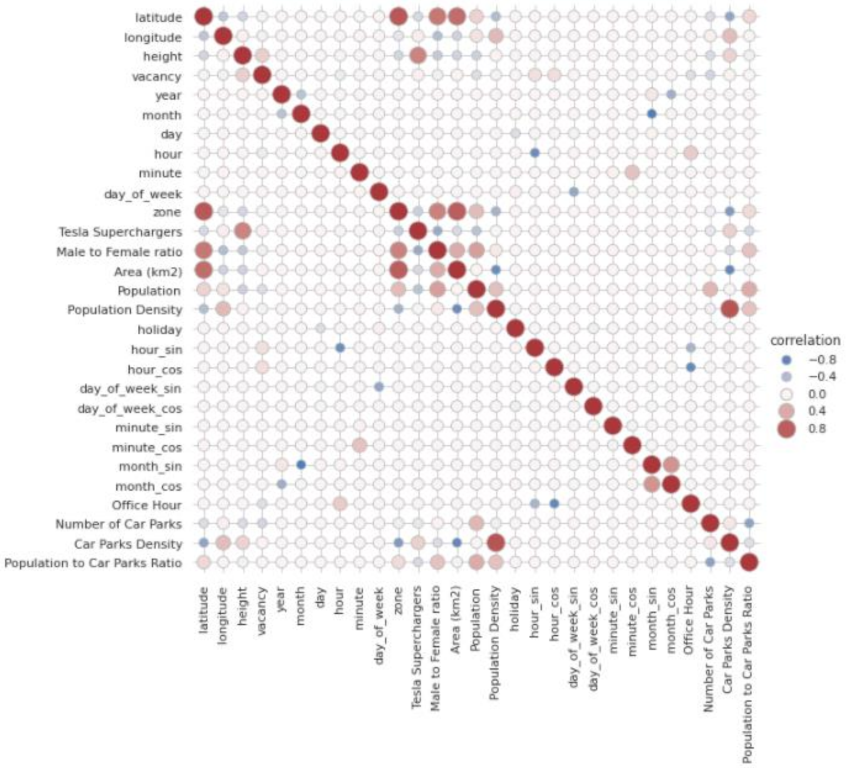


Figure
Pearson's R

Findings:

No feature shares strong corr. with the target.

Strong corr.:

month - month_sin

Population Density - Car Parks Density

latitude - zone

zone - Area (km2)

Car Parks Density - Area (km2)

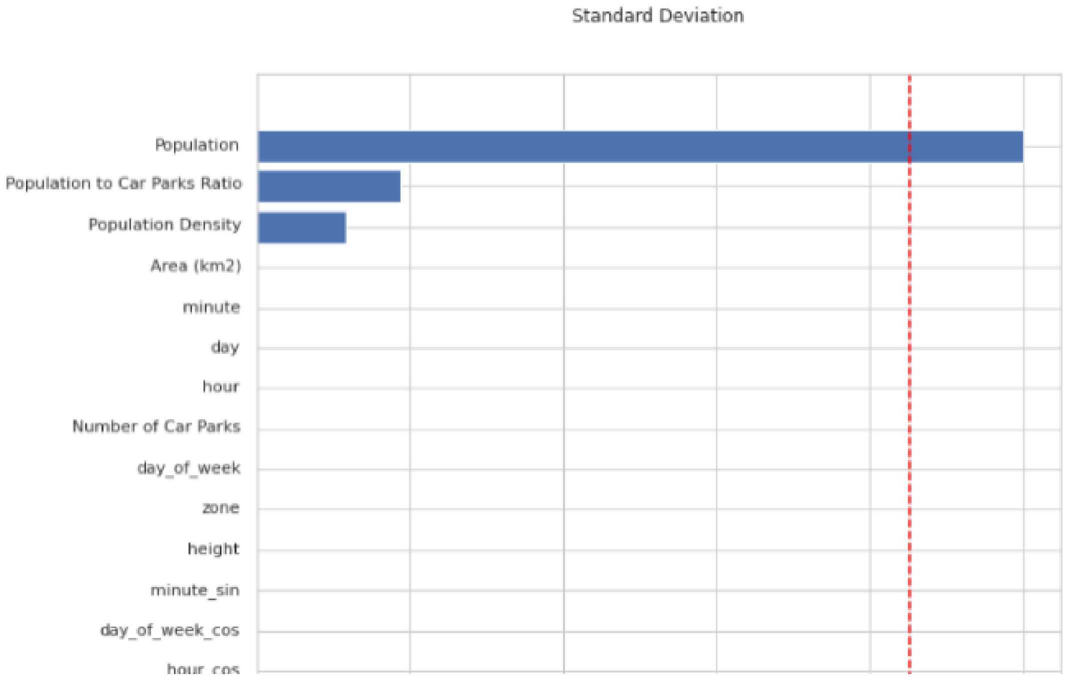
hour_cos - Office Hour

Area (km2) - Population Density

hour - hour_sin

Area (km2) - latitude

Exploratory Data Analysis



Figure

Standard Deviation Plot

Findings:

Only features “Population”, “Population to Car Parks Ratio” and “Population Density” shows obvious level of variance (contain very little information).

Even though most of them contain very little information we still would like to try them out to build our model.

Data Preprocessing - Data Cleaning

- Label Encoding categorical features such as “park_id”
- Imputing Missing Values
  - 9 car parks’ districts are missing
  - Directly recovered via: http://www.gohk.gov.hk/chi/welcome/index.html

Dropping unnecessary columns:

- Highly correlated (Pearson’s R)
- Unbalanced
- Remove outliers (prevent bias)
- Remove duplicated instances
- Remove instances outside the desired time period

Data Preprocessing - Dimensionality Reduction

"Population2020" Dataset

- Features "DC" & "OBS_VALUE" were selected
- "DC" was mapped into "District"
- "OBS_VALUE" was renamed into "Population"

"Population_sex" Dataset

- Features "DC", "Sex", "Age" & "OBS_VALUE" were selected
- "DC" was mapped into "District"
- "OBS_VALUE" was renamed into "Population"

"area" Dataset

- Entire dataset was used

Data Preprocessing - Feature Engineering

Following features are created:

- Grouping districts into zones such as Hong Kong Island, Kowloon
- Converting timestamp into ‘day_of_week’, ‘hours’, ‘minute’
- Cyclical Encoding ‘month’, ‘day_of_week’, ‘hours’, ‘minute’
- Listing car park with ‘Tesla Superchargers’
- Male-to-Female ratio per district
- Area (km^2) of each district
- Population per district
- Population Density per district
- Public holidays

Machine Learning Models

	Speed	Overfitting/Underfitting	Performance
Linear Regressor	Medium	May overfit without regulation	Poor with high dimensional data
Random Forest Regressor	Slow	Prone to overfitting	Usually Steady
AdaBoost Regressor	Medium	Rarely Overfit/Underfit	Usually good
CatBoost Regressor	Medium	Rarely Overfit/Underfit	Usually good
XGB Regressor	Fast	May overfit without regulation	Usually good
LightGMB Regressor	Fast	Rarely Overfit/Underfit	Usually good

Machine Learning Models

	Speed	Overfitting/Underfitting	Performance
Decision Tree Regressor	Medium	May underfit with proper tuning	Worst than **Random Forest Regressor**
Extra Trees Regressor	Medium	Prone to overfitting	Worst than **Random Forest Regressor**
KNN Regressor	Slow	May underfit with proper tuning	Usually perform poorly
Radius Neighbors Regressor	Slow	May underfit with proper tuning	Usually perform poorly
Support Vector Regressor	Extremely Slow	May overfit without proper tuning	Perform well with small amount of data

Machine Learning Models - Limitation of model

Support Vector Regressor

Problem:

- Processing time extremely slow
- As shown in the fig:
    - Executing time was over 3 hours (without considering polynomial kernel function)

Conclusion:

- Model cannot be presented and therefore will be included in future development stage

Training Features

Original:

- Day of week: Categorical
- Hours: Categorical
- Minute: Categorical
- District: Categorical
- Park_id: Categorical

Created:

- Zone: Categorical
- Tesla Supercharger: Categorical
- Holiday: Categorical
- Total population: Numerical
- Population density: Numerical
- Male-to-female ratio: Numerical
- Area: Numerical

To Be Confirmed

Model Evaluation - Result

Model trained:

Dummy Regressor,Decision Tree,Random Forest, xgboost, GridSearchCV

Best accuracy model (Highest R2):

Random forest

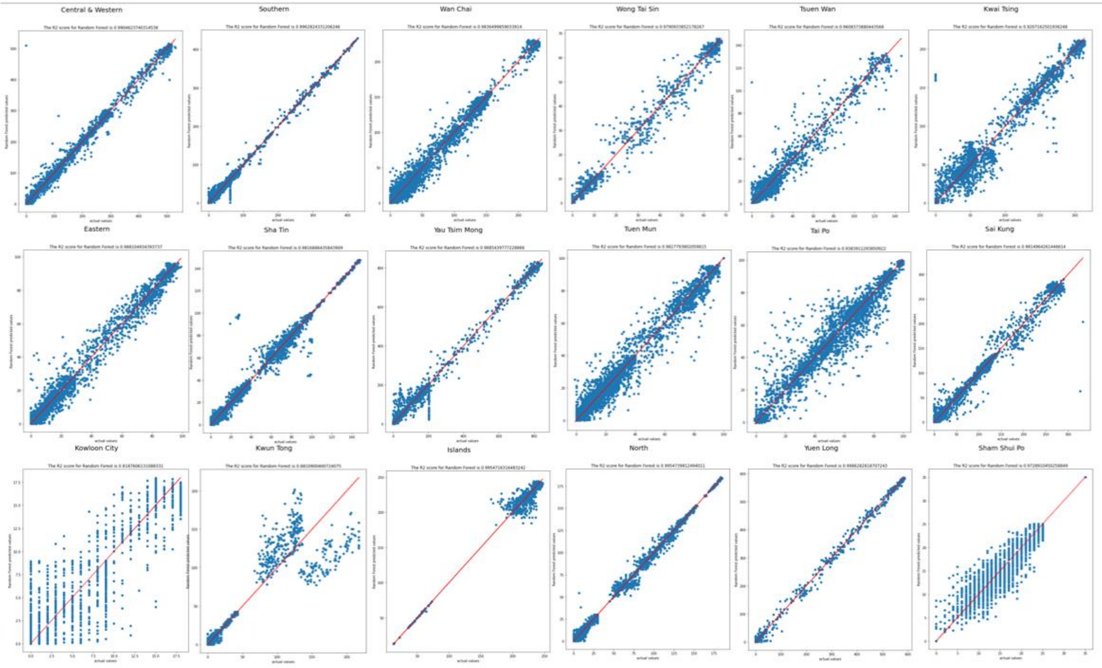
Best result metrics:

R2 Score = 0.989

Mean-absolute error = 8.807

Root-mean-squared error = 3.338

	Model Name	r2_score	RMSE	MAE
0	Dummy Regressor	-0.000022	84.743973	50.166253
1	Decision Tree	0.987799	9.360393	3.473115
2	Random Forest	0.989199	8.807001	3.338286
3	xgboost	0.941619	20.475740	12.227856



Random Forest R2 score for 18 districts

Model Selection

Zone

Created by grouping districts

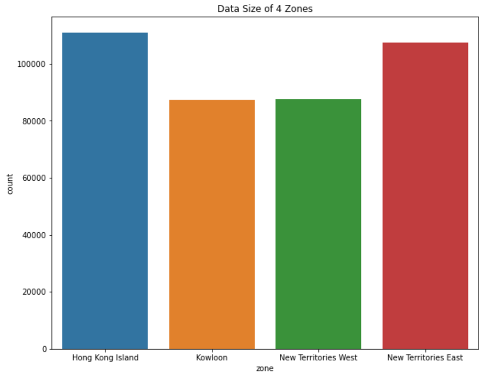
Model Selection as district or zone

Problem:

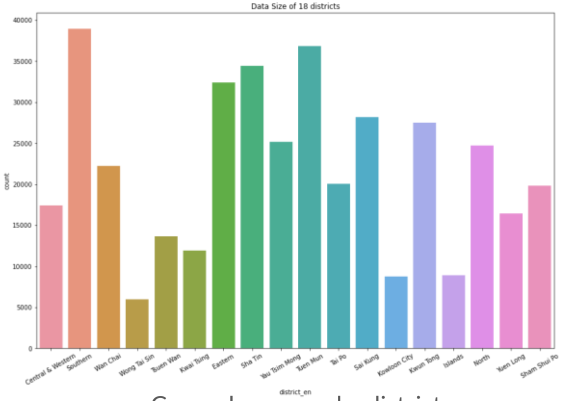
Some district have relatively low R^2 score if model is trained by district (as seen in previous result slide)

Hong Kong Island	Central & Western, Southern, Wan Chai, Eastern
Kowloon	Sham Shui Po, Wong Tai Sin, Kowloon City, Kwun Tong, Yau Tsim Mong
New Territories (West)	Islands,Tsuen Wan, Kwai Tsing, Yuen Long, Tuen Mun
New Territories(East)	Tai Po, Sai Kung, North, Sha Tin

Table of zone grouping

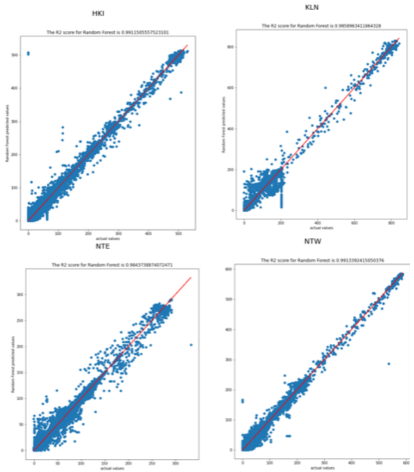


Car park vacancy by Zone



Car park vacancy by district

Future Development: Model Selection



Random Forest R2 score for 4 zones

Model Selection as district or zone

Solution:

1. Train model with zone
2. R2 score will be compared for district inputted by user
3. Zone or district model of random forest will be deployed based on higher R2 score of district selected

Conclusion:

Districts in HKI, NTW, NTE → corresponding zone model will be used
Districts in KLN → Hong Kong (ALL) model

Hong Kong Island	RMSE: 8.611	R2:0.991	MAE:3.542
Kowloon	RMSW: 13.242	R2:0.986	MAE:3.955
N.T. West	RMSE:7.895	R2:0.991	MAE:3.557
N. T. East	RMSE:5.997	R2:0.984	MAE:2.607
Hong Kong (ALL)	RMSE:8.797	R2:0.989	MAE:3.338

Deployment and result

Deployed Model:

Random Forest

User input:

1. Date: DD/MM/YYYY
2. Time: HH:MM
   - Round off to every 15 mins for prediction
3. District: Name of the district

Output return:

List of Car park vacancy in the district

```
input_date = input("Full date (DD/MM/YYY):")
input_time = input("Time (HH:MM):")
input_district = input ("District:")
```

```
Full Date (DD/MM/YYYY):06/05/2021
Time (HH:MM):12:35
District:Kwun Tong
```

User input

```
Hi There. You are now in Kwun Tong at 12:35 on 06/05/2021
Nearby carparks with the corresponding vacancy are listed below:

   - Skye Parking Ho Tin Street: 3.0 spot(s)
   - Castle Peak Beach: 0.0 spot(s)
   - San On Street: 12.0 spot(s)
   - Hoi Wah Road: 12.0 spot(s)
   - Tsing Yin Street: 12.0 spot(s)
   - Tuen Yee Street: 1.0 spot(s)
   - Castle Peak Road - Castle Peak Bay: 1.0 spot(s)
   - Sam Shing Street 1: 1.0 spot(s)
   - Sam Shing Street 4: 1.0 spot(s)
   - The Jockey Club Tuen Mun Butterfly Beach Sports Ce: 1.0 spot(s)
   - Tuen Mun North West Swimming Pool: 1.0 spot(s)
   - Tuen Yee Street: 12.0 spot(s)
   - Tuen Mun Town Plaza Phase 2: 13.0 spot(s)
```

Enquiry result

Model Evaluation

Cross Validation Scores (Training)	Negative Mean Absolute Error	Negative Mean Squared Error	R2
XGBRegressor	-24.378	-1581.384	0.812
AdaBoostRegressor	-54.467	-4577.359	0.465
LGBMRegressor	-9.553	-235.217	0.972
CatBoostRegressor	-6.493	-142.134	0.983
Decision Tree Regressor	-4.674	-133.617	0.984
Random Forest Regressor	-4.105	-94.101	0.989

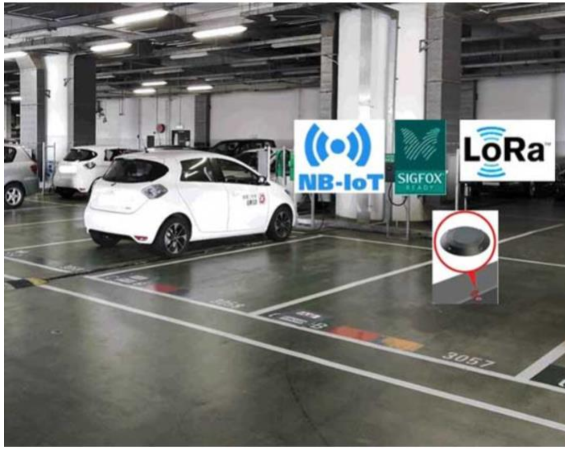
Model Evaluation

Scores (Testing)	Mean Absolute Error	Mean Squared Error	R2
XGBRegressor	22.524	1447.011	0.826
AdaBoostRegressor	57.749	4991.048	0.400
LGBMRegressor	9.227	223.574	0.973
CatBoostRegressor	6.086	127.645	0.985
Decision Tree Regressor	4.556	124.18	0.985
Random Forest Regressor	4.071	92.331	0.989

Future Development: Enhancing Prediction Accuracy

Trying more methods in the data preparation process:

- More data
- More external data for feature engineering
- Normalizing, Logarithmic, Box Cox Transformation , etc.
- Live vacancy data from EMSD smart car park management



Hong Kong EMSD smart car park management

Future Development: Enhancing Prediction Accuracy

Adding more training features:

- Median salary of each district - leading to more car owners
- Weather: Rainy vs sunny - affect traffics
- Pre and post Pandemic figures - influence on car park occupancy
- Number of privately owned parking slot - available public parking slot number in contrast
- Charging fees of each car park - spacing availability lower at cheaper car parks

Future Development: Enhancing Prediction Accuracy

Trying more methods in the data preparation process:

- More data
- More external data for feature engineering
- Normalizing, Logarithmic, Box Cox Transformation , etc.
- Live vacancy data from EMSD smart car park management

Future Development: Hyperparameter Tuning

	Implement	Level of Understanding	Credibility/Accessibility
Sklearn GridSearchCV	Easy	Easy to understand	Provided by Sklearn
Bayesian Search	Difficult	Certain level of understanding to Bayesian Theorem	Accessible but may not be reputable
Genetic Algorithm	Difficult	Must be familiar to genetic algorithm	No reputable packages online

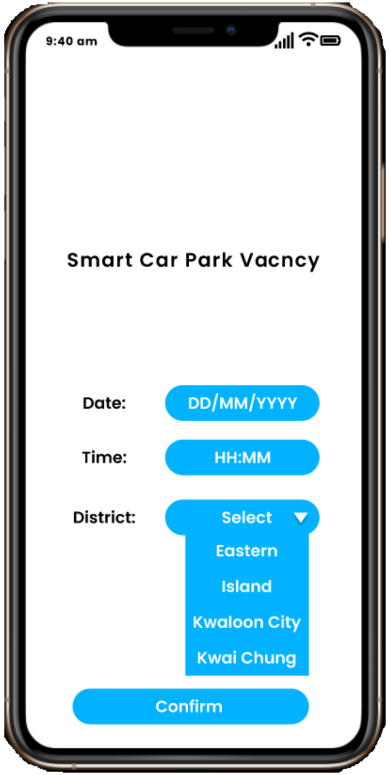
Future Development: Practical Usages

Deploying on:

- Web App
- Mobile App

How it works:

- Applying cloud database storage
- Automatically performing data preprocessing
- Dynamically retraining models with continuous learning
- Managing and monitoring models for model drift, bias and risk on dashboard
- Collect enquiry data on parking demand, which can be a potential feature



The End