

# Understanding Jakarta Floods Using Data

IBM Data Science Capstone Project

# CONTENT

**01**

Introduction

**02**

Understanding Rainfall  
Rate and Flood  
Occurrences

**03**

Predicting Future Floods  
and their  
Potential Collateral  
Damage

**04**

Clustering: Finding Out the  
Districts  
with High and Low Risks of  
Floods

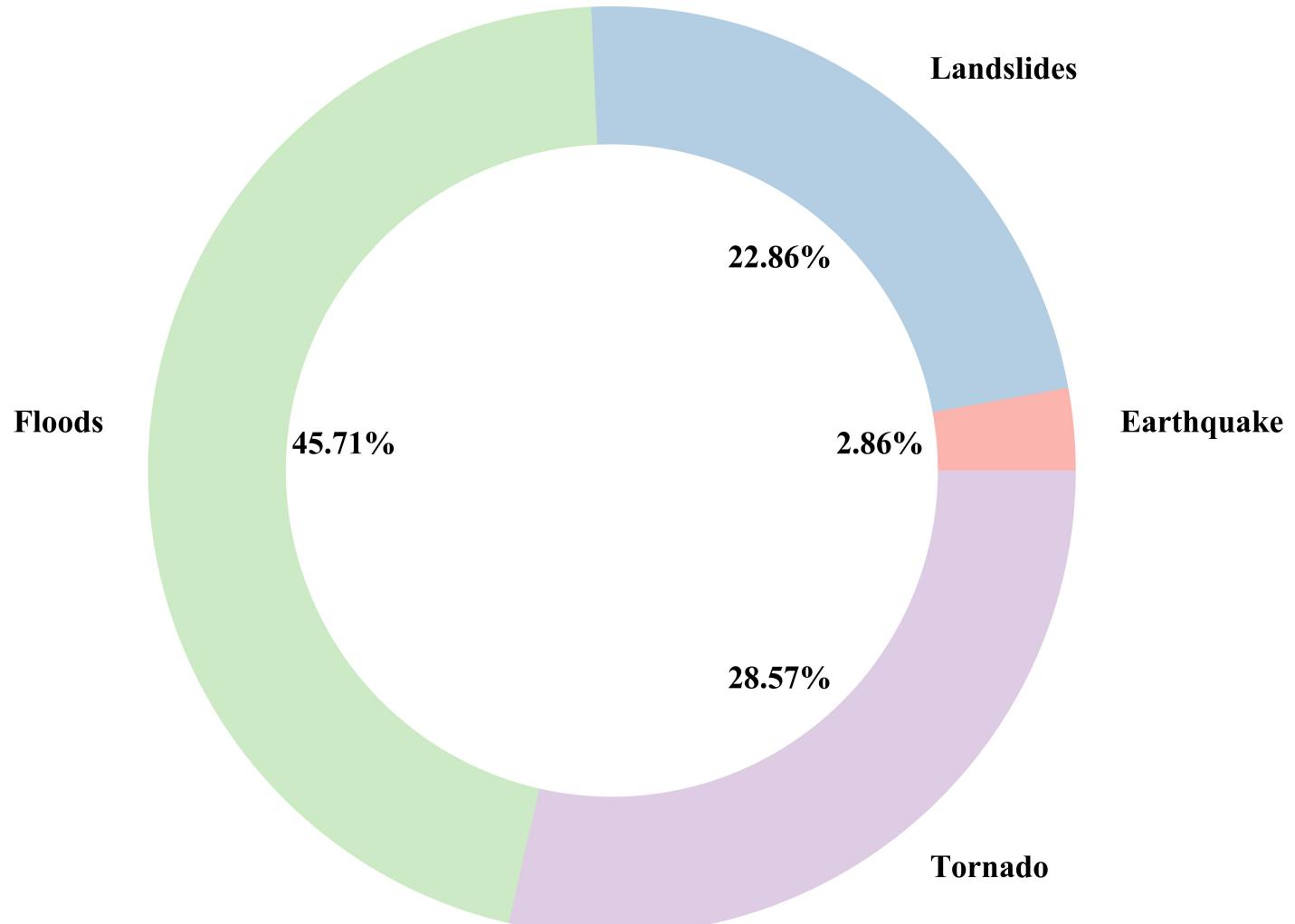
**05**

Influence of Parks to  
Mitigate the  
Floods

# ONE

## Introduction

Why is this topic about floods important?

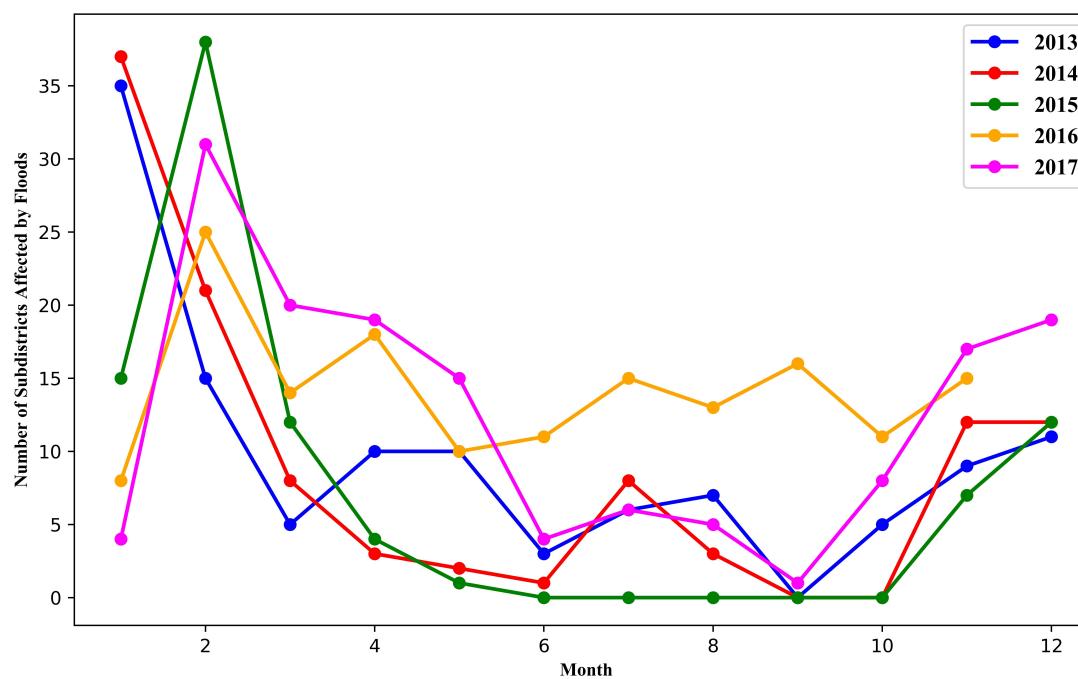
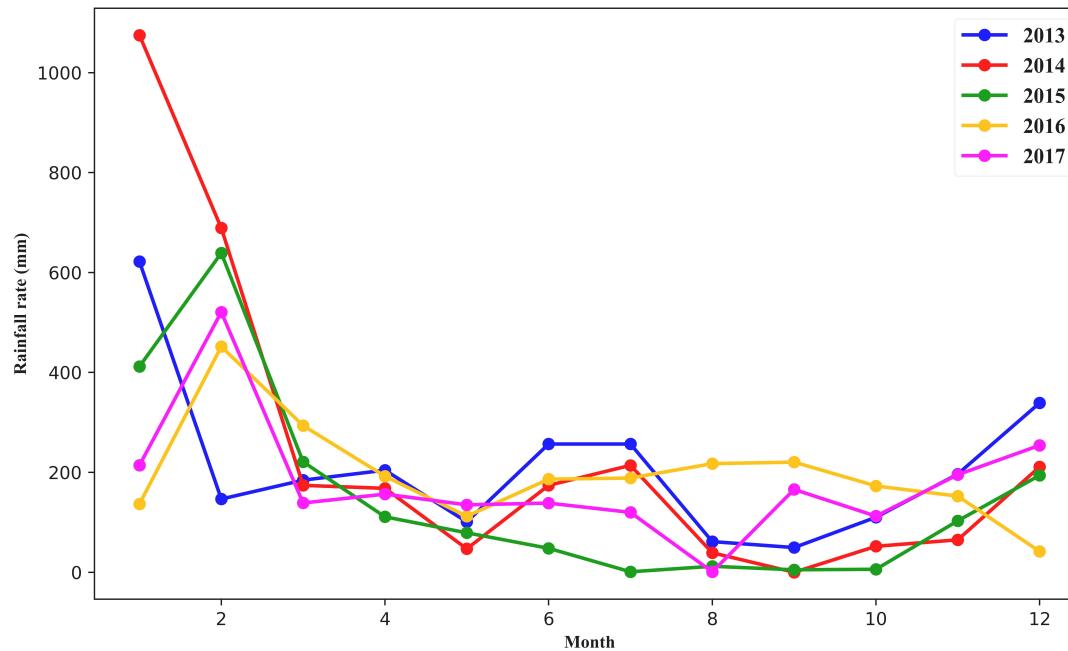


- Flood has been an ever-present phenomenon in Jakarta.
- Flood occupied nearly 50% of natural disasters occurred in Jakarta throughout 2018.
- By using the data, an insight about why, when, or where the floods will occur in any given time can be predicted.

# TWO

## Understanding Rainfall Rate and Flood Occurrences

What is the correlation between rainfall rate and the flood occurrences in Jakarta?

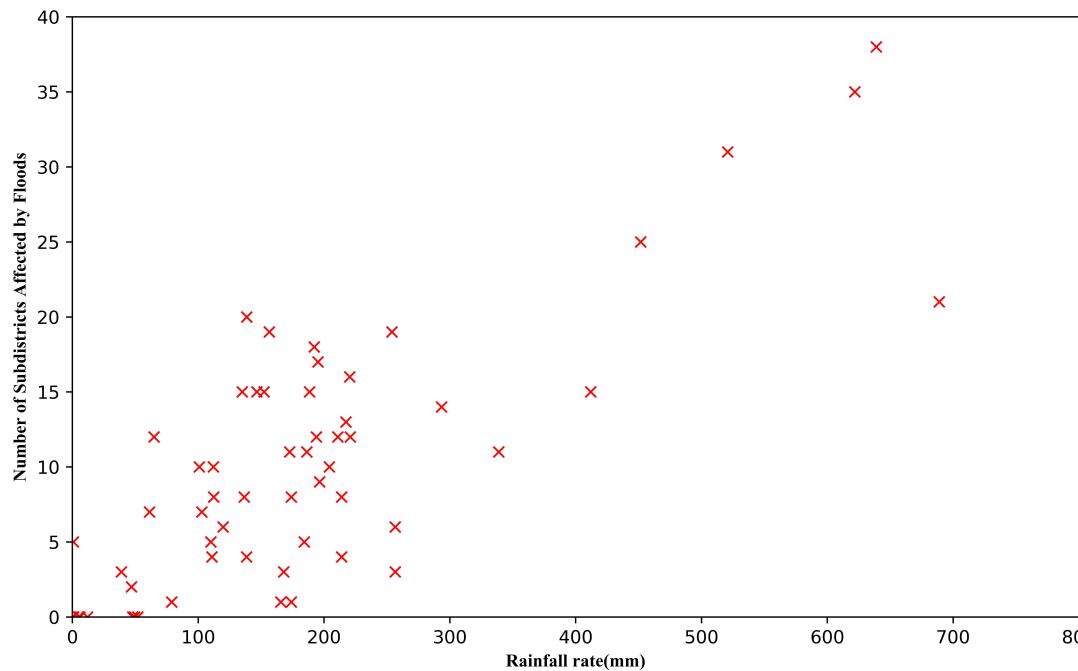


- There is a similarity of pattern between flood occurrences and rainfall rate.
- The flood case and rainfall rate in Jakarta reach their highest during January and February.
- Starting from March and the following months, the flood occurrence and rainfall rate are decreasing until reaching their lowest point in June until October.
- They start to increase again on November and so on.

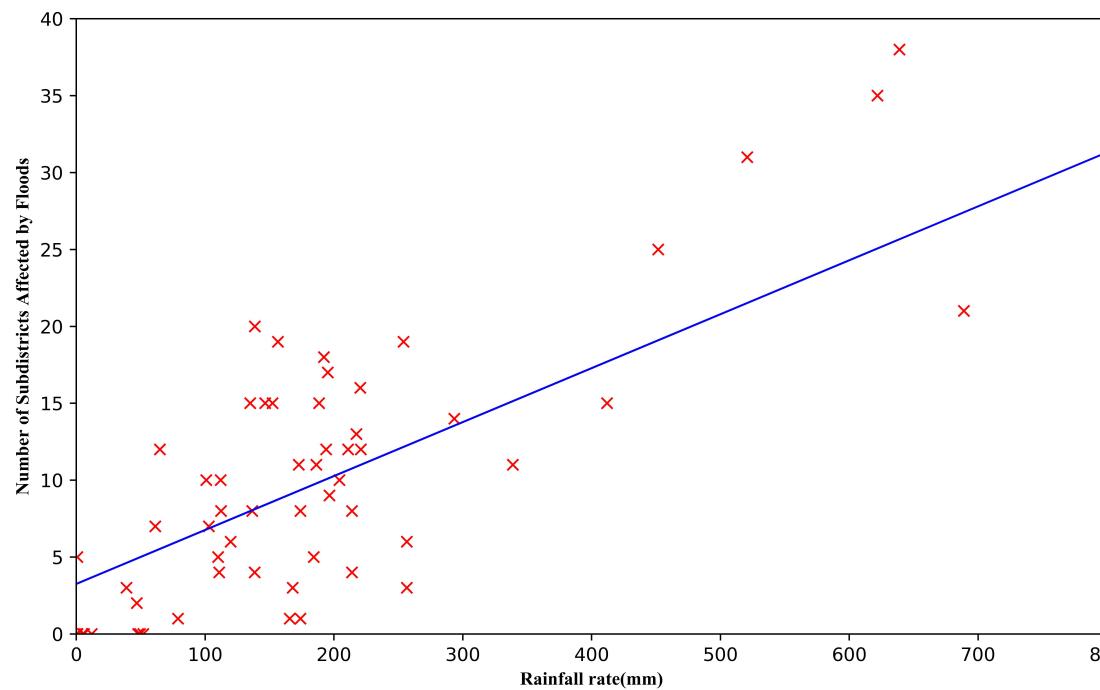
# THREE

## Predicting Future Floods and their Potential Collateral Damage

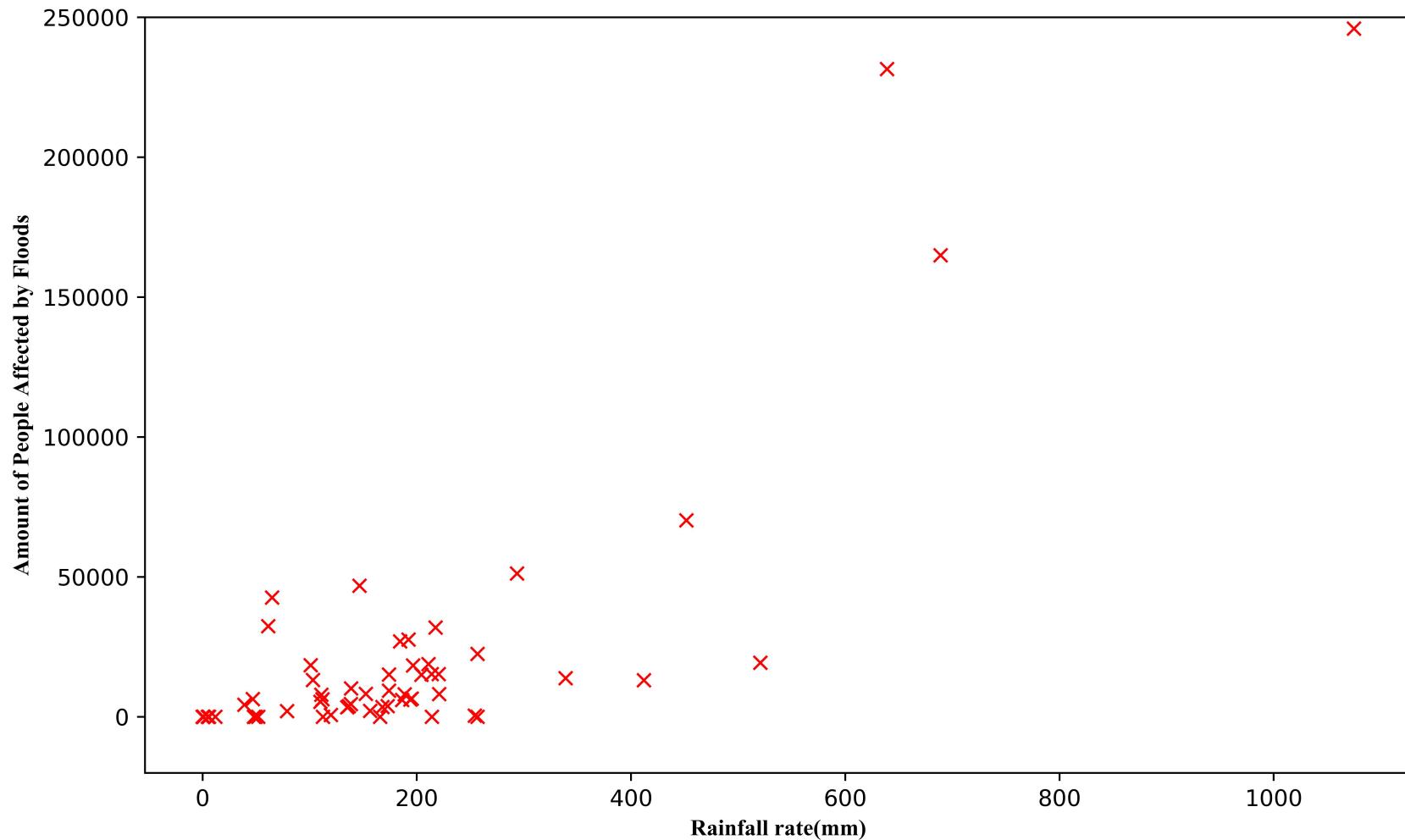
Can the amount of sub-districts and people who will be affected by floods in a given rainfall rate be predicted using predictive modeling algorithm?



There is a positive correlation between the rainfall rate and the amount of sub-districts that will be affected by floods.

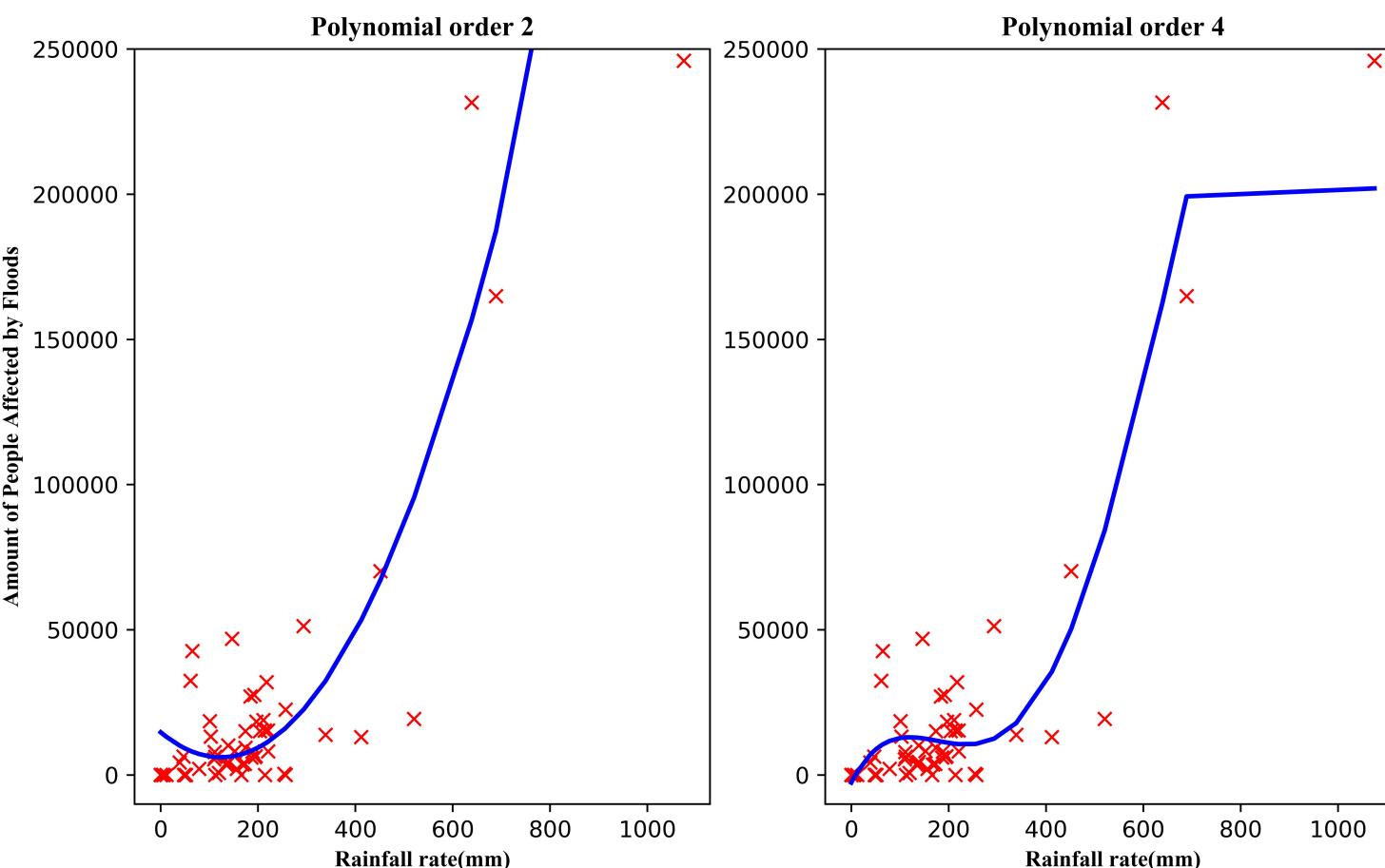
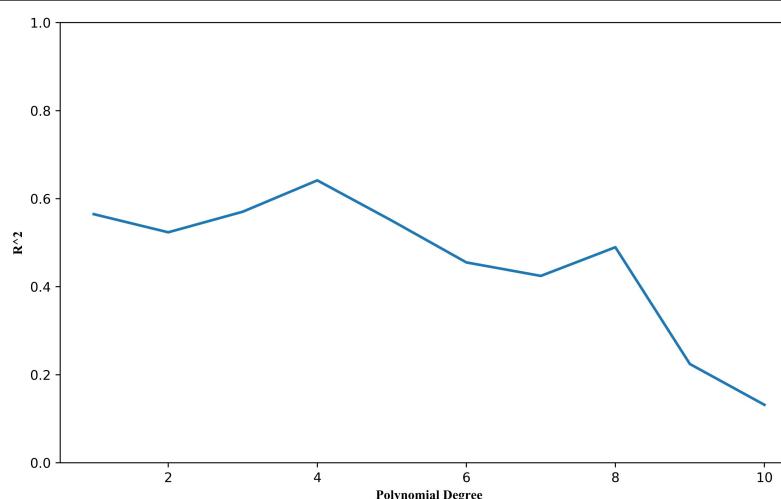


The amount of sub-districts that will be affected by floods in any given rainfall rate can be estimated using linear regression model.



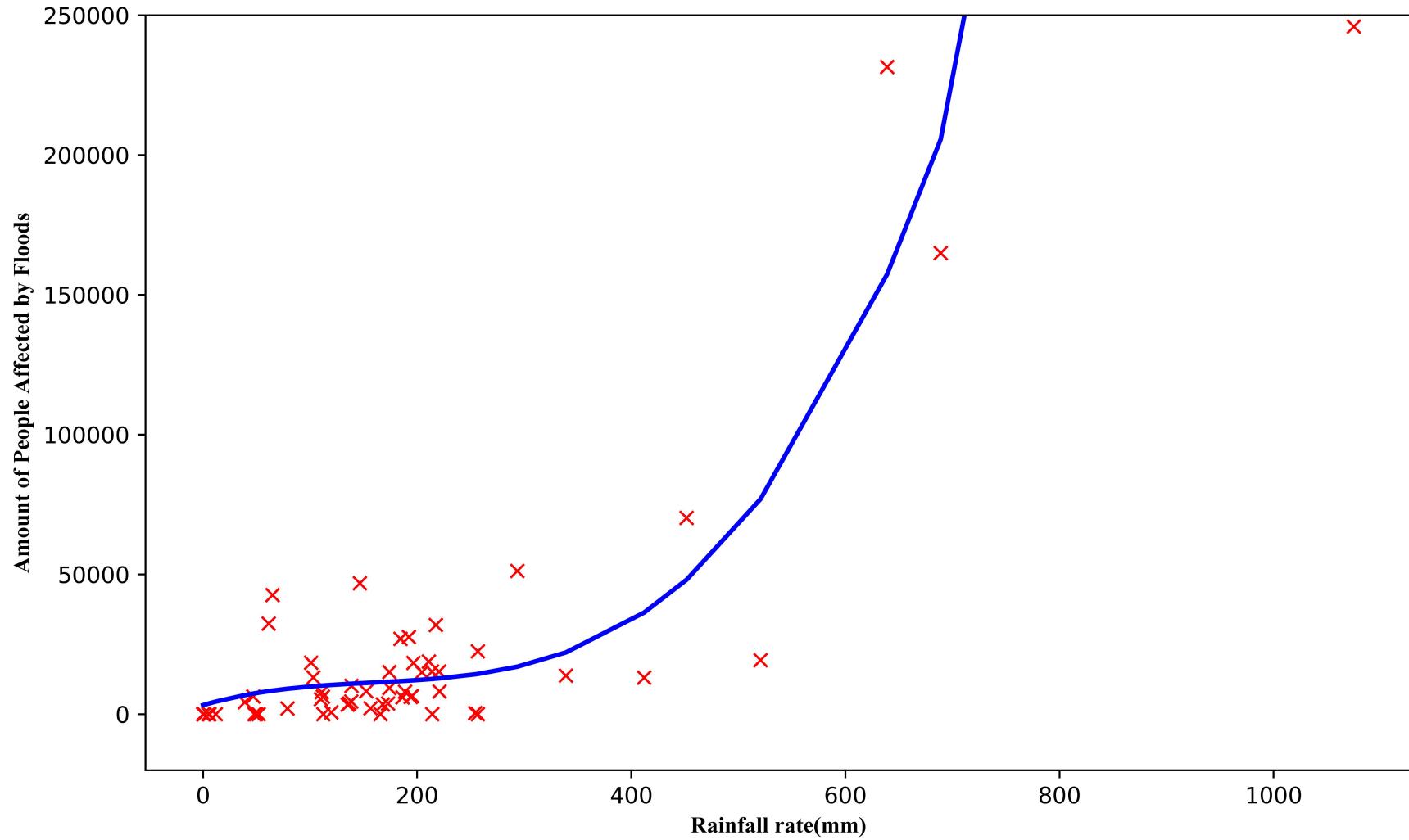
- There is a positive correlation between the rainfall rate and the amount of people who will be affected by floods.
- The relationship between two variables would be best estimated using polynomial regression model.
- But what is the most appropriate order of polynomial model to tackle this problem?

R-squared graph shows that polynomial with order of 2,3, and 4 would be good candidates.



Polynomial model with order 2 shows a sign of slight underfitting.

Polynomial model with order 4 shows some signs of overfitting.

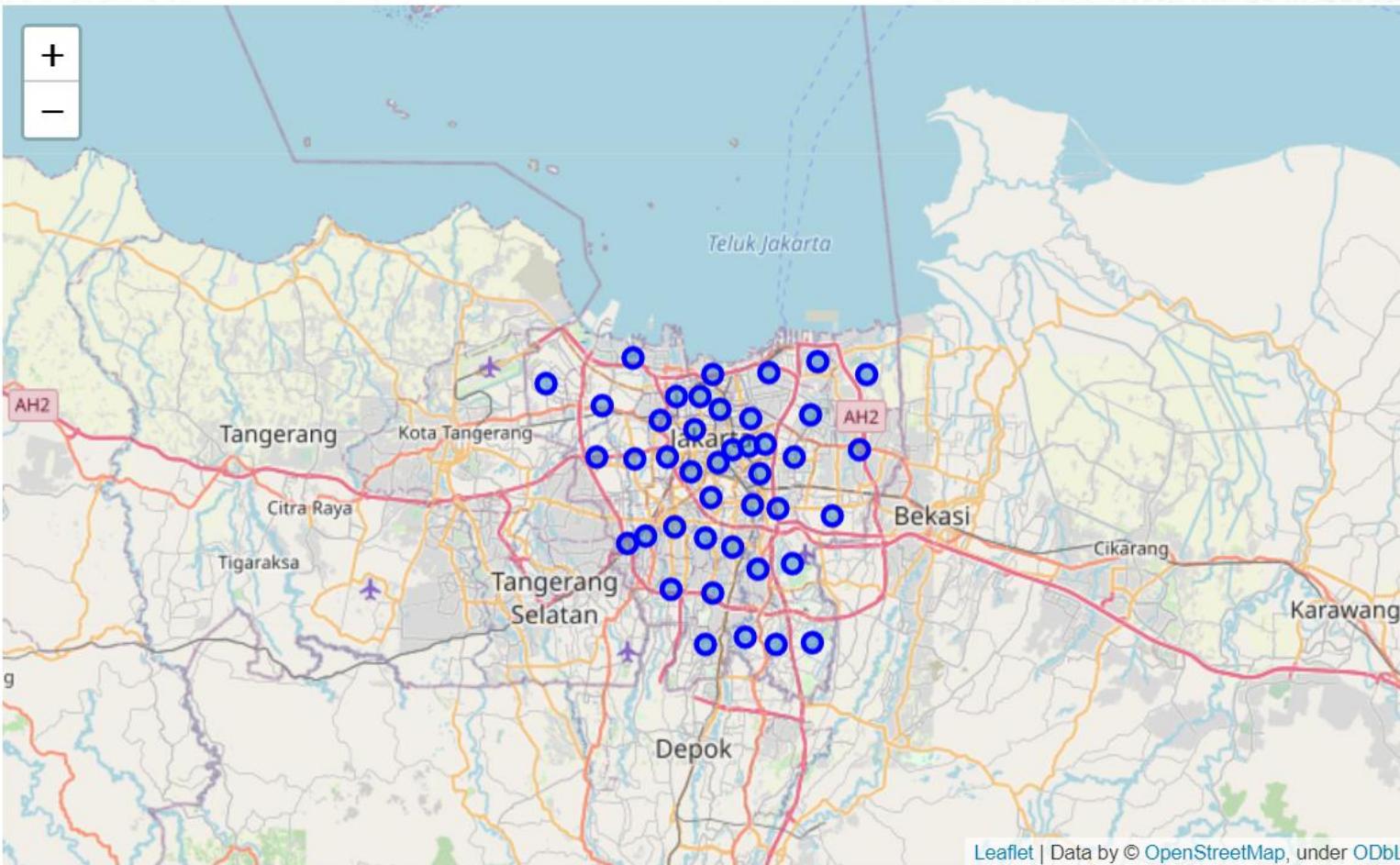


- Third order polynomial model shows 'just about right' estimation.
- The amount of people who will be affected by floods in any given rainfall rate can be predicted best by using third order polynomial regression model.

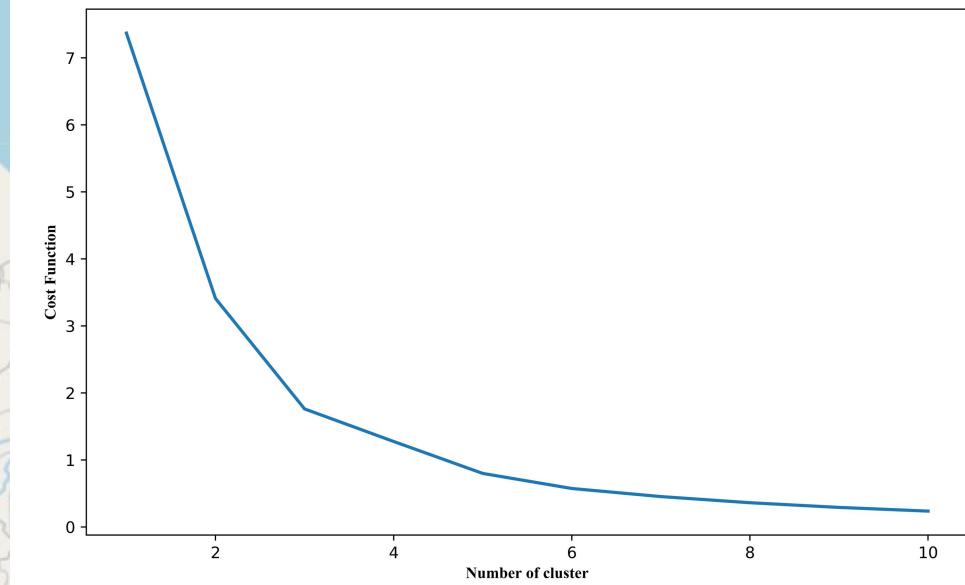
# FOUR

## Clustering: Finding Out the Districts with High and Low Risks of Floods

Which districts should the authorities focus their attention to when a heavy rain pours down the city?

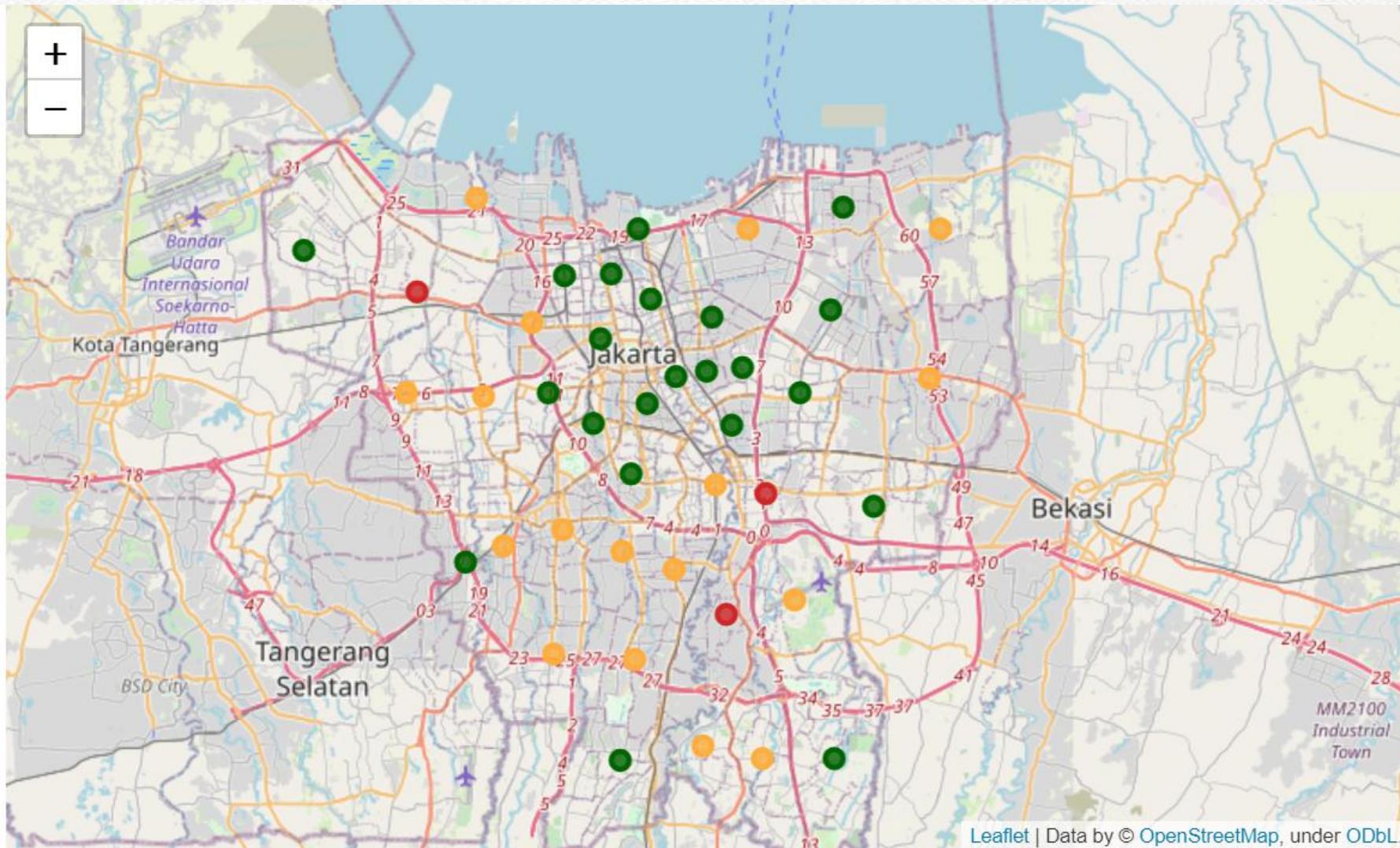


All of the 42 districts of Jakarta can be visualized in a map using folium library.



To choose optimum number of cluster, the elbow method needs to be conducted.

From the graph, it is shown that the optimum number of cluster = 3



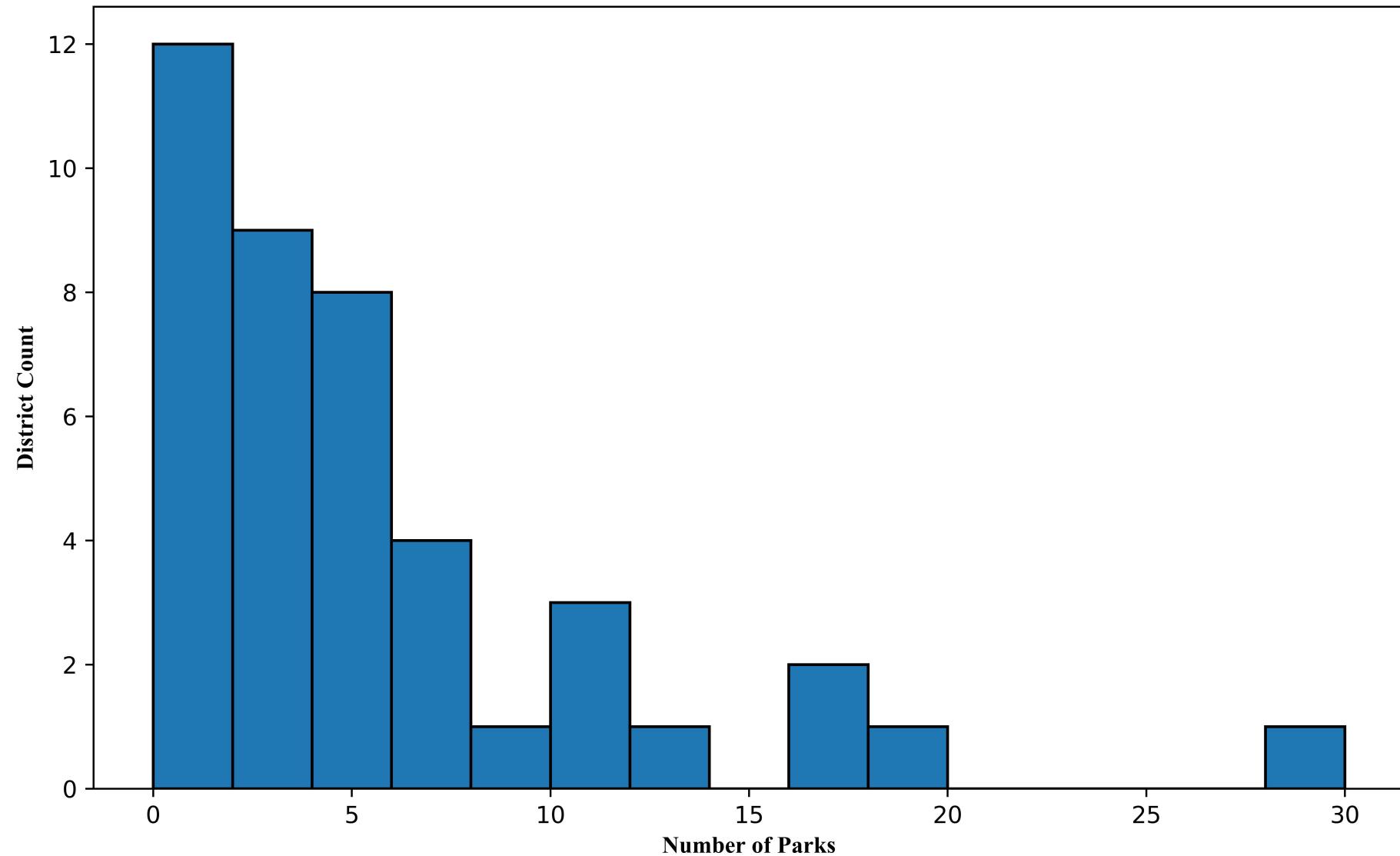
## Clustering result:

- Green: the safest and the most robust districts from floods.
  - Orange: districts which need moderate attention when a heavy rainfall pours down the city.
  - Red: districts with highest risks in terms of potential damage and severity from floods. These districts need intensive attention.

# FIVE

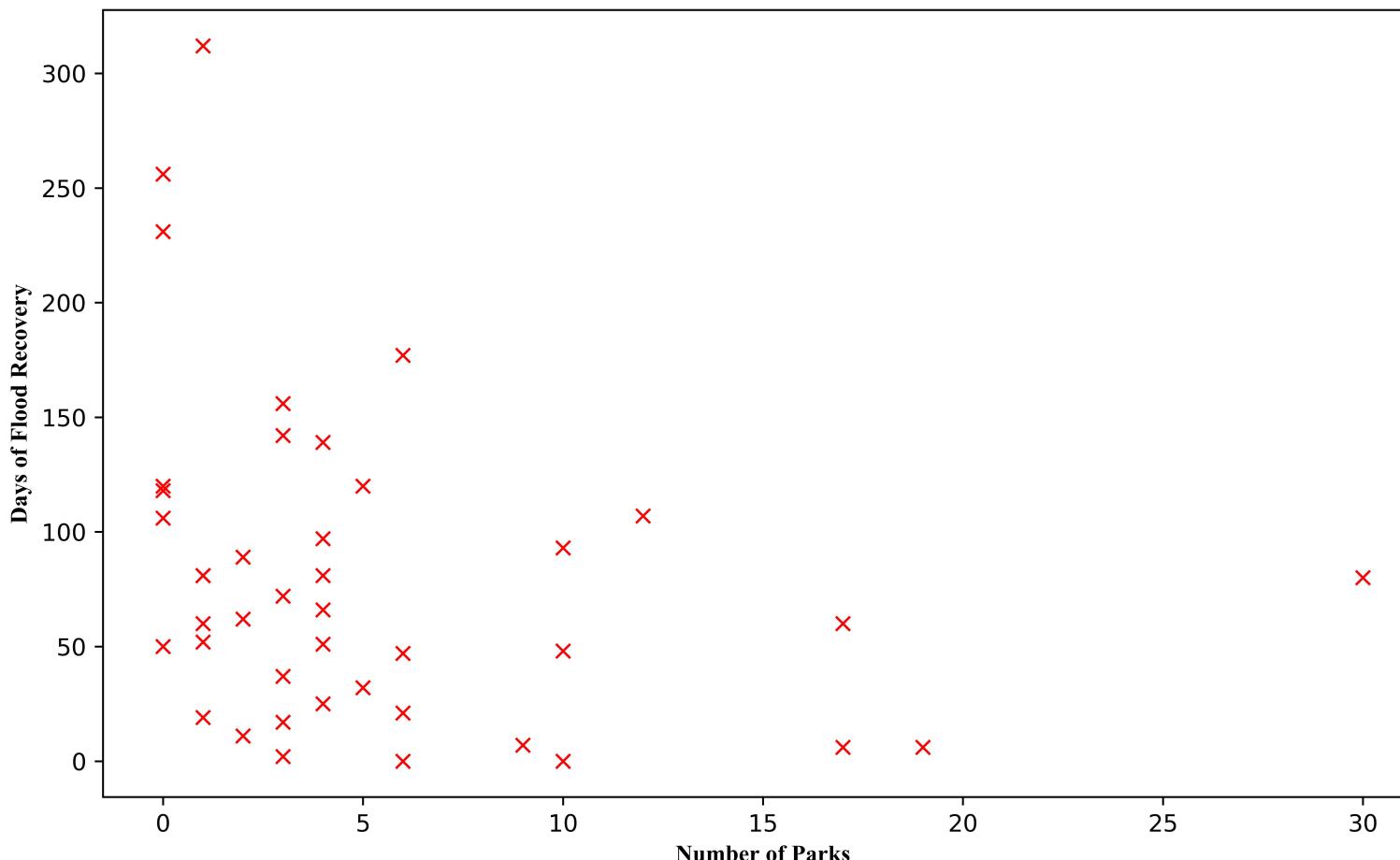
## Influence of Parks to Mitigate the Floods

Is the amount of parks in each district influential to mitigate the floods?



- The distributions regarding the amount of parks in each district is heavily right-skewed.
- Only 8 out of 42 districts have more than 10 parks in the neighborhood.

	<b>no. of cases</b>	<b>no. of People Affected</b>	<b>no. of People Forced to Relocate</b>	<b>Days of Flood Recovery</b>	<b>Park</b>
no. of cases	1.000000	0.574098	0.622585	0.942140	-0.246997
no. of People Affected	0.574098	1.000000	0.854797	0.693484	-0.188454
no. of People Forced to Relocate	0.622585	0.854797	1.000000	0.780099	-0.255296
Days of Flood Recovery	0.942140	0.693484	0.780099	1.000000	-0.281488
Park	-0.246997	-0.188454	-0.255296	-0.281488	1.000000



- The Pearson's correlation table shows that there is no significant correlation between the amount of parks and severity of floods.
- However, they have a negative correlation, which should encourage each district to open more green spaces in around their area.

# THE END

Thank you for the attention!