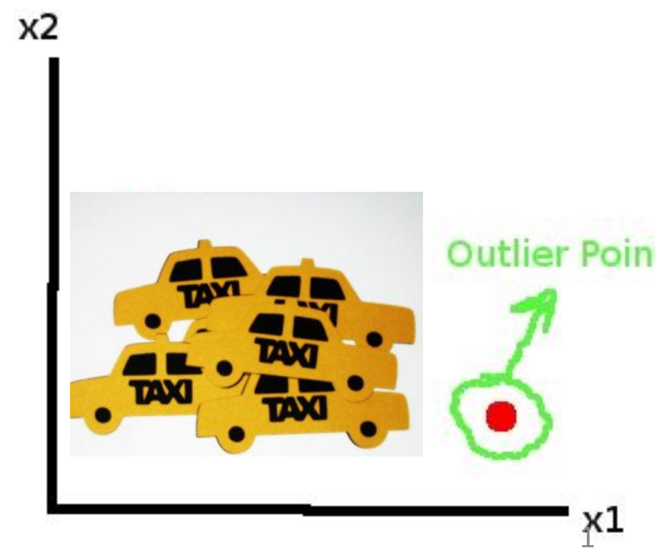
Anomaly Detection in NYC Taxi Data

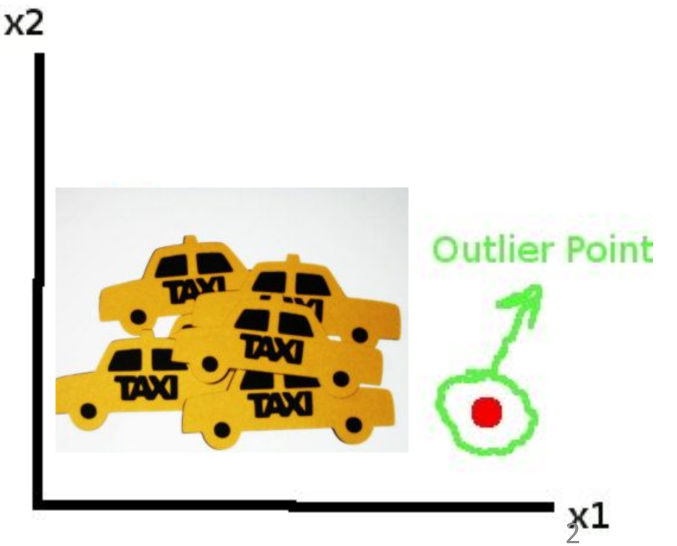
Harish Pullagurla
Hari Krishna Majety
Kenneth Tran



What are Anomalies ??

Anomaly Detection in NYC Taxi Data

Harish Pullagurla
Hari Krishna Majety
Kenneth Tran

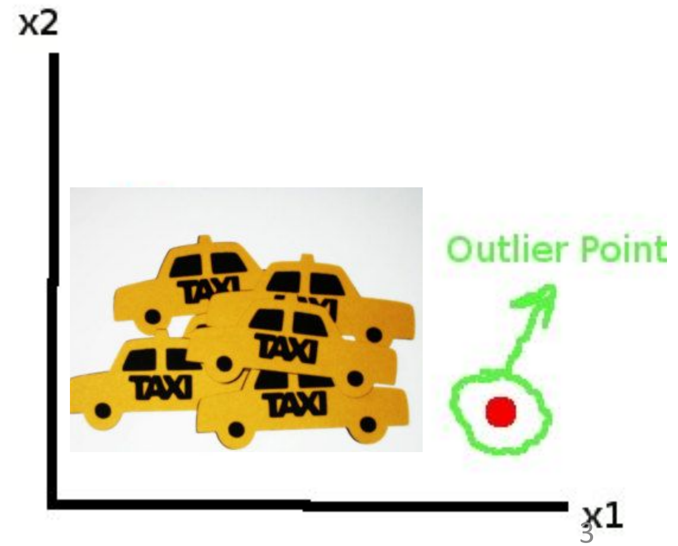


What are Anomalies ??

What is this Data Set about ??

Anomaly Detection in NYC Taxi Data

Harish Pullagurla
Hari Krishna Majety
Kenneth Tran



What are Anomalies ?

- The set of data points that are considerably different than the remainder of the data
- Anomaly is a pattern in the data that does not conform to the expected behaviour
- “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”, (Hawkins 1980)

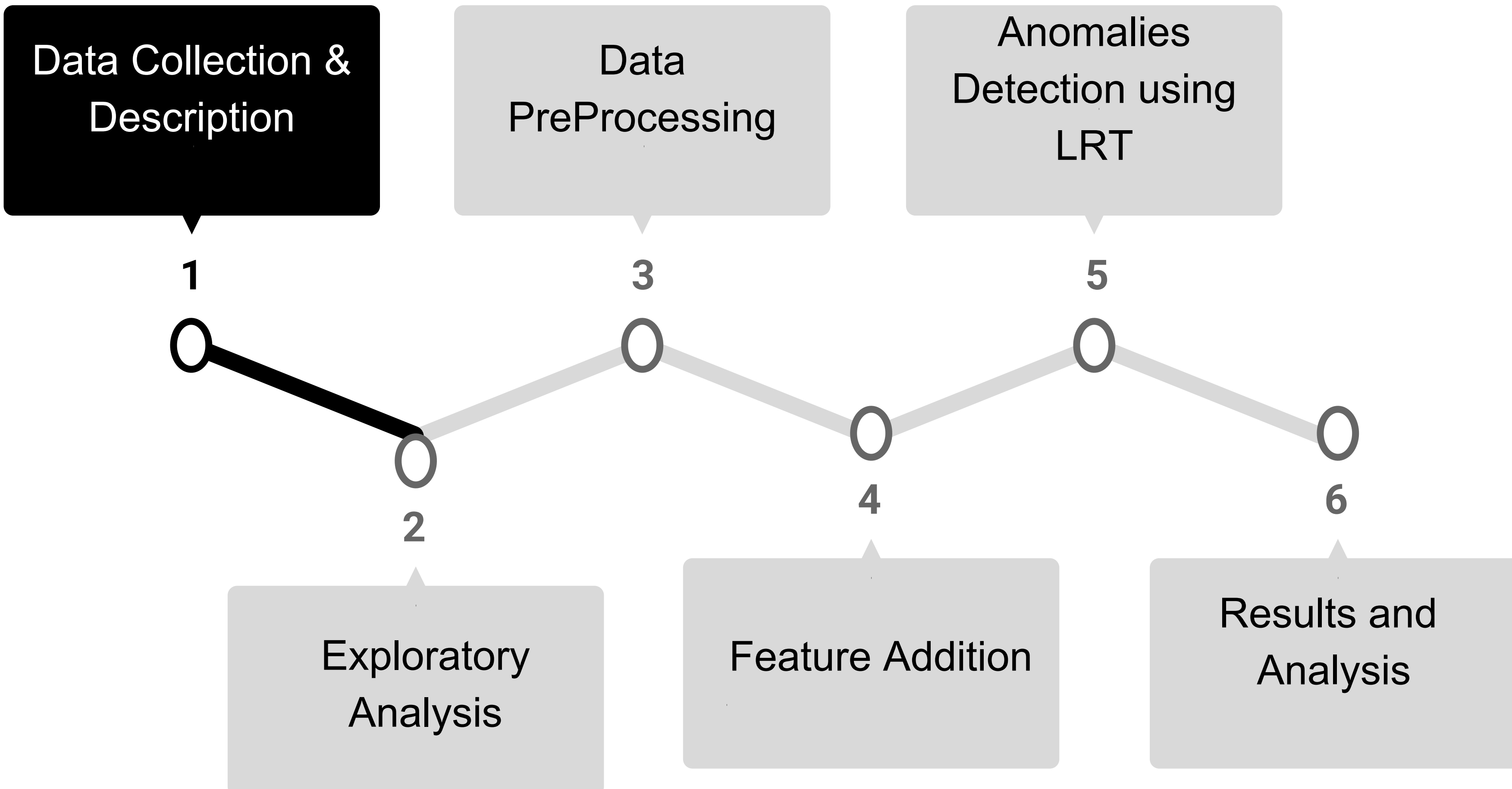
Taken From STDM Lecture Notes

Related Work

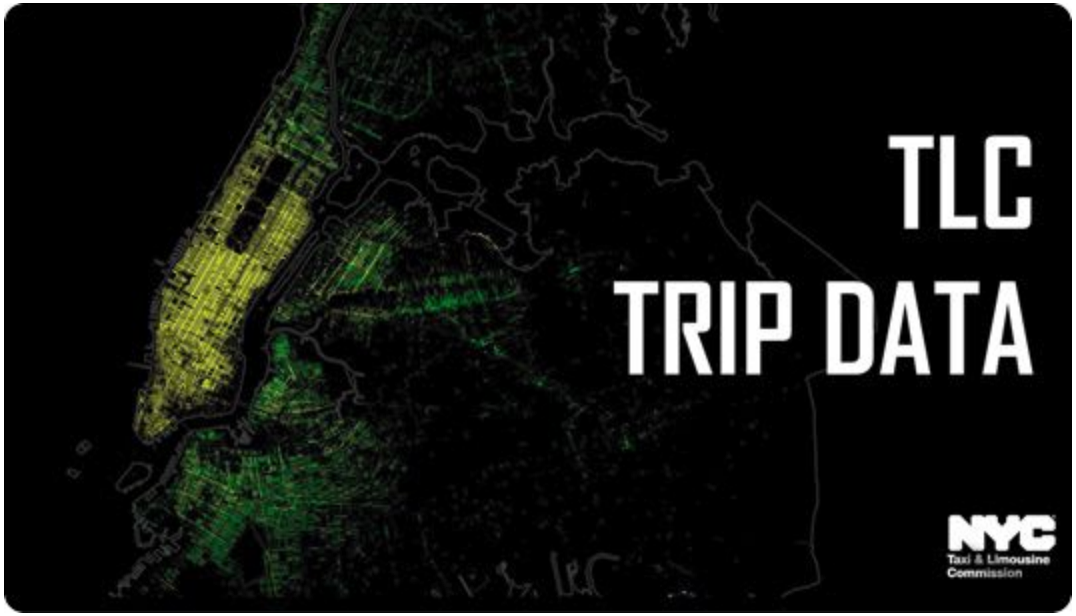
- *Wu, Mingxi, et al. A LRT Framework for Fast Spatial Anomaly Detection. Research Gate, Proceedings of the 15th ACM SIGKDD, Jan. 2009.*
  - Applying LRT to Anomaly Detection
  - Region pruning methods to reduce computation

- *Pang, Linsey Xiaolin, et al. On Detection of Emerging Anomalous Traffic Patterns Using GPS Data. Data Knowledge Engineering, North-Holland,18 May 2013.*
  - Applying Anomaly Detection LRT to specific data sets
  - Includes case study on Beijing taxi data

Pipeline


Data Collection & Description
1
Data PreProcessing
3
Anomalies Detection using LRT
5
2
Exploratory Analysis
4
Feature Addition
6
Results and Analysis

Data Set



Data Set Description

1.5 Million Trip Records
From Jan - July 2016

Data Set Description

1.5 Million Trip Records
From Jan - July 2016

Temporal Attributes - 3 dim
Pick up & drop off Time,
Trip Duration

Data Set Description

1.5 Million Trip Records
From Jan - July 2016

Temporal Attributes - 3 dim
Pick up & drop off Time,
Trip Duration

Spatial Attributes - 4 dim
Latitude , Longitude
Pick up & Drop

Data Set Description

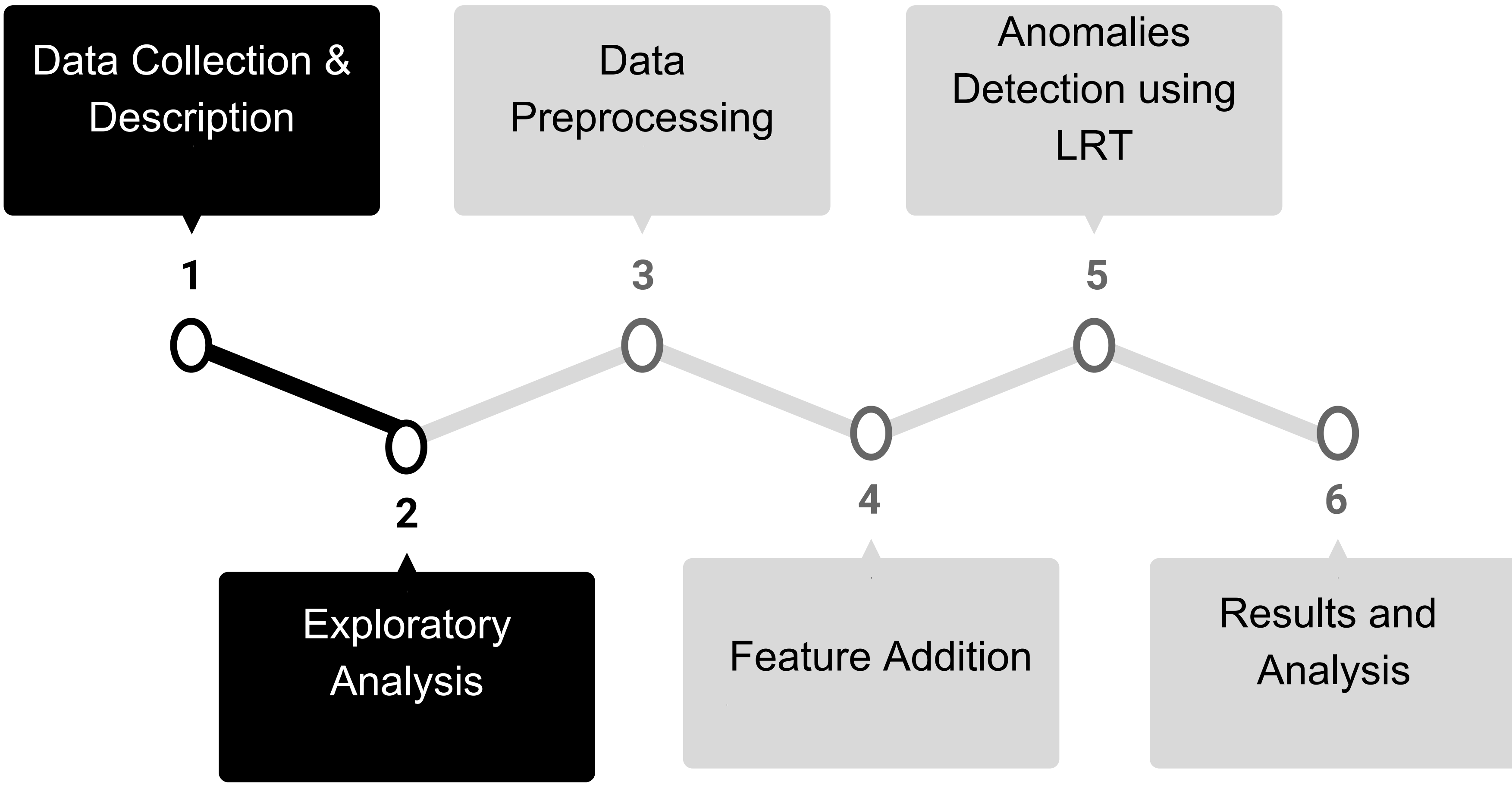
1.5 Million Trip Records
From Jan - July 2016

Temporal Attributes - 3 dim
Pick up & drop off Time,
Trip Duration

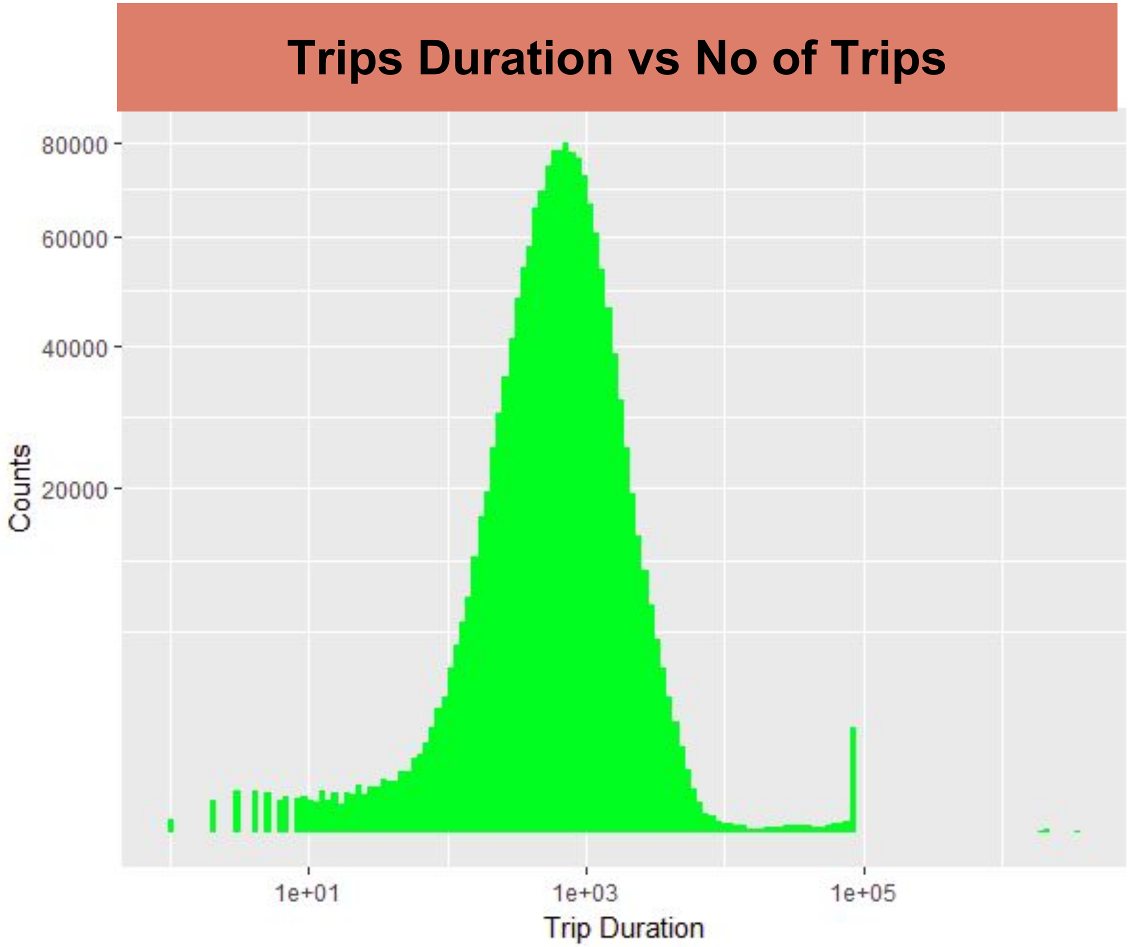
General Attributes - 4 dim
Passenger Count, Vendor Id
Transmission Type

Spatial Attributes - 4 dim
Latitude , Longitude
Pick up & Drop

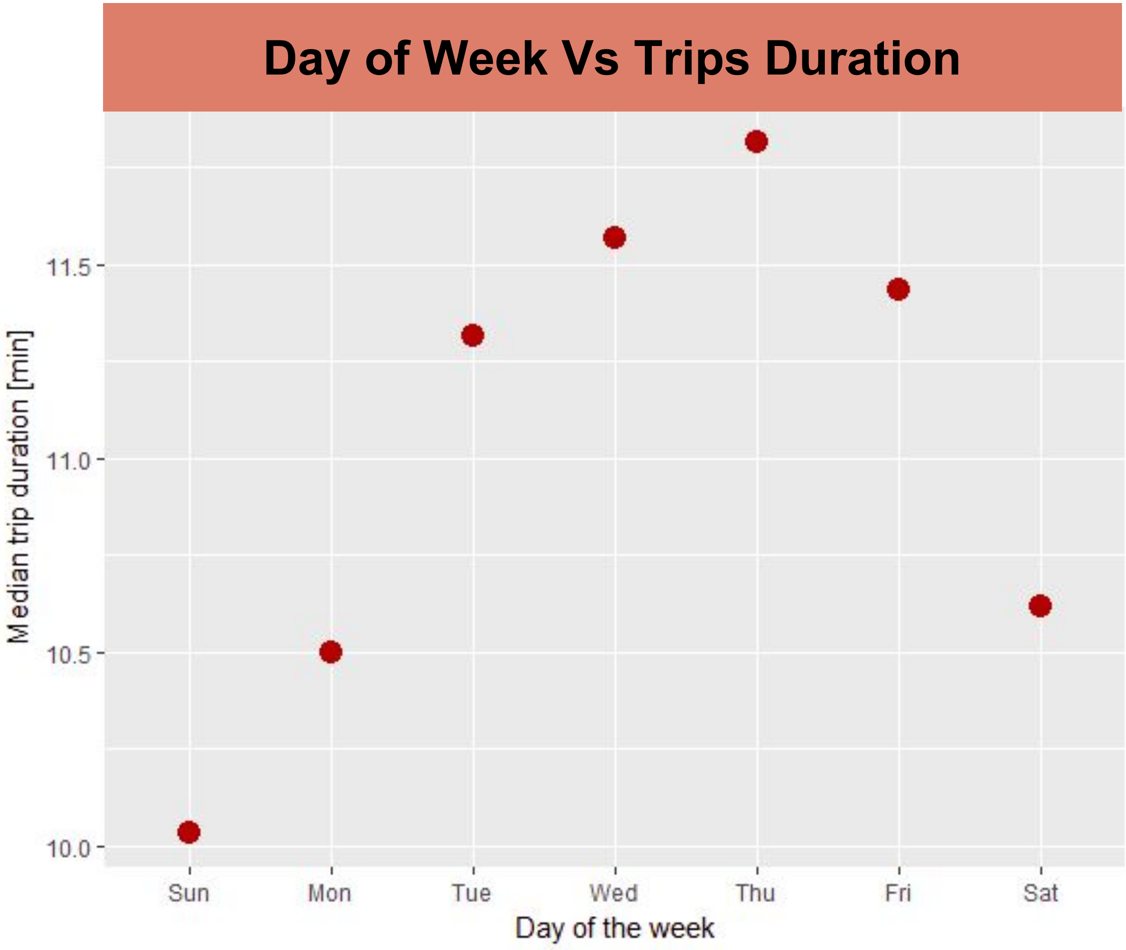
Pipeline


Data Collection & Description
1
2
Exploratory Analysis
Data Preprocessing
3
4
Feature Addition
Anomalies Detection using LRT
5
6
Results and Analysis

Exploratory Analysis

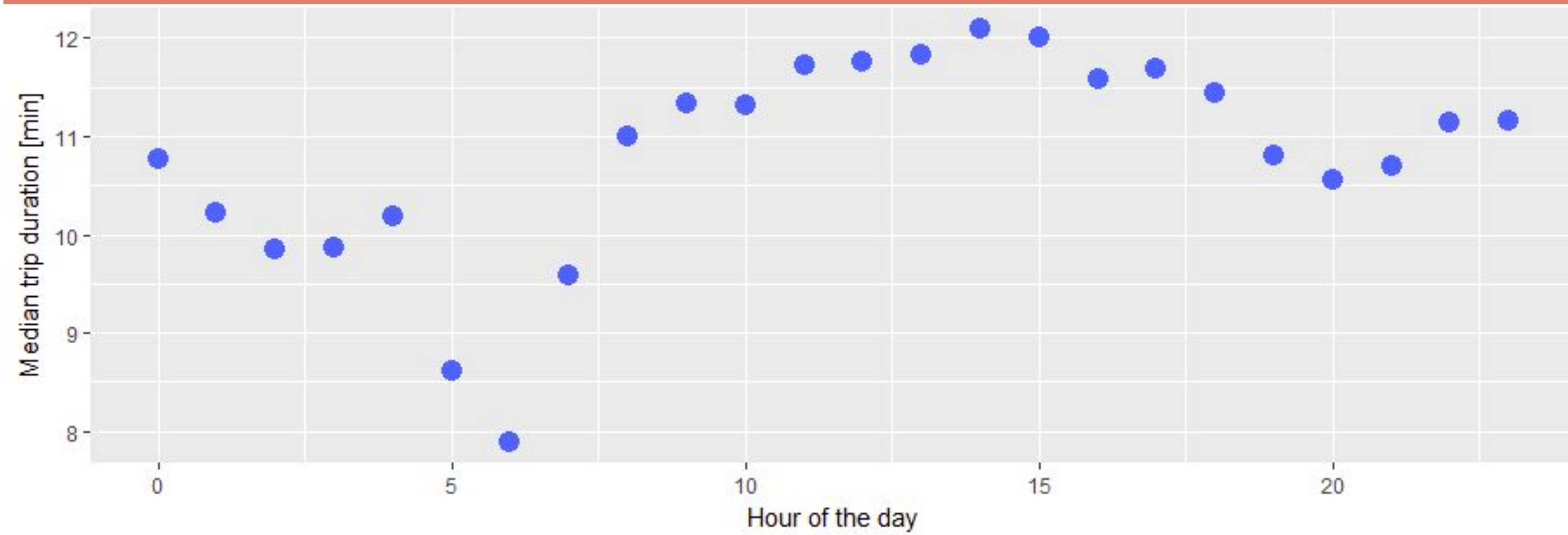


Exploratory Analysis

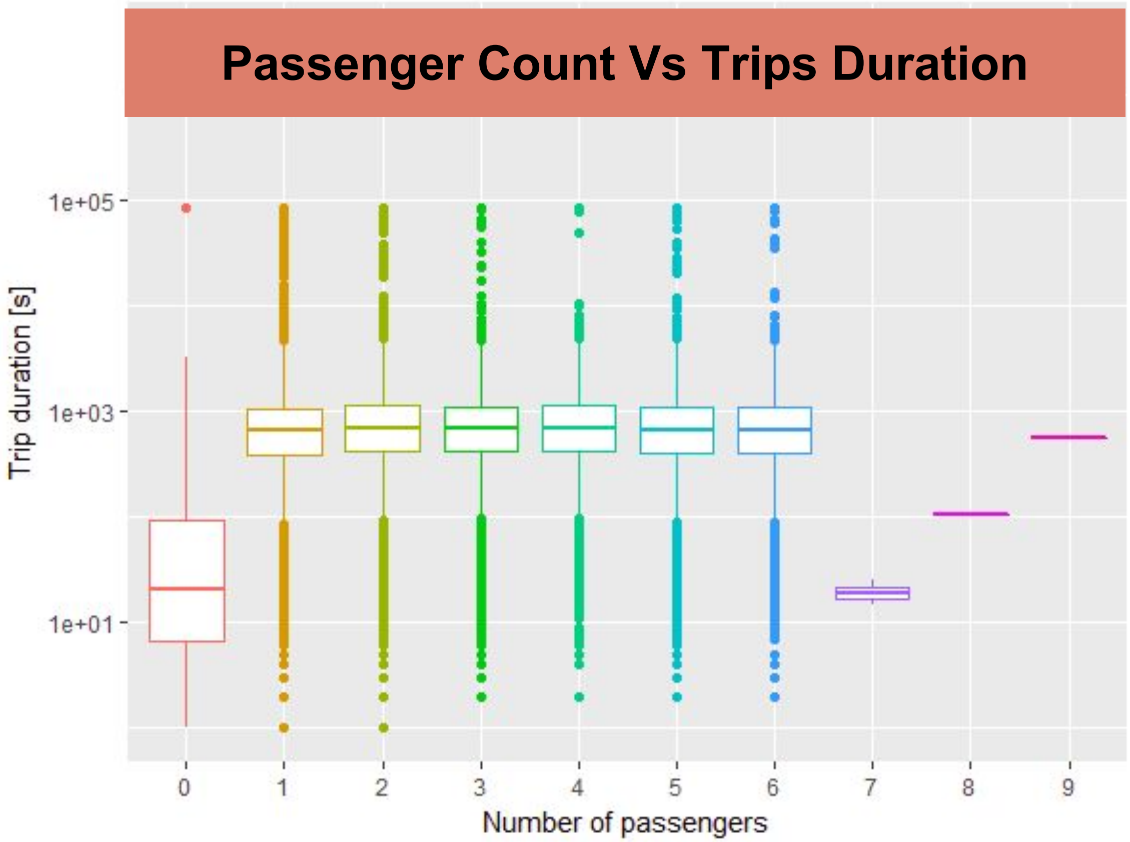


Exploratory Analysis

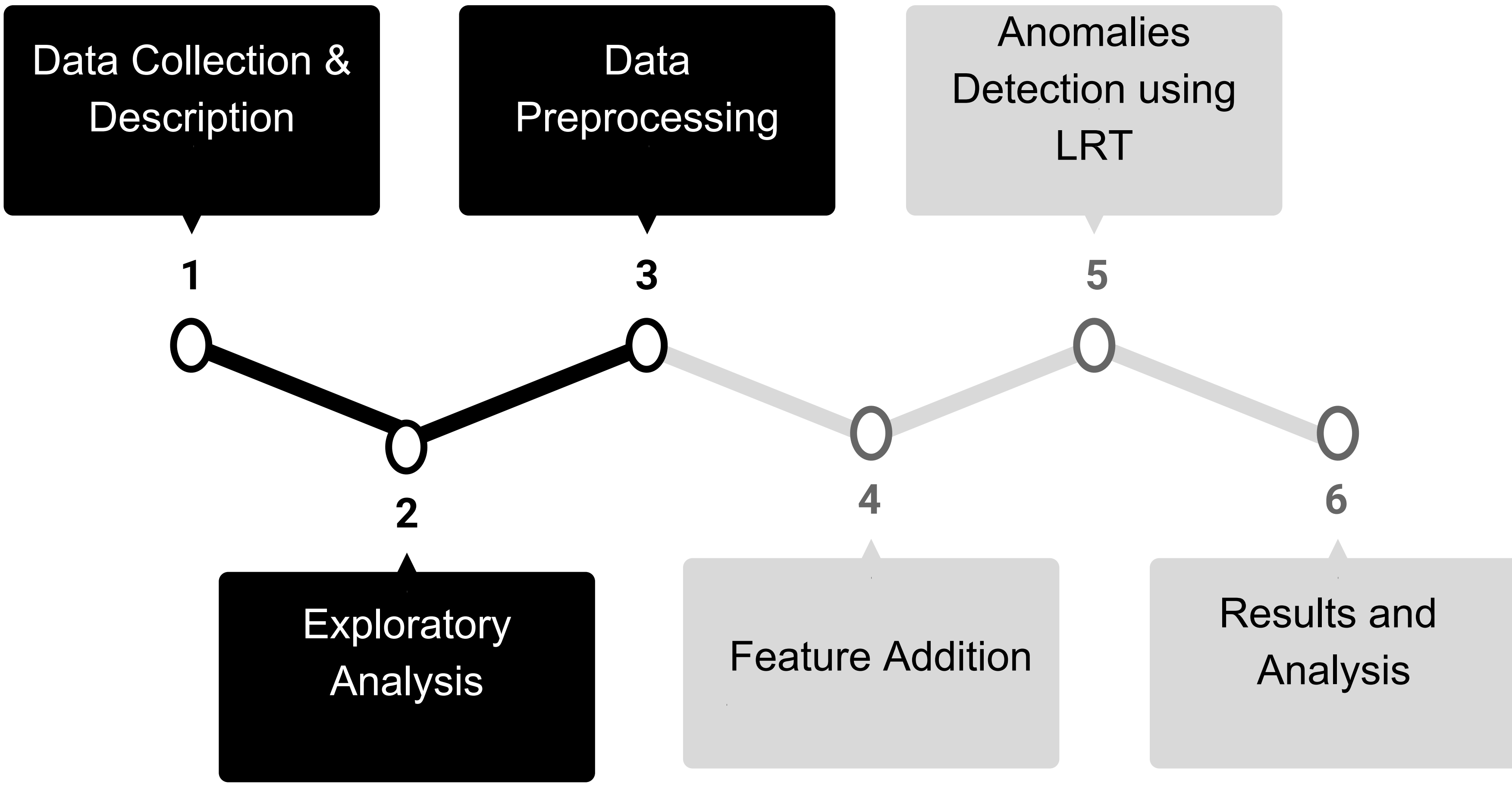
Hour of Day Vs Trips Duration



Exploratory Analysis



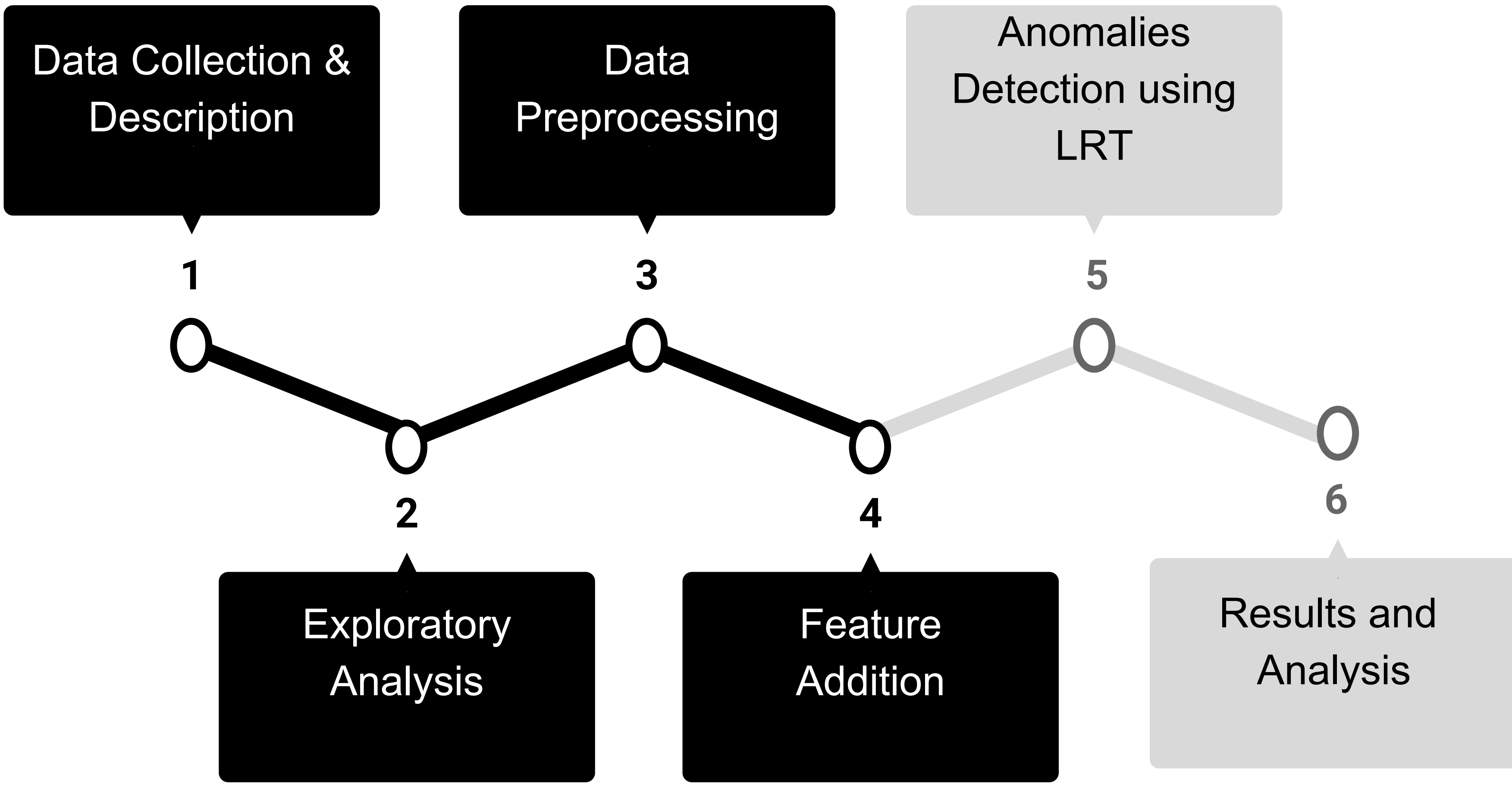
Pipeline


Data Collection & Description
1
Exploratory Analysis
2
Data Preprocessing
3
Feature Addition
4
Anomalies Detection using LRT
5
Results and Analysis
6

Data PreProcessing

- *Following data points were eliminated:*
  - Data points with missing attribute values
  - Duplicates
  - Pick up and drop locations which are not within the New York city limits
  - Data points with passenger counts like 0,7,8,9
  - Trip Durations which are more than 5 standard deviations away from the mean.
- *This reduced the total number of data points from 1.45 million to 1.438 million samples.*

Pipeline


Data Collection & Description
Data Preprocessing
Anomalies Detection using LRT
1
3
5
2
4
6
Exploratory Analysis
Feature Addition
Results and Analysis

Feature Addition

Existing Features

Spatial Attributes - 4 dim
Latitude , Longitude
Pick up & Drop

Temporal Attributes - 3 dim
Pick up & drop off Time,
Trip Duration

General Attributes - 4 dim
Passenger Count, Vendor Id
Transmission Type

New Features

OSRM Data -
Gives real world travel info
like Google Maps

Feature Addition

Existing Features

Spatial Attributes - 4 dim
Latitude , Longitude
Pick up & Drop

Temporal Attributes - 3 dim
Pick up & drop off Time,
Trip Duration

General Attributes - 4 dim
Passenger Count, Vendor Id
Transmission Type

New Features

OSRM Data -
Gives real world travel info
like Google Maps

Date Time
Extract Features such as Day
of Year, week etc

Feature Addition

Existing Features

Spatial Attributes - 4 dim
Latitude , Longitude
Pick up & Drop

Temporal Attributes - 3 dim
Pick up & drop off Time,
Trip Duration

General Attributes - 4 dim
Passenger Count, Vendor Id
Transmission Type

New Features

OSRM Data -
Gives real world travel info
like Google Maps

Date Time
Extract Features such as Day
of Year, week etc

Haversine Distance

Feature Addition

Existing Features

Spatial Attributes - 4 dim
Latitude , Longitude
Pick up & Drop

Temporal Attributes - 3 dim
Pick up & drop off Time,
Trip Duration

General Attributes - 4 dim
Passenger Count, Vendor Id
Transmission Type

New Features

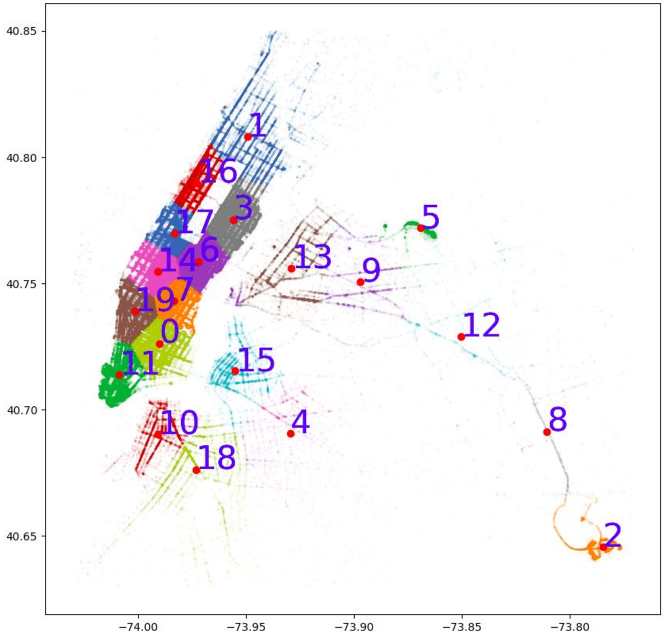
OSRM Data -
Gives real world travel info
like Google Maps

Date Time
Extract Features such as Day
of Year, week etc

Haversine Distance

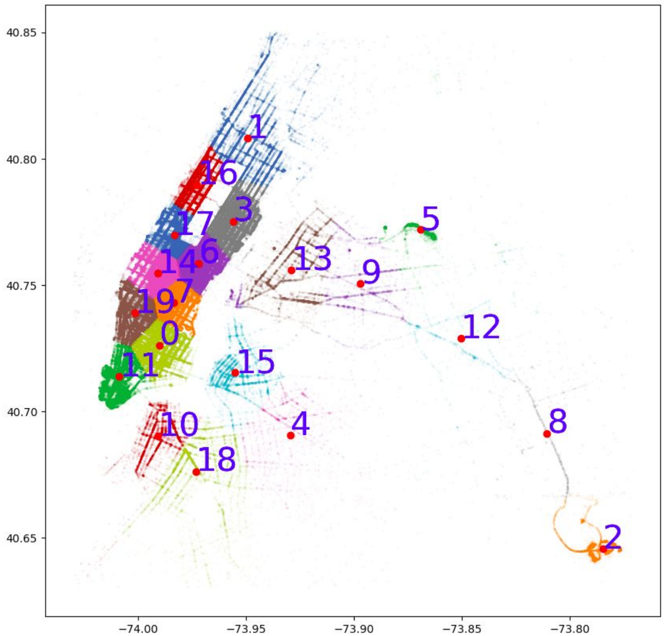
Region Clustering

Region Labeling

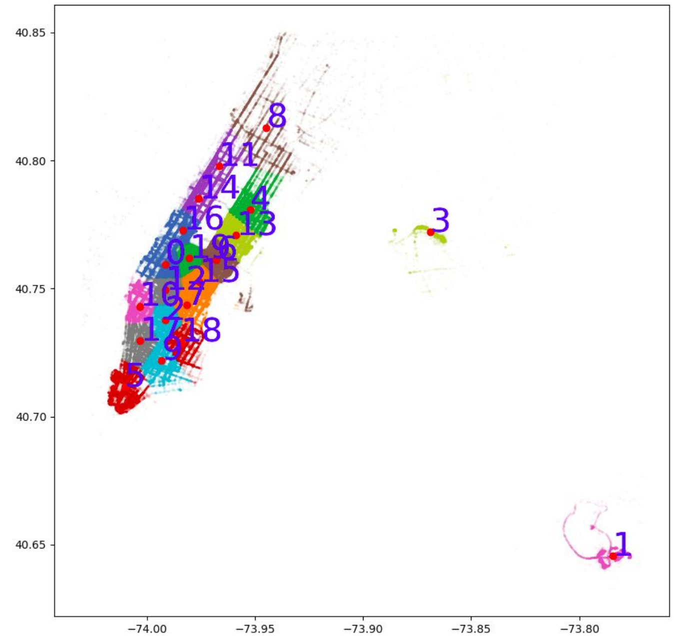


K means Cluster Label Map with Full Data

Region Labeling

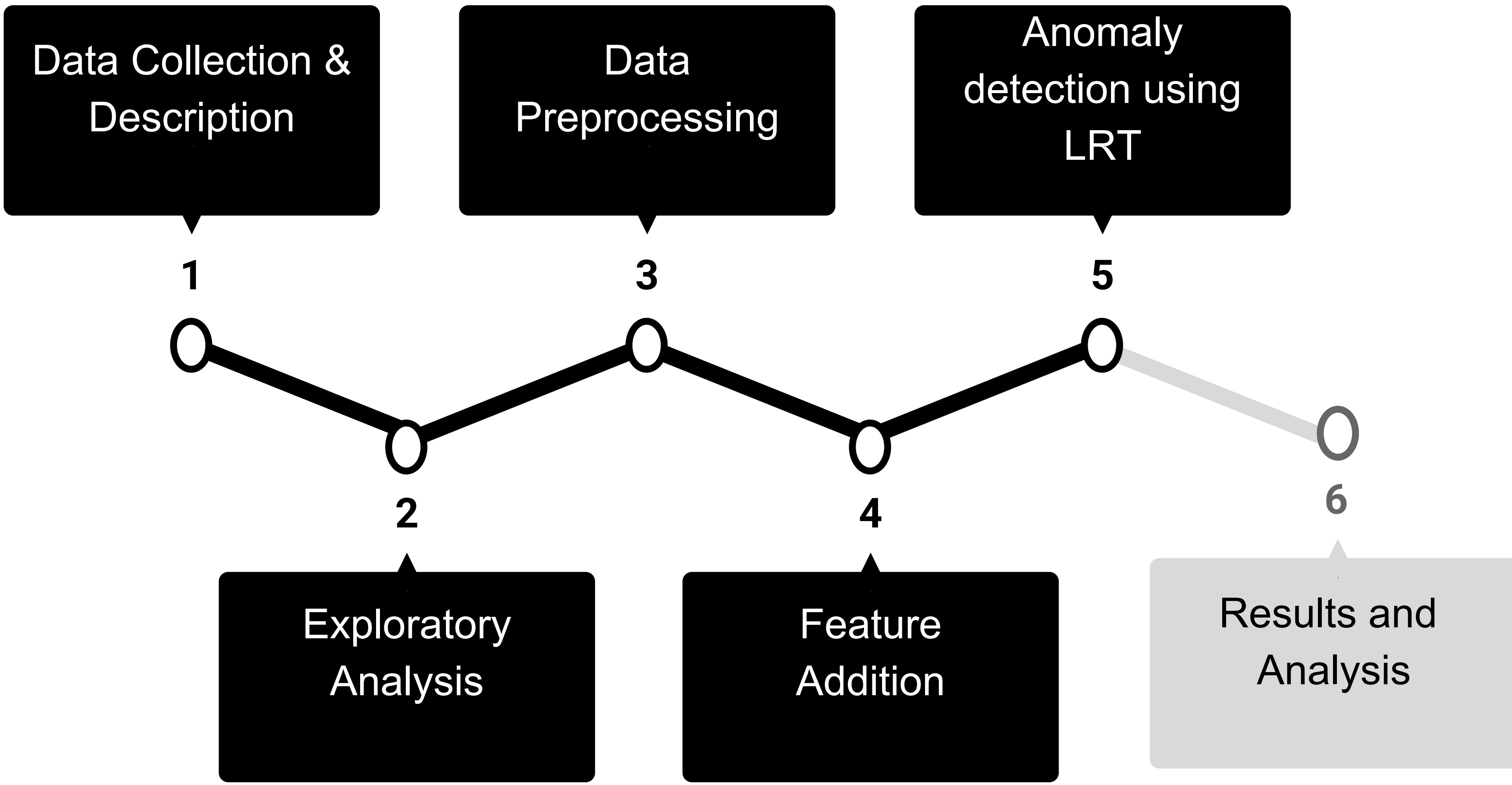


K means Cluster Label Map with Full Data



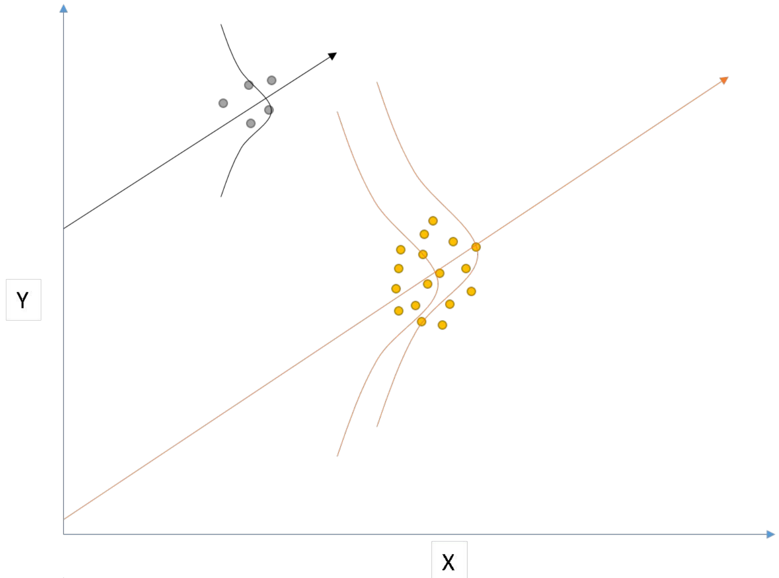
Cluster Label map after removing small clusters

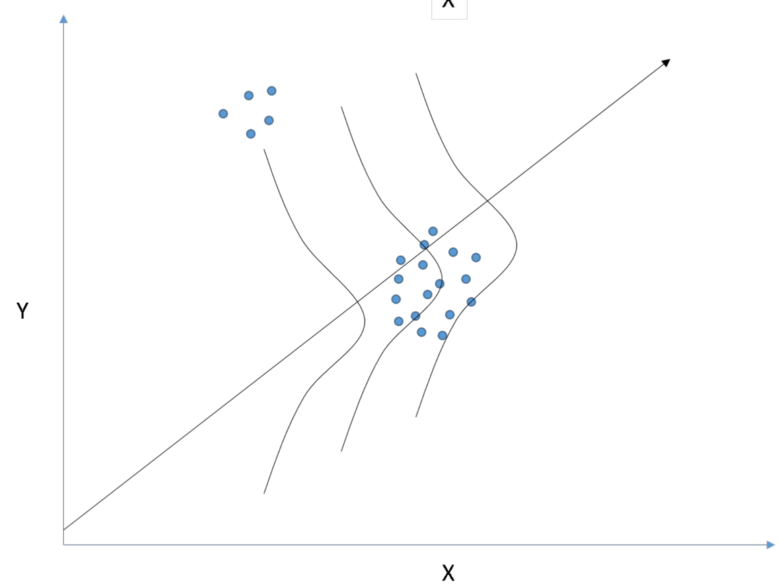
Pipeline


Data Collection & Description
1
Data Preprocessing
3
Anomaly detection using LRT
5
2
Exploratory Analysis
4
Feature Addition
6
Results and Analysis

What is the Likelihood Ratio Test ?

Likelihood Ratio Test





$$\lambda = \frac{L(\theta_s|X_s)L(\theta_{\bar{s}}|X_{\bar{s}})}{L(\theta|X)}$$

Ratio of likelihoods between:

- Product of anomaly specific model and non-anomaly model
- Global model

How do we Model the Data for LRT ?

Generalized Linear Model (GLM)

The model that encompasses a group of regressions including linear regression and logistic regression.

Consists of three parts:

1. Exponential family that predicted value follows
2. Linear predictors (i.e. $b_1x_1 + b_2x_2$)
3. Link function that maps linear predictors to predicted variables

Examples:

Linear Regression: Predicted follows normal with an identity link

Logistic Regression: Predicted follows Bernoulli with a logit (or inverse sigmoid) link

Data Subsampling

Assumption : Data is homogeneous, it models similarly for all subsets

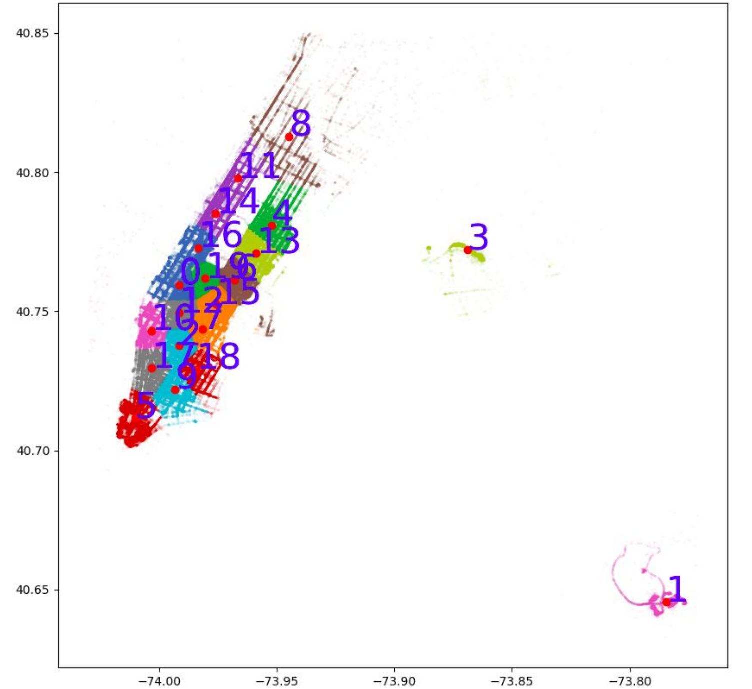
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday

Day Sample

Hour Sample - 8 am - 12 am

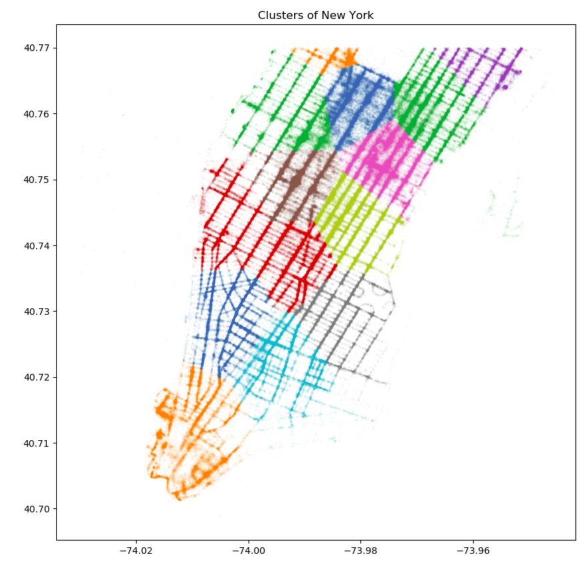
Model 1 - Spatial Anomalies

- Subset data by spatial information
  - Use the K-Means clusters performed in pre-processing
- Model all data using generalized linear model (GLM)
- For each different region, model data within the region and outside of the region, using same GLM configurations
- Compute the likelihood ratio for each region



Model 2 - Temporal Anomalies

- Subset data by Temporal information
  - Consider only Lower Manhattan Region
  - Use variable day of year
- Model all data using generalized linear model (GLM)
- For each different time segment, model data within the time segment and outside of the time segment, using same GLM configurations
- Compute the likelihood ratio for each time segment



Pipeline

1. Data Collection & Description
2. Exploratory Analysis
3. Data Preprocessing
4. Feature Addition
5. Anomaly detection using LRT
6. Results and Analysis

Demo

Results and Conclusion

- *Spatial Analysis*
  - Top two anomalous regions were the JFK and La Guardia Airports in the city

Results and Conclusion

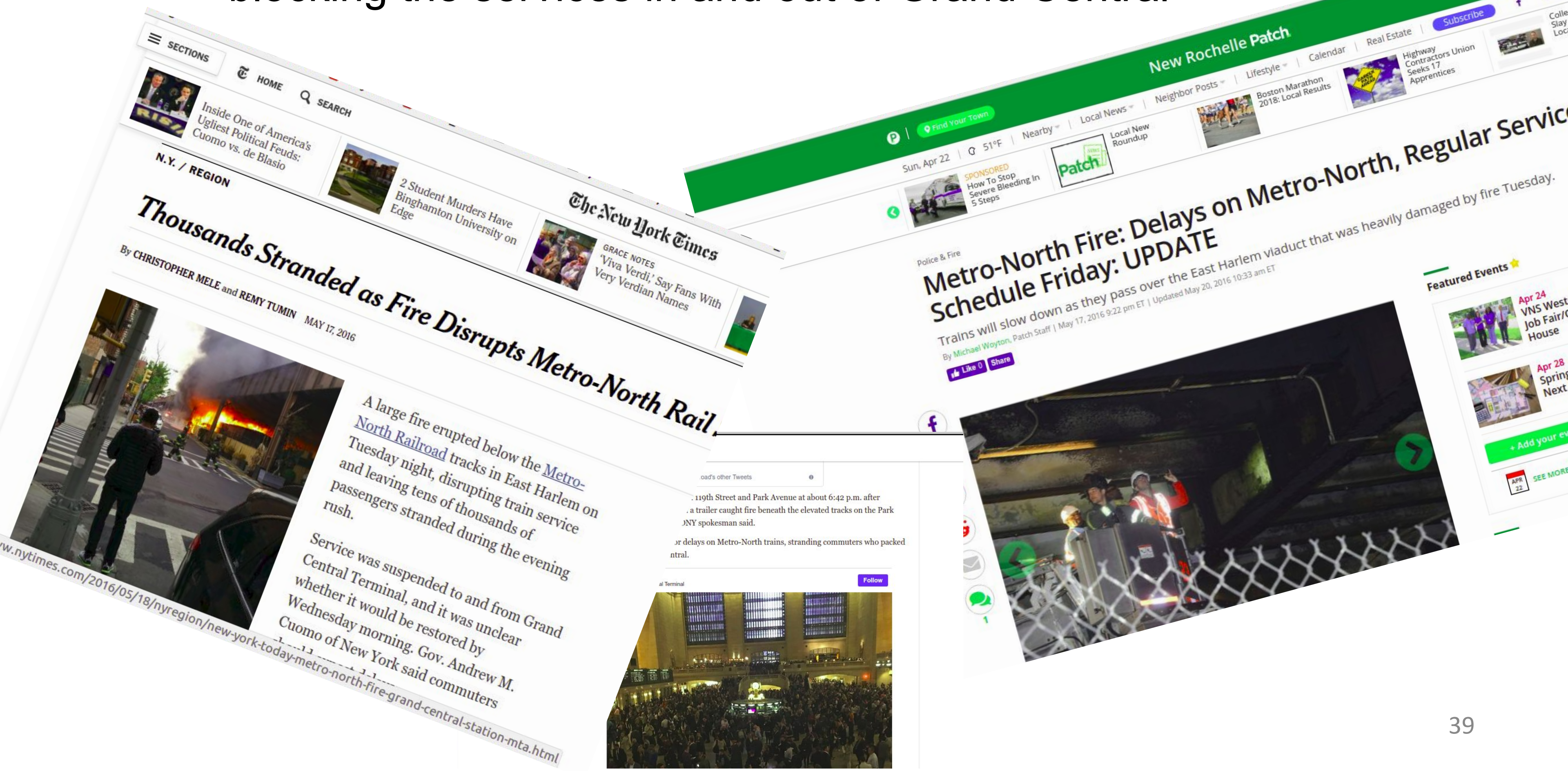
- *Temporal Analysis*
    - Most anomalous - January 26th - After New york was hit by biggest blizzard since 1869.

Results and Conclusion

- *Temporal Analysis*
  - Followed by June 8th - President Obama tapes a show with Jimmy Fallon in Time Square

Results and Conclusion

- *Temporal Analysis*
    - Followed by May 16 - Fire under Metro North track in Manhattan blocking the services in and out of Grand Central



Future Work

- Spatio-temporal analysis(Computationally intensive)
- Considering seasonality during temporal analysis
- Addressing positive correlation between data points and their likelihood values.

Project Learnings

- Different Modeling Strategies - MLE , GLM
- Likelihood Ratio Test and ways to incorporate it with NY Taxi cab data