Discovering Nulls and Outliers

DS-1004 Big Data

Advisor: Prof. Juliana Freire

MINIMIZERS

Diogo Mesquita Mihir Rana Kenil Tanna

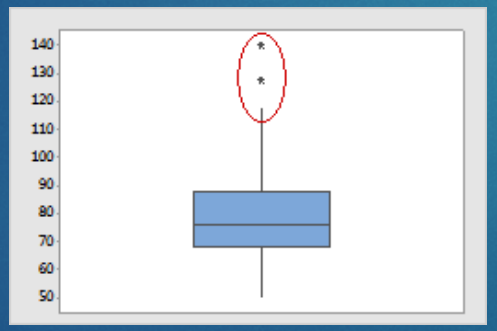
https://github.com/ranamihir/big_data_project

INTRODUCTION

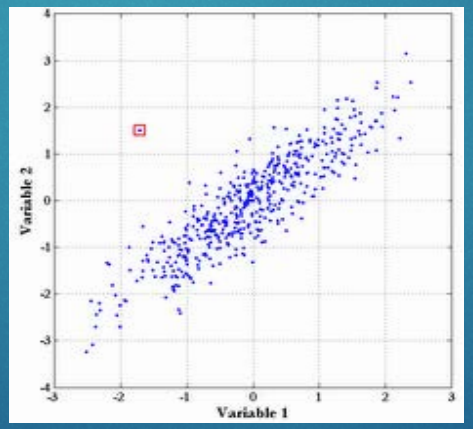
- Problem Statement:
  1. Null Value Detection
  2. Outlier Detection
     - Univariate outliers
     - Multivariate outliers
- Data Set Collection:
  - NYC Open Data
  - 50 data sets

PROBLEM FORMULATION

1. Data Cleaning
   - "$1.99" → 1.99, "1,000" → 1000, 10003 → "10003" (zip code)
2. Missing Value Treatment
   - "None", "N/A", " ", "-", "-999", "999", etc.
3. Outliers

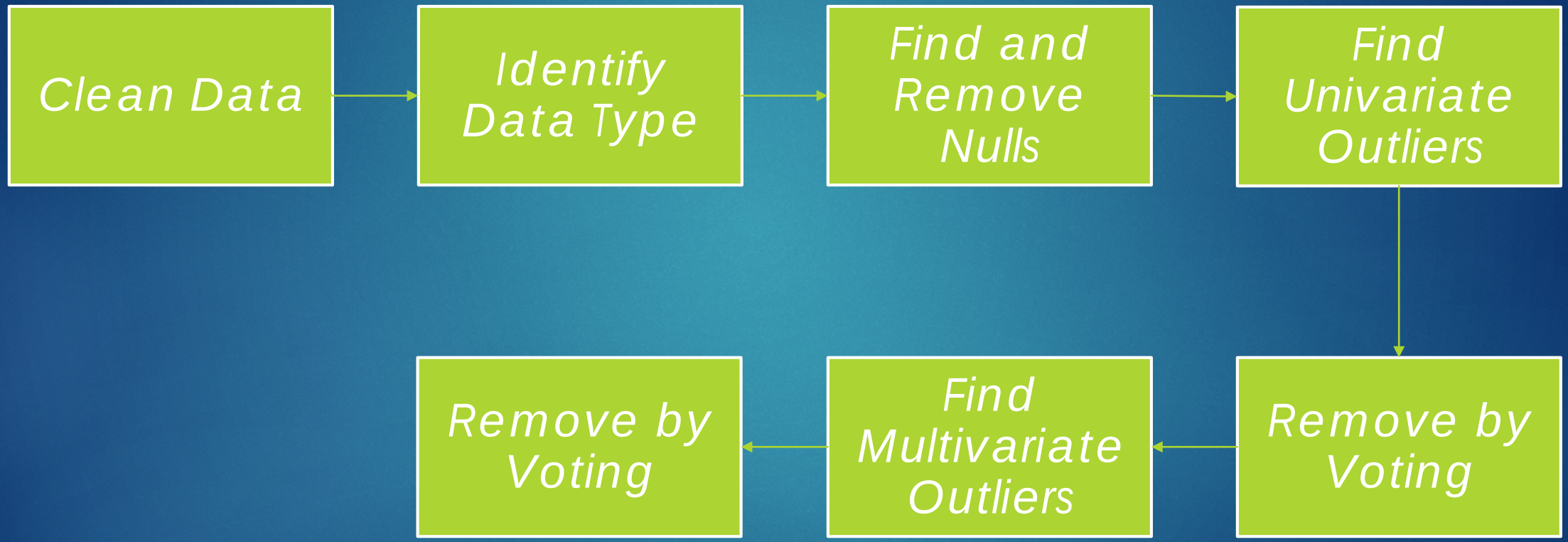


Source: link



Source: link

METHODOLOGY



OUTLIER DETECTION

- *Nearest Neighbor based*
  - *DBSCAN*
- *Clustering based*
  - *k-Means*
- *Mixture of Parametric Distributions*
  - *Gaussian Mixture Models*

OUTLIER DETECTION

- Non-Parametric
  - Histogram / frequency – based
- Statistical Anomaly based
  - Box plot Rule
  - Gaussian model based (z-score)
  - Probabilistic models

KEY STRENGTHS

- End-to-end automated framework
- Box plot Rule at core
  - No input specific to particular data set / column required
- Robust
  - Multiple techniques optimizing different metrics
  - Voting / Intersection of multiple similar techniques
- Efficient
  - Remove univariate outliers before finding multivariate ones

RESULTS

- Data Cleaning
  - "$1.99“ → 1.99, "1,000“ → 1000, 10003 → “10003” (zipcode)
- Missing Value Treatment
- Outliers

REFERENCES

1. *Jason Brownlee. 2016. How To Handle Missing Values In Machine Learning Data With Weka. (Jun 2016). https://machinelearningmastery.com/ how-to-handle-missing-values-in-machine-learning-data-with-weka/*
2. *Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2007. Anomaly Detection: A Survey. (2007).*
3. *Google. 2018. Locate Outliers, Google Cloud Dataprep Documentation, Google Cloud. (2018). https://cloud.google.com/dataprep/docs/html/Locate-Outliers_ 57344572*
4. *Ming Hua and Jian Pei. 2007. Cleaning Disguised Missing Data: A Heuristic Approach. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07). ACM, New York, NY, USA, 950–958. https://doi.org/10.1145/1281192.1281294*
5. *Ming Hua and Jian Pei. 2008. DiMaC: a disguised missing data cleaning tool. In KDD.*
6. *Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. 2000. Efficient Algorithms for Mining Outliers from Large Data Sets. SIGMOD Rec. 29, 2 (May 2000), 427–438. https://doi.org/10.1145/335191.335437*