CITY PROFILING BY BIG DATA ANALYSIS

Ritika[1], Avinandan Mitra[2], Divyanshu Tiwari[3], Shubham Kumar Mandal[4], Dr. Manjusha Pandey[5], Siddharth Swarup Rautaray[6]
[1,2,3,4,5,6] School Of Computer Engineering
[1,2,3,4,5,6] Kalinga Institute Of Industrial Technology
Bhubaneswar,Odisha,India
[1] ritikaa1518@gmail.com ,[2] avi.mitra.59@gmail.com,
[3]divyanshutiwari3016@gmail.com ,[4]mandalshubham2019@gmail.com,
[5] manjushafcs@kiit.ac.in, [6]siddharthfcs@kiit.ac.in

ABSTRACT:
All throughout the world, cities are trying to become smart cities.
The large inflow of data from different sectors of the city needs to be stored and processed in such a way thatit would provide valuable insights which in turn would help in transforming and developing the cities.

Data Analysis is the process of cleaning,analyzing,interpreting,and visualizing data using various techniquesand business intelligence tools.Here, the data collected from various government records are cleaned,analyzed and interpreted to provide necessary information.

The process of converting unprocessed data into comprehensible representations is known as data pre-processing. Initially,the data obtained were in unorganized manner and scattered around,these data are then organized and the missing values and data are required to be filled up.

Advantages of this paper are that this will help in growth and development of different government department by providing trends and patterns in different sectors of city like traffic,water management,slumpopulation etc and will thereby help in developing them.

INTRODUCTION:

DATA: Cities all across the world are aggressively working to become "smart cities," utilizing technology to improve sustainability, efficiency, and the general standard of living for citizens. A key component of this development is the handling of the enormous amount of data that is produced by different industries in the city. This data must be carefully processed and kept, spanning from energy and transportation use to infrastructure and public services. Cities are able to extract useful information from this data mine by utilizing AI and advanced analytic. These insights form the basis for well-informed decision-making, empowering local authorities to take advantage of new possibilities and handle urgent concerns. In the end, this data-driven strategy promotes innovation, resilience, and prosperity for the present and the future by assisting in the comprehensive development and transformation of cities.

DATA ANALYSIS: Using a wide range of methods and business intelligence technologies, data analysis is a complex process that includes careful cleaning, in-depth analysis, perceptive interpretation, and efficient visualization of data. This refers to the methodical cleansing and processing of data gathered from diverse sources in the context of government records. Patterns, trends, and anomalies in the data are found through thorough research, offering priceless insights into public services, societal trends, and governmental operations. These insights—obtained through advanced analytical techniques—are then evaluated to elicit relevant data, which helps decision- and policy-makers make well-informed choices. Furthermore, these results are presented in an understandable and user-friendly way through the use of visualization tools, which aid in understanding and distribution to pertinent parties.In the end, data analysis is essential to maximizing the use of public records to support evidence-based decision-making, promote organizational effectiveness, and improve the provision of public services.

DATA PREPROCESSING: Transforming unprocessed data into an organized and understandable format is a crucial stage in the data pre-processing process. The data is frequently jumbled and fragmented at first, with little clarity or consistency. This data is methodically arranged and formatted by pre-processing, guaranteeing that it is prepared for additional analysis. In order to guarantee that the dataset is comprehensive and representative, filling in missing values and incomplete data points is an essential part of this procedure. Organizations can improve the quality and dependability of their datasets by data pre-processing, which paves the way for more precise and perceptive analysis, decision-making, and strategic planning.

ADVANTAGES: The benefits of this paper are found in its ability to spur growth and development in a number of government agencies by providing insightful information on patterns and trends in several areas of the city, including slum populations, water resources, and traffic management. Through an analysis and interpretation of the data gathered from various sectors, the paper can reveal important information that can guide strategic efforts and decision-making processes targeted at improving urban development. Equipped with this understanding, government agencies can formulate focused initiatives and regulations to effectively tackle issues, maximize resource distribution, and enhance the provision of services to the public. In the end, this document is a useful instrument for promoting innovation and advancement, encouraging sustainable development, and raising metropolitan regions' general standard of living.

TABLE 1

STATE OF ART: The following table represents the state of Work done for different smart city based researches for Big Data from a period of 2015-2023

AUTHOR/YEAR	TITLE	PURPOSE	ALGORITHM USED	REVIEW
Chiehyeon Lim,Paul P. Maglio,Kwang-Jae Kim (2018)[1]	**Smart cities with big data: Referencemodels, challenges,and considerations**	We examine and group application cases for urban data into four reference models. Six obstacles are found in the process of t urning data into information for smart cities. To overcome the difficulties in putting t he models into practice, we offer five points to think about. Based on four government-sponsored action research initiatives, the difficulties and factors are discussed. In an economy rich in data, our insights can help with policy formation and urban planning.	• Finding the reference models can be accomplished with the help of the case analysis method. • An effective strategy for empirically comprehending the issues and concerns is the practice-driven action research approach. • Together, the two approaches created an integrated framework of reference models, difficulties, and factors to take into account.	This paper's primary contribution is the frameworks and knowledge that are developed for the usage of data in smart cities using this application-oriented viewpoint. Using the information gathered from them, the proposed classification approach offers four reference models to add value for local government, businesses, tourists, and residents.
Yunhe Pan,Yun Tian,Xiaolong Liu, Dedao Gu , Gang Hua (2016)[2]	**Urban Big Data and the Development of City Intelligence**	This research defines urban big data and examines its characteristics and uses in China's city intelligence. In order to emphasize and contrast China's unique definition and model for city intelligence, the contrasts between China's city intelligence and other nations' "smart city" concepts are compared in this study.	Important information is gathered via correlating, integrating, cleaning, processing, analyzing, mining, and displaying the enormous volumes of data. gathering, analyzing, and exchanging data In order to guarantee the steady and dependable functioning of the urban big data service architecture, a big data operation assurance mechanism is created.	The creation of city intelligence, which is the perfect starting point for urban growth, is greatly aided by urban big data. Successfully managed and opened urban big data will foster the growth of an urban knowledge-based services sector, open up new markets and commercial prospects, and advance the development of city intelligence.
Jeannette Chin; Vic Callaghan; Ivan Lam (2017)[3]	**Understanding and personalising smartcity services using machine learning, The Internet-of-Things and Big Data**	• This research defines urban big data and examines its characteristics and uses in China's city intelligence. In order to emphasize and contrast China's unique definition and model for city intelligence, the contrasts between China's city intelligence and other nations' "smart city" concepts are compared in this study.Using machine learning approaches to associate weather conditions with short-distance bike trips, the study investigated how Al may leverage IoT and Big Data to promote the development of personalized services in Smart Cities. • The goal of the study was to determine whether weather-related factors and short-cycling behavior are correlated. Four well-known machine learning classification algorithms were used, and the data came from six different datasets.	• Naïve Bayesian Classifier • J48 Tree Classifier • Nearest Neighbour Classifier	The findings suggest that developers of smart city technologies and services have a lot of promise when combining ML, IoT, and big data.
Eleonora D'Andrea,Pietro Ducange, DaniloLoffreno Francesco Marcelloni and Tommaso Zaccone (2018)[4]	**Smart Profiling of City Areas Based onWeb Data**	The framework for describing and profiling city areas using data from websites and online services is presented in the article. Local news, traffic information, events happening in the city, lifestyle, and human behaviors are among the points of interest (restaurants, services, hotels, schools, churches, stores, wi-fi access points, etc.) that are dispersed across the city. The framework enables the selection of various data sources, data preparation, significant feature extraction, clustering algorithm execution to identify the profiles of individual city areas, and visualization of the outcomes on a city map. The creation of a virtual grid of squared cells on the city serves as the foundation for the area definitions.	k-means clustering The current implementation of the proposed framework includes only clustering algorithms for profiling city areas.	outlined the outcomes obtained from a draft version of the framework—which is currently undergoing development—for grouping and characterizing the sections of a city that are located utilizing a virtual grid. Using information about POIs that are publicly available online and that can be obtained from websites and web services, one can use it todefine the areas.

Ibrahim Abaker Targio Hashem, VictorChang, Nor Badrul Anuar , Kayod e Adewole, Ibrar Yaqoob, Abdullah Gani, Ejaz Ahmed, Haruna Chiroma (2016)[5]	**The role of big datain smart city**	The purpose is to: ● give an example of big data analytics in the context of smart cities. ● A future business model that aims to manage massive data for smart cities is offered. ● Determine and talk about the difficulties in business and technology research. ● We offer an overview of the current communication technologies utilized in smart cities.	● The Internet of Things (IoT) facilitates the smooth communication of sensors and actuator devices in smart city environments and allows for more convenient cross-platform information sharing. ● The current implementation of the proposed framework includes only clustering algorithms for profiling city areas ● The interconnectivity of sensing and actuating equipment is emphasized in smart cities, which makes it possible to share information across platforms using a single framework. Cloud computing serves as the unifying platform for data analytics, information representation, and seamless omnipresent sensing to facilitate such sharing.	The outcome suggests that massive data from IoT devices is crucial for the design of smart cities.
Kamran Soomro, Muhammad Nasir Mumtaz Bhutta, Zaheer Khan, Muhammad A.Tahir (2019)[6]	**Smart city big data analytics: An advanced review**	In this paper, the authors present a : ● Thorough examination of the research on big data analytics for smart cities. ● categorization model that examines four facets of this field's research. ● gap analysis and determine potential study areas.	Big data analytics solutions are becoming more and more in demand as ICT plays a role in enabling and supporting smart cities. Numerous AI, data mining, machine learning, and statistical analysis-based solutions have been effectively implemented in a variety of topic areas.	Big data modeling and analysis aid in the creation of smart cities.
M.Mazhar Rathore, Anand Paul, WonHwa Hon g , HyunCheol Seo, Imtiaz Awan, Sharjil Saeed (2018)[7]	**Exploiting IoT andbig data analytics: Defining Smart Digital City using real-time urban data**	The purpose is to: ● Data generation, collection, preprocessing computing, aggregation, filtering, classification, and decision-making. ● Using real-time urban data, a smart IoT-based digital city is implemented. ● Using the Hadoop ecosystem, big data analytics is being used to plan cities.. ● Large-scale graph processing for alerts and traffic data. ● The system's scalability and real-time data processing are assessed.	To build a Smart Digital City, utilize IoT devices, Big Data analytics, and graph processing technology. Preprocessing, analysis, decision-making, and data collection are all included, with an emphasis on scalability and real-time processing capabilities.	Big Data, IoT, and cloud infrastructures are contactless technologies for smart cities. Making decisions based on data:The system gathers and analyzes real-time data from multiple sources (traffic cameras, Internet of Things sensors, etc.) to offer insights that may be applied to enhance resource allocation, city planning, and overall efficiency. Scalability: Big Data, or vast amounts of data produced by Internet of Things (IoT) devices in smart cities, are best handled by the Hadoop environment.Real-time processing: Within the Hadoop ecosystem, frameworks such as Apache Spark allow real-time data streams to be processed for applications such as traffic control.
Mama Nsangou Mouchili,Shadi Aljawarneh,Wette Tchouati (2018)[8]	Smart city data analysis	The purpose is to: highlight the significance of smart cities and the ways that technology, in particular machine learning and big data analytics, may be used to solve urban problems and enhance city development and management.	● Utilizing a blend of data analytics and machine learning algorithms, one can examine past data, forecast future actions, and arrive at well-informed conclusions regarding parking and traffic control. ● The current implementation of the proposed framework includes only clustering algorithms for profiling city areas ● Large-scale dataset processing and analysis are made easier by the usage of big data technologies like Elasticsearch and Hadoop, and a variety of machine learning algorithms for data mining and predictive modeling are available in Python packages.	The outcome suggests that massive data from IoT devices is crucial for the design of smart cities.
Zaheer Khan,Ashiq AnjumKamran Soomro & MuhammadAtif Tahir (2015)[9]	**Towards cloud based big data analytics for smart future cities**	The purpose is to: ● Provide a cloud-based analytical service architecture and compare various data processing infrastructures to further the understanding and progress of smart city data analytics. ● The purpose of the article is to illustrate how big data analytics can be used to address urban issues and support well-informed decision-making in smart cities.	A mix of statistical analysis, machine learning methods, big data processing frameworks, data pretreatment, and visualization tools to evaluate large- scale urban data and extract insights tohelp with smart city initiative decision-making.	Big data analytics combined with cloud-based services enable the provision of smart city applications.

METHODOLOGY:CITY PROFILING

DATA DESCRIPTION:

Understanding the data is essential before beginning any study when using big data for city profiling. Recapitulating and comprehending the features of the dataset are part of the data description process. The steps for describing data in big dataanalysis for city profiling are as follows:

Data Collection: Gathered information on cities from pertinent data sources. These sources may consist of information from satellite photography, social media, government databases, and sensors.In the purpose of this article, the data was obtained fromthe https://data.gov.in/ website, from which we gathered several datasets derived from various catalogs. Bhubaneswar, Odisha, was selected as the study city.

Data Integration: Combined data from different sources into a single dataset, ensuring that the data is in a consistent format andcan be easily analyzed together.

Data Cleaning: Cleaned the data to address any issues such as missing values, inconsistencies, errors, or outliers. This step iscrucial for ensuring the quality and reliability of the analysis.

Exploratory Data Analysis (EDA): carried out exploratory data analysis to learn more about the dataset. This include creating info graphics, summary statistics, and variable connection analysis. You could want to examine a range of characteristics for cityprofiling, including environmental elements, infrastructure, crime rates, economic indicators, and population demographics.

Descriptive Statistics: computed the descriptive statistics for numerical variables, including mean, median, mode, standarddeviation, range, etc. You can calculate the frequencies and percentage of each category for categorical variables.

DATA PREPROCESSING:

A CSV file containing city profile data usually needs to go through a number of processes of data preprocessing in order to cleanand get it ready for modeling or analysis. Here's a quick rundown:
Data loading: With the use of a pandas library, read the CSV file into a Data Frame.

Managing Missing Values: The missing values in the dataset were identified and their handling was determined. Using moresophisticated methods like predictive imputation, or deleting rows with missing values and imputing them using the mean, median, and mode are some of the options.

Coding Categorical Variables: You might need to numerically encode categorical variables in the data using methods like one-hot encoding or label encoding if the data contains them, such as city names or categories.

Feature Scaling: We took into consideration scaling the numerical features with varying scales in order to guarantee that each feature makes an equal contribution to the study. Standardization, which involves scaling to a zero mean and unit variance, andnormalization, which involves scaling characteristics to a range between 0 and 1, are common scaling techniques.

DATA CLUSTERING:

In data analysis, the process of grouping a dataset into clusters—groups of data points that are more similar to one another than tothose in other clusters—is known as data clustering. Within the report's CSV file including city profile data, clustering may be useful in spotting trends or commonalities amongst cities according to different characteristics like population, income, crime rate, etc.
The data can be loaded, preprocessed, clustered, and visualized using Python libraries like matplotlib, scikit-learn, and pandas. For instance, you can load and preprocess the data using pandas, apply clustering methods using scikit-learn, then visualize theclustered data using matplotlib.

Read the CSV file into a pandas DataFrame to load the data.

Select Features: Decided which features (columns) of the dataset to use for clustering. These features should be relevant to theclustering task and could include population density, median income, crime rate, etc.

Choose a Clustering Algorithm: Selected a clustering algorithm suitable for your data. Common algorithms include K-means,hierarchical clustering, and DBSCAN.

Fit the Model: Apply the chosen clustering algorithm to the selected features of the dataset.

Analyze Results: Visualized the clustered data to understand the underlying patterns and relationships between cities. This couldinvolve scatter plots, heatmaps, or other visualization techniques to display the clusters and their characteristics.

Interpretation: Interpreted the results of clustering to gain insights into the city profiles and identify any meaningful clusters orgroups

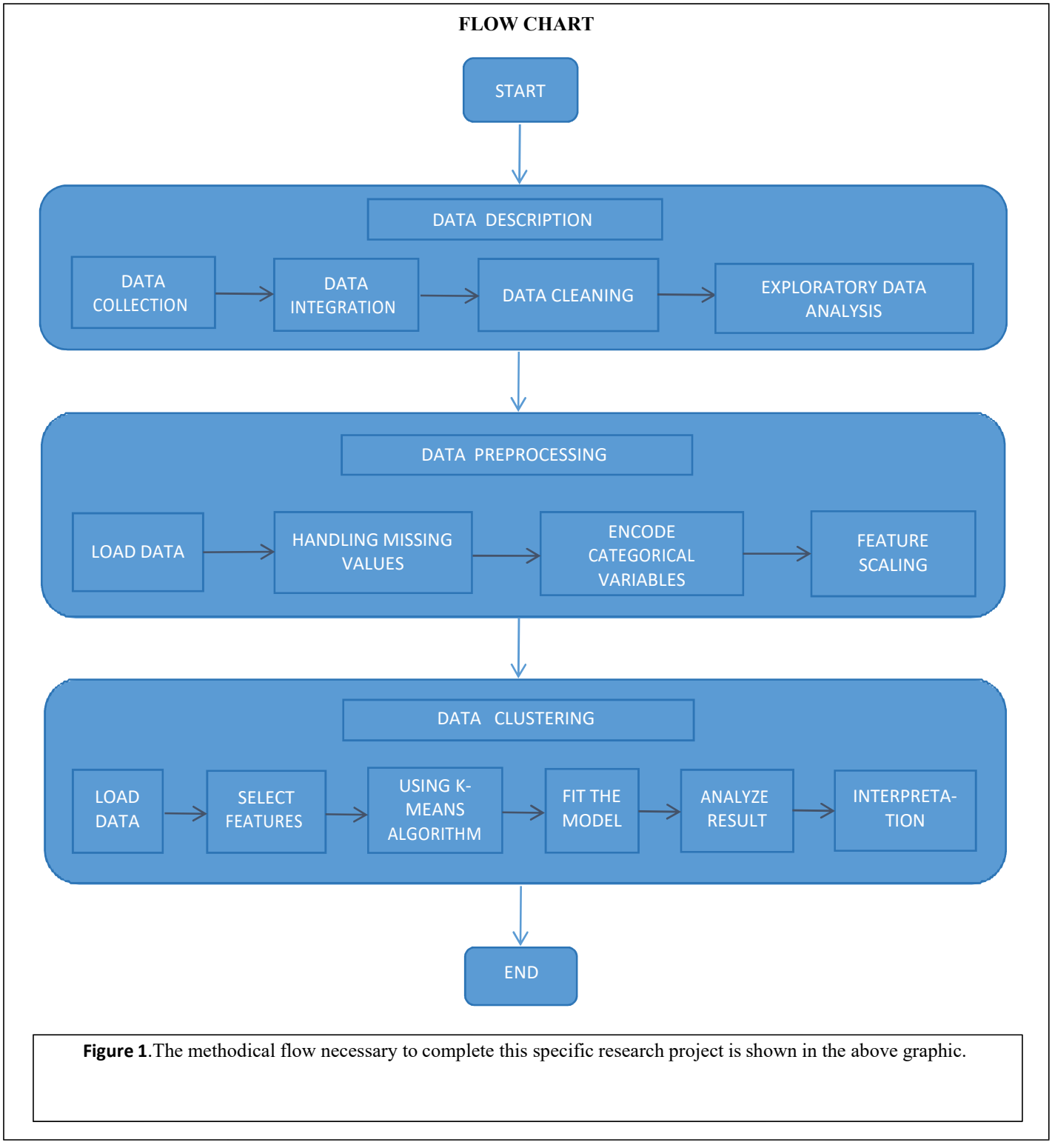
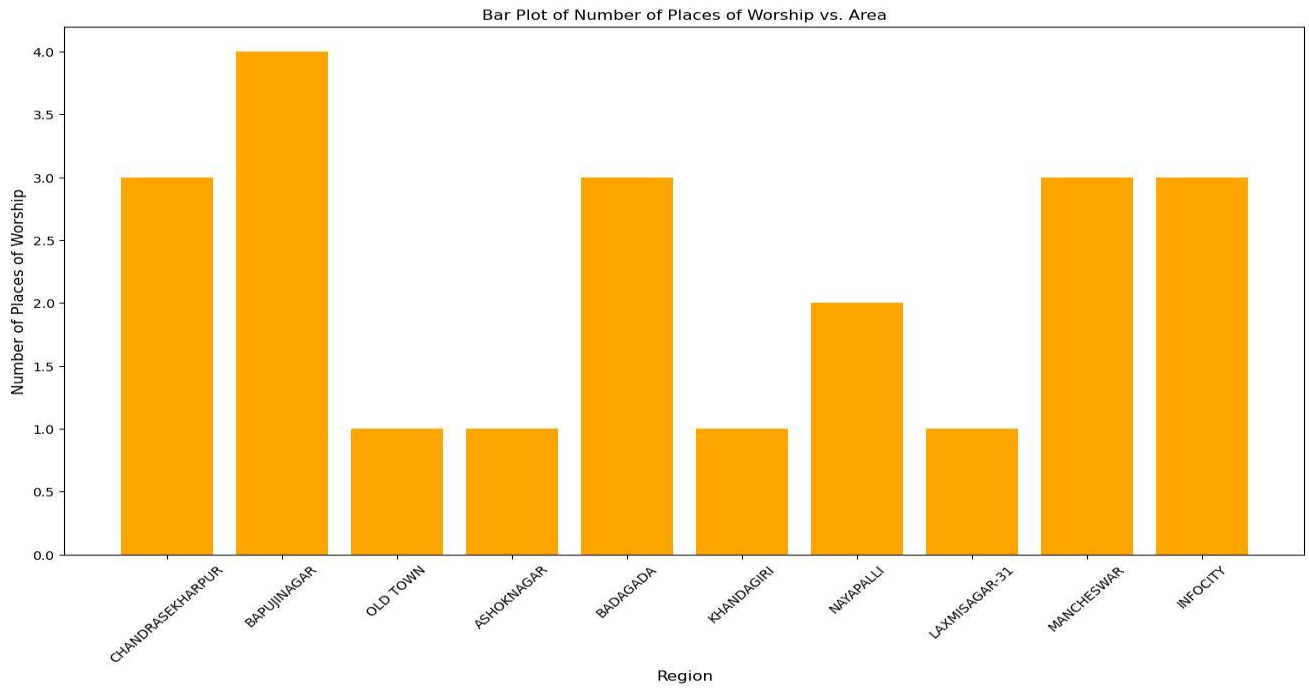


Figure 1.The methodical flow necessary to complete this specific research project is shown in the above graphic.

RESULTS AND ANALYSIS:

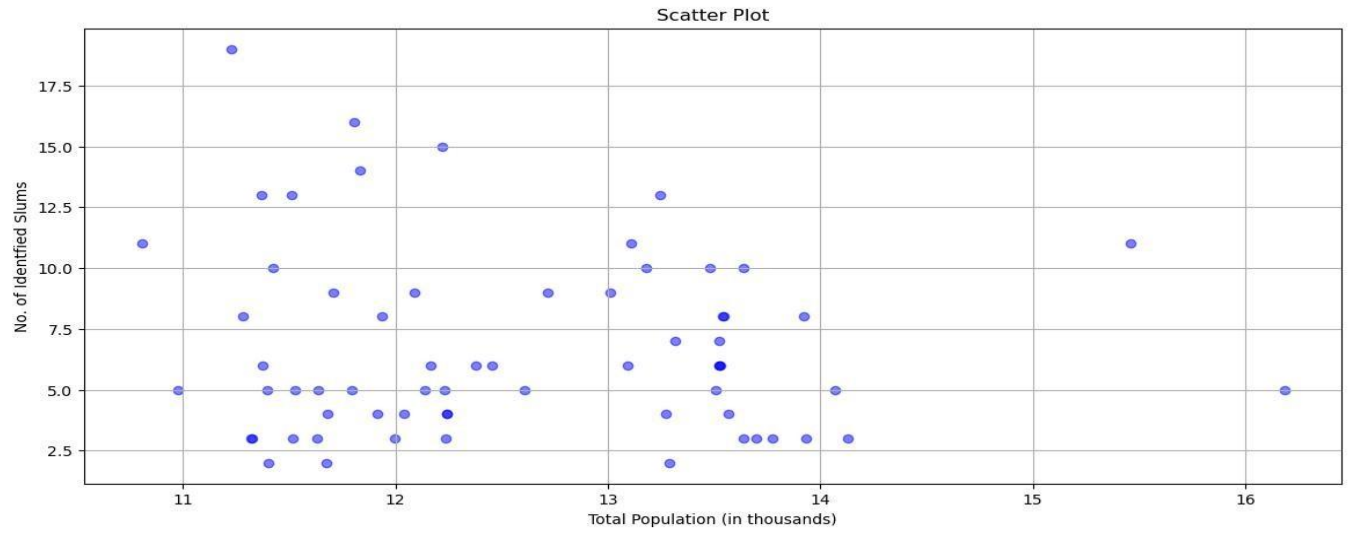
From the above operations we have found the scatter plots,graphs and bar graphs for different sectors of city profiling which would help understand the trends andpattern present and help to provide a insight towards those sectors. Some of the graphs and scatter plots have been thereby mentioned below for better understanding of the datasets trends

FIGURE-2



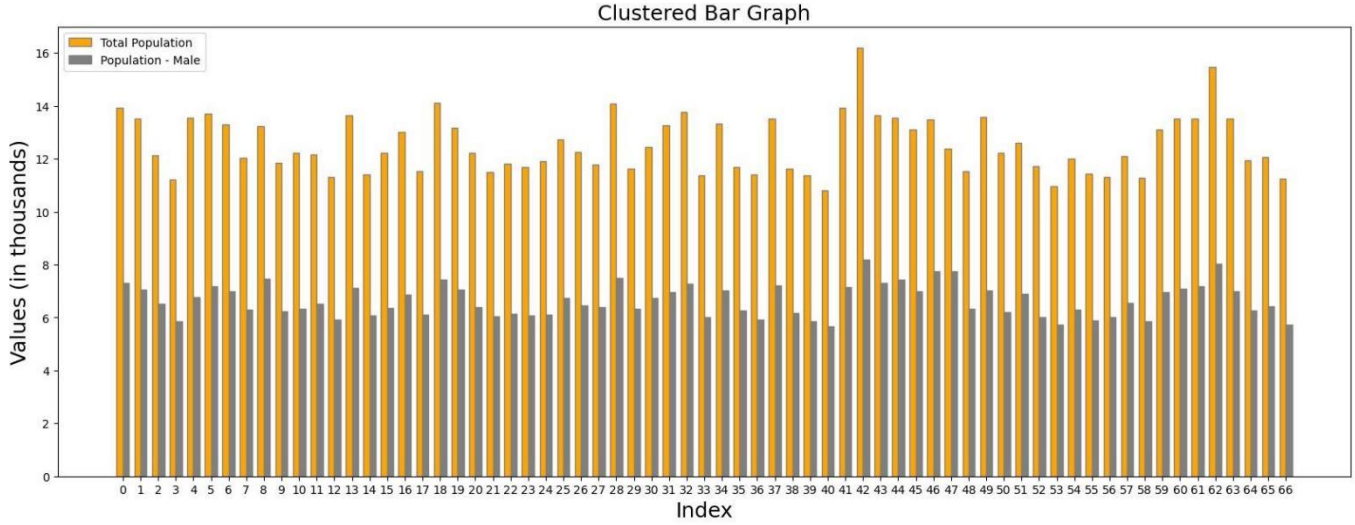
This figure demonstrates the distribution of number of places of worship with respect to the different regions of the city.

FIGURE-3



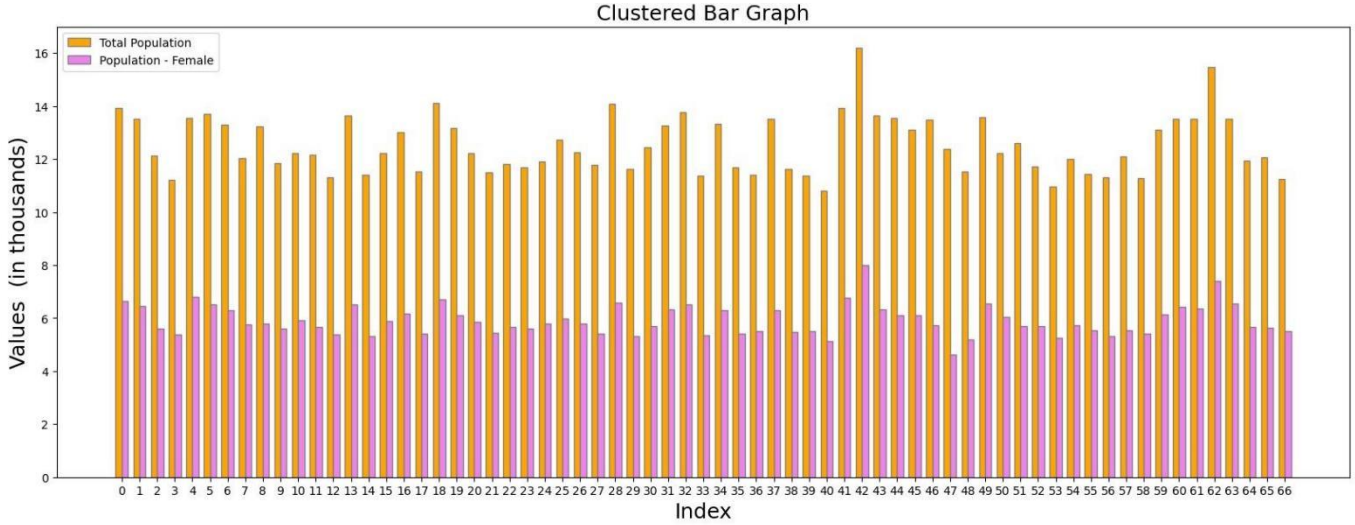
This figure gives an idea about the distribution of the marginalised population across the number of identified slums. .

FIGURE-4



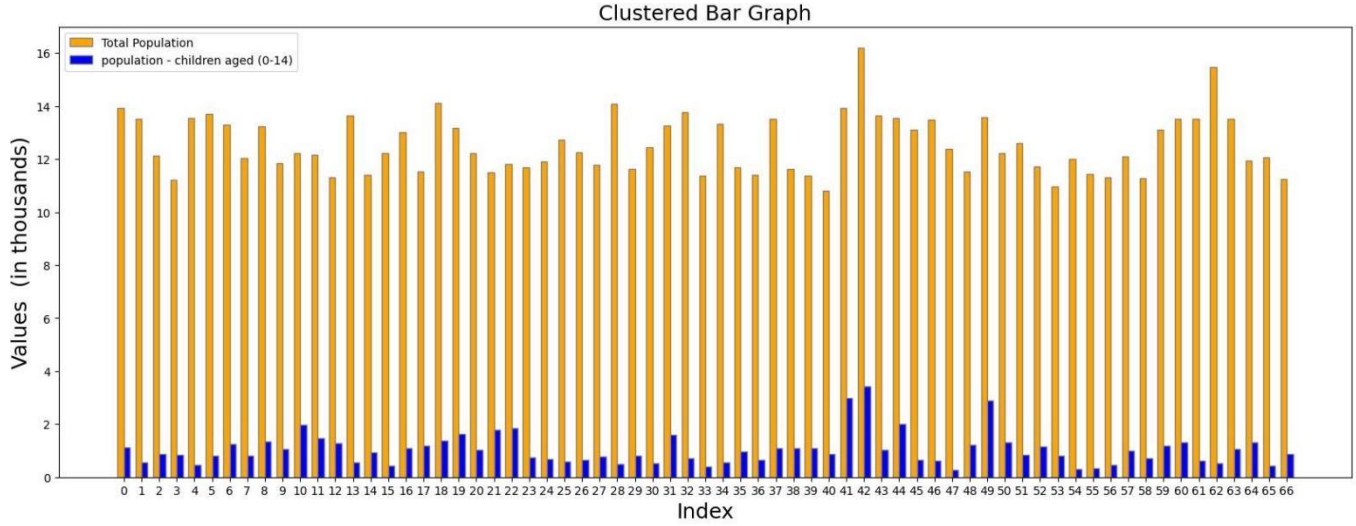
This figure provides idea of the male population in different regions of the city with respect to the total population.

FIGURE-5



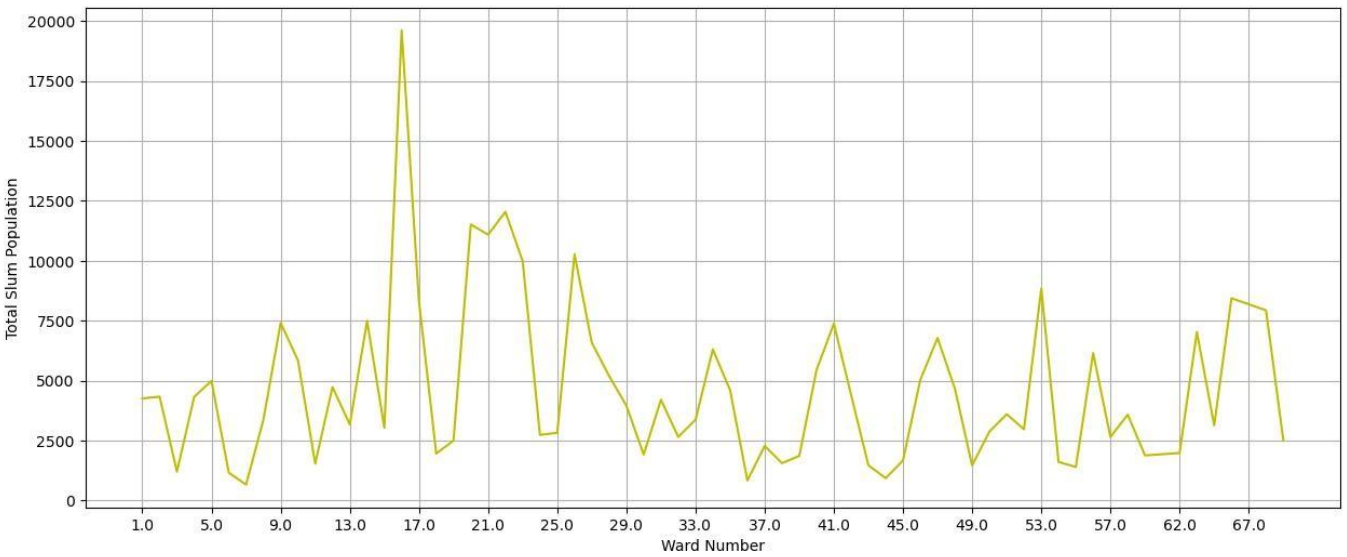
This figure provides idea of the female population in different regions of the city with respect to the total population.

FIGURE-6



This figure provides idea of the population of children aged between 0 to 14 years in different regions of the city with respect to the total population.

FIGURE-7



This figure provides idea of the slum population in different wards of the city.

CONCLUSION

City profiling, when paired with visualization techniques, offers a powerful way to completely understand and communicate theintricate structure of urban environments. City profiling and visualization weave an engrossing tale that appeals to a variety of stakeholders by fusing data-driven research with aesthetically pleasing representations.

In conclusion, city profiling and visualization work together to create a synergy that extends beyond basic data aggregation andfacilitates deeper insights, meaningful interaction, and well-informed decision-making.A clear depiction of population patterns,infrastructure distribution, economic dynamics, and social trends helps stakeholders understand the nuances of urban environments. Additionally, visualizations serve as catalysts for dialogue, collaboration, and creativity, empowering locals to shape the future of their towns.

REFERENCES:

1.Lim, Chiehyeon, Kwang-Jae Kim, and Paul P. Maglio. "Smart cities with big data: Reference models, challenges, andconsiderations." Cities 82 (2018): 86-99.

2.Pan, Yunhe, et al. "Urban big data and the development of city intelligence." Engineering 2.2 (2016): 171-178.

3.Chin, Jeannette, Vic Callaghan, and Ivan Lam. "Understanding and personalising smart city services using machine learning,The Internet-of-Things and Big Data." 2017 IEEE 26th international symposium on industrial electronics (ISIE). IEEE, 2017.

4.D'Andrea, Eleonora, et al. "Smart profiling of city areas based on web data." 2018 IEEE International Conference on SmartComputing (SMARTCOMP). IEEE, 2018.

5.Hashem, Ibrahim Abaker Targio, et al. "The role of big data in smart city." International Journal of information management36.5 (2016): 748-758.

6.Soomro, Kamran, et al. "Smart city big data analytics: An advanced review." Wiley Interdisciplinary Reviews: Data Miningand Knowledge Discovery 9.5 (2019): e1319

7.Rathore, M. Mazhar, et al. "Exploiting IoT and big data analytics: Defining smart digital city using real-time urban data."Sustainable cities and society 40 (2018): 600-610.

8.Mouchili, Mama Nsangou, Shadi Aljawarneh, and Wette Tchouati. "Smart city data analysis." Proceedings of the FirstInternational Conference on Data Science, E-learning and Information Systems. 2018.

9. Khan, Zaheer, et al. "Towards cloud based big data analytics for smart future cities." Journal of Cloud Computing 4 (2015): 1-11.

10. Shrivastava, P., Sahoo, L., Pandey, M., Agrawal, S. (2018). AKM—Augmentation of K-Means Clustering Algorithm for Big Data. In: Bhateja, V., Coello Coello, C., Satapathy, S., Pattnaik, P. (eds) Intelligent Engineering Informatics. Advances in Intelligent Systems and Computing, vol 695. Springer, Singapore. https://doi.org/10.1007/978-981-10-7566-7_11

11. Jena, B., Gourisaria, M. K., Rautaray, S. S., & Pandey, M. (2017). A survey work on optimization techniques utilizing map reduce framework in hadoop cluster. International Journal of Intelligent Systems and Applications, 9(4), 61.

12. Pathak, A.R., Pandey, M. & Rautaray, S.S. Approaches of enhancing interoperations among high performance computing and big data analytics via augmentation. Cluster Comput 23, 953–988 (2020). https://doi.org/10.1007/s10586-019-02960-y

13.A. Sen, M. Pandey and K. Chakravarty, "Random Centroid Selection for K-means Clustering: A Proposed Algorithm for Improving Clustering Results," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 2020, pp. 1-4, doi: 10.1109/ICCSEA49143.2020.9132921. keywords: {Computer science;Computational modeling;Clustering algorithms;Genetic algorithms;K-means clustering;Genetic Algorithm;Text mining;Topic modelling},

14. Moharana, M., Pandey, M. & Rautaray, S.S. Clustering Based BMI Indexing for Child Disease Prone-Probability Prediction. SN COMPUT. SCI. 4, 413 (2023). https://doi.org/10.1007/s42979-023-01823-z