Telangana Tourism Insights Analysis using Data Engineering System

Dr T Sankara Rao[1], A Bhanu Mythreyi[2], Ch Gayatri Sri Sowmya[3], B Yasaswini[4], N Lohita[5]

[1]Associate Professor, Dept. of Computer Science Engineering, GITAM University, Andhra Pradesh, India
[2,3,4,5]Student, Dept. of Computer Science Engineering, GITAM University, Andhra Pradesh, India

---***---

Abstract - Recognizing the transformative power of data analysis in fostering economic growth, this project endeavors to enhance revenue generation in Telangana by focusing on its tourism and culture sector. The proposed system aims to equip the Telangana government with actionable insights to inform strategic decision-making in the tourism and culture domain. By analyzing trends and patterns related to tourist spots, domestic and foreign visitors, as well as revenue generated through taxes, the system facilitates the development of tailored strategies to enrich tourism experiences, promote cultural attractions, and optimize resource allocation. To achieve these objectives, the project advocates for the implementation of a robust data engineering system. Data Engineering encompasses a set of operations designed to efficiently utilize data for business purposes. The system's core focus lies in designing and building data gathering and storage mechanisms, ensuring that raw data is meticulously prepared for in-depth analysis. Activities within the data engineering process include configuring data sources, integrating analytical tools, and managing the architecture of the entire system. This project positions data analysis as a strategic enabler for Telangana to harness the vast potential of its tourism and culture sector. By leveraging insights derived from effective data analysis, the state can attract more tourists, encourage increased spending within its borders, and realize sustainable economic growth. Establishing a robust data analysis system is deemed crucial for Telangana to capitalize on the wealth of opportunities within its tourism and culture sector, paving the way for a prosperous and economically vibrant future.

Key Words: Data Engineering, Tourism, Data Pipeline, Data Warehouse, Data Visualization, Cloud Computing, Structured Query Language (SQL)

1.INTRODUCTION

Data Engineering is a set of operations that make data efficiently used by businesses. It is required to design and build systems for gathering and storing data at stake and preparing it for further analysis. It involves gathering raw data to analyze valuable insights from the gathered data. It involves processes such as configuring data sources to integrating analytical tools. All these systems are to be architecture, built, and managed.

The data engineering process involves activities that enable us to use vast raw data for practical purposes. Stages in Data Engineering: 1. Data Ingestion 2. Data Transformation 3. Data Serving 4. Data flow orchestration

Data Ingestion - Moves data from multiple sources to a target system which is later processed for further analysis, Data Transformation - Makes data into a valuable form of data which involves removing duplicates, and errors, normalizing the data, and converting it into the form that is required for us to perform further processes, Data Serving - Delivers transformed data to end users, Data flow Orchestration, It provides visibility into the entire process and ensures that all the processes are successful

Data Pipeline - In simple terms, it is a mechanism that automates the ingestion, transformation, and serving steps of the data engineering process. It can also be considered as a series of automated processes that move data from one system or stage to another. It combines the integration tools and connects sources to a data warehouse, and it also helps in loading information from one place to another. The processes in a data pipeline can include: Extraction, Validation, Transformation, Loading, Monitoring. The data pipeline is beneficial because it would have been complicated to manually transfer data and perform extractions, transformations, and track changes in data without it.

Data Warehouse - It is a central repository for storing data in query able forms. It can also be considered a regular database which is enhanced for reading and querying huge amounts of data. The main advantage of a data warehouse is the historical data, as the general transactional databases do not store historical data. They use data sources like flat files, relational databases, and other forms of data. General databases normalize data by eliminating data redundancies and making them into different tables. Such processes might involve heavy computations as each simple query demands to combine various tables. We use simple queries with fewer tables in data warehouses, improving performance. Data Analytics It involves analyzing the data to find valuable insights and draw valid conclusions from the information. It involves streaming analytical results from the data processed and stored. It improvises business intelligence and helps businesses grow revenue and use data efficiently.

2. LITERATURE REVIEW

This comprehensive literature review covers the key aspects of data engineering and performance dashboards, providing a foundation for understanding their significance in analyzing tourism insights within the state of Telangana, India. Data Integration, Transformation, and Conversions - Effective data engineering begins with data integration, a critical foundation that facilitates the management of data from diverse sources. In the context of tourism insights analysis, data from various stakeholders, including hotels, tour operators, and government agencies, must be consolidated. Integration technologies and Enterprise Software Systems are vital tools in this process. Data Warehousing - Data warehousing plays a pivotal role in the tourism sector's data infrastructure. Understanding the distinction between Online Transaction Processing (OLTP) system and Online Analytical Processing (OLAP) system is essential. OLAP systems enable complex, multidimensional queries, making them invaluable for business intelligence and decision support in the tourism industry.

[1] Data Integration Patterns (DIP) and Enterprise Integration Patterns (EIP) contribute to effective data engineering. DIP involves mapping data elements from various sources to target data elements, providing a framework for modeling and transferring data elements in the context of tourism data. EIP, on the other hand, focuses on managing messaging and communication aspects within enterprise systems, complementing data integration with communication management in service or enterprise application integration.

[2] Data pipelines automate data transfer, manipulation, and validation, enhancing efficiency and minimizing manual intervention. In tourism insights analysis, these pipelines are crucial for data extraction, transformation, and loading processes, ensuring that data is available for end-to-end analysis.

[3] Data visualization is an essential component in tourism insights analysis. It allows stakeholders to transform raw data into meaningful visual representations, revealing concealed insights. Static visualizations, such as bar charts and pie charts, are useful for representing aggregated data, while dynamic visualizations and geospatial representations enable real-time monitoring and decision-making.

[4] The adoption of cloud data storage is a cost-effective shift from traditional disk-based storage. Public, private, and hybrid cloud options are available, offered by leading cloud computing providers like Amazon, Azure, and Google Cloud Platform. These services ensure scalability, reliability, and data accessibility for tourism insights analysis in Telangana.

[5] Performance dashboards serve as vital tools for analyzing tourism insights in the state of Telangana. Originally developed for the business industry, they have found diverse applications, including tourism. Cleverley and Cleverly (2005) highlight their significance in monitoring and improving an organization's performance, offering real-time visibility into critical performance indicators. Development and Application of Dashboards - Performance dashboards create a graphical representation summarizing decision-related data through the use of visual elements such as charts and graphs. They structure information, highlight factors for consideration, and simplify data evaluation. Dashboards are adaptable for diverse sectors, and their main objective is to support in the process of making decisions.

Types of Performance Dashboards - When examining insights related to tourism, three types of performance dashboards are particularly relevant: 1. Strategic Dashboards: These dashboards cater to top-level management, focusing on monitoring the execution of strategic objectives and organizational goals. 2. Tactical Dashboards: Tailored for departmental managers, tactical dashboards emphasize process tracking, analysis, and monitoring against budget and goals. 3. Operational Dashboards: These dashboards are vital for frontline professionals involved in tourism insights analysis, offering real-time monitoring of core operational processes.

Key Performance Indicators (KPIs) KPIs are essential in tourism insights analysis, helping measure progress toward organizational goals. They can encompass various categories, including visitor arrivals, popular tourist destinations, revenue generated from tourism, and visitor satisfaction.

In conclusion, in Telangana's tourism sector, the interplay between data engineering and performance dashboards is indispensable for gaining valuable insights. Data integration, transformation, warehousing, pipelines, visualization, and cloud storage provide the infrastructure necessary to process and analyze tourism data effectively. Performance dashboards offer the means to visualize and monitor these insights in real-time. Together, these components empower Telangana's tourism industry to make data-driven, informed decisions, ensuring competitiveness in the market and contributing to the state's tourism success. This literature review provides a comprehensive foundation for understanding the role of data engineering and performance dashboards in Telangana's tourism insights analysis.

3. OBJECTIVES

- Analyzing the district that has the highest number of tourists spots.

- Trends in tourism are based on the number of units of TSTDC (Telangana State Tourism Development Corporation) in each district and their effect on the number of TSS artists and tourist spots.

- Analyzing the tariff (tax) obtained by the government based on the hotels in different cities and districts.

- Analyzing the trends in tariff based on the days of week.

- Analyzing the tariff based on hotel price ranges of the districts.

- Analyzing the number of visitors month wise in each district

- Figuring out the district which has top number of visitors.

- Analyzing the number of visitors based on the number of tourists places.

- Figuring out the month in which attracts a smaller number of visitors.

- The district that contributes more to attracting tourists and generating more revenue

- The districts that have more TSS artists than tourist spots, have a greater number of visitors and other valuable trends.

4. SCOPE

The data collected to build the system is taken from open data Telangana website (data.telangana.gov.in), which was launched by KTR, former IT Minister of Telangana to open source the Telangana data and let others access the data to use for commercial or non-commercial purposes. All the insights that we generate are from the available data on the website and will be accurate as per the available data. In the entire project we would use the data from the mentioned website and draw useful insights which can help the government take decisions related to tourism and culture. This project utilizes the below data.

1. Total Tariff (Tax) collected from each hotel in different cities based on days of week (data updated as of 2021).

2. Data containing information about number of tourist spots, tourism development units and TSS Artists.

3. Types of visitors like foreign visitors and in country (domestic visitors) and number of visitors based on months from year 2016 to 2020.

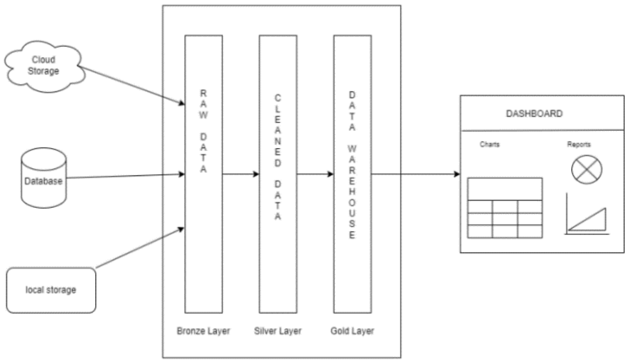
5. SYSTEM METHODOLOGY

System Analysis - The system consists of three essential parts, they are described as follows:

Data Storage: The first part serves as the foundation of the system, where data is stored in various formats across different locations, such as databases, local systems, and designated storage points.

Data Pipeline and Data Warehouse: The second component involves managing the data. The Data Warehouse plays a crucial role in structuring data models, carefully crafted through a multi-step process. This includes gathering data from various sources, making necessary changes, and combining the most important attributes from all data sources. Simultaneously, the ETL (Extract, Transform, Load) pipeline automates the process of moving data from one place to another, ensuring its smooth flow. It extracts data from multiple sources, transforms it to meet specific requirements, and loads it into the data warehouse, simplifying the data engineering process.

Data Visualization: The third component completes the system's functionality. After the data is collected, processed, and transformed, it undergoes analysis using visualization tools. This component focuses on creating user-friendly dashboards with various visual representations, such as graphs, charts, tables, and informative reports. In essence, this component makes the data accessible and easy to understand, aiding in decision-making and generating insights.



The Data Layers:

Bronze Layer: Raw data, which is directly obtained from the source, much of this data might not be useful for us in our process of extracting information and insights.

Silver Layer: This includes Transformed data, where we eliminate the unnecessary data and replace unusual data.

Gold Layer: This is the critical data on which we work to generate typical insights.

6. TECHNOLOGY OVERVIEW

Docker: We utilized Docker to establish an isolated environment, enabling the installation of Apache Airflow alongside its requisite dependencies, thereby maintaining separation from the host system. Also, created a Python virtual environment within the Docker container for further encapsulation. This approach ensures a clean and self-contained setup for the Airflow project, safeguarding the local environment from disruptions.

Apache Airflow: We employed Apache Airflow to create and manage our ETL (Extract, Transform, Load) pipelines. This helped us automate the process of gathering, transforming, and loading data from various sources, making our data workflows more efficient and reliable. Airflow made it easier to schedule and monitor these tasks, improving our data processing capabilities.

Mysql Workbench: We utilized Mysql workbench to store our database, making it possible for our data pipeline to access and work with structured data. This allowed us to store and retrieve data in an organized way, making it readily available for our data processing tasks within the pipeline.

Amazon S3: We stored our data in Amazon S3 to ensure easy access and retrieval, especially when our data is updated. Amazon S3 serves as a secure and scalable storage solution, allowing us to keep our data safe and available at all times. This means that whenever our data is refreshed or changed, we can still easily access it and fetch the latest results without any hassle.

Snowflake: We chose to use Snowflake for our data warehousing needs. Snowflake provides a place where we can store our data in an organized and efficient manner. It's like a big storage facility where we can keep all our data safe and easy to access when we need it for analysis and other purposes. This helps us manage our data effectively.

Power Bi: We made use of Power BI to create visuals for our data. Power BI is a helpful tool that allows us to turn our data into meaningful charts, graphs, and reports. These visuals make it much easier to understand and interpret our data, helping us make informed decisions and communicate our findings effectively. Power BI's user-friendly interface and powerful features enable us to present our data in a visually appealing and insightful way, making it a valuable asset for our project.

7. IMPLEMENTATION

Datasets description

Four key datasets, namely *hoteltariff*, *domesticvisitors*, *foreignvisitors*, and *tourismandculture*, are central to this project.

hoteltariff

Attributes	***Data Type***
Hotel	*text*
City	*text*
District	*text*
Contact	*int*
Roomtype	*text*
TotalRooms	*int*
SundayTariff	*int*
MondayTariff	*int*
TuesdayTariff	*int*
WednesdayTariff	*int*
ThursdayTariff	*int*
FridayTariff	*int*
SaturdayTariff	*int*

domesticvisitors and foreignvisitors

Attributes	***Data Type***
District	*text*
Month	*text*
Visitors_2016	*int*
Visitors_2017	*int*
Visitors_2018	*int*
Visitors_2019	*int*
Visitors_2020	*int*

tourismandculture

Attributes	Data Type
District	text
tstdcunits	int
touristspots	int
tssartists	int

ETL Pipelines:

The execution process involves ETL pipelines, which Extract, Transform, and Load the data. The extracted data is initially considered as the "bronze layer," and after transformation, it becomes the "silver layer." Finally, a table with essential attributes is created, forming the "gold layer."

Execution Workflow:

The workflow is orchestrated through Airflow using the Python programming language. Four Directed Acyclic Graphs (DAGs) have been implemented:

1. First DAG:

. Extraction of *hoteltariff* data in CSV format from the local system.

. Transformation includes replacing null values with strategic replacements. Weekends' tariff values are set to the highest value of the week, optimizing profits. Nulls on weekdays are substituted with the mode value.

. Loaded into the Snowflake data warehousing tool using SQL.

2. Second DAG:

. Extracts data from the MySQL Workbench database corresponding to *tourismandculture*.

. Transformation involves replacing null values with 0 for units and artists, ensuring maximum efficiency.

. Loaded into Snowflake.

3. Third and Fourth DAGs:

. Both DAGs share the functionality of extracting data from AWS S3.

. Transformation involves the mean method. For *domesticvisitors* and *foreignvisitors*, the average number of visitors per district is calculated for each month throughout the year. Null values are replaced with the calculated mean.

. Loaded into Snowflake.

Transformation Methods:

Efficient data replacement methods, such as Mean, Median, and Mode, are employed based on dataset characteristics:

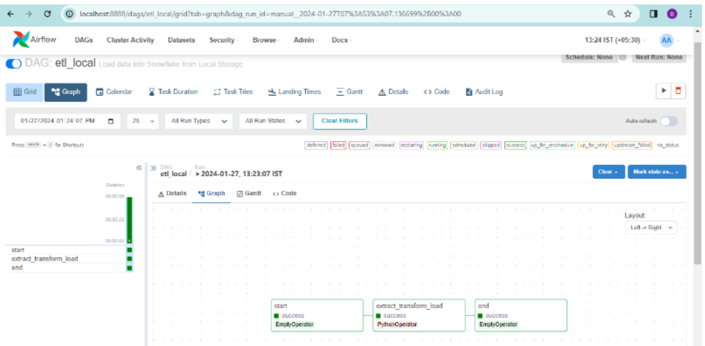
. Mode is chosen for *hoteltariff*, leveraging repeated values for efficiency.

. Mean is chosen for visitor data where districts attract a similar number of visitors, ensuring efficiency.

Data Integration and Visualization:

After loading into Snowflake, a new table with significant attributes is created. This "gold layer" data is then connected to Power BI, enabling the creation of diverse visualizations for comprehensive data analysis.

The airflow DAGs execution is illustrated in the following images.



localhost:8888/dags/etl_local/grid?tab=graph&dag_run_id=manual__2024-01-27T07%3A53%3A07.136699%2B00%3A00

DAG: etl_local Load data into Snowflake from Local Storage

Schedule: None Next Run: None

01/27/2024 01:24:07 PM 25 All Run Types All Run States

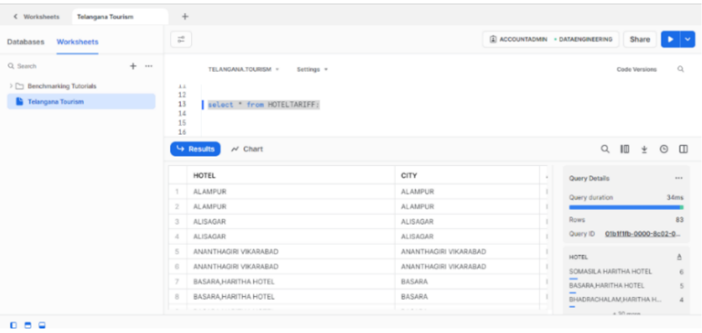
DAG etl_local / Run 2024-01-27, 13:23:07 IST

start success EmptyOperator

extract_transform_load success PythonOperator

end success EmptyOperator

and the resulting data is stored in Snowflake as depicted below.



```
select * from HOTELTARIFF;
```

	HOTEL	CITY
1	ALAMPUR	ALAMPUR
2	ALAMPUR	ALAMPUR
3	ALISAGAR	ALISAGAR
4	ALISAGAR	ALISAGAR
5	ANANTHAGIRI VIKARABAD	ANANTHAGIRI VIKARABAD
6	ANANTHAGIRI VIKARABAD	ANANTHAGIRI VIKARABAD
7	BASARA,HARITHA HOTEL	BASARA
8	BASARA,HARITHA HOTEL	BASARA

Query Details

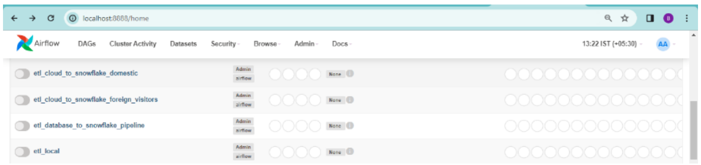
Query duration 34ms

Rows 83

Query ID 01b1f1fb-0000-8c02-0...

HOTEL	
SOMASILA HARITHA HOTEL	6
BASARA,HARITHA HOTEL	5
BHADRACHALAM,HARITHA H...	4

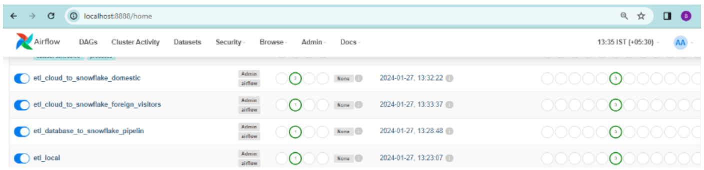
Similarly, all four DAGs undergo execution.



localhost:8888/home

Airflow DAGs Cluster Activity Datasets Security Browse Admin Docs 13:22 IST (+05:30) AA

etl_cloud_to_snowflake_domestic	Admin airflow	None	
etl_cloud_to_snowflake_foreign_visitors	Admin airflow	None	
etl_database_to_snowflake_pipeline	Admin airflow	None	
etl_local	Admin airflow	None	

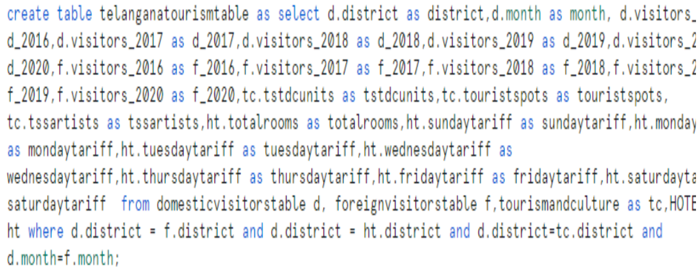


localhost:8888/home

Airflow DAGs Cluster Activity Datasets Security Browse Admin Docs 13:35 IST (+05:30) AA

etl_cloud_to_snowflake_domestic	Admin airflow	None	2024-01-27, 13:32:22
etl_cloud_to_snowflake_foreign_visitors	Admin airflow	None	2024-01-27, 13:33:37
etl_database_to_snowflake_pipelin	Admin airflow	None	2024-01-27, 13:28:48
etl_local	Admin airflow	None	2024-01-27, 13:23:07

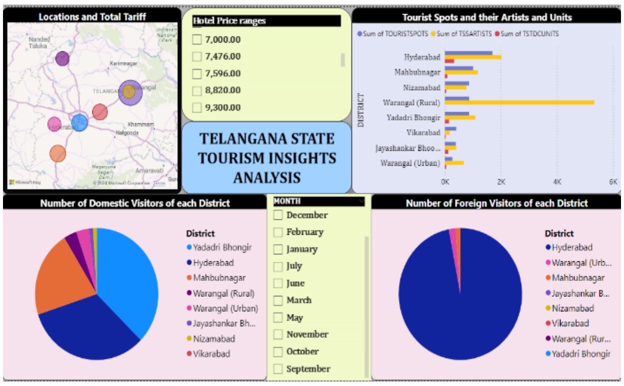
The generation of the Gold Layer is demonstrated below.



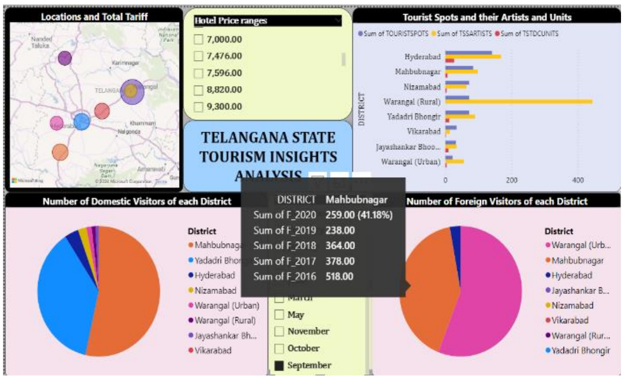
```
create table telanganatourismtable as select d.district as district,d.month as month, d.visitors_
d_2016,d.visitors_2017 as d_2017,d.visitors_2018 as d_2018,d.visitors_2019 as d_2019,d.visitors_2
d_2020,f.visitors_2016 as f_2016,f.visitors_2017 as f_2017,f.visitors_2018 as f_2018,f.visitors_2
f_2019,f.visitors_2020 as f_2020,tc.tstdcunits as tstdcunits,tc.touristspots as touristspots,
tc.tssartists as tssartists,ht.totalrooms as totalrooms,ht.sundaytariff as sundaytariff,ht.monday
as mondaytariff,ht.tuesdaytariff as tuesdaytariff,ht.wednesdaytariff as
wednesdaytariff,ht.thursdaytariff as thursdaytariff,ht.fridaytariff as fridaytariff,ht.saturdayta
saturdaytariff  from domesticvisitorstable d, foreignvisitorstable f,tourismandculture as tc,HOTE
ht where d.district = f.district and d.district = ht.district and d.district=tc.district and
d.month=f.month;
```

7. CONCLUSION

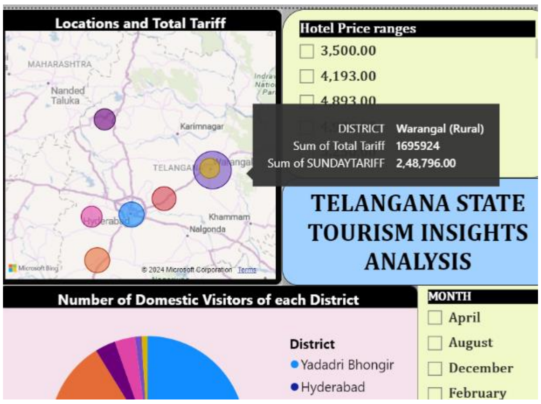
The picture shown below is the overall dashboard.



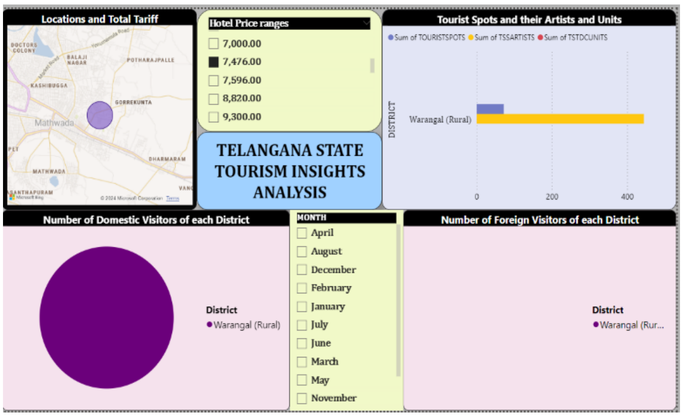
We can find the number of domestic and foreign visitors of each district at every month.



We can get to know which district generates more tariff.

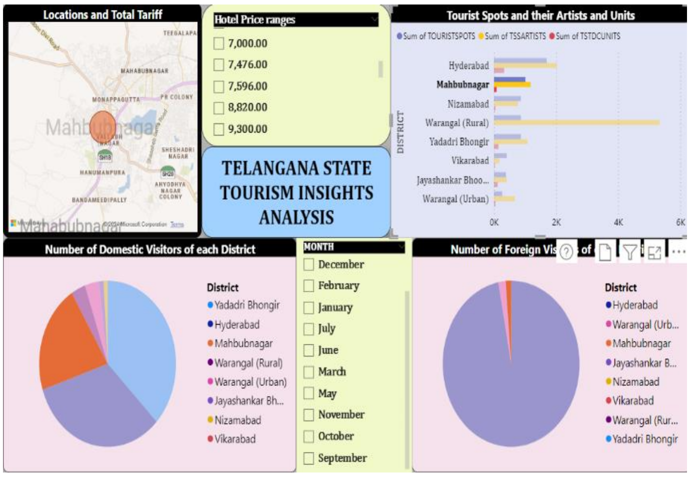

Locations and Total Tariff
MAHARASHTRA
Nanded Taluka
Karimnagar
TELANGANA
Hyderabad
Khammam
Nalgonda
Microsoft Bing
© 2024 Microsoft Corporation Terms
Hotel Price ranges
3,500.00
4,193.00
4,893.00
DISTRICT Warangal (Rural)
Sum of Total Tariff 1695924
Sum of SUNDAYTARIFF 2,48,796.00
TELANGANA STATE TOURISM INSIGHTS ANALYSIS
Number of Domestic Visitors of each District
District
Yadadri Bhongir
Hyderabad
MONTH
April
August
December
February

Analyzing the tariff obtained based on the hotel price ranges of that district.

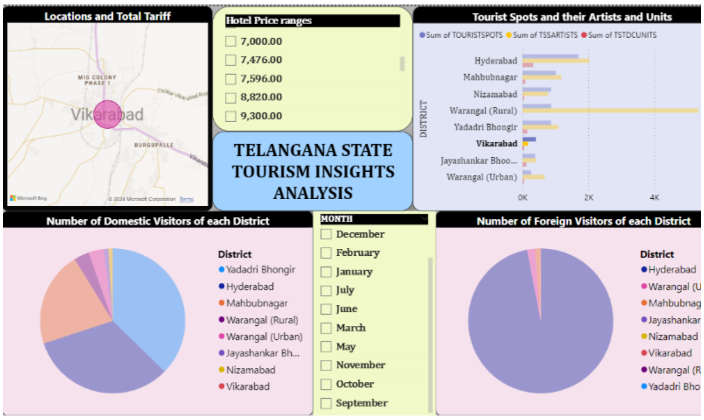

Locations and Total Tariff
Hotel Price ranges
7,000.00
7,476.00
7,596.00
8,820.00
9,300.00
TELANGANA STATE TOURISM INSIGHTS ANALYSIS
Tourist Spots and their Artists and Units
Sum of TOURISTSPOTS Sum of TSSARTISTS Sum of TSTDCUNITS
Warangal (Rural)
DISTRICT
0
200
400
Number of Domestic Visitors of each District
District
Warangal (Rural)
MONTH
April
August
December
February
January
July
June
March
May
November
Number of Foreign Visitors of each District
District
Warangal (Rur...

The districts that have more TSS artists than tourist spots, have a greater number of visitors and other valuable trends.

For example, if we take Mahbubnagar, it has more TSS artists than tourist spots. Hence the number of visitors is more.



For example, if we take Vikarabad, it has a smaller number of TSS artists than tourist spots. Hence the number of visitors is negligible.



REFERENCES

[1] Roland J. Petranch and Richard R. Petranch, “Data Integration and Interoperability: Towards a ModelDriven and Pattern-Oriented Approach”, 2022, MDPI.

[2] Aiswarya raj, Jan Bosch, Helena Holmstrom Olsson, Tian J.Wang, “Modelling Data Pipelines”, 2020, IEEE.

[3] Matthew N O Sadiku, Adebowale E. Shadare, Sarham M. Musa, Cajetan Akujuobi, “Data visualization”, 2016.

[4] Jiyi WU, jianquing FU, Zhijie LIN, jianlin ZHANG, “A survey on cloud storage”, Journal of computers, 2011.

[5] Sandra C. Buttigieg, Cheryl Rathert, Adriana Pace. "Hospital Performance Dashboards: A literature review", Journal of Health Organization and Management, 2017.