

Homework1-DataMining-CMPE255

Colab notebook:


 Homework1_CMPE255.ipynb

Tableau notebook and other project details are uploaded in [Github](#)

Data:

Data is acquired from Opendata Telangana, which is an open-source data.[\(link\)](#)

Problem Description:

I have data regarding revenue of Telanagana .We need to derive insights on state development from the data we have.

I have data from state of Telangana regarding the state revenue.

1. dim_districts
2. dim_date
3. fact_stamps
4. fact_TS_iPASS

MetaData:

Column description for dim_districts:

- The table contains information about districts.
- dist_code: This column represents the district code or identifier for each district.
- district: This column represents the name of the district.

Column description for dim_date:

This table contains the dates at the monthly level. Please be aware that the fiscal year of Telangana spans from April to March.

- month: This column contains the starting date of each month.
- Mmm: This column contains the name of the month.
- quarter: This column contains the associated quarter for each particular month.
- fiscal_year: This column contains the corresponding fiscal year of each month.

Column description for fact_stamps:

The table provides data on the revenue generated from document registrations and estamp challan payments aggregated at the district and monthly level.

- dist_code: This column represents the district code.
- month: This column represents the starting date of each month.
- documents_registered_cnt: This column represents the total count of documents registered.
- documents_registered_rev: This column represents the total revenue generated from the registered documents which include like stamp duty, other taxes etc .
- estamps_challans_cnt: This column represents the count of e-stamps challans.
- estamps_challans_rev: This column represents the revenue generated by online stamp duty submissions.

Column description for fact_TS_iPASS:

The TS-iPASS dataset in Telangana comprises data concerning units or businesses established within the state under the "Industrial Project Approval and Self-Certification System" (iPASS). This government initiative aims to foster industrial growth and investment by streamlining project approvals and enabling self-certification for businesses.

For further details, visit: <https://ipass.telangana.gov.in/>

- dist_code: This column represents the district code.
- month: This column represents the starting date of each month.
- sector: This column represents the industry category. Examples of sectors include 'Automobiles', 'Beverages', 'Engineering', 'Food Processing', etc.
- investment_in_cr: The column represents the investment made in the specific sector, measured in crores (a unit of currency), for the corresponding district and month.
- number_of_employees: This column represents the number of employees associated with that sector for given district and respective month.

Preprocessing :

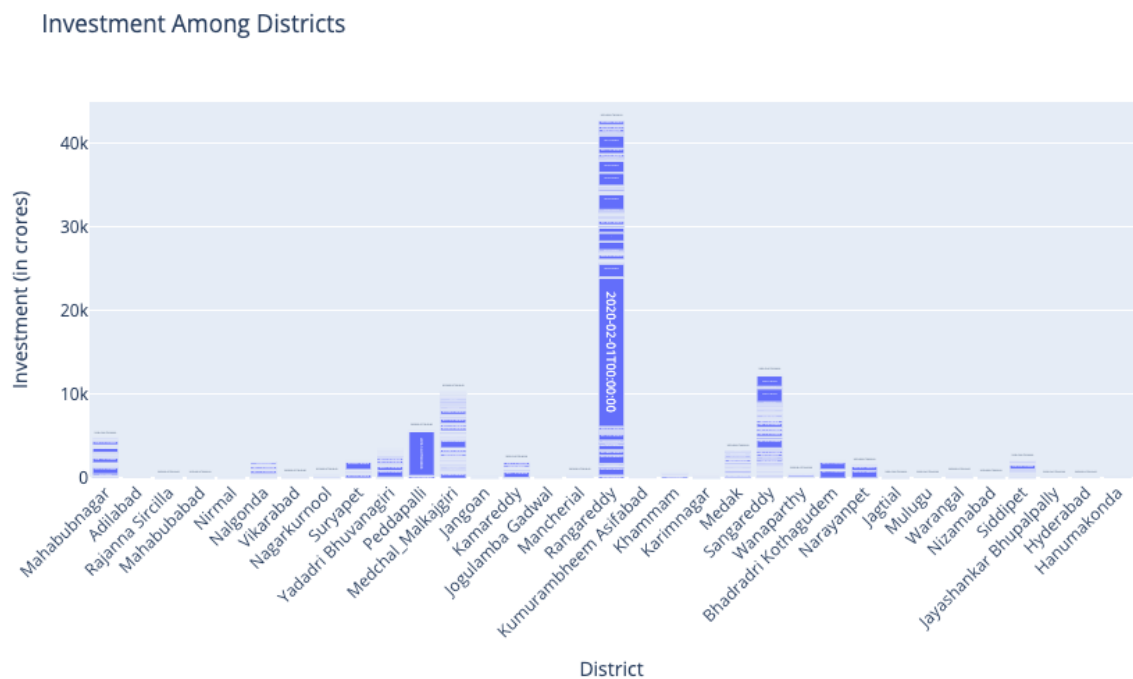
In table fact_TS_iPASS there are no district name ,I merged them from the 'dim_districts' file. After joining the CSV files, the output is saved as 'mer_iPass'. I did this to make it easier since reading district names can be a bit difficult, and everyone would be more familiar with them.

```
df1=pd.read_csv("/content/drive/MyDrive/C7/fact_TS_iPASS.csv")
df2=pd.read_csv("/content/drive/MyDrive/C7/dim_districts.csv")
mf2=pd.merge(df1,df2,on="dist_code",how="outer")
mf2.to_csv("/content/drive/MyDrive/C7/mer_iPass.csv",index=False)
mf2
```

	dist_code	month	sector	investment in cr	number_of_employees	district
0	14_1	01-04-2019	Engineering	2.32	15	Mahabubnagar
1	14_1	01-04-2019	Paper and Printing	14.40	305	Mahabubnagar
2	14_1	01-05-2019	Pharmaceuticals and Chemicals	66.90	190	Mahabubnagar
3	14_1	01-05-2019	Granite and Stone Crushing	0.00	1000	Mahabubnagar
4	14_1	01-05-2019	Food Processing	3.78	90	Mahabubnagar
...
5748	21_1	01-03-2023	Wood and Leather	0.20	1	Hanumakonda

Analysis 1:

What are the top districts where most of the money is invested in?



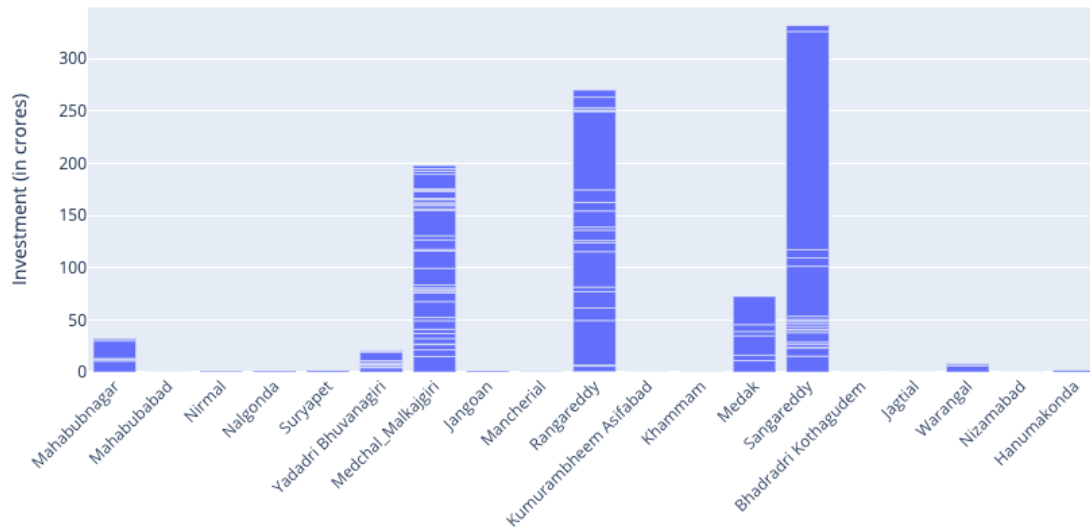
Clearly Rangareddy , Sangareddy, Medchal are the top three districts for investment.

Analysis 2 :

What if stakeholder needed for different sectors?

Select Sect... Electrical and Electronic Produc ▼

Investment in Electrical and Electronic Products



I have implemented a dropdown menu for selecting sectors. This allows stakeholders to view individual sectors and identify the top-performing districts in each specific sector. For example, in Electrical and Electronics Production, Sangareddy ranks as the top district.

Analysis 3:

Is there a relation between the Stamps revenue and Investment?

I have calculated a correlation between e-stamps revenue and investments in crores. Initially, we considered data from 2021 onwards and grouped it by date. The correlation between them was found to be 0.77. This suggests that when there is an increase in investment, there is a corresponding increase in e-stamps revenue, and vice versa.

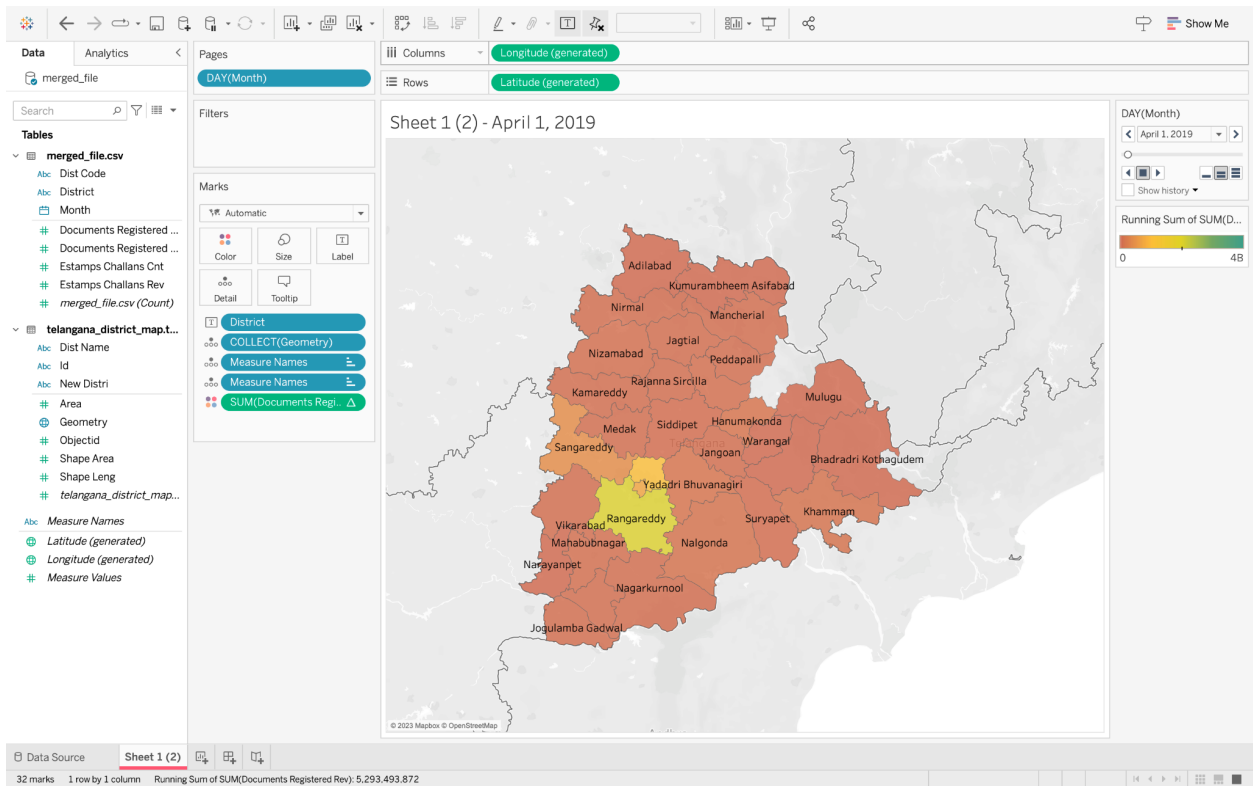


Analysis 4:

Which district had the most revenue in Documents registering?

This is regarding IStamps dataset. I have created a GeoJSON file to depict the evolving financial landscape of Telangana over time. Using distinct colors to denote varying revenue levels, this visualization is hosted on a local server and accessible through Tableau after logging in. Think of it as a time-lapse of the state, showcasing the generated revenue.

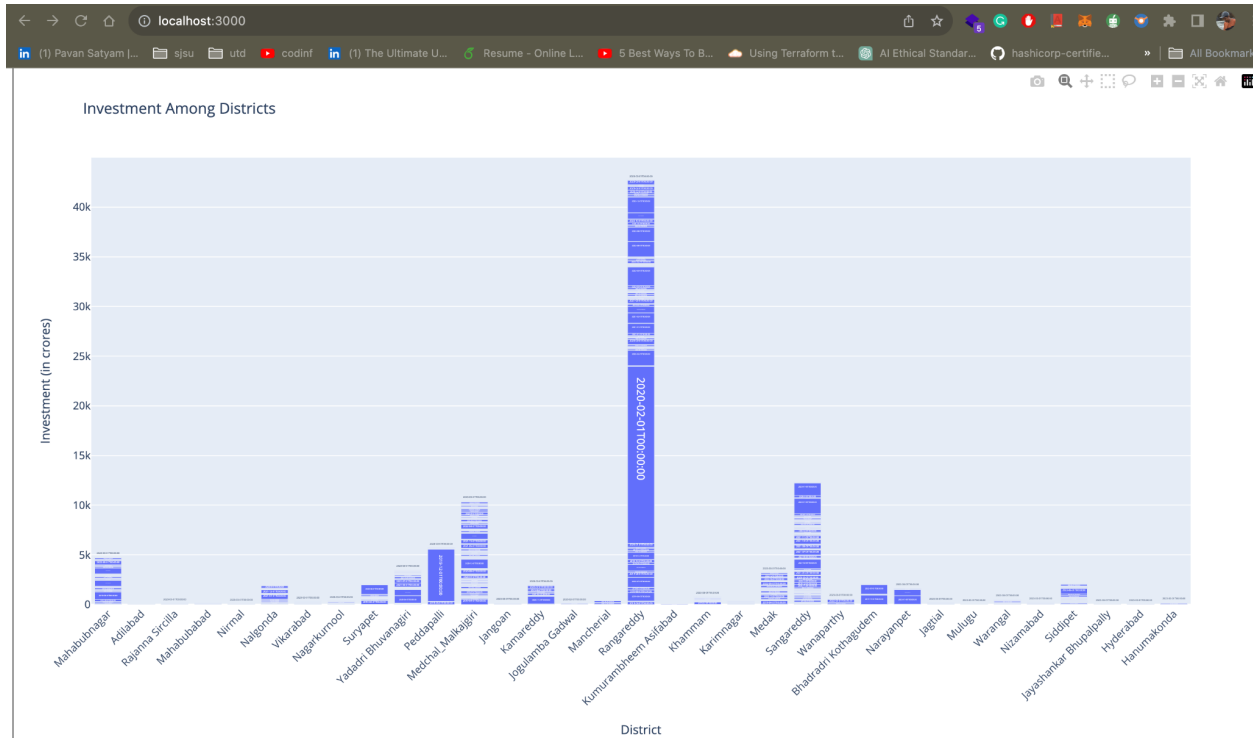
This visualization serves multiple purposes. It enables us to grasp revenue trends in Telangana across time, pinpoint regions with high and low revenue, and track the impact of governmental policies and initiatives. Furthermore, it facilitates comparisons between the revenue performance of different districts and regions.



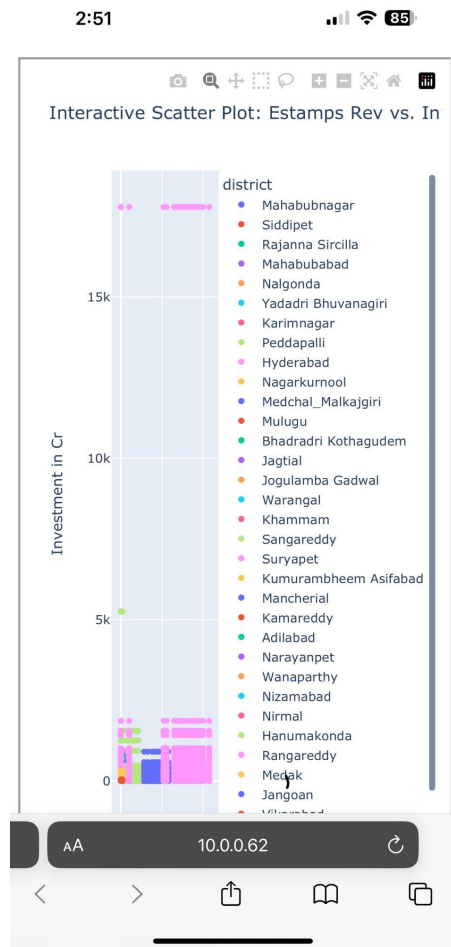
In this visualization, we can observe a significant increase in revenue per district over time, as represented on the timeline.

Visualising in Webserver

I have used the node server for embedding the Visualization. For interactive visualization I have save as html file and embedded in js file.



The benefit of the server is we can access this webpage whoever in this network .I have accessed this website through mobile also



All visualizations are including tableau visualization is also embedded in this webpage.