Final Capstone Project: By Infinite Analytics

Follow the social handles of Infinite Analytics, where hints for the hackathons will be released on a daily basis! There will also be bonus evaluation points to be won for additional quickfire quizzes which will be held on Wednesday & Friday on Linkedin and Instagram. Answers for the quizzes will need to be sent to https://forms.gle/6m8EiFgo3dJF2xPRA

LinkedIn: https://www.linkedin.com/company/infinite-analytics/
Twitter: https://twitter.com/infanatweets
Facebook: https://www.facebook.com/infiniteanalytics
Instagram ID: @infinite_analytics
Website: https://infiniteanalytics.com/

Company Overview:

Who are we?
Infinite Analytics is a Boston headquartered AI startup that specializes in using online and offline consumer data to help our clients understand their consumers and drive up their customer acquisitions effectively. It works with companies across India, US and the Middle East.

Our team of engineers come from MIT, Stanford and IITs, and we're backed by stalwarts such as Sir Tim Berners-Lee (Inventor of the World Wide Web) and Ratan Tata (Chairman Emeritus of the Tata Group). With an exceptional team and exemplary backers, we've built a reputation over the past 10 years to help businesses across verticals such as Automobile, FMCG, Retail and Banking including multiple Fortune 500s to resonate with new consumers and nurture their existing consumer base.

What do we do?
Sherlock.AI analyzes over 350 million anonymized consumers everyday across 40+ datasets. Our machine language algorithms provide a deeper understanding of online and offline consumer behavior to acquire consumers efficiently.

Leveraging our expertise working with global multinationals, we developed Sherlock.AI, our premier data analytics platform with a single mission: to democratize consumer behavior insights and make data-driven marketing accessible to all businesses irrespective of their size. From data exploration to digital marketing campaign management, our platform will provide the tools and insights you need to identify and resonate with your target audience. Best of all? We've designed this platform so anyone can do it,"not just data analysts!
Take a look at how Sherlock AI makes deductions ELEMENTARY!
https://youtu.be/z8xlu5-9DzU

Problem Statements

Broadly, we have 2 problem statements for you. You may choose to work on any one of them as per your wish i.e. you can choose between either of the two problem statements.

Option 1: Segmentation of places -

You will be given POI (Point of Interest) data from OSM (openstreetmaps). These POIs will include locations ranging from grocery stores, shopping malls to car dealerships within the city. You may also enrich your POI (point of interest, meaning, what type of place- showroom/ building/ outlet/ shop etc.) data using any location data available on the web (that you can extract/scrape). Your goal is to divide the city geographically into various blocks/localities. Next you are supposed to create clusters of similar localities and characterize each cluster so that these clusters are human interpretable. Note that we will be evaluating your solution based on actionable insights you are able to draw from EDA (exploratory data analysis) and the clusters. As a side note feel free to use the IFA data into your solution if it seems useful.
Examples of insights are given below:
Cluster 1 (name should be given to each cluster) has a high range of rates for commercial spaces (~Rs. 40000/per sqft) and also has presence of premium cafes and pubs. Overall the correlation between commercial rates and number of pubs/premium cafes in Mumbai is 0.8.

Cluster 5 has a high volume of grocery stores and saloons indicating the basic characteristics of residential areas.

The schema of OSM data is given below:

source: source from where the data was collected
poi_code: unique identifier of the POI
name: name of the POI
poi_type: type of POI (e.g car dealership, shopping mall, etc)
lat: latitude of the POI
long: longitude of the POI
address: address of POI
city: city of POI
state: state of POI
country: country of POI
pin_code: pincode of POI
brand: brand information of POI

Expected Output: Dashboard & Code in the form of Jupyter Notebook. Additionally, insights and graphical format of any data insights or correlation can be presented as PPT's as well.

Submission Link: https://forms.gle/y9CHjjb8dY58HYDr6

Option 2: Segment of people/IFAs -

IFA are unique advertisement IDs of each mobile device which can be considered as a proxy for one person. These IFAs are collected by mobile applications from the users while using the application. The raw IFA contains timestamp and location details of numerous IFAs. We have processed this raw data to visitations of each location of OSM data. The site is identified by the POI code. You are supposed to use OSM and IFA data to analyze the behavior of people/IFA based on their types of places and frequency of their visitations. Your end goal is to get the clusters of IFAs having similar behavior. Submit the clusters in the form of a csv file with a cluster label for each IFA. Also characterize each cluster so that they should be human interpretable and thus bring insights on the table.Few examples are:

- Cluster 1 IFAs have a high frequency of visiting parks, gym and thus represent the population of fitness enthusiasts.
- Cluster 2 IFAs have a high frequency of visiting pubs and nightlife bars thus representing a population of 22-30 age group.

The schema of this data is given below:

1. UID: unique identifier for each mobile device a.k.a IFA ID
2. poi_code : unique identifier of the POI visited (from OSM data)
3. poi_type : type of POI visited
4. date : date of visitation

Expected Output: in the form of csv, with cluster labels for each IFA with proper cluster definitions based on different visitation
Sheet 1: Column 1- IFA ID, Column 2- Cluster label (comma separated, in case of multiple tags for same IFA)
Sheet 2: Column A (name of cluster): Cluster 1, Column B (Cluster name): Health conscious people, Column C (cluster brief): People who are trying to be healthy, Column 4 (mapping): Gyms, Yoga Centers, No junk food, etc.),

Submission Link: https://forms.gle/jVmpC5DfFMsn2Bf17

Evaluation Metrics (for both Problem statements):

- Weightage:
  - 35% Identifying valuable correlations and insights (Quantity and quality of correlations & EDA)
  - 25%: Data Enrichment from other sources
  - 20% Tech Skills- Lean and scalable code blocks
  - 15% Design & Presentation of insights
  - 5% Social Media quizzes.
- Novel Data enrichment techniques from publicly available data sources will be given a boost
- The Quizzes on social media will be graded and carries a 5% weightage on the total score.

All the best!