Mamta Khatri

UMIP23752

DS-Internship

Projects

FIFA World-Cup Analysis
Big Game Sensus Analysis
Hospitality Analysis
Crop Production Analysis
Chatbots
Climate Change Modeling
Healthcare(Heart Disease Analysis)

Projects

FIFA World-Cup Analysis

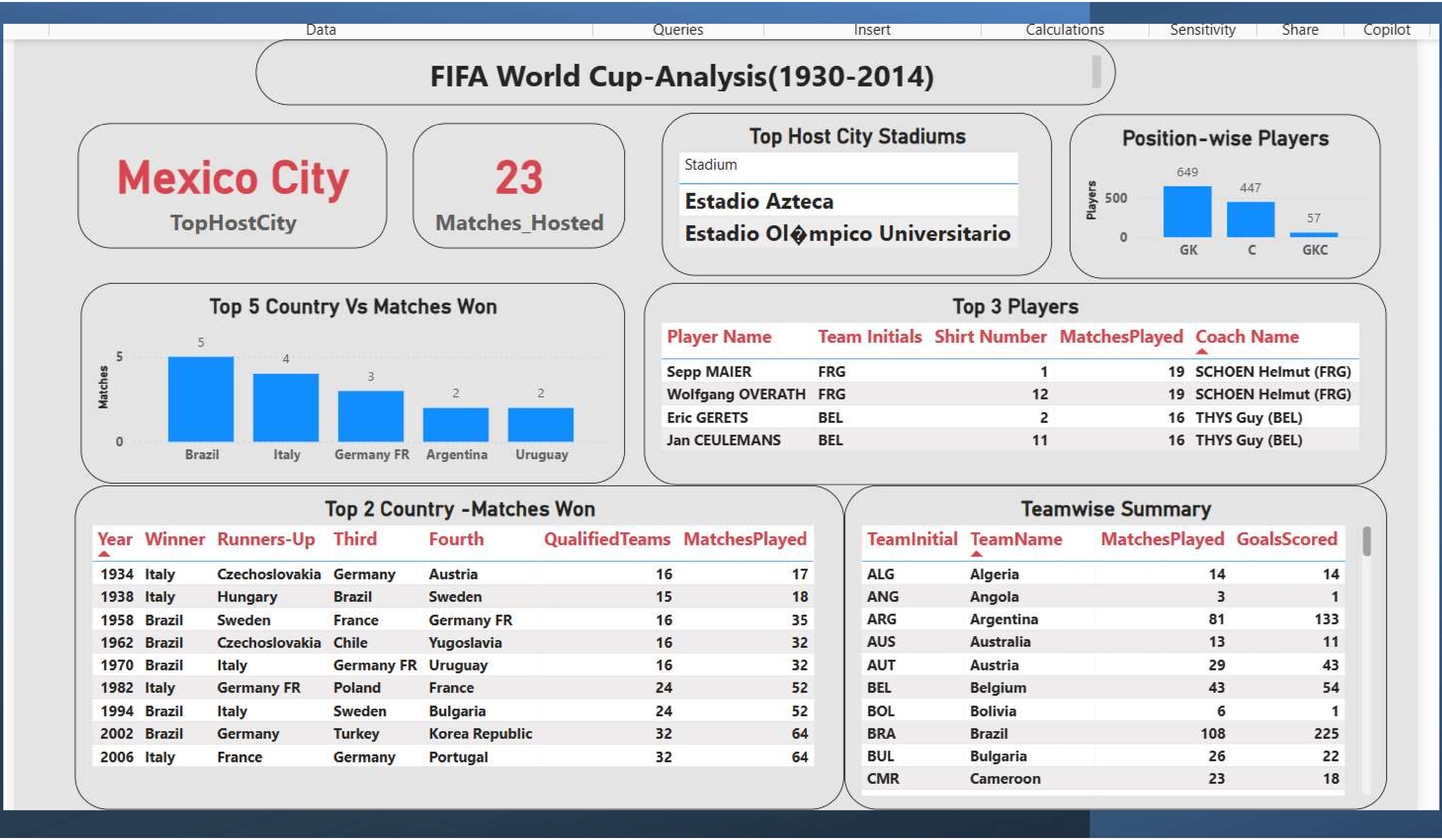
FIFA World-Cup Analysis

Introduction :
The FIFA World Cup is a global football competition contested by the various football-playing nations of the world. It is contested every four years and is the most prestigious and important trophy in the sport of football.

Task :
Find key metrics and insights from the dataset

DataBase

The World Cups dataset shows all information about all the World Cups in history, while the
World Cup Matches dataset shows all the results from the matches contested as part of the
cups



FIFA World Cup-Analysis(1930-2014)

Mexico City

TopHostCity

23

Matches_Hosted

Top Host City Stadiums

Stadium
Estadio Azteca
Estadio Ol�mpico Universitario

Position-wise Players

Top 5 Country Vs Matches Won

Top 3 Players

Player Name	Team Initials	Shirt Number	MatchesPlayed	Coach Name
Sepp MAIER	FRG	1	19	SCHOEN Helmut (FRG)
Wolfgang OVERATH	FRG	12	19	SCHOEN Helmut (FRG)
Eric GERETS	BEL	2	16	THYS Guy (BEL)
Jan CEULEMANS	BEL	11	16	THYS Guy (BEL)

Top 2 Country -Matches Won

Year	Winner	Runners-Up	Third	Fourth	QualifiedTeams	MatchesPlayed
1934	Italy	Czechoslovakia	Germany	Austria	16	17
1938	Italy	Hungary	Brazil	Sweden	15	18
1958	Brazil	Sweden	France	Germany FR	16	35
1962	Brazil	Czechoslovakia	Chile	Yugoslavia	16	32
1970	Brazil	Italy	Germany FR	Uruguay	16	32
1982	Italy	Germany FR	Poland	France	24	52
1994	Brazil	Italy	Sweden	Bulgaria	24	52
2002	Brazil	Germany	Turkey	Korea Republic	32	64
2006	Italy	France	Germany	Portugal	32	64

Teamwise Summary

TeamInitial	TeamName	MatchesPlayed	GoalsScored
ALG	Algeria	14	14
ANG	Angola	3	1
ARG	Argentina	81	133
AUS	Australia	13	11
AUT	Austria	29	43
BEL	Belgium	43	54
BOL	Bolivia	6	1
BRA	Brazil	108	225
BUL	Bulgaria	26	22
CMR	Cameroon	23	18

Insights

Mexico City

Hosted **23** matches

Brazil -5

Italy -4

Top 2 Countries Won most Matches

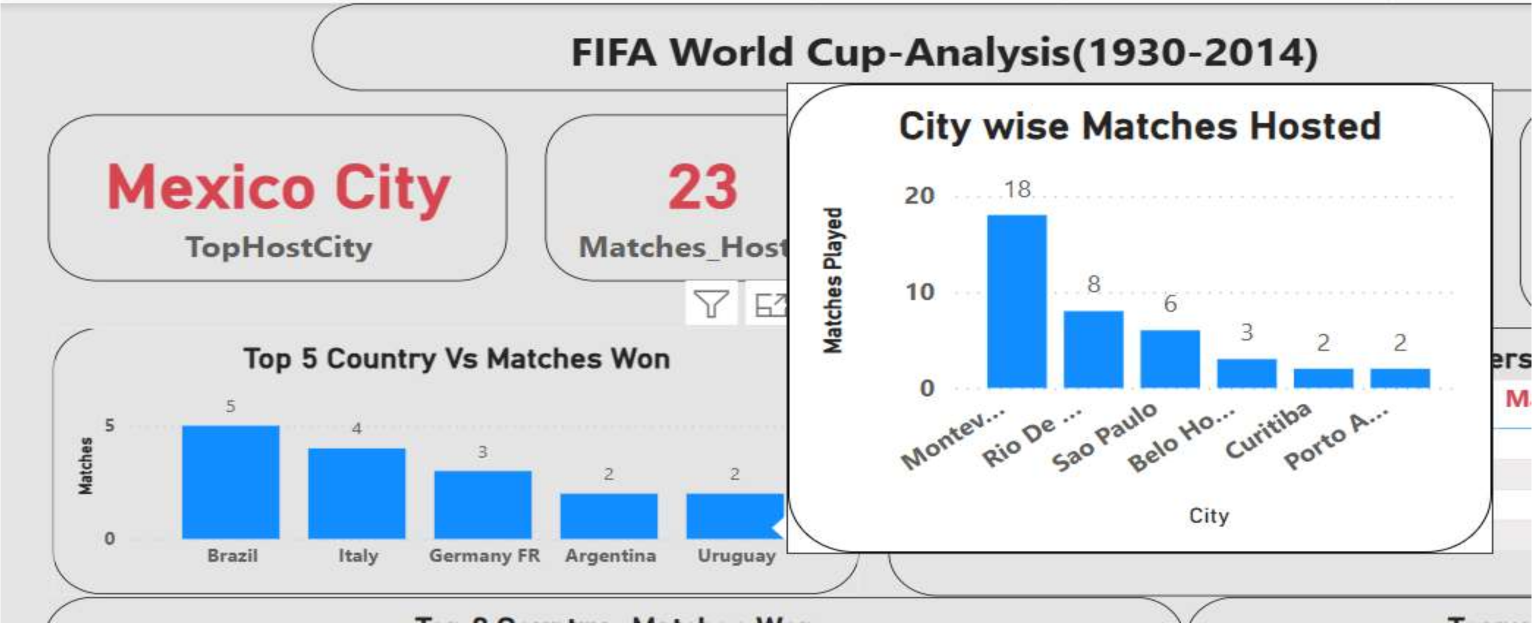
Seep Maier

Wolfgang OverWrath

Played **19** Matches

Coach Schoen Helmet

ToolTip for Top 5 Countries most matches won – citywise Matches Hosted

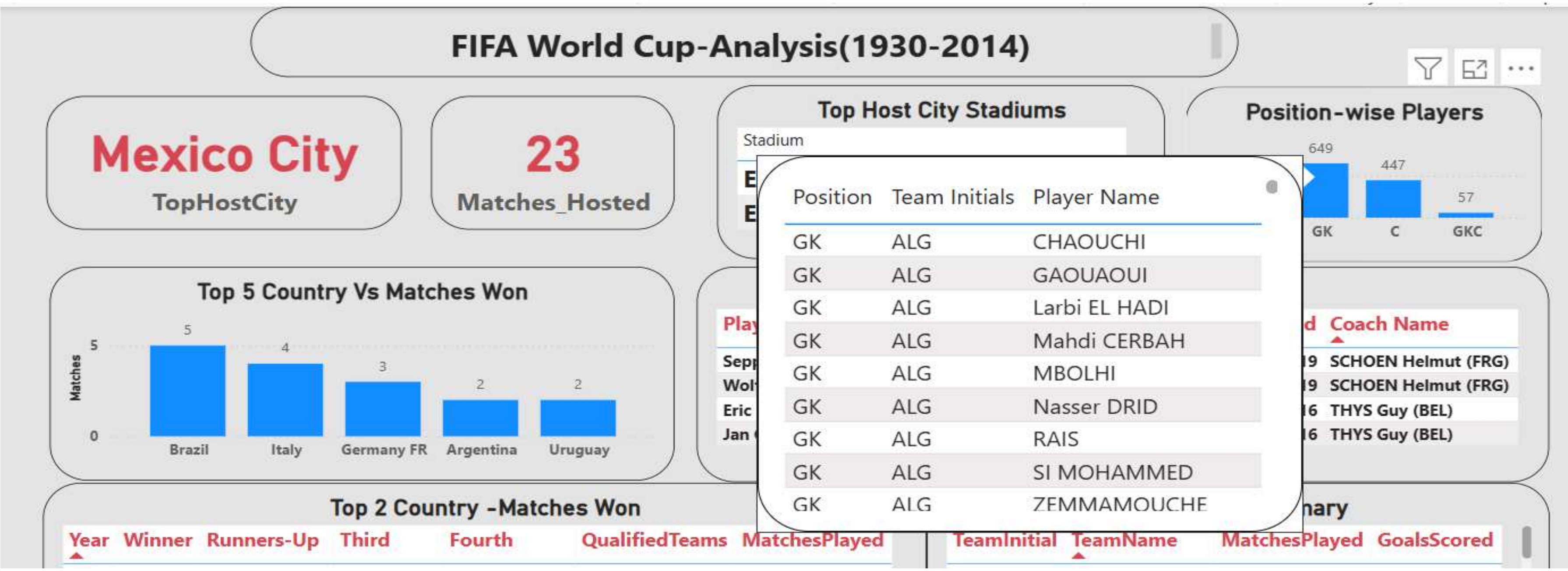


FIFA World Cup-Analysis(1930-2014)

Mexico City
TopHostCity

23
Matches_Host

ToolTip for List of Position-wise Players



FIFA World Cup-Analysis(1930-2014)

Mexico City

TopHostCity

23

Matches_Hosted

Top Host City Stadiums

Stadium

E

E

Position	Team Initials	Player Name
GK	ALG	CHAOUCHI
GK	ALG	GAOUAOUI
GK	ALG	Larbi EL HADI
GK	ALG	Mahdi CERBAH
GK	ALG	MBOLHI
GK	ALG	Nasser DRID
GK	ALG	RAIS
GK	ALG	SI MOHAMMED
GK	ALG	ZEMMAMOUCHE

Position-wise Players

Top 5 Country Vs Matches Won

Play

Sepp

Wol

Eric

Jan

d	Coach Name
9	SCHOEN Helmut (FRG)
9	SCHOEN Helmut (FRG)
6	THYS Guy (BEL)
6	THYS Guy (BEL)

Top 2 Country -Matches Won

Year	Winner	Runners-Up	Third	Fourth	QualifiedTeams	MatchesPlayed

nary

TeamInitial	TeamName	MatchesPlayed	GoalsScored

Projects

Big Game Sensus Analysis

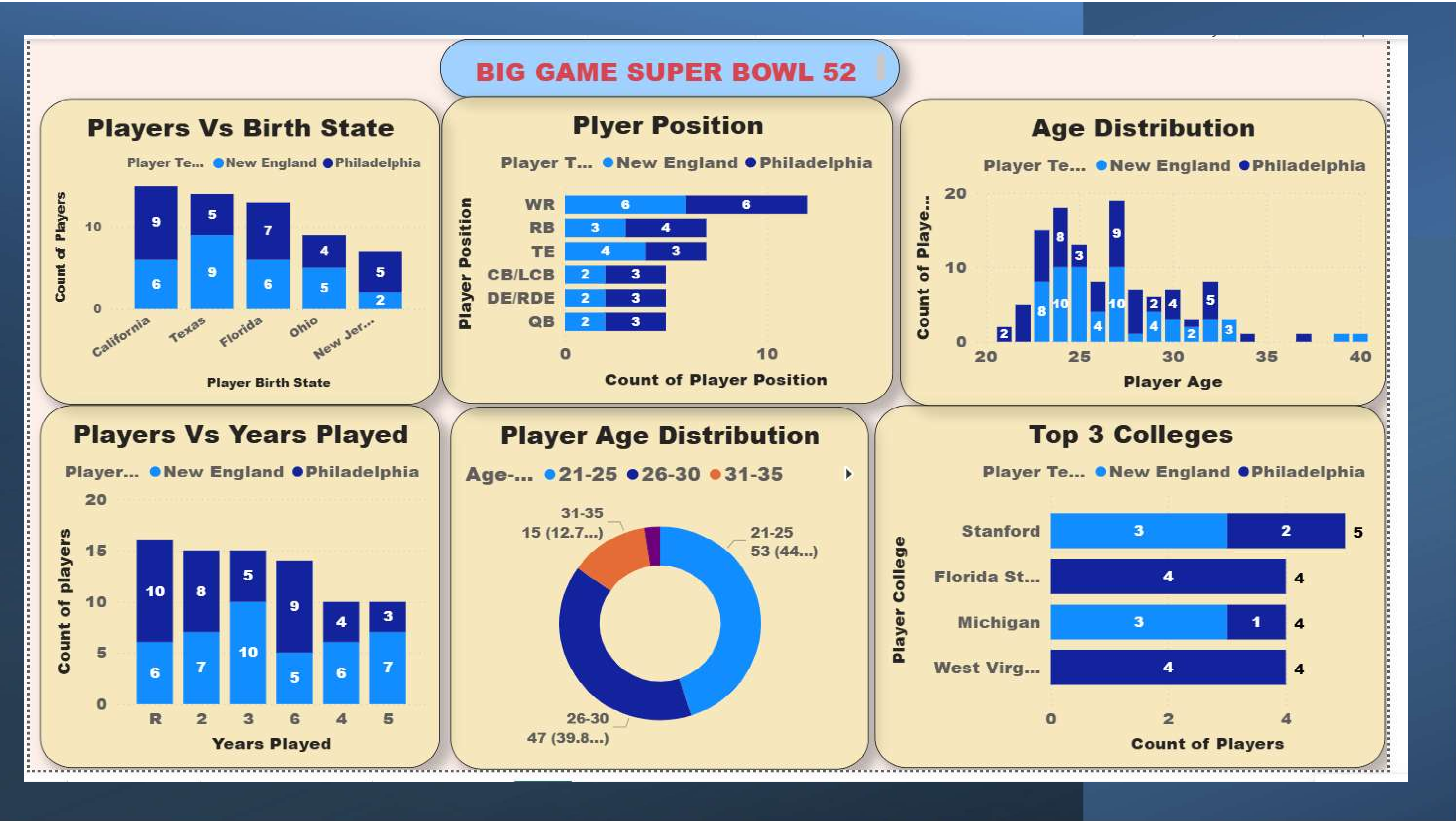
BigGame Cencus Analysis

Introduction :
This Big Game Census data visualization where Super Bowl 52 players census dataset is used. Super Bowl LII, or Super Bowl 52, was the championship game of the National Football League (NFL) for the 2017 season. It was played on February 4, 2018, at U.S. Bank Stadium in Minneapolis, Minnesota. The Philadelphia Eagles defeated the New England Patriots with a score of 41-33, winning their first Super Bowl title.

Task :
Find key metrics and insights from the dataset

DataBase

The dataset has rosters for both, competing teams, with the corresponding roster information and birthplace and state population information. The developers utilized census data pulled from census.gov, and roster information from Yahoo Sports.



BIG GAME SUPER BOWL 52

Players Vs Birth State

Plyer Position

Age Distribution

Players Vs Years Played

Player Age Distribution

Top 3 Colleges

Insights

New England – AFC
Philadelphia – NFC

AFC – Texas
NFC - California

Most Players

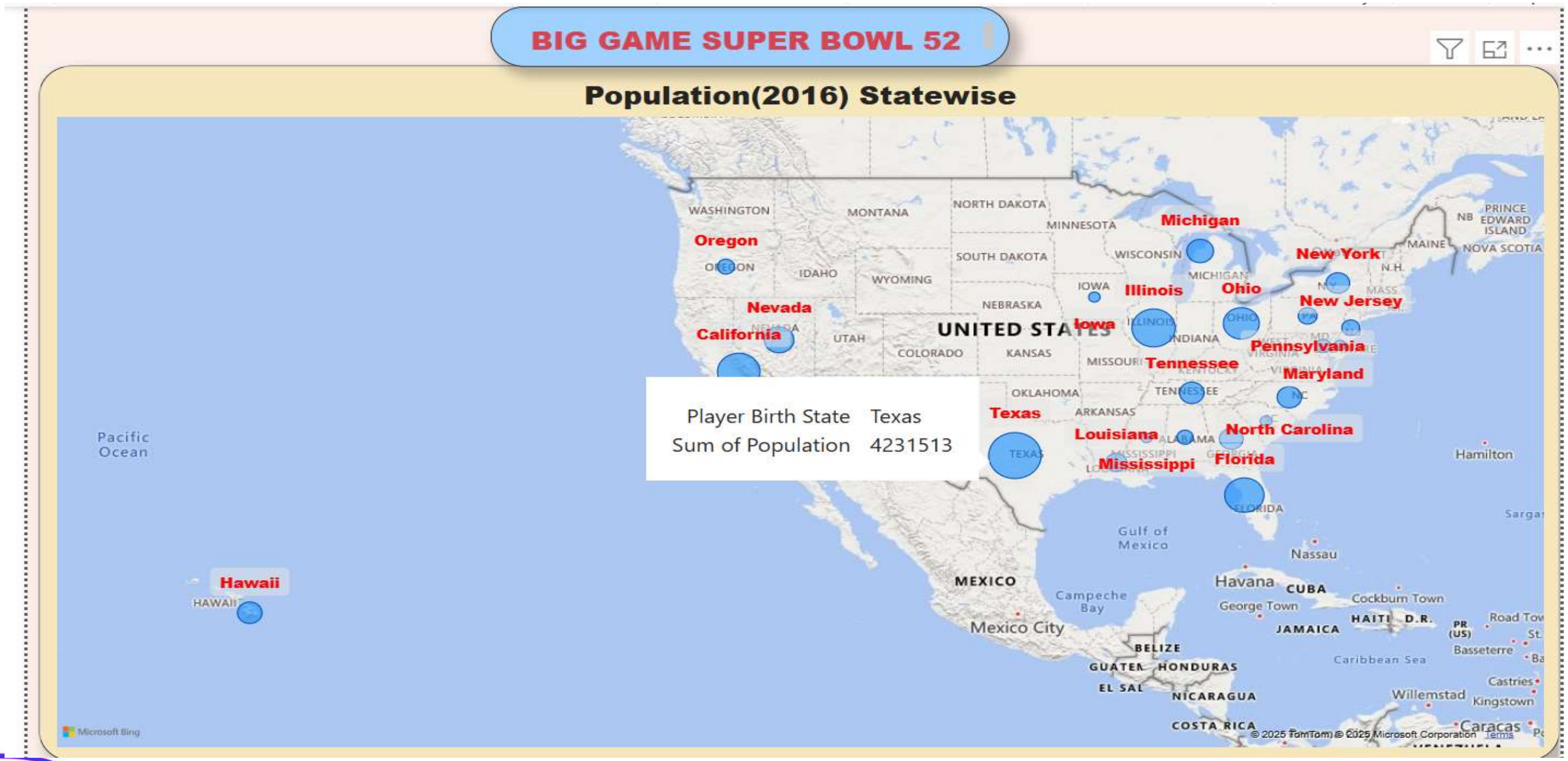
AFC – 13 to 40 yrs
NFC- - 21 to 37 yrs

Age Distribution

Stanford
Florida State
Michigan

Top 3 Most Players Colleges

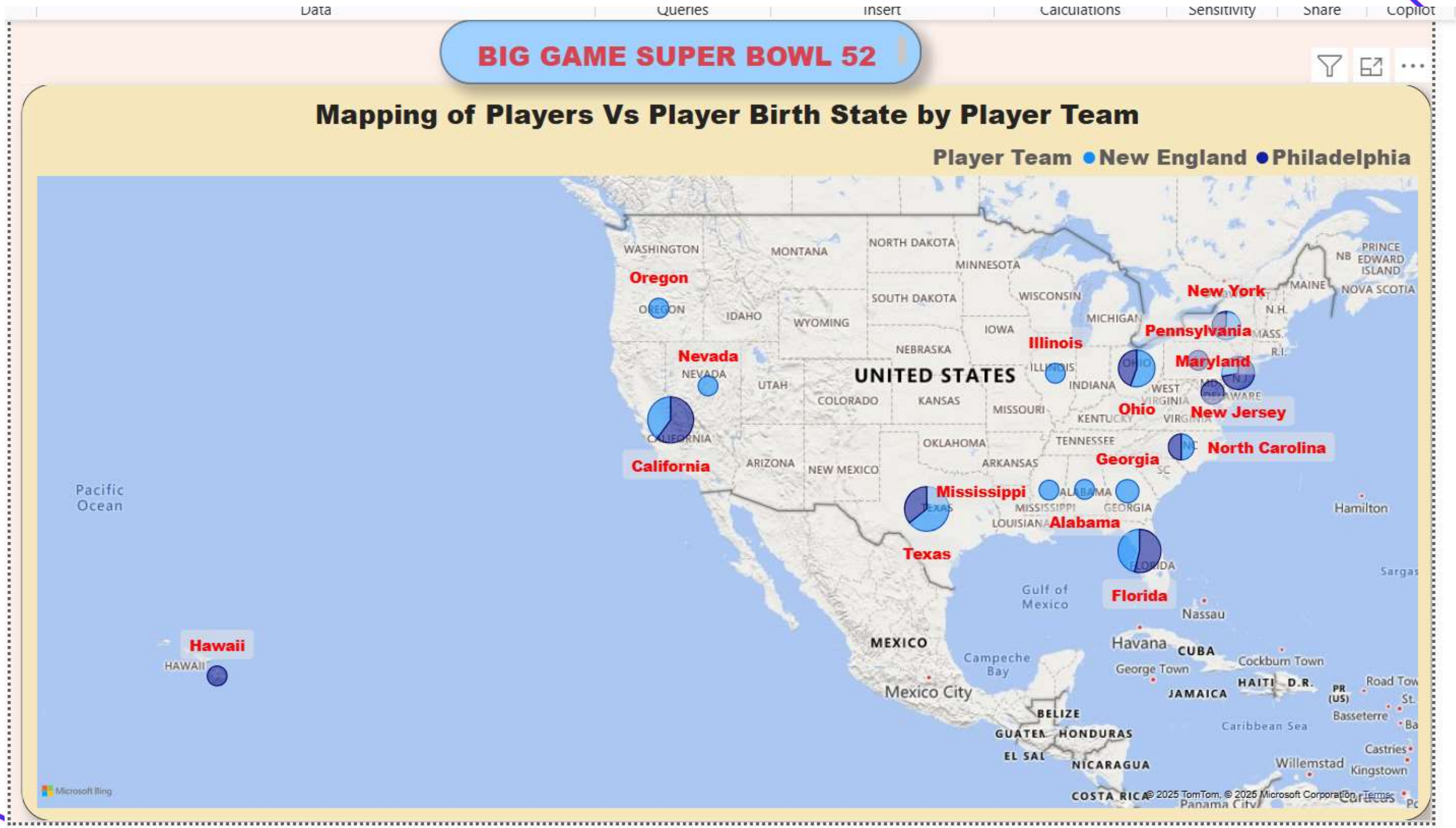
Mostly populated State -Texas



BIG GAME SUPER BOWL 52

Population(2016) Statewise

Player Birth State	Texas
Sum of Population	4231513



BIG GAME SUPER BOWL 52

Mapping of Players Vs Player Birth State by Player Team

Player Team ● New England ● Philadelphia

Projects

Hospitality Analysis

Hospitality Analysis

Introduction :

Atliq Grands owns multiple five-star hotels across India. They have been in the hospitality industry for the past 20 years. Due to strategic moves from other competitors and ineffective decision-making in management, Atliq Grands are losing its market share and revenue in the luxury/business hotels category.

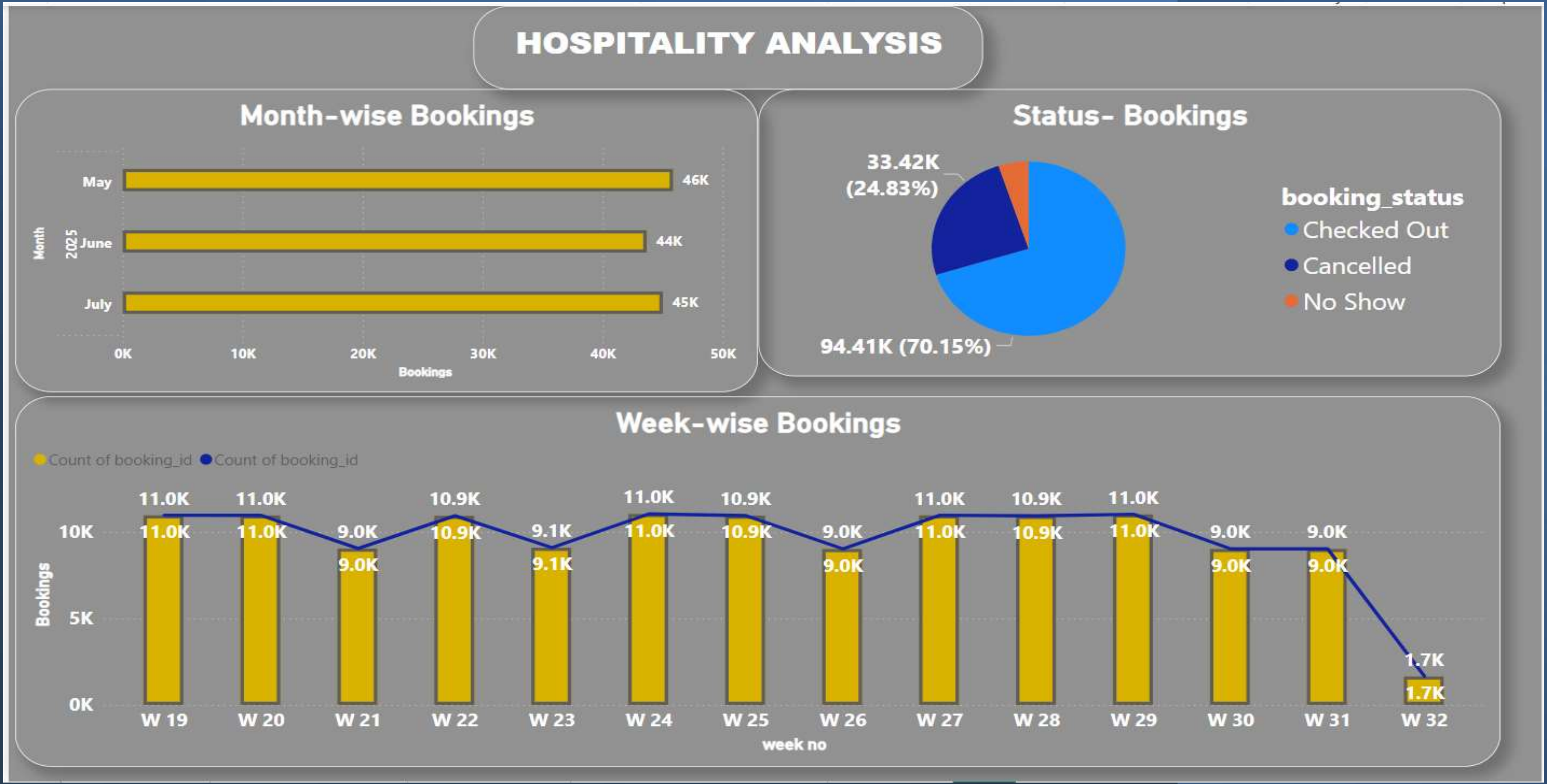
Task :

They do not have an in-house data analytics team to provide them with these insights. Their revenue management team had decided to hire a 3rd party service provider to provide them with insights from their historical data.

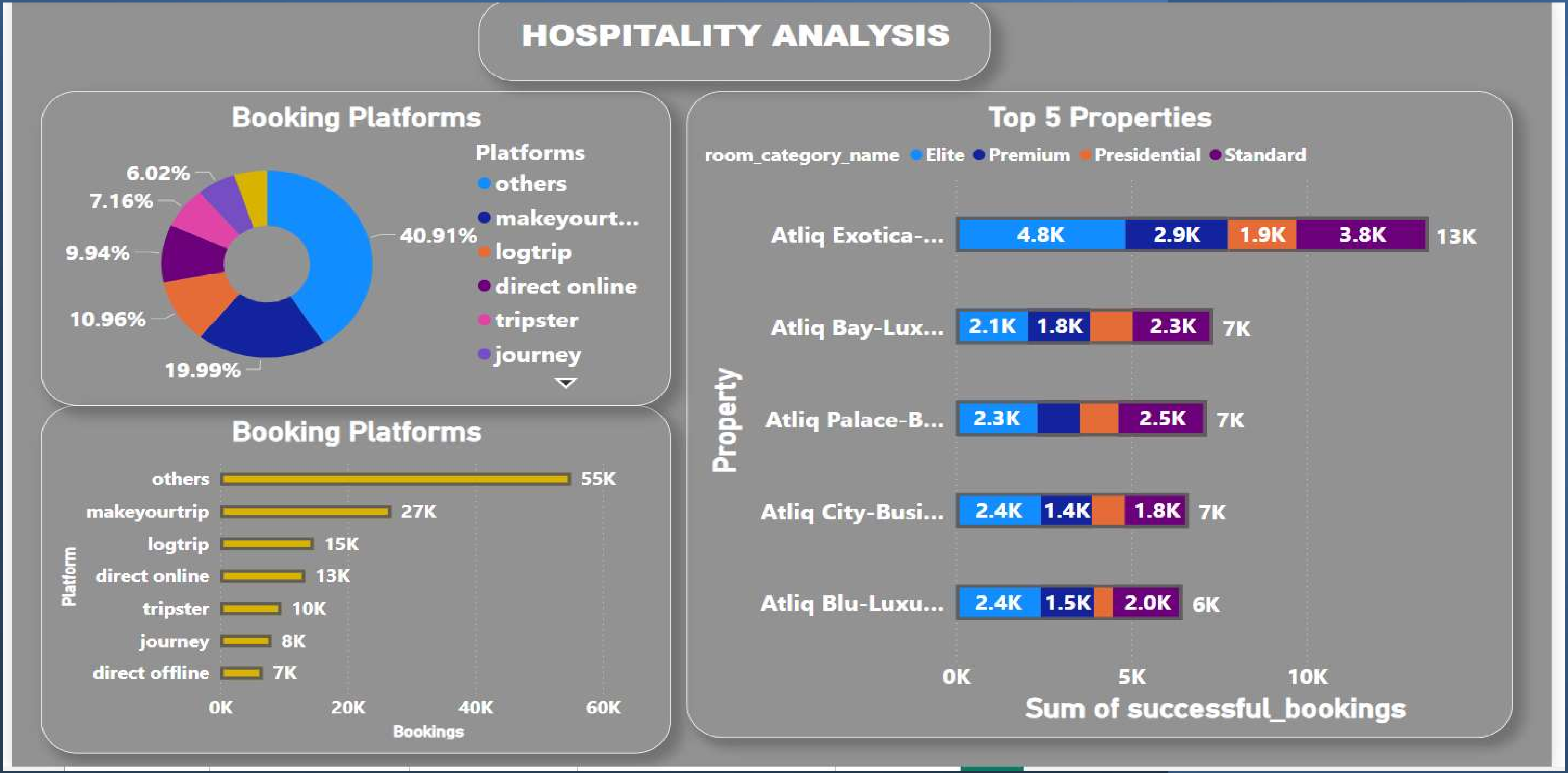
DataBase

3 dimension tables – date, hotels, rooms
2 Fact Tables – aggregated Bookings, Bookings
Data Modeling was done and new measures and summary tables were built

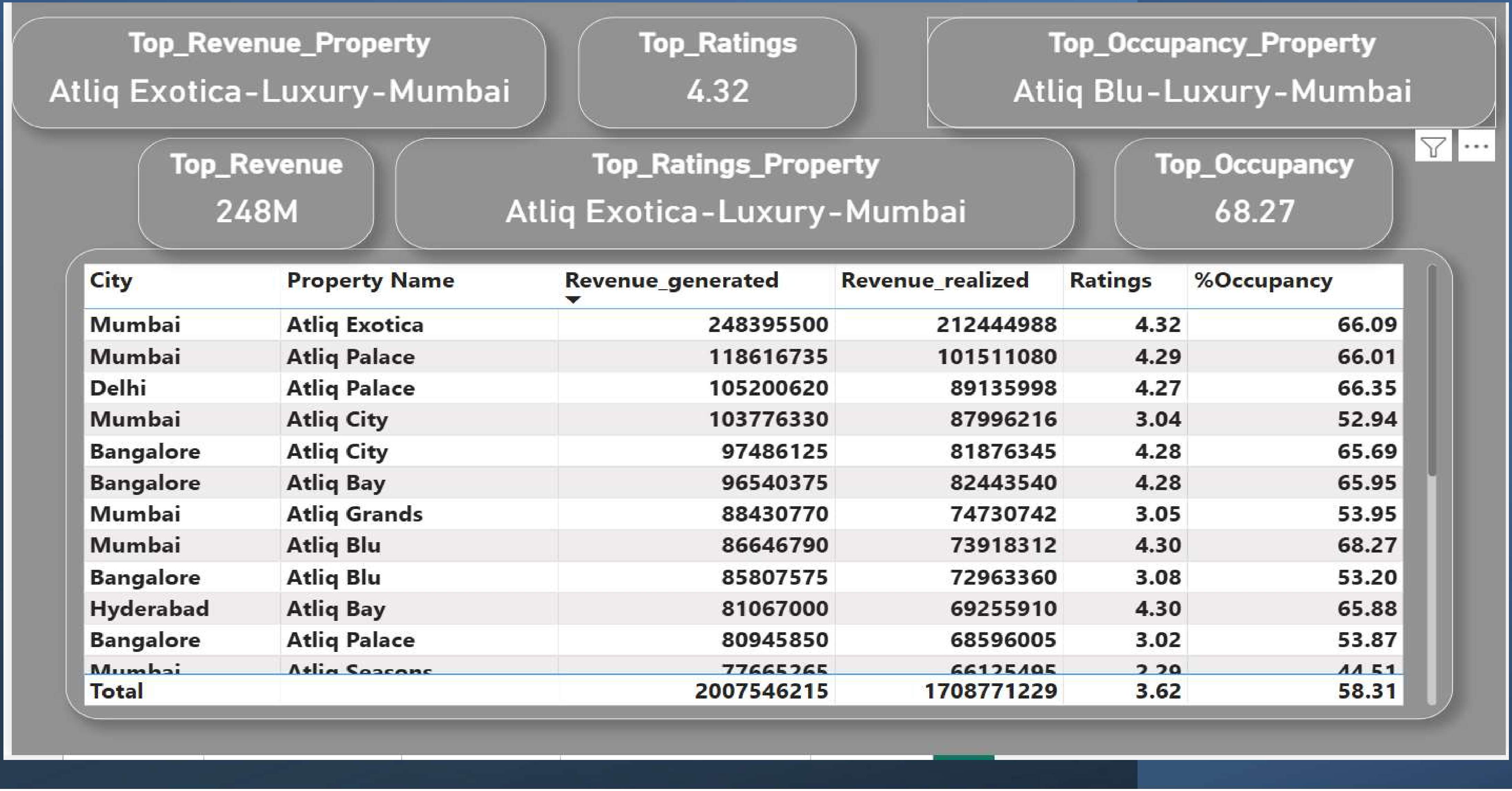
Overview 1


HOSPITALITY ANALYSIS
Month-wise Bookings
Month
2025
May
June
July
46K
44K
45K
0K
10K
20K
30K
40K
50K
Bookings
Status- Bookings
33.42K
(24.83%)
94.41K (70.15%)
booking_status
● Checked Out
● Cancelled
● No Show
Week-wise Bookings
● Count of booking_id ● Count of booking_id
Bookings
10K
5K
0K
11.0K
11.0K
11.0K
11.0K
9.0K
9.0K
10.9K
10.9K
9.1K
9.1K
11.0K
11.0K
10.9K
10.9K
9.0K
9.0K
11.0K
11.0K
10.9K
10.9K
11.0K
11.0K
9.0K
9.0K
9.0K
9.0K
1.7K
1.7K
W 19
W 20
W 21
W 22
W 23
W 24
W 25
W 26
W 27
W 28
W 29
W 30
W 31
W 32
week no

Overview 2


HOSPITALITY ANALYSIS
Booking Platforms
6.02%
7.16%
9.94%
10.96%
19.99%
40.91%
Platforms
● others
● makeyourt...
● logtrip
● direct online
● tripster
● journey
Booking Platforms
others
55K
makeyourtrip
27K
logtrip
15K
direct online
13K
tripster
10K
journey
8K
direct offline
7K
Platform
0K
20K
40K
60K
Bookings
Top 5 Properties
room_category_name
● Elite
● Premium
● Presidential
● Standard
Property
Atliq Exotica-...
4.8K
2.9K
1.9K
3.8K
13K
Atliq Bay-Lux...
2.1K
1.8K
2.3K
7K
Atliq Palace-B...
2.3K
2.5K
7K
Atliq City-Busi...
2.4K
1.4K
1.8K
7K
Atliq Blu-Luxu...
2.4K
1.5K
2.0K
6K
0K
5K
10K
Sum of successful_bookings

KPI and Metrics



Top_Revenue_Property
Atliq Exotica-Luxury-Mumbai

Top_Ratings
4.32

Top_Occupancy_Property
Atliq Blu-Luxury-Mumbai

Top_Revenue
248M

Top_Ratings_Property
Atliq Exotica-Luxury-Mumbai

Top_Occupancy
68.27

City	Property Name	Revenue_generated	Revenue_realized	Ratings	%Occupancy
Mumbai	Atliq Exotica	248395500	212444988	4.32	66.09
Mumbai	Atliq Palace	118616735	101511080	4.29	66.01
Delhi	Atliq Palace	105200620	89135998	4.27	66.35
Mumbai	Atliq City	103776330	87996216	3.04	52.94
Bangalore	Atliq City	97486125	81876345	4.28	65.69
Bangalore	Atliq Bay	96540375	82443540	4.28	65.95
Mumbai	Atliq Grands	88430770	74730742	3.05	53.95
Mumbai	Atliq Blu	86646790	73918312	4.30	68.27
Bangalore	Atliq Blu	85807575	72963360	3.08	53.20
Hyderabad	Atliq Bay	81067000	69255910	4.30	65.88
Bangalore	Atliq Palace	80945850	68596005	3.02	53.87
Mumbai	Atliq Seasons	77665265	66125495	2.29	44.51
Total		2007546215	1708771229	3.62	58.31

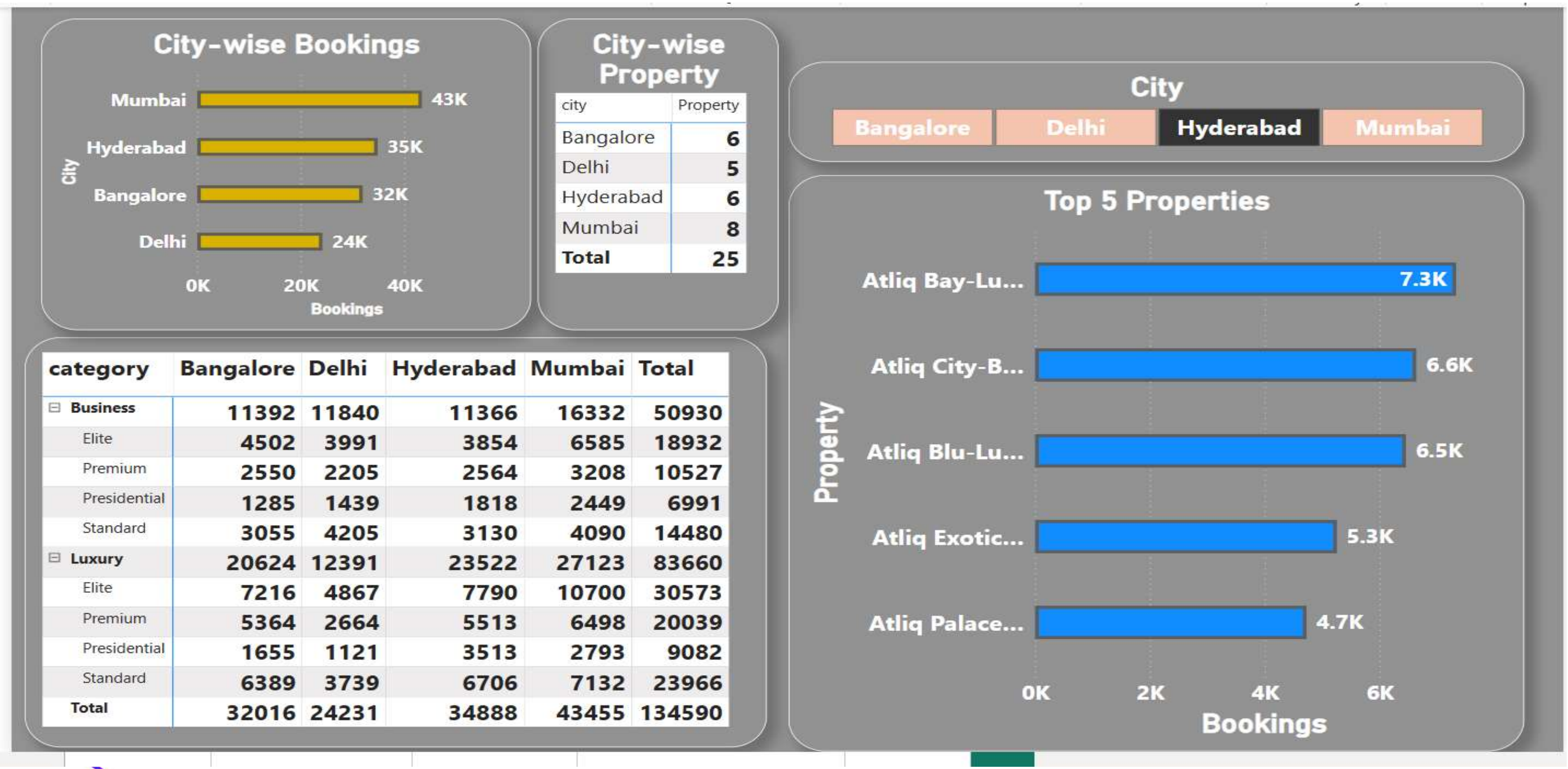
Insights

Bookings
Mumbai – 43K
Delhi -23K
Delhi business can be reworked for improvement

Cancellation – 34K
No show – 7K
Booking platforms and partners strategy to be revised to reduce this

Max Occupany % = 68
This can be improved by matching with demand of standard and elite rooms is more

City-wise Details

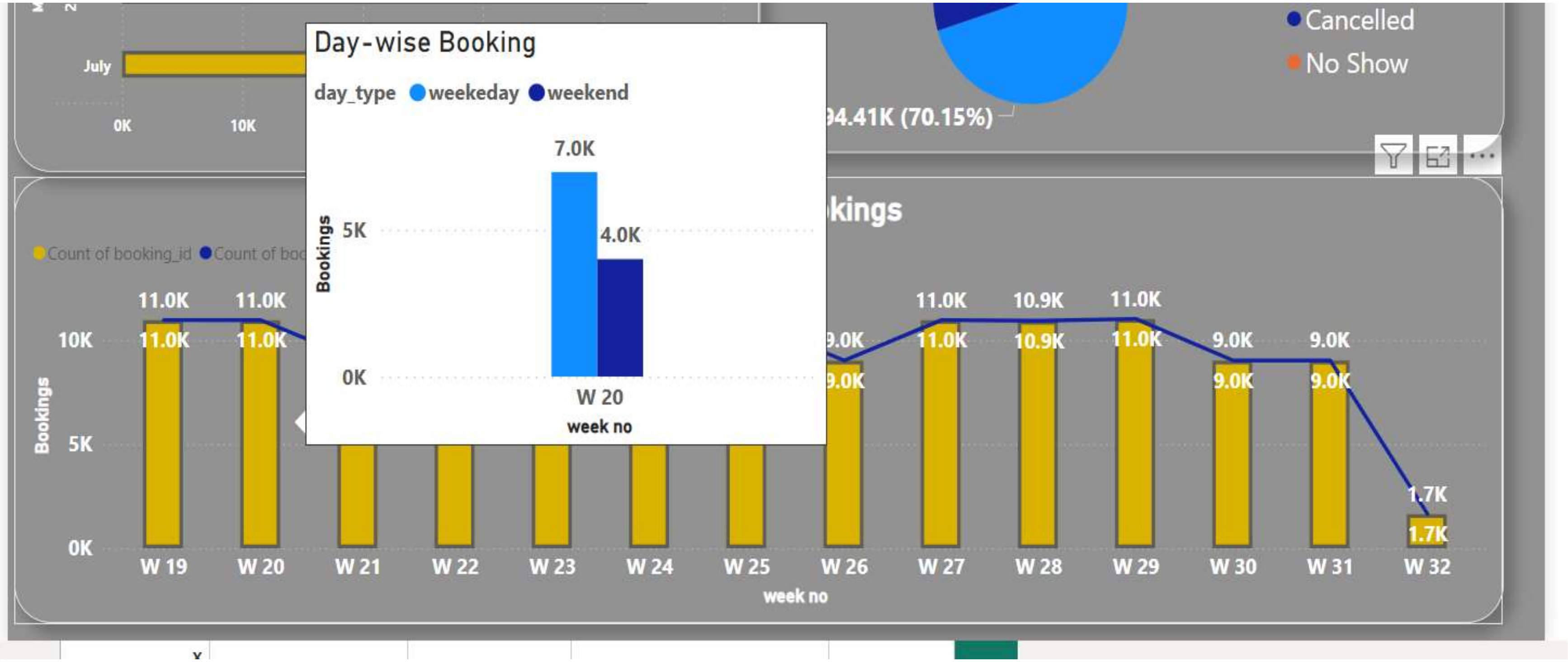


City-wise Property

city	Property
Bangalore	6
Delhi	5
Hyderabad	6
Mumbai	8
Total	**25**

category	Bangalore	Delhi	Hyderabad	Mumbai	Total
⊟ **Business**	11392	11840	11366	16332	50930
Elite	4502	3991	3854	6585	18932
Premium	2550	2205	2564	3208	10527
Presidential	1285	1439	1818	2449	6991
Standard	3055	4205	3130	4090	14480
⊟ **Luxury**	20624	12391	23522	27123	83660
Elite	7216	4867	7790	10700	30573
Premium	5364	2664	5513	6498	20039
Presidential	1655	1121	3513	2793	9082
Standard	6389	3739	6706	7132	23966
Total	32016	24231	34888	43455	134590

ToolTip for Weedays/weekends bookings for each week



Projects

Crop Production Analysis

Crop Production Analysis

Introduction :

The Agriculture business domain, as a vital part of the overall supply chain, is expected to highly evolve in the upcoming years via the developments, which are taking place on the side of the Future Internet. This paper presents a novel Business-to-Business collaboration platform from the agri-food sector perspective, which aims to facilitate the collaboration of numerous stakeholders belonging to associated business domains, in an effective and flexible manner

Task :

Make views and dashboards first and also make a story out of it

DataBase

This dataset provides a huge amount of information on crop production in India ranging from several years.

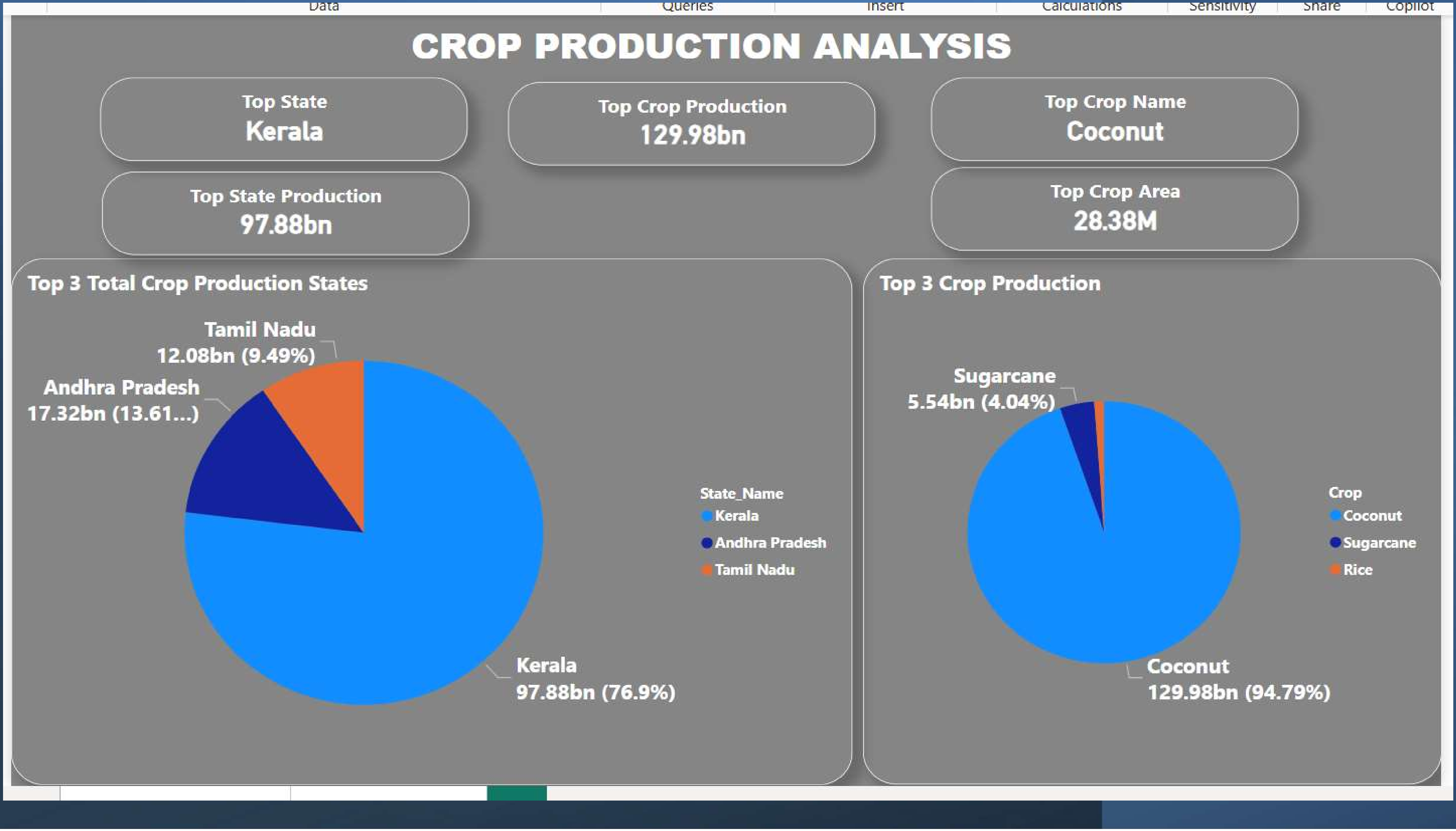
33 states

646 districts

19 years

6 seasons

124 Crops



CROP PRODUCTION ANALYSIS

Top State
Kerala

Top Crop Production
129.98bn

Top Crop Name
Coconut

Top State Production
97.88bn

Top Crop Area
28.38M

Top 3 Total Crop Production States

Top 3 Crop Production

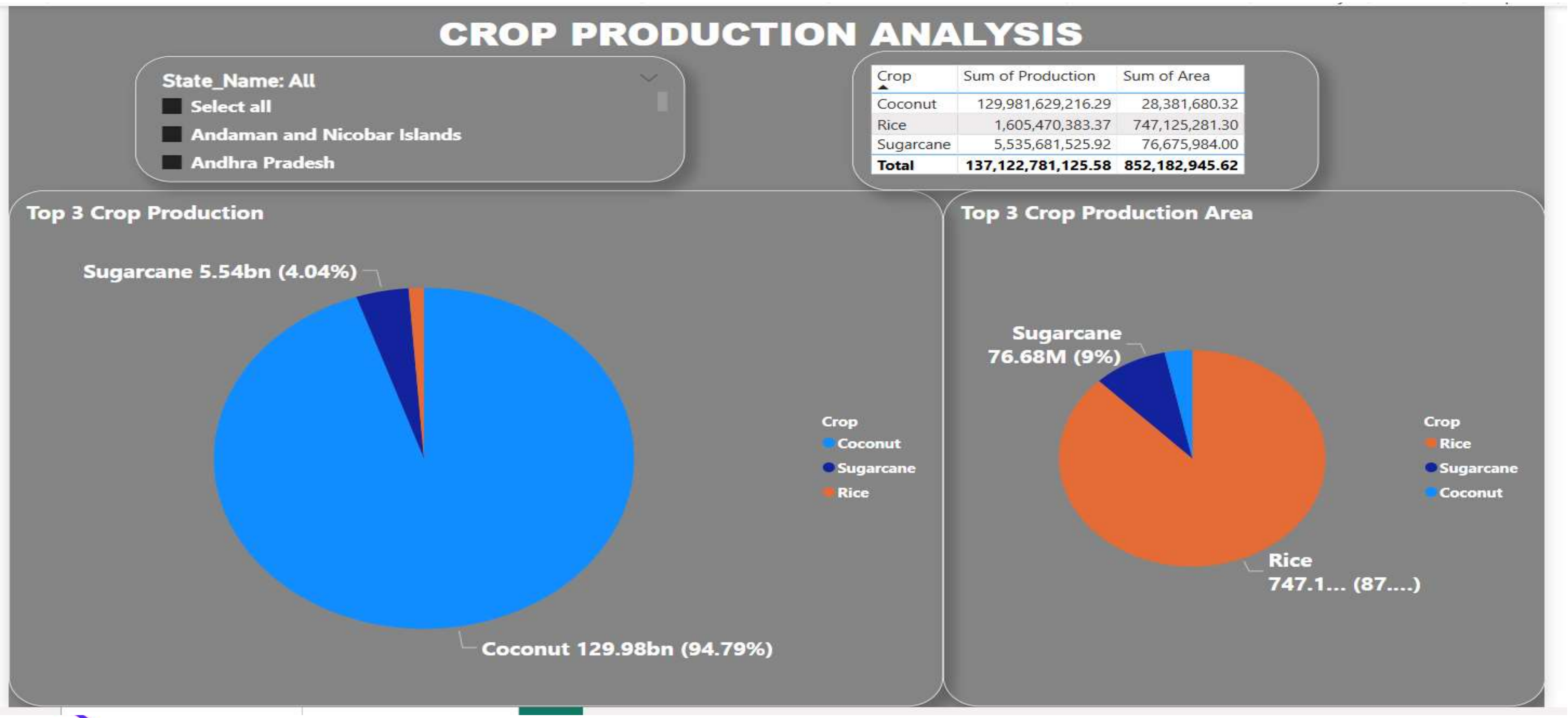
Insights

Kerala – 77 %
Andra Pradesh – 13.5 %
Tamil Nadu – 9.5%
Top **3** Crop Production States

Kerala – 97.88 bn
Top Crop Production State

Coconut – 129.98 bn
Area – 28.38M
Top production crop in Country

State-wise Details



CROP PRODUCTION ANALYSIS

State_Name: All

- ■ Select all
- ■ Andaman and Nicobar Islands
- ■ Andhra Pradesh

Crop	Sum of Production	Sum of Area
Coconut	129,981,629,216.29	28,381,680.32
Rice	1,605,470,383.37	747,125,281.30
Sugarcane	5,535,681,525.92	76,675,984.00
Total	**137,122,781,125.58**	**852,182,945.62**

Projects

Chatbots

Research Papers QAbot

Introduction :

The Chatbots Machine Learning project involves developing a conversational agent (chatbot) capable of interacting with users in natural language. This can include answering questions, providing information, performing tasks, or holding a conversation.

Task :

An intelligent Question-Answering Bot that lets you upload research papers (PDFs) and ask natural language queries. The bot uses semantic search and text summarization to retrieve the most relevant content from your uploaded papers.

DataBase

I have used Research Papers regarding Autism Syndrome Disease(ASD) in children and how AI helps in early detection of it.

Any no. of research papers(PDFs) can be uploaded and the Bot will answer your related queries.

🧠 How It Works

1.**Upload PDFs** → Extracts (PyPDF2) and splits text into chunks.

2.**Huggingface Sentence Transformers/all-MiniLM-L6-v2 tokenizer/Embedding** → Each chunk is converted to a vector.

3.**Pinecone Indexing** → Embeddings are stored in Pinecone.

4.**Query Input** → User asks a question.

5.**Similarity Search** → Bot finds closest matching chunks with pinecone index cosine metrics

6.**Summarize/Answer** → Cohere llm – command nightly generates the final answer.

7. App : https://research-appbot-d9qnnrbgn9piznzet28rcw.streamlit.app/

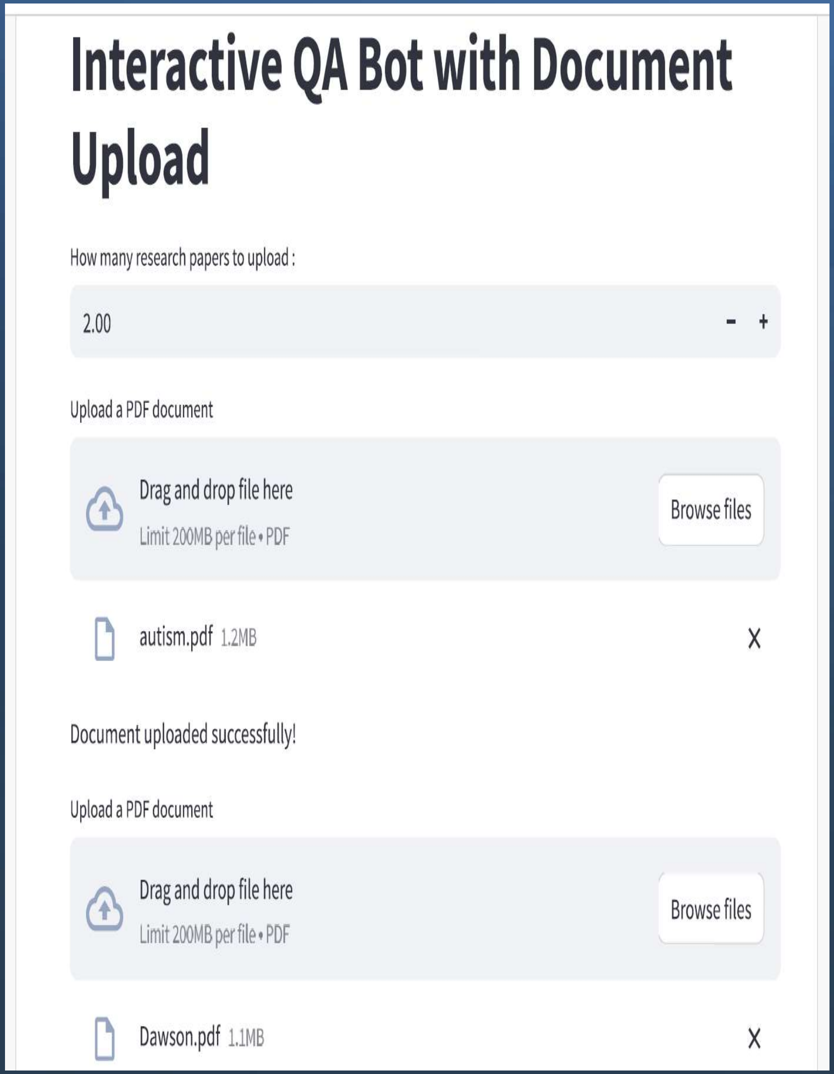
8. Project : https://github.com/gitmamtahub/Research-QAbot

Architecture

All-MiniLM-L6-v2
Huggingface Sentence-transformers tokenizer, embedding

Pinecone
Vector DB with dimensions=384 and metric=cosine

Command-nightly
Cohere model to generate/summarize answer



Interactive QA Bot with Document Upload

How many research papers to upload :

2.00

Upload a PDF document

Drag and drop file here

Limit 200MB per file • PDF

autism.pdf 1.2MB

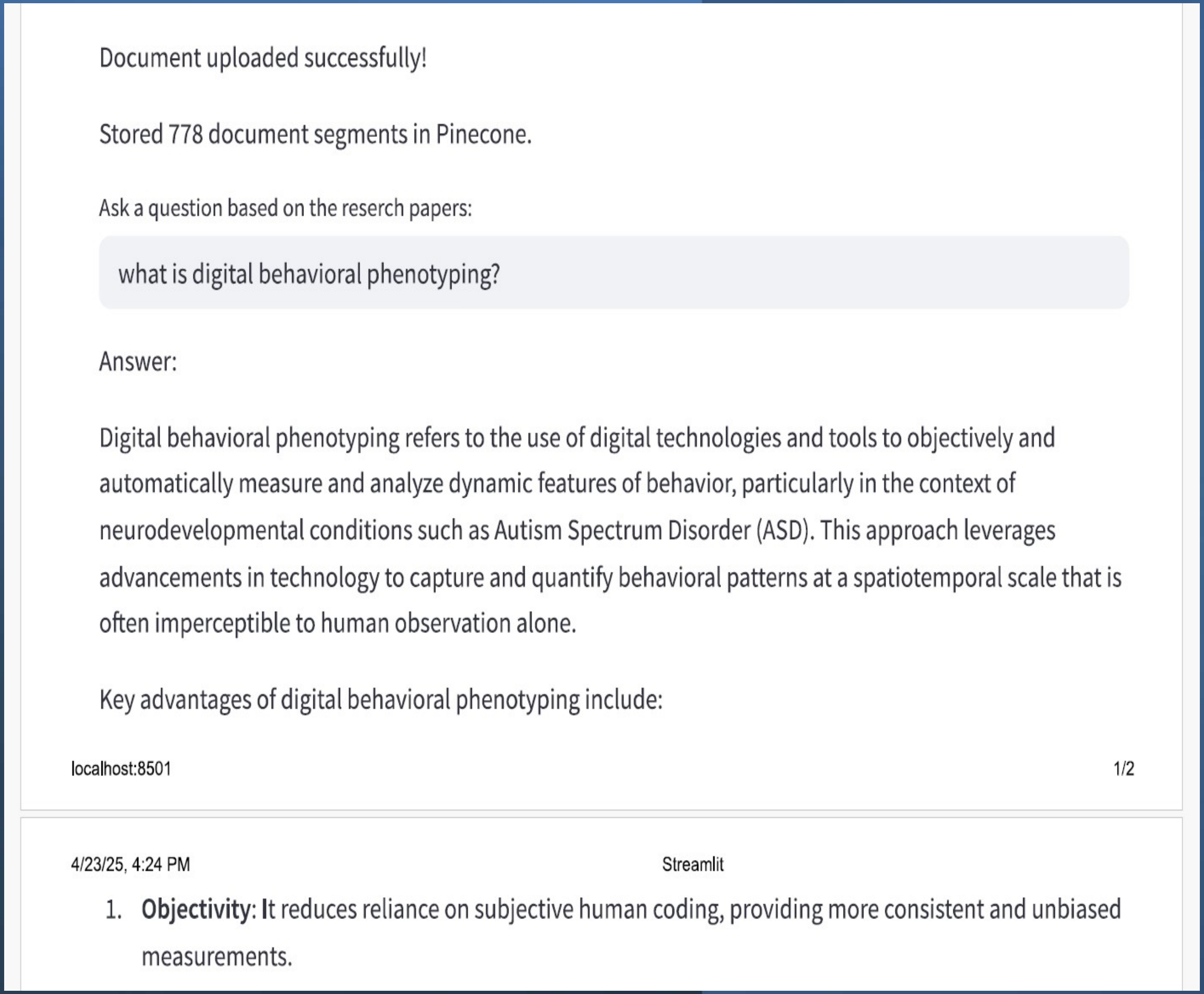
Document uploaded successfully!

Upload a PDF document

Drag and drop file here

Limit 200MB per file • PDF

Dawson.pdf 1.1MB



Document uploaded successfully!

Stored 778 document segments in Pinecone.

Ask a question based on the reserch papers:

what is digital behavioral phenotyping?

Answer:

Digital behavioral phenotyping refers to the use of digital technologies and tools to objectively and automatically measure and analyze dynamic features of behavior, particularly in the context of neurodevelopmental conditions such as Autism Spectrum Disorder (ASD). This approach leverages advancements in technology to capture and quantify behavioral patterns at a spatiotemporal scale that is often imperceptible to human observation alone.

Key advantages of digital behavioral phenotyping include:

localhost:8501 1/2

4/23/25, 4:24 PM Streamlit

1. **Objectivity:** It reduces reliance on subjective human coding, providing more consistent and unbiased measurements.

Projects

Climate Change Modeling

Climate Change Modeling

Introduction :

The Climate Change Modeling project aims to develop a machine learning model to predict and understand various aspects of climate change. This can include Sentiment Analysis, Trend Analysis, Engagement Analysis, Topic Modeling, etc.

Task :

Sentiment Analysis: Gauge public opinion on climate change and NASA's communication strategies.

Topic Modeling: Discover prevalent themes in public discourse about climate change.

DataBase

This dataset encompasses over 500 user comments collected from high-performing posts on NASA's Facebook page dedicated to climate change (https://web.facebook.com/NASAClimateChange/). The comments, gathered from various posts between 2020 and 2023, offer a diverse range of public opinions and sentiments about climate change and NASA's related activities.

DataBase

Column Descriptors

1. **Date:** The date and time when the comment was posted.
2. **LikesCount:** The number of likes each comment received.
3. **ProfileName:** The anonymized name of the user who posted the comment.
4. **CommentsCount:** The number of responses each comment received.
5. **Text:** The actual text content of the comment.

🧠 How It Works

1. Libraries → keybert, transformers, torch, langchain, wordcloud

2. Sentiment Analysis – Autotokenizer/AutoModelForSequenceClassification, open source model - "cardiffnlp/twitter-roberta-base-sentiment"

3.Topic Modeling - KeyBERT(model='all-MiniLM-L6-v2')

4. **Colab code file :** https://colab.research.google.com/drive/1mE0WqxBTZwXzpi3lrIbcYndVZhdqA5aK?usp=sharing

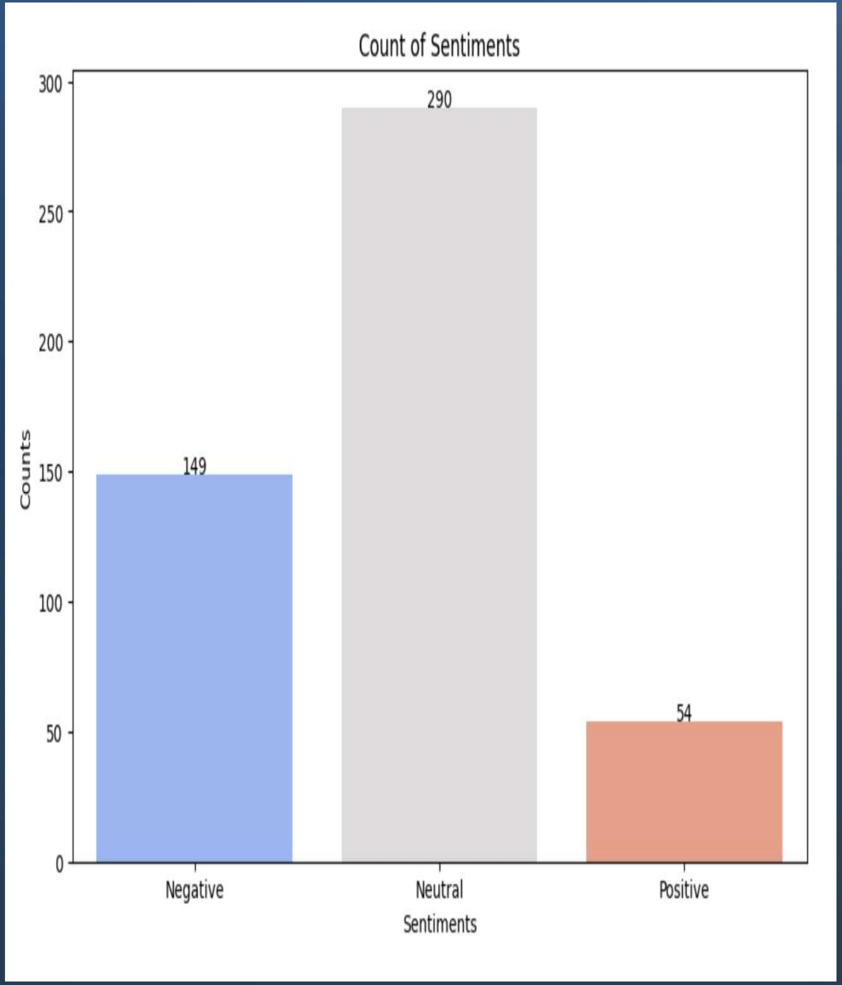
Architecture

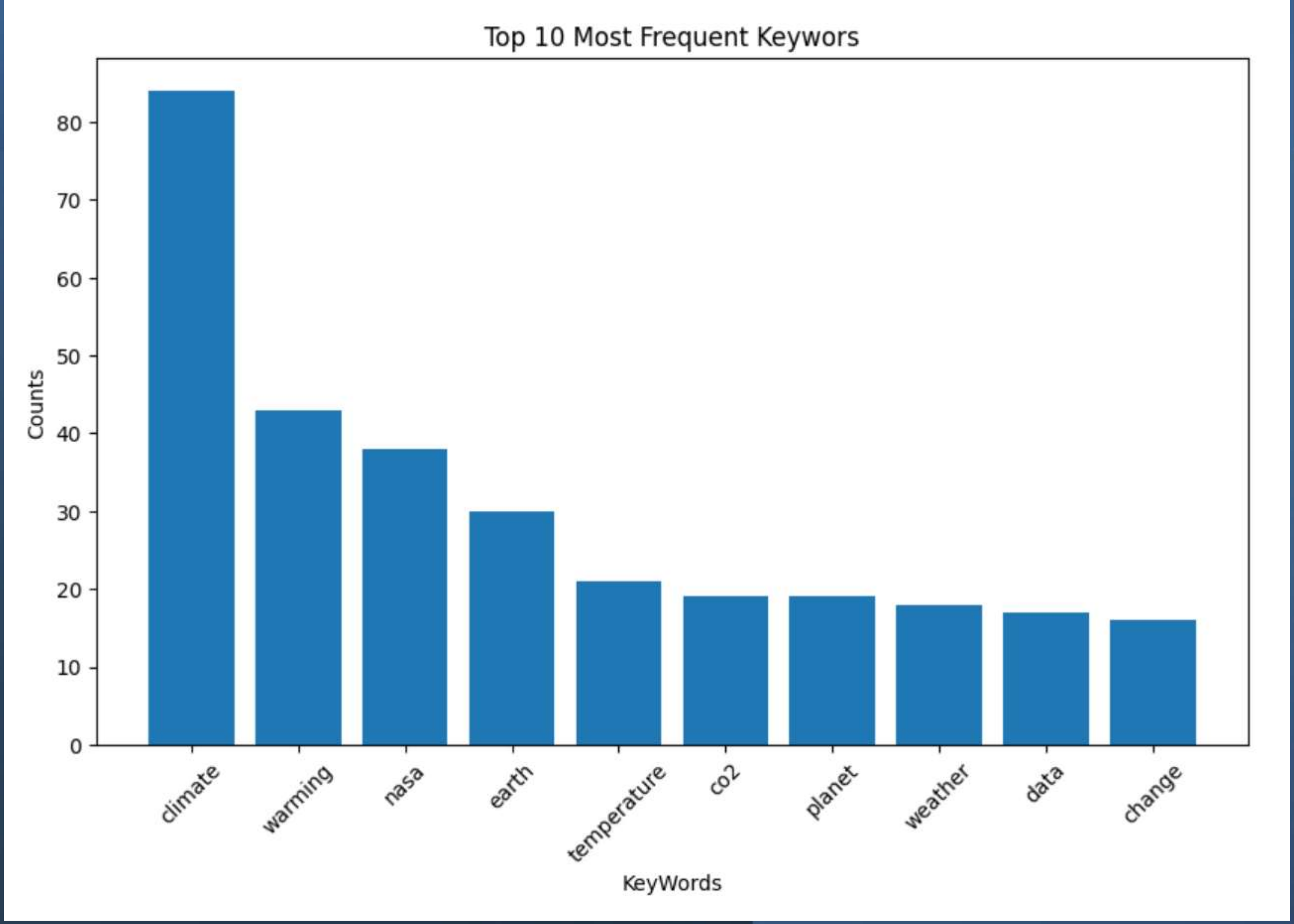
AutoTokenizer
AutoModelForSequenceClassification
transformers

"cardiffnlp/twitter-roberta-base-sentiment"
Sentiment Analysis

KeyBERT(All-MiniLM-L6-v2)
Topic Modeling

Sentiment Analysis and Toping Modeling Plots





WordCloud

climate
temperature
warming
earth
nasa
planet
year
data
scientist
change
weather
greenhouse
ice
co2
science
global
people
emission
carbon
excellent
hemisphere
gases
stop
warmer
facts
better
decaf
graph
oil
post
argue
difference
cold
dont
love
pattern
denial
youtube
energy
methane
extinction
feel
december
day
trend
forest
florida
accurately
lie
sea
average
plastic
level
wow
disaster
watch
eruptions
anthropogenic
fear
stratosphere
glacier
temps
anomaly
know
chart
increase
heat
cooling
just
halloween
1930s
reason
born
radiation
ignorant
fake
think
president
pollution
understand
changing
sustainable
magnetic
ppm
caused
clouds
agree
colder
drink
sun
right
start
world
nature
degree
growing
environment
cycle
government
yes
trolls
plan
gas
record
solar
medieval
southern
mean
going
politicians
tree
atmosphere
scientific
air
human
northern
thermometer
historical
new
talking
horrendous
outdoors
political
thought
deniers
summer
fires
way
line
hotter
happen
rain
scary
wake
nut
hope
thats
cooler
coffee
ocean
cause
propaganda
need
extreme
species
pandemic
humanity
live
sunspot
hot
like
fuel
winter
moved
oh
heating
good
did
ignorance
garbage
information
life
agenda
volcanic
industry
effect
warm
natural
reduce
dioxide
hottest
ago
reducing
interesting
correlation
aware
fossil
read
measure
anthropocene
appening
climatic
water
happened
care
time
thank
believe
great
70s
building
comments
does
rctic

Projects

Healthcare(Heart Disease Analysis)

HealthCare (Heart Disease)

Introduction :

Blood datasets typically encompass a broad array of information related to hematology, blood chemistry, and related health indicators. These datasets often include data points such as blood cell counts, hemoglobin levels, hematocrit, platelet counts, white blood cell differentials, and various blood chemistry parameters such as glucose, cholesterol, and electrolyte levels. Machine learning techniques are often applied to blood datasets to develop predictive models for diagnosing Diseases.

Task :

Diagnosis of Heart Disease

DataBase

Diagnosis of Heart Disease from health details like BMI, Race, Age, Smoking, Asthma, Diabetic, etc. Also physically Active, Skin cancer, Kidney Disease, Sleep time, etc.

Key-Concern

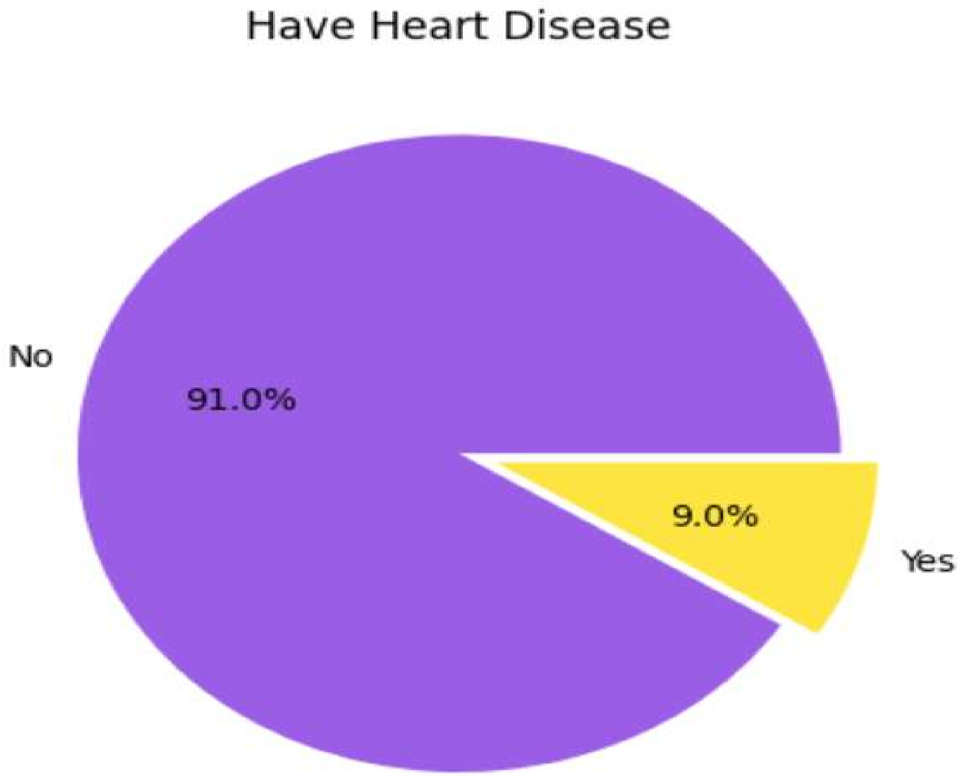
Main Focus is no Heart Disease patient should be left undetected. In ML terms False Negative cases should be as low as possible near to null Means If Patient is having heart disease but Reports are negative i.e. False Negative

Note:
The Confusion Matrix created has four different quadrants: (for Python) TN FP FN TP True Negative (Top-Left Quadrant) False Positive (Top-Right Quadrant) False Negative (Bottom-Left Quadrant) True Positive (Bottom-Right Quadrant)

Process

1. Data Understanding
2. Data Cleaning
3. Exploratory Data Analysis
4. Data Modeling and Evaluation
5. Uncover Insights

UnBalanced Dataset



1)More data collection
2)SMOTE
3)Explicitly Balance data

<u>EDA for Patients having Heart Disease</u>

1) Male – 58.9%, Female – 41.4%
2) Smokers - 63% of Male, 52.4% of Female
3) Alcohol – 4.4% of Male, 3.8% of Female
4) Stroke – 14.8% of Male, 17.9% of Female
5) Difficult Walking – 30.5% of Male, 45.8 % of Female
6) Physical Active – 68.5% of Male, 57% of Female
7) Diabetic – 33.4% of Male, 32% of Female
8) Asthma – 13.9% of Male, 24.1% of Female
9) Skin cancer – 20.5% of Male, 14.9% of Female
10) Kidney Disease – 11.4% of Male, 14.5% of Female
11) Age > 40, BMI>18.5
12) General Health and Sleep time doesnot give any clear idea for heart disease

Insights

Colab code file : https://colab.research.google.com/drive/1h0OFc0a7wGy2MNvXz1zLojEvNmUwgfrr?usp=sharing

LR, DT, RF, XGB
ML models trained/tested

89 to 95%
Accuracy

FN -Poor Recall and Precision
Drawback of all models

Acknowledgement

PowerBI skills :
Learnt making interactive visuals Dashboards, Extract, Transform and Load Data

Exploratory Data Analysis skills:
Python with libraries like pandas, numpy, matplotlib, pyplot

ML-AI Skills :
ML models LR, DT, RF, XGB.
Transformers- tokenizer, embedding, vector DB(pinecone), LLM models – KeyBert, Roberta, cohere, RAG

Thank you!