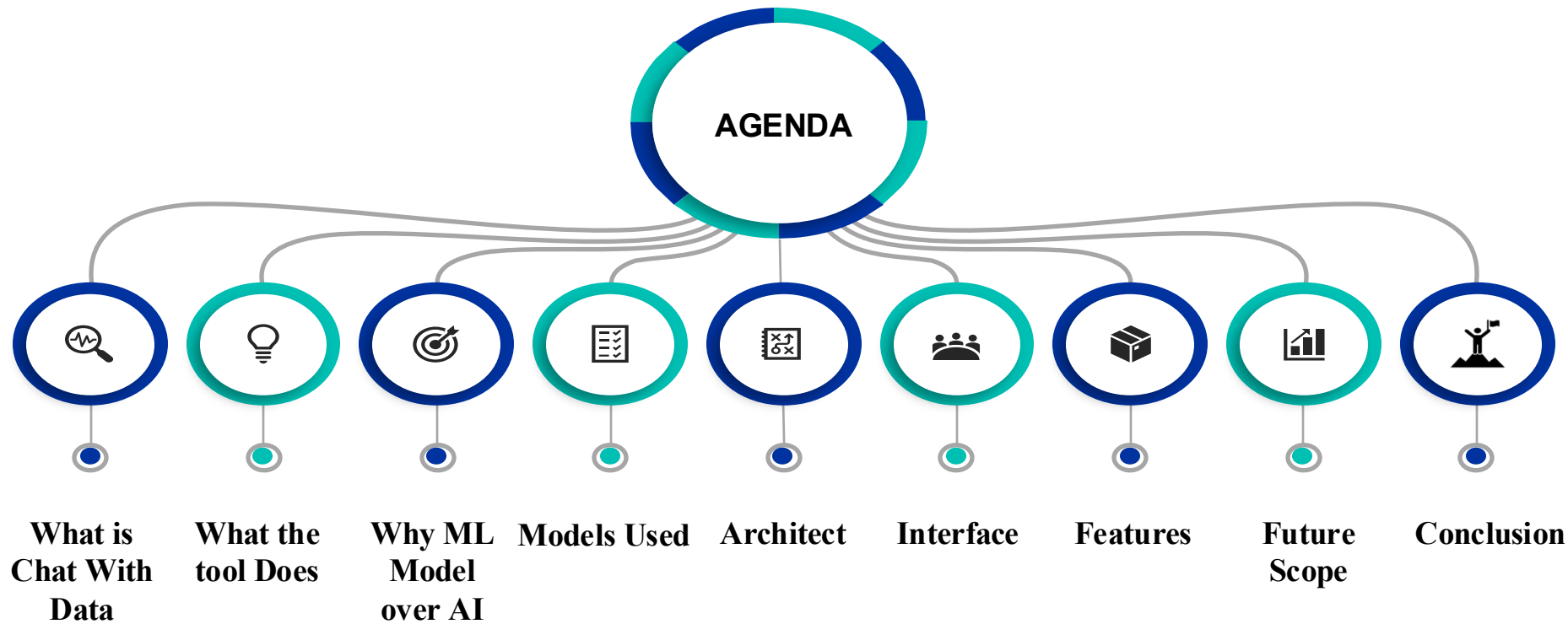


Single Day Event

Chat With Data

Mokesh U,
Tony Mathew
Pfizer



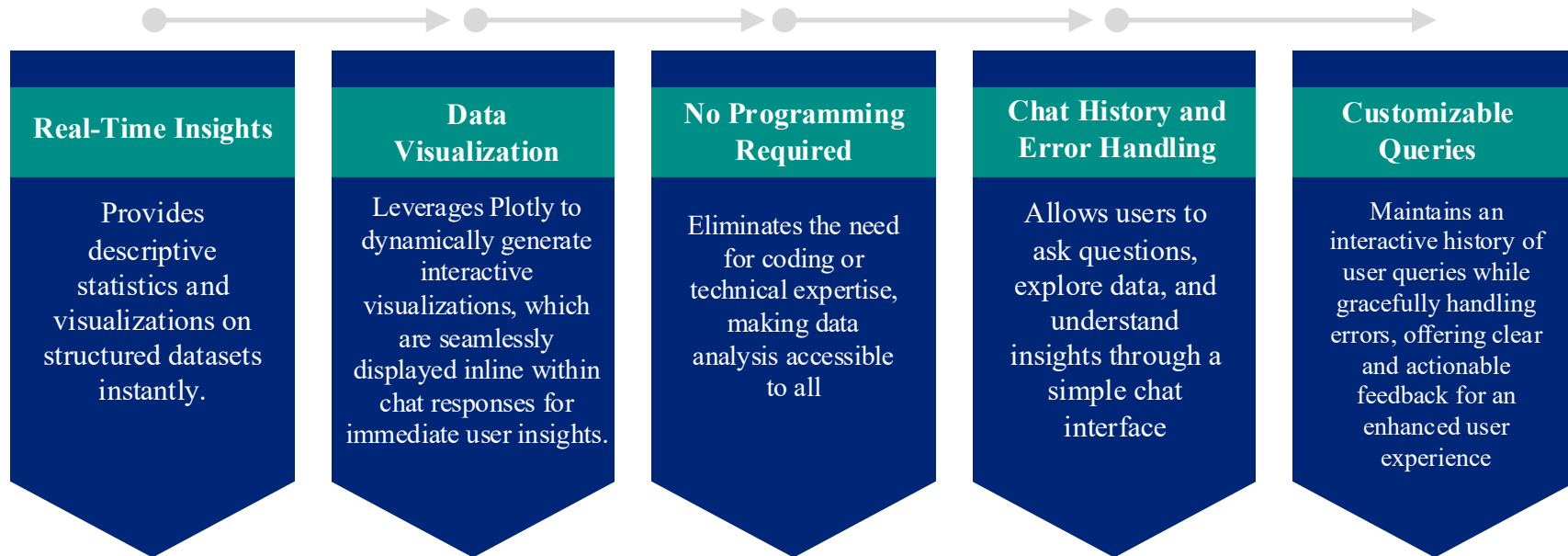


WHAT IS Chat With Data

- **Chat with Data** is a chatbot designed to make data exploration and analysis accessible to everyone, regardless of their technical background. This tool allows users to interact with structured datasets through natural language queries, removing the need for manual coding or advanced statistical knowledge.
- By simplifying complex data analysis tasks into easy, conversational interactions, **Chat with Data** empowers users to efficiently explore and gain insights from their data without requiring specialized skills



WHAT THE TOOL DOES



WHY ML MODEL OVER AI

AI Model



Data Integrity & Security

Often rely on external APIs, increasing risk of data exposure and compromising security.



Scalability

Struggle to scale effectively with large datasets or new query types due to reliance on predefined rules and external dependencies.



Efficiency

requiring significant computational power and storage, especially when utilizing external databases or complex algorithms.

VS

ML Model



Data Integrity & Security

ensuring sensitive information remains secure, preserving data integrity



Scalability

Scale efficiently as data volumes grow, adapting to new data with minimal intervention



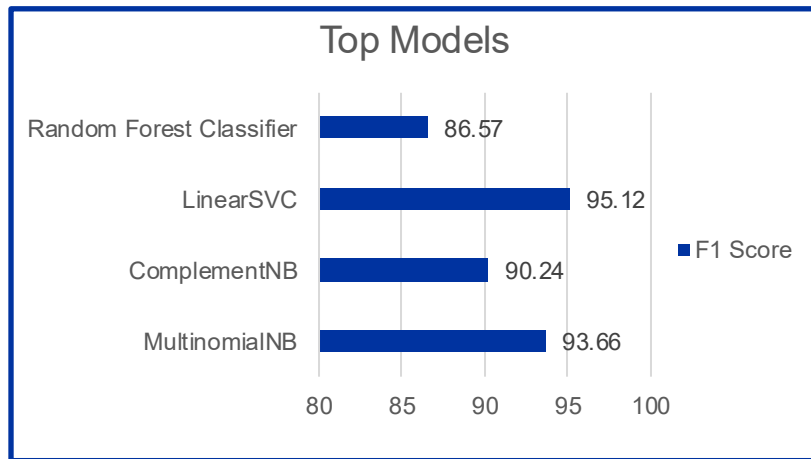
Efficiency

Efficient in terms of storage and computation.



MODELS USED

- For intent prediction, several models were evaluated, including **Multinomial Naive Bayes (MultinomialNB)**, **Complement Naive Bayes (ComplementNB)**, **Linear Support Vector Classifier (LinearSVC)**, and **Random Forest Classifier**.
- **MultinomialNB** is a simple and fast model that works well for text classification, particularly when features are conditionally independent.
- **ComplementNB** a variant of Naive Bayes, is better suited for imbalanced datasets.
- **Random Forest Classifier** Strong for complex relationships
- **Linear Support Vector Classifier (LinearSVC)** Best suited for high-dimensional text data and provided the most balanced performance.



LinearSVC performed the best, achieving a high **95% accuracy** and the highest **F1 score**, making it the most effective model



MODEL TRAINING

- To train the model for our tool, we developed a custom dataset consisting of user queries related to descriptive statistics & Graphs. The dataset was designed to enhance the chatbot's ability to understand and process natural language queries. It contains three key components:

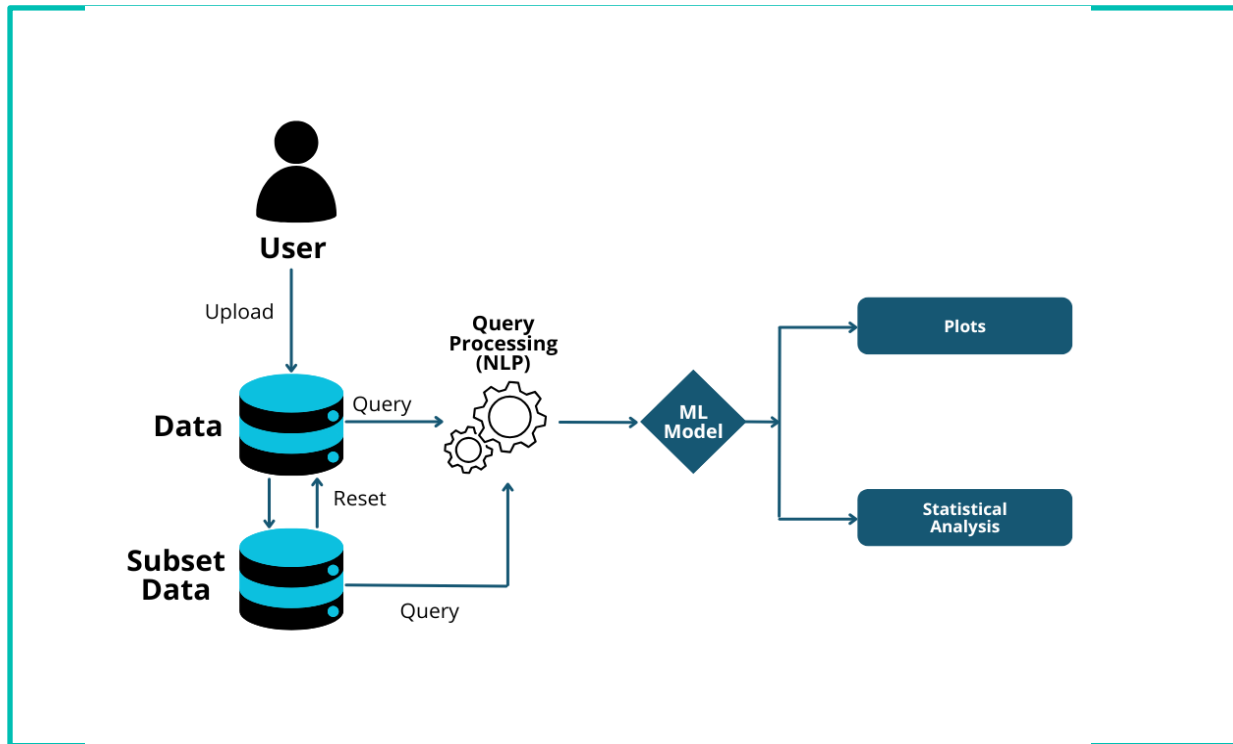
Query: The user's natural language question related to statistical analysis.

Intent: The specific statistical operation requested.

Code: The Python code corresponding to the query, designed to compute the requested statistic.

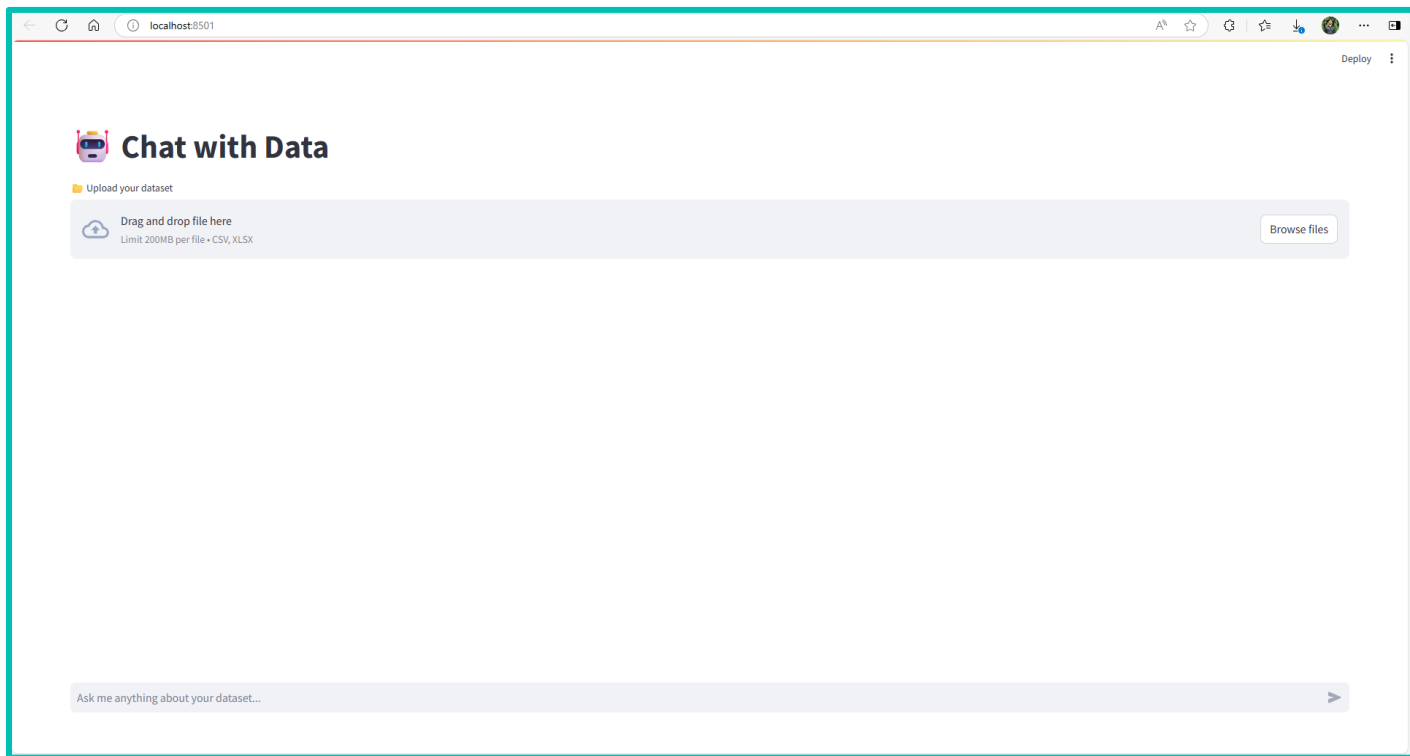
- As we expand and diversify the training dataset, the chatbot's performance improves, enabling more accurate and efficient user interactions.

query	intent	code
Tell me the mean	mean	df['column'].mean()
Find the average	mean	df['column'].mean()
What is the mean value	mean	df['column'].mean()
Calculate the mean	mean	df['column'].mean()
Give me the mean	mean	df['column'].mean()
What is the average	mean	df['column'].mean()
Give me the average	mean	df['column'].mean()
What's the middle value	median	df['column'].median()
Find the median	median	df['column'].median()
Could you determine the median	median	df['column'].median()
Calculate the middle value	median	df['column'].median()
What is the median	median	df['column'].median()
Find the median	median	df['column'].median()
Calculate the median	median	df['column'].median()
Give me the median	median	df['column'].median()
What is the middle value	median	df['column'].median()
Calculate the mode	mode	df['column'].mode()[0]
Find the mode	mode	df['column'].mode()[0]
What is the mode	mode	df['column'].mode()[0]
Tell me the mode	mode	df['column'].mode()[0]
Give me the most frequent value	mode	df['column'].mode()[0]





INTERFACE



FEATURES



LOAD DATA

← ↻ 🏠 local:8501

⌵ ⭐ 🔍 📄 🌙 ⚙️ ⋮

Deploy ⋮

 **Chat with Data**

Upload your dataset

 Drag and drop file here
Limit 200MB per file • CSV, XLSX

Browse files

 adae.xlsx 224.5KB

×

✔ Dataset uploaded!

	studyid	siteid	usubjid	trta	trtan	age	agegr1	agegr1n	race	racen	sex	saffl	trtsdt	trtedt	astdt	astdtf	astdy	aendt	aendy	adurn	aduru	aeterm
0	CDISCPIL0T01	701	01-701-1015	Placebo	0	63	<65	1	WHITE	1	F	Y	2014-01-02 00:00:00	2014-07-02 00:00:00	2014-01-03 00:00:00	None	2	None	None	None	None	APPLICATION SITE ERYTHEMA
1	CDISCPIL0T01	701	01-701-1015	Placebo	0	63	<65	1	WHITE	1	F	Y	2014-01-02 00:00:00	2014-07-02 00:00:00	2014-01-03 00:00:00	None	2	None	None	None	None	APPLICATION SITE PRURITUS
2	CDISCPIL0T01	701	01-701-1015	Placebo	0	63	<65	1	WHITE	1	F	Y	2014-01-02 00:00:00	2014-07-02 00:00:00	2014-01-09 00:00:00	None	8	2014-01-11 00:00:00	10	3	DAY	DIARRHOEA
3	CDISCPIL0T01	701	01-701-1023	Placebo	0	64	<65	1	WHITE	1	M	Y	2012-08-05 00:00:00	2012-09-01 00:00:00	2012-08-07 00:00:00	None	3	2012-08-30 00:00:00	26	24	DAY	ERYTHEMA
4	CDISCPIL0T01	701	01-701-1023	Placebo	0	64	<65	1	WHITE	1	M	Y	2012-08-05 00:00:00	2012-09-01 00:00:00	2012-08-07 00:00:00	None	3	None	None	None	None	ERYTHEMA

Ask me anything about your dataset... ▶



EXAMPLE QUERY

localhost:8501

Deploy

Chat with Data

Upload your dataset

Drag and drop file here

Limit 200MB per file • CSV, XLSX

Browse files

adae.xlsx 224.5KB

X

	studyid	siteld	usubjid	trta	trtan	age	agegr1	agegr1n	race	racen	sex	saffl	trtsdt	trtedt	astdt	astdtf	astdy	aendt	aendf	aendy	adurn	aduru	aeterm
0	CDISCILOT01	701	01-701-1015	Placebo	0	63	<65	1	WHITE	1	F	Y	2014-01-02 00:00:00	2014-07-02 00:00:00	2014-01-03 00:00:00	None	2	None	None	None	None	None	APPLICATION SITE ERYTHEMA
1	CDISCILOT01	701	01-701-1015	Placebo	0	63	<65	1	WHITE	1	F	Y	2014-01-02 00:00:00	2014-07-02 00:00:00	2014-01-03 00:00:00	None	2	None	None	None	None	None	APPLICATION SITE PRURITUS
2	CDISCILOT01	701	01-701-1015	Placebo	0	63	<65	1	WHITE	1	F	Y	2014-01-02 00:00:00	2014-07-02 00:00:00	2014-01-09 00:00:00	None	8	2014-01-11 00:00:00	10	3	DAY	DIARRHOEA	
3	CDISCILOT01	701	01-701-1023	Placebo	0	64	<65	1	WHITE	1	M	Y	2012-08-05 00:00:00	2012-09-01 00:00:00	2012-08-07 00:00:00	None	3	2012-08-30 00:00:00	26	24	DAY	ERYTHEMA	
4	CDISCILOT01	701	01-701-1023	Placebo	0	64	<65	1	WHITE	1	M	Y	2012-08-05 00:00:00	2012-09-01 00:00:00	2012-08-07 00:00:00	None	3	None	None	None	None	None	ERYTHEMA

find the unique values in race column

Query: find the unique values in race column

- Intent: Unique Values List
- Column: race
- Code Executed: `st.session_state.datain['race'].unique()`
- Output: `['WHITE' 'BLACK OR AFRICAN AMERICAN' 'AMERICAN INDIAN OR ALASKA NATIVE']`

Ask me anything about your dataset...



MORE QUERIES



maximum of ADURN variable



✓ Query: maximum of adu

- Intent: Maximum
- Column: adurn
- Code Executed: `st.session_state.datain['adurn'].max()`
- Output: 444.0



find the average age



✓ Query: find the average age

- Intent: mean
- Column: age
- Code Executed: `st.session_state.datain['age'].mean()`
- Output: 74.82451721242653



number of missing values in aendt

values in aendt

values

`state.datain['aendt'].isnull().sum()`



standard deviation of ADU



✓ Query: standard dev

- Intent: Standard Dev
- Column: adurn
- Code Executed: `st.session_state.datain['adurn'].std()`
- Output: 40.24451464364423

TERM

values in aeterm

- Column: aeterm
- Code Executed: `st.session_state.datain['aeterm'].nunique()`
- Output: 242



SUBSET DATA

subset the data



Subset Mode: Define your filters below.

Select number of variables to subset:

2



Select variable 1:

race



Select condition for race:

in



Select values for race:

AMERICAN INDIA... x

BLACK OR AFRIC... x



Select variable 2:

age



Select condition for age:

>=



Enter value for age:

79.00



Apply Filters

Subsetting complete. Chat will now use the filtered data.

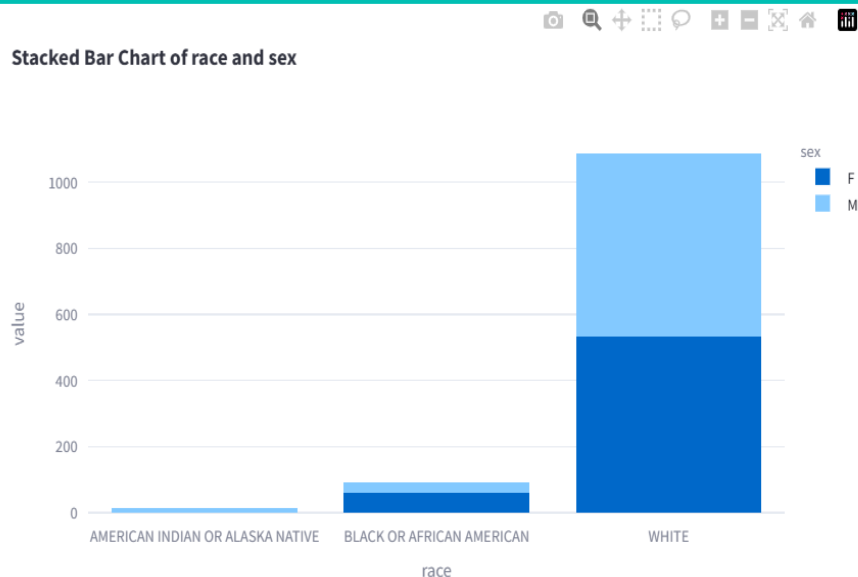
	studyid	siteid	usubjid	trta	trtan	age	agegr1	agegr1n	race	racen	sex	saffl	trtsdt	trtedt	asstdt	asstdf	astdy	astdyf
111	CDISCPLOT01	701	01-701-1203	Placebo	0	81	>80	3	BLACK OR AFRICAN AMERICAN	2	F	Y	2013-02-02 00:00:00	2013-08-03 00:00:00	2013-03-21 00:00:00	None	48	2
201	CDISCPLOT01	701	01-701-1363	Placebo	0	81	>80	3	BLACK OR AFRICAN AMERICAN	2	F	Y	2013-05-30 00:00:00	2013-11-27 00:00:00	2013-06-14 00:00:00	None	16	2
202	CDISCPLOT01	701	01-701-1363	Placebo	0	81	>80	3	BLACK OR AFRICAN AMERICAN	2	F	Y	2013-05-30 00:00:00	2013-11-27 00:00:00	None	None	None	1
203	CDISCPLOT01	701	01-701-1363	Placebo	0	81	>80	3	BLACK OR AFRICAN AMERICAN	2	F	Y	2013-05-30 00:00:00	2013-11-27 00:00:00	2013-07-16 00:00:00	None	48	1



CREATING GRAPHS

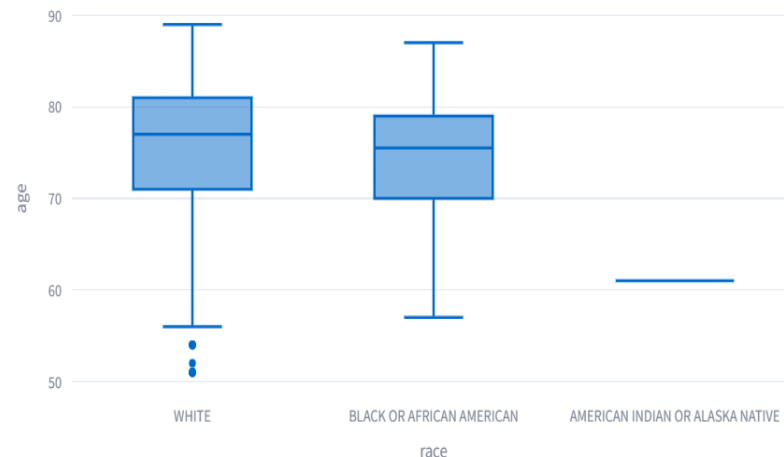
plot stacked bar chart for race and sex

Stacked Bar Chart of race and sex



create a box plot for age and race

Box Plot of age by race

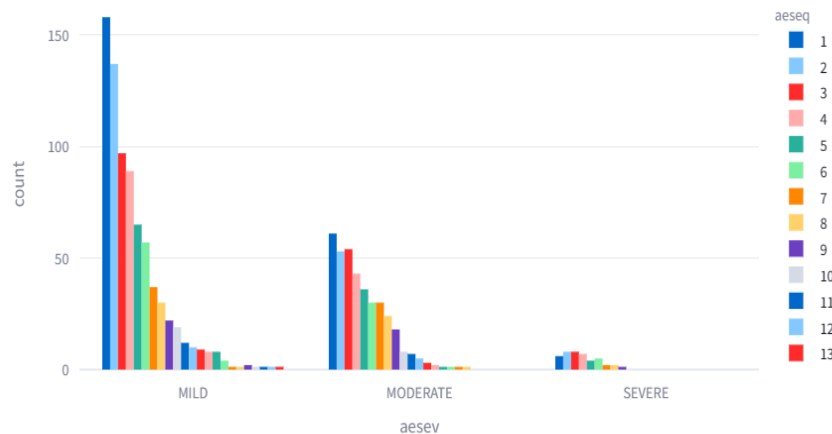




CREATING GRAPHS

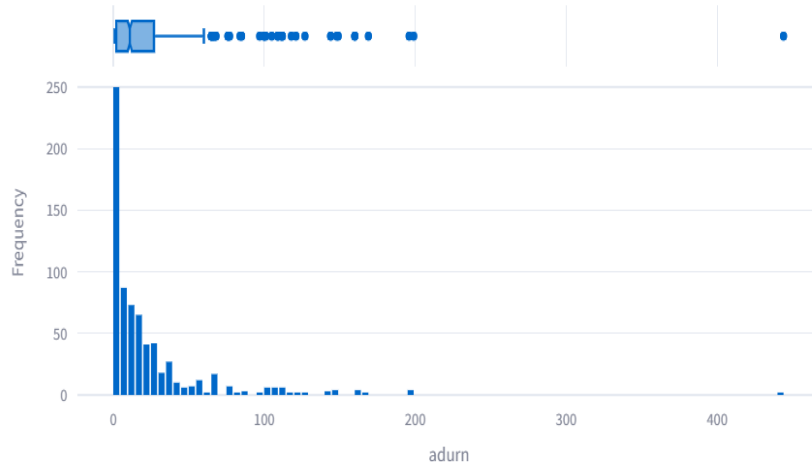
give me multiple bar chart for AESEV and AESEQ variables

Multiple Bar Chart of aesev vs aseq (Categorical)



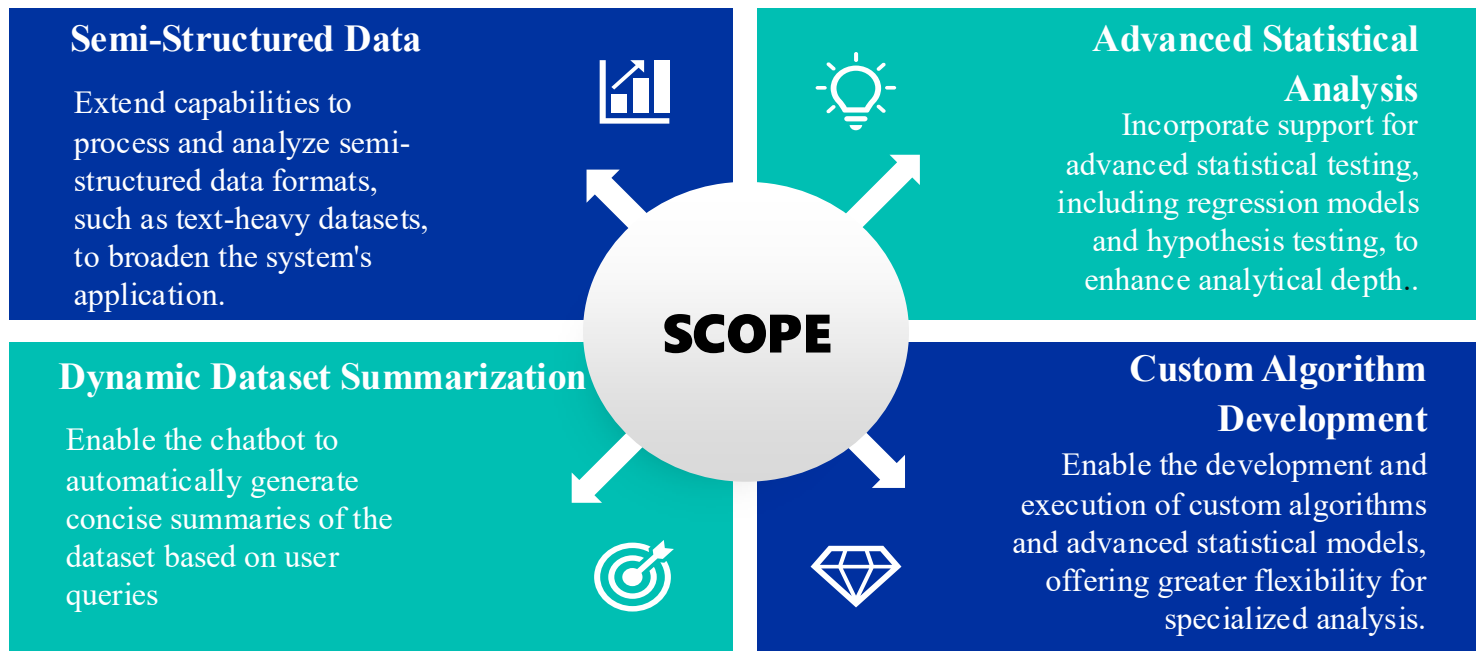
Plot a histogram for ADURN

Histogram of adurn





FUTURE SCOPE





CONCLUSION

Chat with Data empowers users to explore and understand their datasets through natural language, making data analysis more accessible and interactive. By translating user queries into meaningful visualizations, the tool streamlines insights without requiring technical expertise. As the platform evolves, it aims to support a wider range of graphs, enable intelligent data summarization, perform complex statistical analysis, and handle semi-structured data—making it an even more powerful assistant for data-driven decision-making.

THANK YOU



**The Global Healthcare
Data Science Community**

Contact Channels

📞 UK +44 1843 609600
✉ office@phuse.global
🌐 phuse.global

Social Media

🐦 [@phusetwitta](https://twitter.com/phusetwitta)
📘 [/phusebook](https://facebook.com/phusebook)
📺 [/phusetube](https://youtube.com/phusetube)
🌐 [/company/phuse](https://linkedin.com/company/phuse)