Clinical Trial Analysis
Data-Driven Insights for CNS Drug Trials

This project explores how data-driven methods can help identify high-risk trials and guide smarter resource allocation across CNS drug development pipelines.

Overview

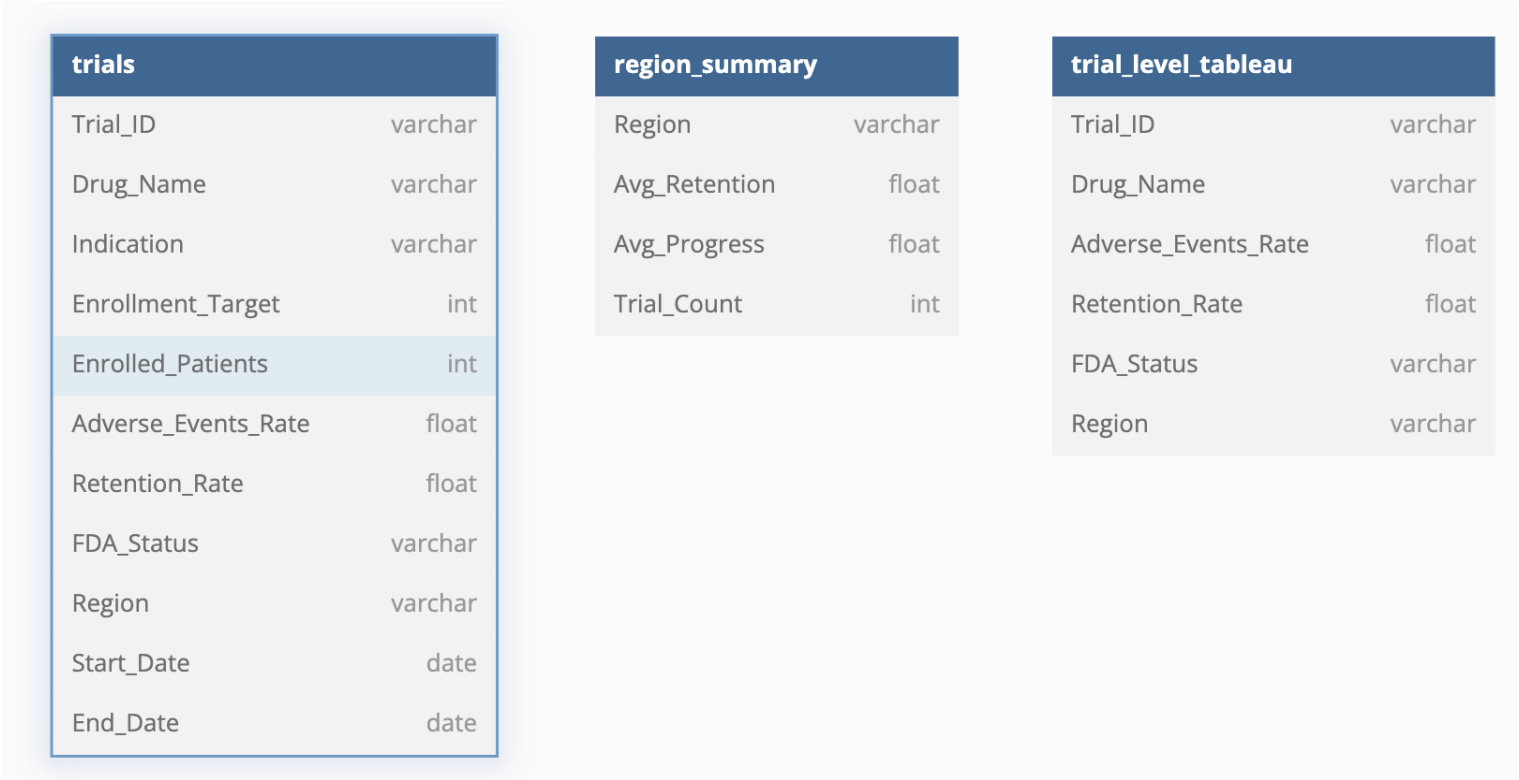
This project investigates late-stage clinical trials for central nervous system (CNS) drugs, focusing on key performance indicators such as adverse events, retention rate, and enrollment progress.

The goal is to identify high-risk trials and provide data-driven guidance for resource allocation and FDA approval readiness.

Tools Used: SQL, Python, Tableau, Google sheet

Dataset Overview

The dataset simulates Phase 3 clinical trials of CNS drugs across different regions.
It was custom-built to reflect real-world variables such as adverse events, retention rate, and FDA status.



trials	
Trial_ID	varchar
Drug_Name	varchar
Indication	varchar
Enrollment_Target	int
Enrolled_Patients	int
Adverse_Events_Rate	float
Retention_Rate	float
FDA_Status	varchar
Region	varchar
Start_Date	date
End_Date	date

region_summary	
Region	varchar
Avg_Retention	float
Avg_Progress	float
Trial_Count	int

trial_level_tableau	
Trial_ID	varchar
Drug_Name	varchar
Adverse_Events_Rate	float
Retention_Rate	float
FDA_Status	varchar
Region	varchar

	Region	Avg_Retention	Avg_Progress	Trial_Count
0	Asia	0.72	0.57	33
1	North America	0.73	0.55	39
2	Europe	0.75	0.46	28

```
query2 = """
SELECT Region,
       ROUND(AVG(Retention_Rate), 2) AS Avg_Retention,
       ROUND(AVG(CAST(Enrolled_Patients AS FLOAT) / Enrollment_Target), 2) AS Avg_Progress,
       COUNT(*) AS Trial_Count
FROM trials
GROUP BY Region
ORDER BY Avg_Progress DESC;
"""
pd.read_sql(query2, conn)
```

Regional Trials Summary via SQL

- 📍 **North America** has the **largest number of trials** (39), with a balanced profile: retention rate at 0.73 and progress at 0.55
- 📍 **Asia** shows the **highest enrollment progress** (0.57)
- 📍 **Europe** leads in **average retention rate** (0.75), but has the **lowest enrollment progress** (0.46)

Identifying High-Risk Trials Using SQL + Python

```
query3 = """
SELECT Trial_ID, Drug_Name, Indication, Adverse_Events_Rate, Retention_Rate
FROM trials
WHERE Adverse_Events_Rate > 0.25 AND Retention_Rate < 0.65
ORDER BY Adverse_Events_Rate DESC;
"""
pd.read_sql(query3, conn)
```

TRIAL_002	Drug_B1	Schizophrenia	0.17	0.95	Approved	North America	233
TRIAL_003	Drug_C2	Bipolar Disorder	0.07	0.97	Pending	North America	213
TRIAL_004	Drug_D3	Bipolar Disorder	0.28	0.64	Pending	North America	332
TRIAL_005	Drug_E4	Bipolar Disorder	0.18	0.65	Approved	Asia	354
TRIAL_006	Drug_F5	Schizophrenia	0.21	0.73	Approved	North America	349
TRIAL_007	Drug_G6	Bipolar Disorder	0.27	0.72	Pending	Asia	340
TRIAL_008	Drug_H7	Bipolar Disorder	0.19	0.98	Pending	Europe	243
TRIAL_009	Drug_I8	Bipolar Disorder	0.1	0.58	Pending	Asia	246
TRIAL_010	Drug_J9	Schizophrenia	0.04	0.51	Approved	Asia	160
TRIAL_011	Drug_K10	Bipolar Disorder	0.14	0.74	Approved	Europe	117
TRIAL_012	Drug_L11	Bipolar Disorder	0.07	0.59	Pending	Asia	185
TRIAL_013	Drug_M12	Bipolar Disorder	0.13	0.68	Denied	North America	199
TRIAL_014	Drug_N13	Bipolar Disorder	0.27	0.86	Approved	North America	320
TRIAL_015	Drug_O14	Schizophrenia	0.1	0.85	Pending	Asia	327
TRIAL_016	Drug_P15	Bipolar Disorder	0.05	0.65	Pending	North America	129

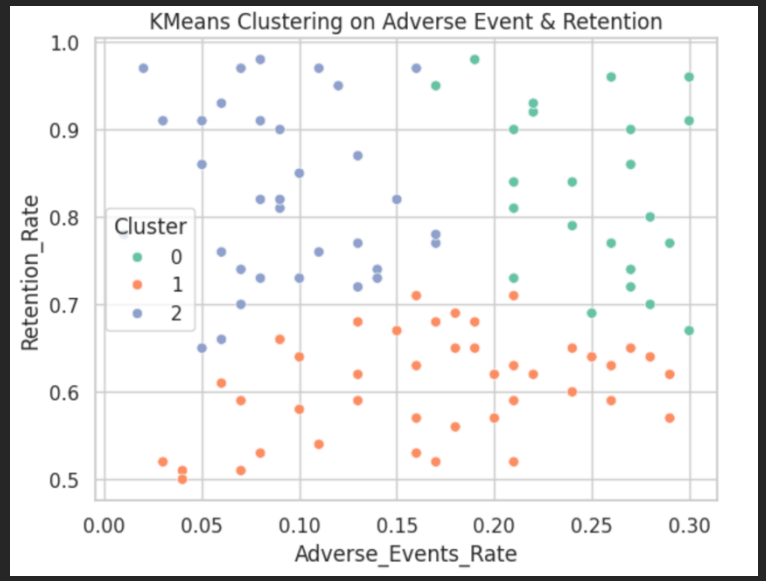
- Most trials with both high adverse event rates (>0.25) and low retention rates (<0.65) remain in the pending stage, rather than being directly denied by the FDA.
- Among the denied trials, a few exhibited relatively low adverse event rates but still failed to meet retention benchmarks. This suggests that while safety concerns may influence FDA outcomes, low retention rates may play a stronger role in denial decisions.
- Overall, FDA denials tend to align more closely with poor retention than with adverse event rates alone, indicating that participant engagement and trial execution quality could be critical factors in final approval outcomes.

```python
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans

features = df[["Adverse_Events_Rate", "Retention_Rate"]].dropna()
scaler = StandardScaler()
X = scaler.fit_transform(features)

kmeans = KMeans(n_clusters=3, random_state=42)
df["Cluster"] = kmeans.fit_predict(X)

sns.scatterplot(data=df, x="Adverse_Events_Rate",
y="Retention_Rate", hue="Cluster", palette="Set2")
plt.title("KMeans Clustering on Adverse Event & Retention")
plt.show()
```



K-means Clustering CNS Drug Trials to Identify High-Potential Candidates

🎯 Goal

Segment trials to guide prioritization and reduce risk.

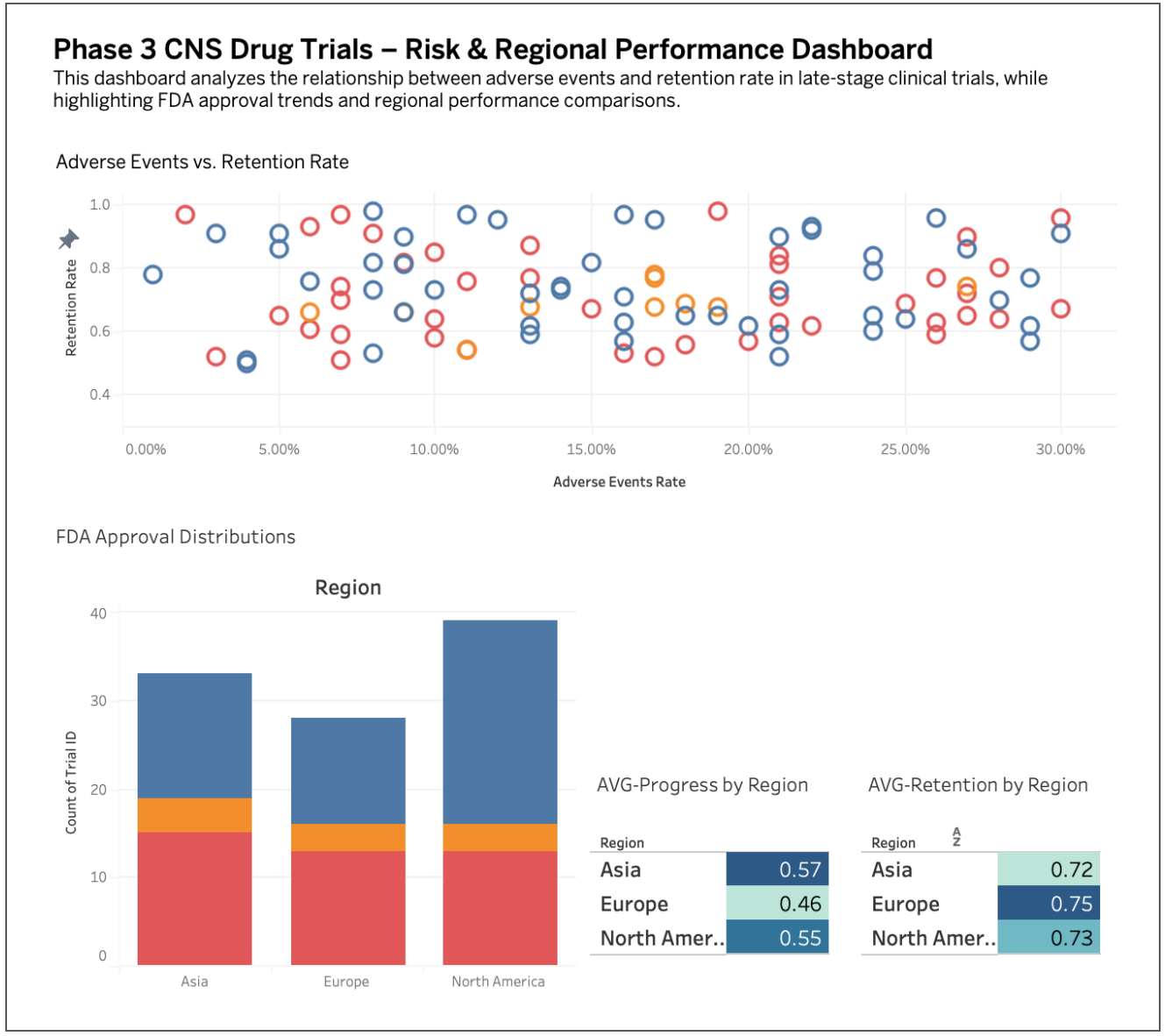
🔍 Method

K-Means (k=3) on Retention Rate + AE Rate

🧠 Key Finding

- Cluster 0: High-retention, low-AE → prioritize

- Cluster 1: Low-retention → high-risk



Phase 3 CNS Drug Trials Interactive Risk & Performance Dashboard

Dashboard Features:

- Scatterplot of Retention vs. Adverse Events, colored by FDA status

- Region-wise stacked bar chart of FDA approvals and denials

- Heatmaps showing average progress and retention by region

Conclusion & Reflection

Key Takeaways

- - Low retention is a stronger predictor of FDA denial than adverse event rate alone
- - Region-level data reveals execution strengths and weaknesses
- - Clustering adds value in prioritizing trial investment

What I Learned

- - How to structure a real-world analytics pipeline: from SQL → Python → Tableau
- - How to turn raw data into stakeholder-ready insights
- - The importance of testing assumptions before drawing conclusions

Thank you for checking out my project!

-Minfei He