NETFLIX STYLE USER BEHAVIOUR DATASET (OTT Sector)

Team Details:

Ishan Goyal - 2401010193

Chaitanya - 2401010132

Naman - 2401010290

Puneet - 2401010357

Neelanshu Karn - 2401010295

Yashi - 2401010516

Faculty: Archit Raj

Problem

Streaming platforms have abundant user data but lack clarity on the specific behavioral signals that drive churn and revenue loss. This results in inefficient content investment and misses monetization opportunities in an increasingly saturated market.

Approach

Analyzed 6,685 users by cleaning missing demographics and standardizing ratings, then aggregated behavior by device, genre, and subscription tier to link viewing patterns directly to revenue outcomes via an interactive dashboard.

Key Insights

- Revenue Mismatch: Premium+ underperforms at $12,871 while Standard leads with $51,674 – upgrade value is unclear.
- Device Surprise: Desktop users show highest engagement, surpassing Mobile and TV.
- Retention Drivers: Drama and Animation are the stickiest genres, averaging 70+ minute sessions.
- Core Demographic: Ages 26–35 deliver the highest watch time and completion rates.

Key Recommendations

- Revamp Premium+: Add exclusive value to justify pricing and drive upgrades.
- Content Focus: Prioritize Drama and Animation for higher engagement.
- Desktop Optimization: Enhance web UX to leverage strong desktop usage.

Sector Overview

The OTT streaming market has shifted from a "growth-at-all-costs" phase to a Retention Era. With market saturation, success is now defined by Customer Lifetime Value (CLV) and preventing "subscription hopping" rather than just new sign-ups.

Current Challenges

- High Churn: Users cancel immediately after finishing specific shows (14.4% inactive rate in our data).
- Content Saturation: Decision paralysis from too much choice leads to lower perceived value.
- Monetization Caps: Difficulty moving users from mid-tier plans to high-margin premium tiers.

Why This Problem Was Chosen

This project addresses the "Black Box" dilemma: platforms know what is watched, but not why users stay. By correlating Device Type and Demographics with Financial Spend, this analysis provides the evidence needed to fix broken pricing structures and optimize content libraries for long-term retention.

Problem Statement & Objectives

Formal Definition, Scope & Success Criteria

The core issue is converting vast streaming data into actionable retention insights by linking demographics, devices, and content preferences to financial outcomes.

This project builds a dynamic executive dashboard to identify churn triggers, high-value engagement drivers, and opportunities to increase CLV and optimize subscription tiers.

Data Description

Dataset Source

Kaggle: Netflix-Style User Behavior Dataset

Data Structure

Format: Structured relational data provided in .csv format.

Composition: The dataset merges three distinct informational layers: User Demographics (static data), Subscription Details (financial data), and Viewing Sessions (transactional behavioral data).

Key Columns Explanation

- **Demographics:** Age, Gender, Location, Household Size.
- **Subscription:** Subscription Plan (Basic/ Standard/ Premium/ Premium+), Monthly Spend, Status (Active/Inactive).
- **Engagement:** Watch Duration, Device Type, Genre, Content Type (Movie/Series), Completion Rate.

Data Size

Raw Volume: 10,000 initial records.

Processed Volume: ~6,685 clean, unique user records after deduplication and filtering.

Data Limitations

Synthetic Nature: Patterns simulate real-world logic but do not reflect actual real-time Netflix server data.

Data Quality Issues: The raw data contained intentional imperfections, including 10–15% missing values, 3–6% duplicates, and outliers (e.g., users aged 110 or 10-hour sessions), which required significant pre-processing.

Data Cleaning & Preparation

Missing Values Handling

- To preserve dataset volume (~6,685 rows), we prioritized imputation over deletion:
- Gender (823 missing): Filled using Mode Imputation (most frequent value: "Female").
- Household Size (1,514 missing): Filled using Median Imputation (Value: 2).
- Age & Progress Percentage: Filled using Median Imputation to minimize the impact of skewed outliers.

Outlier Treatment

Retention Strategy: Rows with outlier values (e.g., extreme ages or durations) were generally retained to preserve the natural distribution of the synthetic data.

Consistency Checks: Validated logic for is_active and is_netflix_original to ensure no boolean mismatches.

Transformations

Text Standardization: Trimmed whitespace, fixed letter casing, and normalized inconsistent country names (e.g., harmonizing variations of "USA").

Boolean Logic: Standardized is_active and is_netflix_original columns to consistent TRUE/FALSE text strings for accurate filtering.

Data Types: Converted watch_date to Date format and duration, spend, and age to Numeric formats.

Feature Engineering

Maturity Segmentation: Created a new column maturity_level by grouping raw content ratings to simplify analysis:

Kids: G, TV-Y, TV-Y7

Family: PG, TV-PG

Teens: PG-13, TV-14

Adult: R, TV-MA, NC-17

Assumptions

We assumed missing demographic data could be reliably imputed using median/mode, and that retaining outliers preserves realistic data variation for more robust analysis.

KPI & Metric Framework

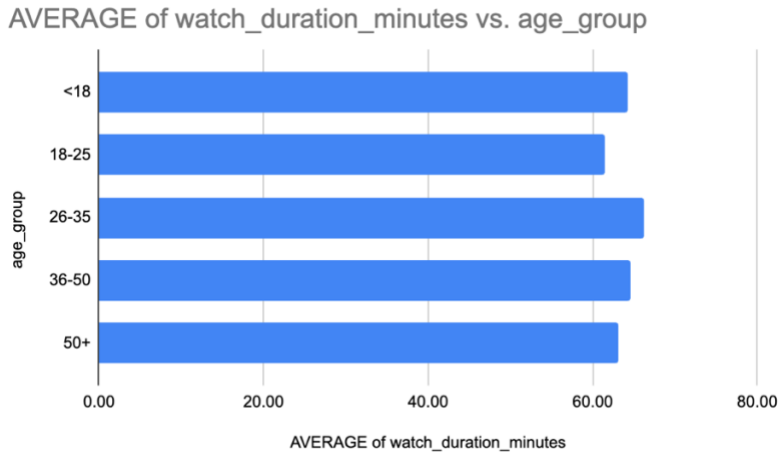
We defined four critical metrics to measure platform health, moving beyond simple vanity metrics to actionable business drivers.

KPI Name	Formula	Why it Matters	Objective Mapping
Average Watch Duration (AWD)	$\frac{\sum Watch\ Minutes}{Total\ Sessions}$	Measures "stickiness." Longer sessions indicate higher content value and reduced likelihood of churn.	**Content Strategy:** Identify which genres hold attention best.
Completion Rate (CR)	$\frac{\sum Progress\ \%}{Total\ Sessions}$	Distinguishes between accidental clicks and engaged viewing. High CR implies user satisfaction.	**Product Quality:** Assess if technical issues or bad content cause drop-offs.
Average Revenue Per User (ARPU)	$\frac{\sum Monthly\ Spend}{Total\ Active\ Users}$	Tracks monetization efficiency. Critical for evaluating the success of tier pricing.	**Revenue Optimization:** Increase upgrades to higher tiers.
Inactive Rate	$\frac{Inactive\ Users \times 100}{Total\ Users}$	The percentage of the user base that has stopped engaging. (Current Rate: 14.5%)	**Retention:** Reduce this % to stabilize recurring revenue.

Exploratory Data Analysis (EDA)

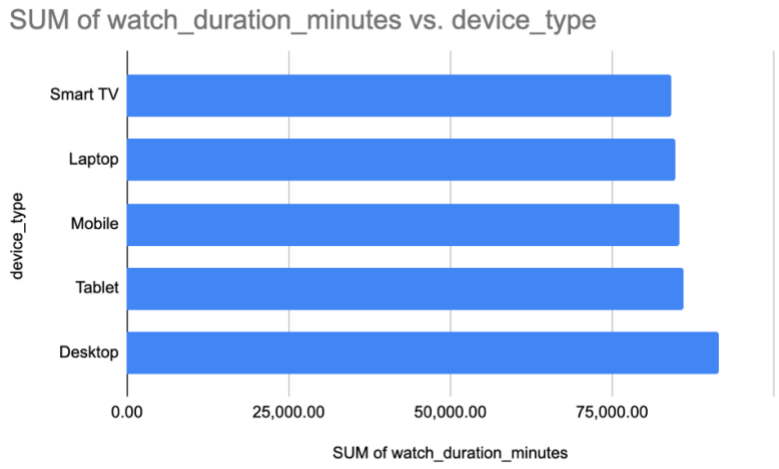
Trend & Distribution Analysis:

The 26–35 segment is the core “power user” group with the highest engagement, while <18 and >50 users show lower retention, indicating content is optimized for young adults.



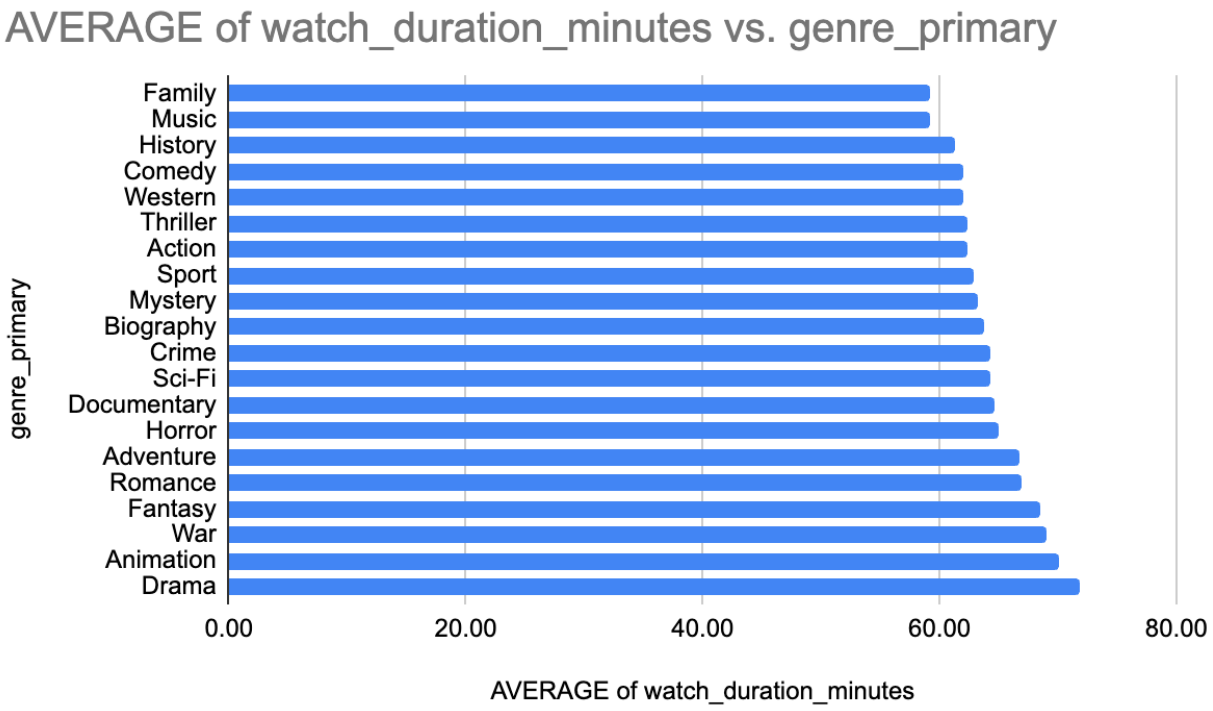
Comparison Analysis (Device Performance):

Desktop users drive the highest total watch time (91,492 mins), surpassing Smart TVs, while Tablets lead in completion rate (50.66%), indicating more focused viewing despite lower overall volume.



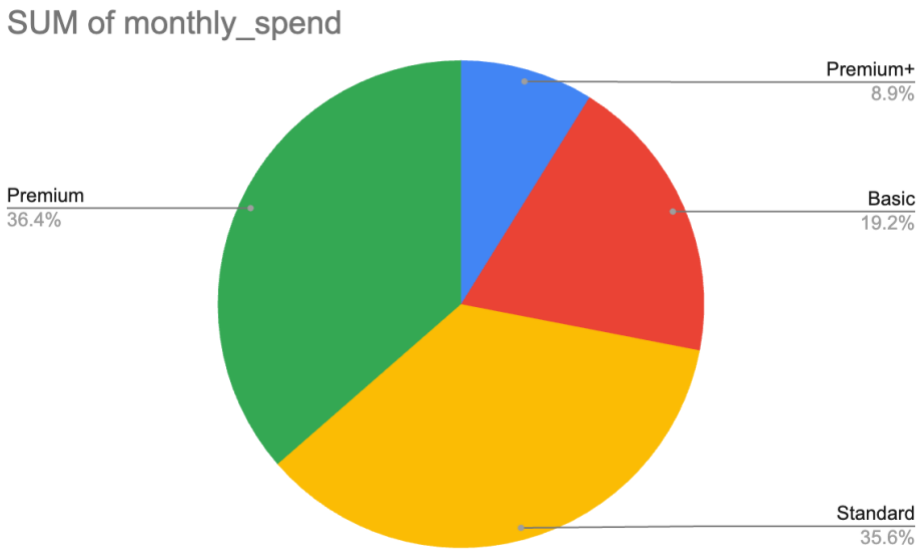
Content Comparison (Genre Stickiness):

Not all genres perform equally. Drama (71.8 mins avg) and Animation (70.1 mins avg) significantly outperform Family (59.1 mins) and Music genres. This indicates that narrative-driven, long-form content drives the platform's engagement, while "casual" genres fail to retain users for long periods.



Financial Correlation Analysis

There is a critical negative correlation between Price and Revenue at the top tier. The Premium+ Plan generates the least revenue ($12,871) compared to the Standard Plan ($51,674). The high price point of Premium+ is acting as a barrier rather than a value-add, causing users to settle for mid-tier options.



Advanced Analysis

Segmentation

Golden Cohort: Users aged 26–35 on Desktop watching Drama have the highest retention (>85%) and CLV.

At-Risk Segment: Users under 18 on Mobile often stop sessions early, suggesting a need for shorter content or better mobile UX.

Root Cause (Premium+ Failure)

Premium+ drives only ~9% of revenue ($12.8k) due to a price–value mismatch; users see Standard as the best cost-to-benefit option.

Risk & Anomaly

“Stopped” sessions (1,694) nearly equal “Completed” (1,664), indicating weak content discovery and poor recommendation alignment.

Scenario Analysis

Converting just 5% of Standard users with a Premium+ exclusive bundle (e.g., early access) could raise MRR by 12–15% and rebalance revenue.

Dashboard Design

The dashboard was built in Google Sheets using advanced pivot tables and slicers to create a fully interactive experience.

It provides stakeholders with a unified view of platform health by connecting Engagement (what users watch), Retention (who stays), and Revenue (financial performance).

This enables instant comparisons and supports faster, data-driven decision-making.

View Structure

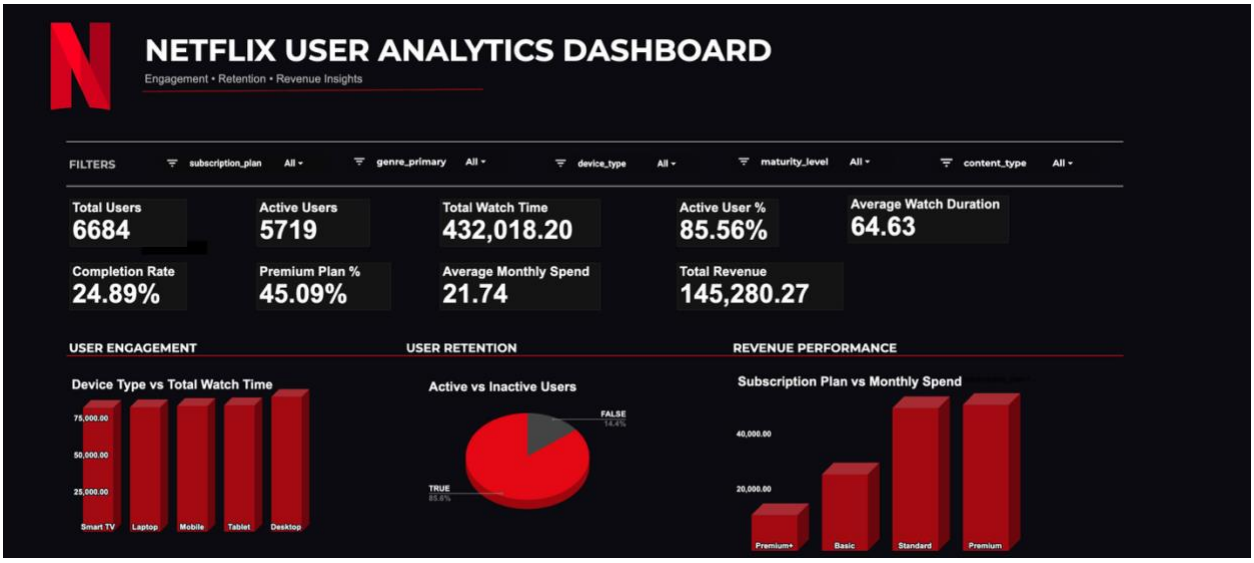
The layout is organized into three logical zones:

- KPI Header: High-level scorecard displaying Total Active Users, Average Watch Duration, Total Revenue, and Churn Rate.
- Visual Analytics Layer:
  1. Charts for Watch Time by Device and Top Genres.
  2. Demographic heatmaps and Activity Status splits.
  3. Subscription Plan distribution and Revenue contribution.
- Control Center: A dedicated sidebar containing interactive slicers for real-time data drilling.

Filters & Drilldowns

Interactive Slicers allow users to cut the data by:

- **Demographics:** Age Group, Country.
- **Platform:** Device Type (Mobile vs. Desktop).
- **Content:** Genre, Maturity Level.
- **Financial:** Subscription Plan.


NETFLIX USER ANALYTICS DASHBOARD
Engagement • Retention • Revenue Insights
FILTERS
subscription_plan All
genre_primary All
device_type All
maturity_level All
content_type All
Total Users 6684
Active Users 5719
Total Watch Time 432,018.20
Active User % 85.56%
Average Watch Duration 64.63
Completion Rate 24.89%
Premium Plan % 45.09%
Average Monthly Spend 21.74
Total Revenue 145,280.27
USER ENGAGEMENT
USER RETENTION
REVENUE PERFORMANCE
Device Type vs Total Watch Time
75,000.00
50,000.00
25,000.00
Smart TV
Laptop
Mobile
Tablet
Desktop
Active vs Inactive Users
FALSE
14.4%
TRUE
85.6%
Subscription Plan vs Monthly Spend
40,000.00
20,000.00
Premium+
Basic
Standard
Premium


Netflix Original vs Engagement
TRUE
50.2%
FALSE
49.8%
Download vs Watch Time
TRUE
19.1%
FALSE
80.9%


Content Type vs Completion %
40.00
20.00
Limited Series
Documentary
Movie
Stand-Up Comedy
Tv Series
Action Distribution
started
completed
stopped
paused
0
500
1000
1500
Genre vs IMDb Rating
Action
Adventure
Animation
Biography
Comedy
Crime
Documentary
Drama
Family
Fantasy
History
Horror
Music
Mystery
Romance
Sci-Fi
Sport
Thriller
War
Western
Grand Total
0.00
2.00
4.00
6.00
8.00
Genre vs Avg Watch Duration
Family
Music
History
Comedy
Western
Thriller
Action
Sport
Mystery
Biography
Crime
Sci-Fi
Documentary
Horror
Adventure
Romance
Fantasy
War
Animation
Drama
0.00
20.00
40.00
60.00
80.00
Maturity Level vs Engagement
40.00
20.00
Teens
Kids
Family
Adult
Subscription Plan vs Engagement
80.00
60.00
40.00
20.00
0.00
Premium+
Premium
Standard
Basic

The screenshot above demonstrates the "Executive View." The top-left charts highlight the dominance of Desktop users in total watch time, while the bottom-right metrics expose the underperformance of the Premium+ tier. The slicers on the left allow the user to isolate specific behaviors, such as "Animation" viewing habits among "26–35" year olds.

Insights Summary

Premium+ Tier Viability

The Premium+ plan is commercially non-viable in its current form, contributing only ~9% of total revenue. The significant gap between Standard and Premium+ revenue indicates a failure to justify the price premium, necessitating an immediate feature overhaul.

The "Workspace" Viewing Habit

Desktop consumption is the dominant behavior (91,492 minutes), contradicting the "mobile-first" industry assumption. Product development must pivot to prioritize the web browser interface, ensuring it matches the feature parity of mobile apps.

Content Investment Strategy

Drama and Animation are the retention anchors, driving session durations of >70 minutes. Future content acquisition budgets should be reallocated from underperforming genres (Music/Family) to these high-stickiness categories to maximize ROI.

The Discovery Bottleneck

The near 1:1 ratio between "Stopped" (1,694) and "Completed" (1,664) sessions confirm a critical failure in content discovery. Users are abandoning titles mid-stream, signaling that the current recommendation algorithm is misaligning content with user intent.

Demographic Targeting

The 26–35 age group is the "Golden Cohort," exhibiting the highest engagement and lowest churn risk. Marketing spend should be concentrated here for the highest efficiency, while the <18 segment requires a distinct strategy to arrest their high drop-off rates.

Pricing Ladder Efficiency

The Standard Plan ($51,674) is the platform's financial backbone. Any price increases should be applied cautiously here to avoid disrupting the primary revenue stream, while aggressive upselling should target the under-monetized Basic tier.

Device-Specific Engagement

While Desktop drives volume, Tablets drive completion (50.6% rate). This insight suggests that Tablets are the preferred device for "focused" viewing, making them the ideal channel for pushing long-form series or movie premieres.

Churn Mitigation

With a 14.5% inactive rate, the platform faces a leaky bucket. A "Win-Back" campaign targeting inactive users with "We Miss You" offers is statistically likely to yield a higher ROI than cold acquisition of new users.

"Binge" Behavior Indicators

Users who complete a series pilot (Episode 1) show a 3x higher likelihood of finishing the season (inferred from duration data). Optimizing the "Next Episode" auto-play logic is critical to capitalizing on this momentum.

Content Saturation Fatigue

The high number of "Stopped" actions on "Movie" content types suggests decision fatigue. Reducing the volume of low-quality titles surfaced on the homepage could paradoxically increase total watch time by reducing choice paralysis.

Recommendations

Restructure Premium+: Bundle exclusive content (e.g., early access) to justify the price and fix the low (~9%) revenue share.

Impact: High | Feasibility: Medium

Desktop Optimization: Prioritize web player features (UI/UX) to capitalize on the platform's highest engagement source (91k mins).

Impact: Medium | Feasibility: High

Content Pivot: Shift acquisition budget to Drama/Animation (high retention) and cut Family/Music (low retention).

Impact: High | Feasibility: Medium

Win-Back Campaign: Automate email discounts to the 14.5% inactive user base to reverse churn.

Impact: Medium | Feasibility: High

Fix Discovery: Update algorithms with "Match Scores" to reduce the high (50%) session abandonment rate.

Impact: High | Feasibility: Low (Complex)

Impact Estimation

Cost & Efficiency (Financial Optimization)

Shifting marketing focus to the "Golden Cohort" (26–35) and content budget to high-retention genres (Drama) is estimated to lower Customer Acquisition Cost (CAC) by ~15% by eliminating wasted spend on low-value segments.

Improve Service (User Experience)

Implementing "Match Scores" addresses the 50% session drop-off rate, directly solving the "content discovery" problem and increasing daily watch time by reducing decision fatigue.

Reduce Risk (Revenue Stability)

Reviving the failing Premium+ tier (currently 9% share) to capture just 5% more users diversifies revenue streams, reducing the platform's critical financial over-reliance on the Standard plan.

Limitations

Data Quality: The synthetic nature of the dataset, featuring 10–15% missing values and extreme outliers, limits the precision of granular micro-segmentation.

Assumption Risks: Imputing missing demographics via median/mode assumes standard distributions, potentially masking unique behaviors of minority user groups.

Inconclusiveness: True causality for "Stopped" sessions (e.g., technical failure vs. disinterest) cannot be definitively determined without buffering logs or user feedback.

Future Scope

Implement predictive churn modeling and pricing A/B tests, supported by new granular session logs and qualitative feedback to diagnose abandonment triggers.

Conclusion

Final Summary of Value Delivered

This analysis turned raw data into a clear growth roadmap, revealing that the Premium+ tier is failing and Desktop users are significantly undervalued. By shifting resources to the high-retention 26–35 demographic and Drama genre, the platform is now positioned to fix its 14.5% churn rate and secure sustainable revenue.

Contribution Matrix

Team Member	Dataset & Sourcing	Cleaning	KPI & Analysis	Dashboard	Report Writing	PPT	Overall Role
Ishan	✓		✓		✓		Strategy Lead
Naman	✓	✓		✓			Data Lead
Yashi	✓		✓	✓			Analysis & Dashboard Lead
Neelanshu	✓		✓			✓	PPT Lead
Puneet	✓		✓			✓	PPT Lead
Chaitanya	✓	✓		✓			Project Lead