BÀI TEST ĐÁNH GIÁ NĂNG LỰC
Vị trí: Intern Data Analyst

Họ và Tên ứng viên: Cao Đỗ Gia Khanh

SĐT: 0963187837

Email: caokhanh.gt2004@gmail.com

Bài làm

Table of Contents

*Kiểm tra và làm sạch dữ liệu đầu vào

Trước khi tiến hành phân tích và tính toán các chỉ số, việc kiểm tra dữ liệu là rất quan trọng để đảm bảo tính đầy đủ và nhất quán của dữ liệu. Một số bước kiểm tra cơ bản đã được thực hiện bằng Python và excel như sau:

-Kiểm tra các giá trị NULL ở tất cả các cột

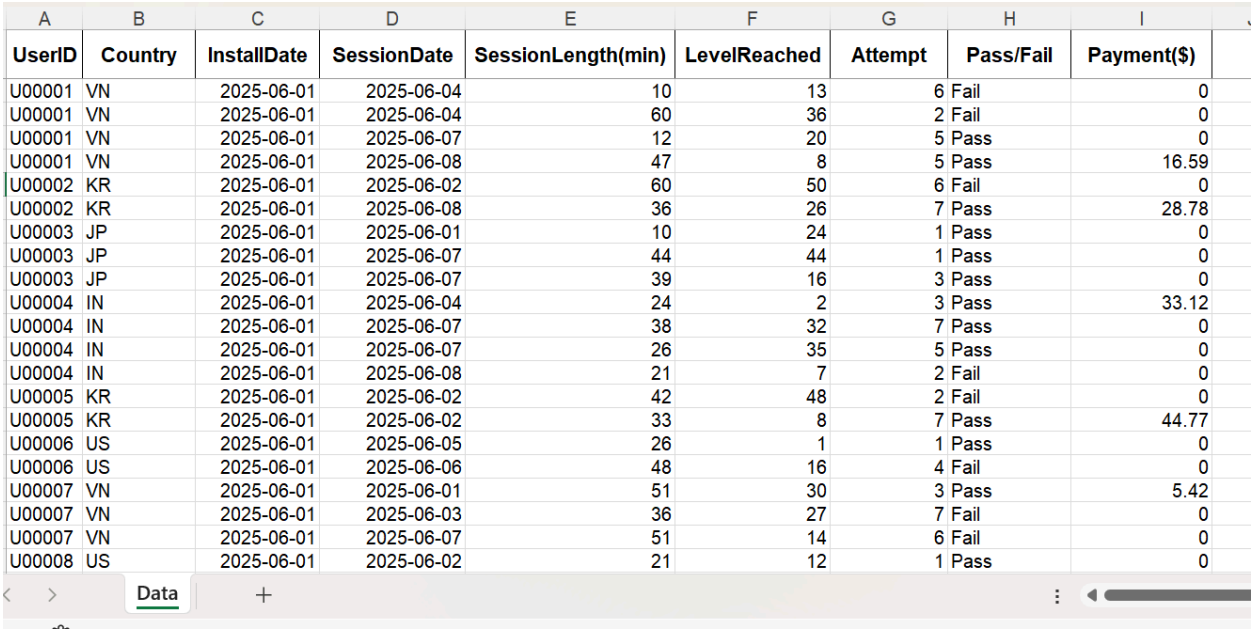
-Kiểm tra các giá trị trùng

-Kiểm tra định dạng dữ liệu ngày tháng, đảm bảo được chuẩn hóa cũng 1 dạng

-Đối chiếu logic thời gian giữa Install Date và Session Date

-Xác minh sự hợp lệ của các kiểu dữ liệu còn lại

-Sắp xếp và rà soát thủ công: Sau khi sắp xếp theo UserID, kết quả cho thấy mỗi người dùng có thể có nhiều lần đăng nhập khác nhau, kéo dài trong nhiều ngày.



A	B	C	D	E	F	G	H	I
UserID	**Country**	**InstallDate**	**SessionDate**	**SessionLength(min)**	**LevelReached**	**Attempt**	**Pass/Fail**	**Payment($)**
U00001	VN	2025-06-01	2025-06-04	10	13	6	Fail	0
U00001	VN	2025-06-01	2025-06-04	60	36	2	Fail	0
U00001	VN	2025-06-01	2025-06-07	12	20	5	Pass	0
U00001	VN	2025-06-01	2025-06-08	47	8	5	Pass	16.59
U00002	KR	2025-06-01	2025-06-02	60	50	6	Fail	0
U00002	KR	2025-06-01	2025-06-08	36	26	7	Pass	28.78
U00003	JP	2025-06-01	2025-06-01	10	24	1	Pass	0
U00003	JP	2025-06-01	2025-06-07	44	44	1	Pass	0
U00003	JP	2025-06-01	2025-06-07	39	16	3	Pass	0
U00004	IN	2025-06-01	2025-06-04	24	2	3	Pass	33.12
U00004	IN	2025-06-01	2025-06-07	38	32	7	Pass	0
U00004	IN	2025-06-01	2025-06-07	26	35	5	Pass	0
U00004	IN	2025-06-01	2025-06-08	21	7	2	Fail	0
U00005	KR	2025-06-01	2025-06-02	42	48	2	Fail	0
U00005	KR	2025-06-01	2025-06-02	33	8	7	Pass	44.77
U00006	US	2025-06-01	2025-06-05	26	1	1	Pass	0
U00006	US	2025-06-01	2025-06-06	48	16	4	Fail	0
U00007	VN	2025-06-01	2025-06-01	51	30	3	Pass	5.42
U00007	VN	2025-06-01	2025-06-03	36	27	7	Fail	0
U00007	VN	2025-06-01	2025-06-07	51	14	6	Fail	0
U00008	US	2025-06-01	2025-06-02	21	12	1	Pass	0

Data

=> Kết quả cho thấy dữ liệu đã hợp lệ, đạt chuẩn, không chứa các giá trị NULL, hay các giá trị trùng lặp và các lỗi định dạng. Do đó, dữ liệu đã rất sạch, đạt chuẩn để chuyển sang các bước tiếp theo.

-Chi tiết phân tích được trình bày ở đường liên kết sau:Link kiểm tra dữ liệu

A. Tính các chỉ số cơ bản

1) Retention D1, D3, D7

-Công thức

Retention Dn=(số lượng user quay lại app vào Dn/Tổng số user cài app ngày D0)*100%, với

+Dn là số ngày sau ngày cài đặt(D1, D3, D7)

+Số lượng user quay lại app vào Dn: Được tính 1 lần/người dù đăng nhập nhiều lần 1 ngày.

+D0: ngày người dùng cài đặt app

+Tổng số user cài app ngày D0:Tính unique user

-Quy trình thực hiện trên excel:

+Tạo cột mới DaySinceInstall=SessionDate-InstallDate. Đây là số ngày kể từ khi người dùng cài đặt app.

+Tạo cột mới IsDn với công thức =[IF(DaySinceInstall=n, 1, 0)] với n là 1,3, 7; nó sẽ trả về giá trị 1 với DaySinceInstall=n, còn lại thì bằng 0

+Bởi vì người dùng có thể quay lại nhiều lần trong 1 ngày, nên dùng PivotTable để tính số người dùng duy nhất .Kéo UserID vào Rows để hiện thị ra số người dùng, kéo lần lượt IsD1, IsD3, IsD7 vào Values và chọn hàm Max, chọn hàm Max vì đảm bảo mỗi user sẽ được tính duy nhất 1 lần cho mỗi mốc thời gian, mặc dù có nhiều phiên truy cập trong cùng 1 ngày. Dùng hàm Sum để đếm số người

-Qua tính toán ta thấy

Retention D1=35.98%

Retention D3=30.88%

Retention D7=34.44%

User ID	Max of IsD7	Max of IsD3	Max of IsD1			Retention	Day
U00001	1	1	0	users of D7	425	34.44%	D7
U00002	1	0	1	users of D3	381	30.88%	D3
U00003	0	0	0	users of D1	444	35.98%	D1
U00004	1	1	0				
U00005	0	0	1				
U00006	0	0	0				
U00007	0	0	0				
U00008	1	0	1				
U00009	1	0	0				
U00010	0	1	1				

+Chi tiết được thể hiện ở: Sheet 1) Retention D1, D3, D7 trong file excel được đính kèm

2) ARPU D7

-Công thức

ARPU D7 = Tổng doanh thu D7 / Số lượng người dùng duy nhất(D0)

+Tổng doanh thu D7:Tổng doanh thu tính đến ngày D7, là tổng tiền mà người dùng tạo ra trong 7 ngày đầu kể từ ngày cài app (từ D0 đến D7).

+Số lượng người dùng duy nhất(D0): Số lượng người dùng cài đặt app vào D0, là số người cài đặt app tại ngày đầu tiên D0, thường tính bằng số UserID duy nhất trong ngày đó.

-Quy trình tính toán:

+Vì chỉ tính tổng doanh thu đến ngày 7, nên chỉ lọc dữ liệu với điều kiện DaySinceInstall <=7, rồi cộng tất cả giá trị doanh thu .

+Giá trị Max của cột DaySinceInstall, =7 cho thấy không có người dùng nào vượt quá ngày D7, nên toàn bộ dữ liệu đều nằm trong khoảng cần phân tích.

+Tổng doanh thu D7= Sum(payment).

+Tính số lượng người dùng D0 bằng cách đếm số UserID duy nhất có InstallDate = D0.

-Kết quả ARPU=12.34

Kiểm tra xem D0(ngày cài đặt) có phải ngày 1/6/2025 không?	TRUE
Kiểm tra xem DaySinceInstall > 7 không	7
Total Revenue D7:	15228.98
Number of Unique Users	1234
ARPU D7=(Total Revenue D7/UserCountD0)	12.34

+Chi tiết được thể hiện ở: sheet 2) ARPU D7 trong file excel được đính kèm

3) Pass Rate Level 1, 5, 10

-Mục tiêu:Tính tỷ lệ vượt qua của người chơi tại các mốc level 1, 5, 10; nhằm đánh giá độ khó, khả năng giữ chân người chơi tại các giai đoạn sớm trong game.
-Công thức:
Pass Rate level N=(số lần LevelReached tại level N có kết quả là Pass/tổng số phiên chơi với attempt =N)
+LevelReached = N: Là các phiên chơi (session) mà người dùng đạt được Level N trong session đó, bất kể số lần thử (Attempt) là bao nhiêu.
+Pass/Fail: Kết quả của phiên chơi đó.
-Quy trình thực hiện trong excel
+Tạo PivotTable với bộ lọc Attempt=n ứng với từng level N cần tính
+Kéo cột Pass/Fail vào values và rows để có thể đếm được số lần attempt tại level N có kết quả là Pass và tổng số phiên chơi với attempt =N.

-Kết quả :
Pass Rate level 1: 50.77%
Pass Rate level 5: 55.22%
Pass Rate level 10: 56.63%
+Chi tiết được thể hiện ở: sheet 3) Pass Rate Level trong file excel được đính kèm

4) Session Length trung bình mỗi level

-Công thức:

Session length trung bình (level=N)=(Tổng sesion length của tất cả session tại level N/Số lượng session tại level N)=Average(SessionLength)

-Quy trình tính toán:

+Sử dụng pivot table để lọc ra các levelreached=N bằng cách kéo row

-Kết quả

Level	Average of SessionLength(min)
1	33.86153846
2	32.57647059
3	32.20253165
4	34.04347826
5	33.82089552
6	36.15492958
7	33.76712329
8	33.17857143
9	28.61428571
10	30.77108434
11	33.32941176
12	30.86666667
13	31.53947368
14	29.88405797
15	32.22857143
16	35.50724638
17	36.775
18	32.59459459
19	31.89361702
20	31.5125

+Chi tiết được thể hiện ở: sheet 4) Session Length trung bình mỗi level trong file excel được đính kèm

B) Phân nhóm & đánh giá

5) High-Value Users (Payment>5$), UserID, Total Payment, Total Session, Country

-Mục tiêu: Xác định các người chơi có tổng thanh toán (Payment($)) > 5$, liệt kê UserID, Total Payment, Total Session, và Country.

-Dữ liệu sử dụng:

+UserID: Mã định danh người chơi.

+Payment($): Số tiền thanh toán trong mỗi session.

+Country: Quốc gia của người chơi.

+Total Session: Đếm số dòng (session) cho mỗi UserID, vì không có SessionID, mỗi dòng trong dữ liệu đại diện cho một session.

-Logic tính toán:

+Nhóm dữ liệu theo UserID để tính:

Total Payment: Tổng Payment($) của tất cả session.

Total Session: Đếm số session (số dòng) cho mỗi UserID.

Country: Lấy giá trị Country (giả định mỗi UserID chỉ thuộc một quốc gia).

Lọc các UserID có Total Payment > 5$.

-Kết quả xem chi tiết trong sheet5)High-value users

6) Nhóm kỹ năng tốt (Pass Rate>80%, Attempt TB<2)

-Mục tiêu: Xác định các UserID có:

+Pass Rate > 80%: Tỷ lệ session có Pass/Fail = Pass so với tổng session.

+Attempt trung bình < 2: Trung bình Attempt mỗi session nhỏ hơn 2.

-Logic tính toán:

+Nhóm theo **UserID** để tính:

Pass Rate:(số session pass/tổng số sesion)*100%

Attempt trung bình: (Tổng attempt của tất cả các sesion/số session)

+Lọc UserID thỏa mãn: Pass Rate > 80% và Attempt TB < 2.

+Trình bày: **UserID**, **Pass Rate**, **Attempt TB**, **Country**.

-Kết quả:

Country	(All)										
	Column Labels										
User ID	Average of Attempt	Total Session	Count of Pass/Fail	Average of Attempt	Total Session	Count of Pass/Fail				Pass Rate	Country
U00003			0.00%	1.666666667	3	100.00%	1.666666667	3		100.00%	JP
U00154			0.00%	1.5	2	100.00%	1.5	2		100.00%	VN
U00224			0.00%	1.666666667	3	100.00%	1.666666667	3		100.00%	KR
U00381			0.00%	1.5	2	100.00%	1.5	2		100.00%	JP
U00619			0.00%	1.5	2	100.00%	1.5	2		100.00%	US
U00634			0.00%	1	2	100.00%	1	2		100.00%	KR
U00678			0.00%	1.5	2	100.00%	1.5	2		100.00%	KR
U00736			0.00%	1.5	2	100.00%	1.5	2		100.00%	IN
U00965			0.00%	1.333333333	3	100.00%	1.333333333	3		100.00%	JP
U01011			0.00%	1.5	2	100.00%	1.5	2		100.00%	JP
U01143			0.00%	1	2	100.00%	1	2		100.00%	US
U01149			0.00%	1.5	2	100.00%	1.5	2		100.00%	VN
U01161			0.00%	1	3	100.00%	1	3		100.00%	US

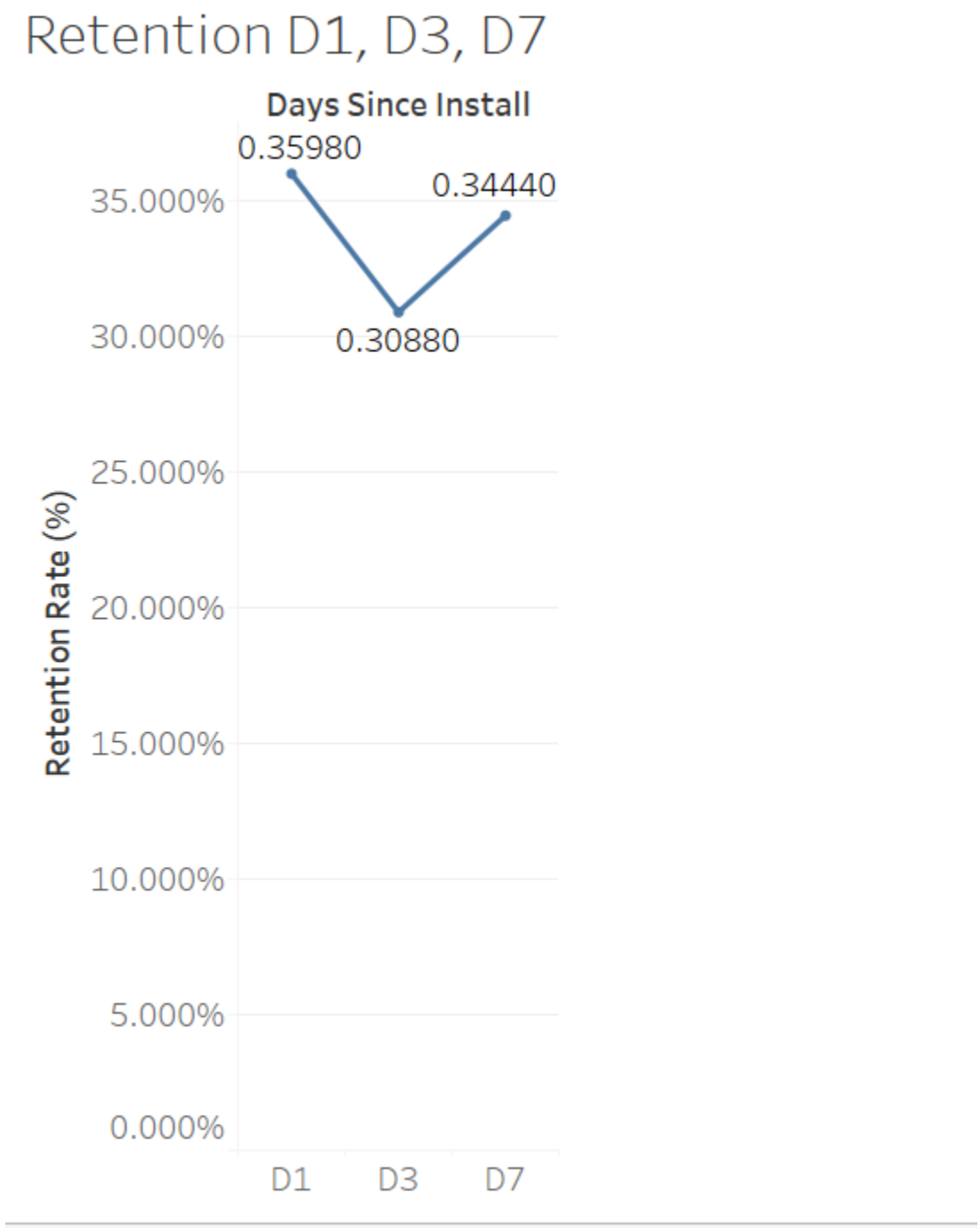
Xem chi tiết ở sheet 6)skilled users

C) Biểu đồ & Insight

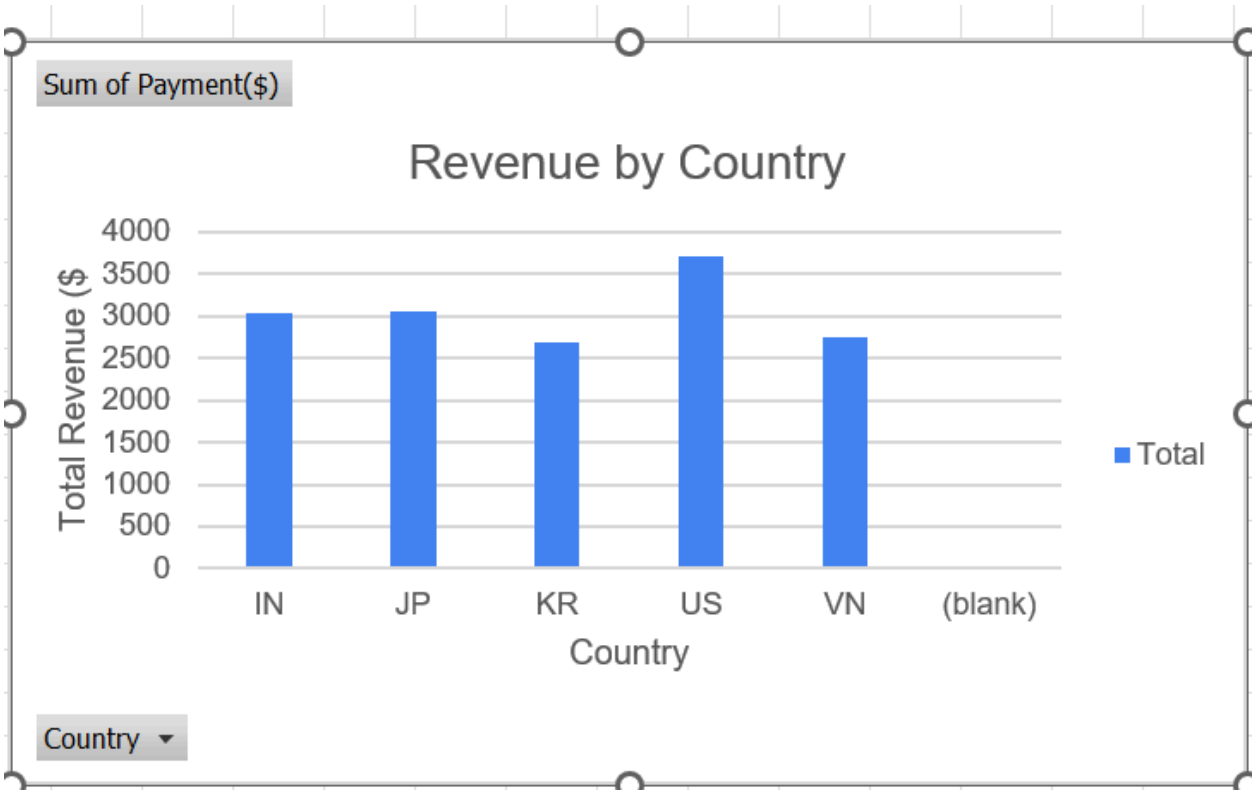
7) Vẽ Retention Curve, Revenue by Country, Histogram of Session Length

-Retention Curve

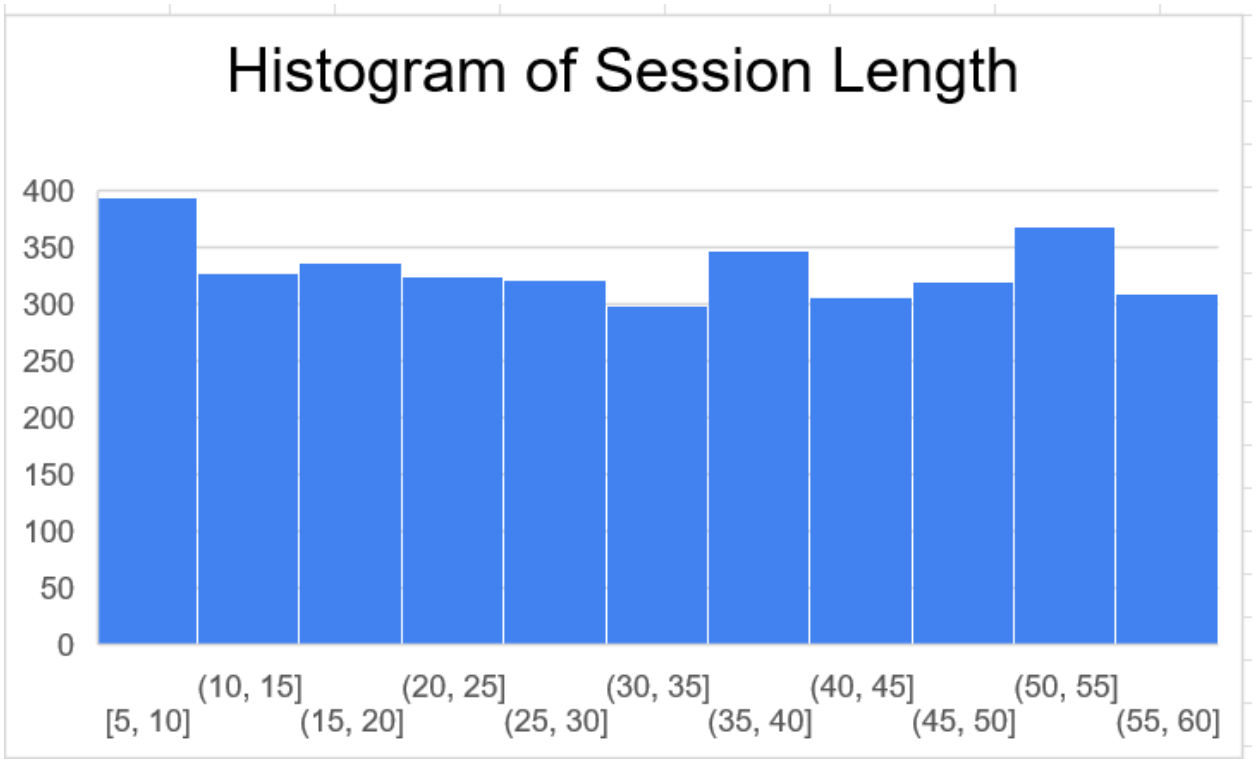
Vì phiên bản excel cũ không hỗ trợ tính toán distinct number, do đó không tạo được pivot table với rentention, nên em đã sử dụng tableau để vẽ curve



-Revenue by Country

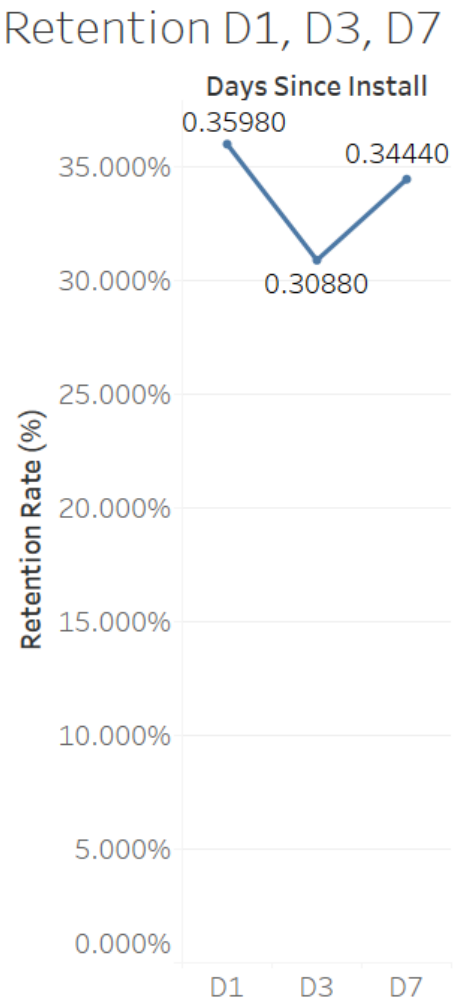


-Histogram of Session Length



8) Viết tối thiểu 5–7 insight, nêu giả thuyết và đề xuất hành động

a. Insight về tỷ lệ giữ chân khách hàng



Chi tiết ở file Retention tableau

Ngày	Phân Tích	Gỉa thuyết	Hành động đề xuất
D1=35.98%	Tỷ lệ giữ chân ở ngày đầu tiên sau khi cài đặt là 35.98%, cao nhất trong số các ngày được đo lường.	Sự hứng thú ban đầu hoặc các khuyến khích (như phần thưởng ngày đầu) có thể là yếu tố thúc đẩy người	Tăng cường các chiến dịch quảng bá hoặc phần thưởng trong ngày đầu để tận dụng thời điểm này, ví

	Điều này cho thấy phần lớn người dùng quay lại ứng dụng ngay sau khi cài đặt.	dùng quay lại sớm.	dụ: tặng coin hoặc ưu đãi đặc biệt để giữ chân người dùng lâu dài.
D3=30.88%	Tỷ lệ giữ chân giảm xuống còn 30.88% vào ngày thứ 3, cho thấy một phần người dùng ngừng tương tác sau ngày đầu tiên.	Người dùng có thể mất hứng thú do thiếu nội dung mới hoặc không tìm thấy giá trị tiếp tục sử dụng sau ngày đầu.	Cung cấp thông báo nhắc nhở hoặc nội dung mới (như cấp độ mới, sự kiện đặc biệt) vào ngày thứ 2 hoặc thứ 3 để duy trì sự quan tâm.
D7=34.44%	Tỷ lệ giữ chân ở ngày thứ 7 là 34.44%, cao hơn một chút so với D3 nhưng vẫn thấp hơn D1. Sự giảm tốc độ giảm cho thấy một nhóm người dùng trung thành bắt đầu hình thành.	Những người dùng quay lại vào D7 có thể là những người đã tìm thấy giá trị lâu dài trong ứng dụng, trong khi số khác đã từ bỏ.	Xác định đặc điểm của nhóm người dùng giữ chân đến D7 (ví dụ: quốc gia, thói quen chơi) và nhắm mục tiêu họ với các ưu đãi để mở rộng nhóm này.

b.Insight về Session Length tb mỗi level

Level	Average of SessionLength(min)
1	33.86153846
2	32.57647059
3	32.20253165
4	34.04347826
5	33.82089552
6	36.15492958
7	33.76712329
8	33.17857143
9	28.61428571
10	30.77108434
11	33.32941176
12	30.86666667
13	31.53947368
14	29.88405797
15	32.22857143
16	35.50724638
17	36.775
18	32.59459459
19	31.89361702
20	31.5125

			thời gian trung bình của mỗi phiên đăng nhập
min	28.36842	level 24	
max	36.775	level 17	

Chi tiết ở sheet 4) Session Length tb mỗi level

Ảnh trên cung cấp thông tin về thời gian trung bình của phiên (Average of SessionLength(min)) theo từng cấp độ (Level) từ 1 đến 50, với giá trị tối thiểu (min) là 28.36842 phút ở Level 24 và giá trị tối đa (max) là 36.775 phút ở Level 17.

c. Insight về 5) High-Value users((Payment>5$), UserID, Total Payment, Total Session, Country

-Với tính toán ở sheet 5) High-Value users, tỷ lệ 38.09% người dùng thuộc nhóm High-Value (có tổng thanh toán > 5$) cho thấy một phần đáng kể người dùng sẵn sàng chi tiêu trong ứng dụng, phản ánh tiềm năng doanh thu lớn từ nhóm này.

-Giả thuyết: Những người dùng này có thể là những người chơi trung thành, thường xuyên tương tác (Total Session cao), hoặc bị thu hút bởi các gói mua trong ứng dụng có giá trị cao.

Đề xuất hành động: Tập trung xây dựng các chiến dịch tiếp thị nhắm đến nhóm High-Value Users, chẳng hạn như ưu đãi cá nhân hóa hoặc gói VIP, để tăng cường chi tiêu và giữ chân họ.

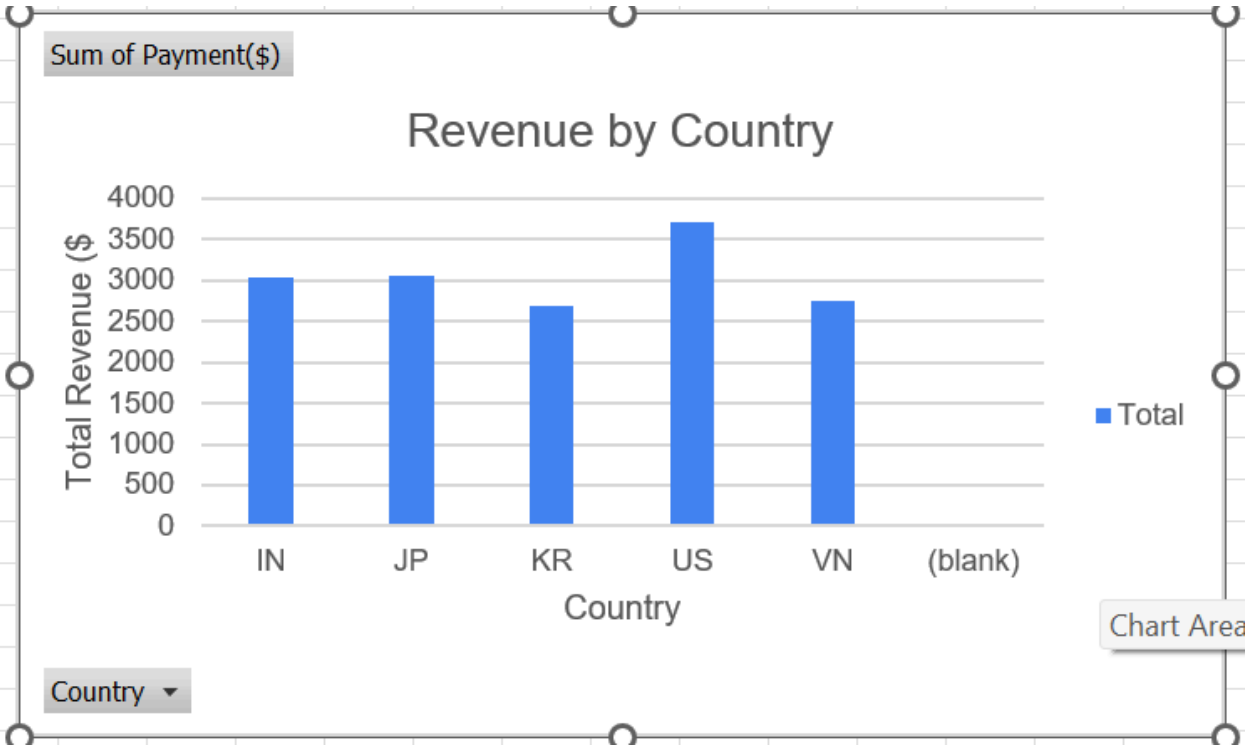
d. Insight về Skilled user của từng nước

Qua sheet 6) Nhóm kỹ năng tốt (Pass Rate > 80%, Attempt TB < 2), tỷ lệ người dùng có kĩ năng tốt theo quốc gia, các giá trị cụ thể như sau:

- Indonesia: 7.69%
- Japan: 30.77%
- Korea: 23.08%
- Myanmar: 7.69%
- Vietnam: 15.38%

Biểu đồ tỷ lệ Skilled Users theo quốc gia cho thấy Nhật Bản và Hàn Quốc là các thị trường chủ lực với tỷ lệ cao (30.77% và 23.08%), trong khi Indonesia, Myanmar, và một phần Việt Nam (15.38%) có tiềm năng cải thiện. Sự khác biệt này có thể xuất phát từ kỹ năng người dùng, thiết kế game, hoặc chiến lược tiếp cận. Các đề xuất hành động, bao gồm điều chỉnh độ khó, địa phương hóa nội dung, và tăng cường hỗ trợ, sẽ giúp tối ưu hóa hiệu suất người dùng và nâng cao tỷ lệ Skilled Users trên toàn cầu.

e.revenue by country

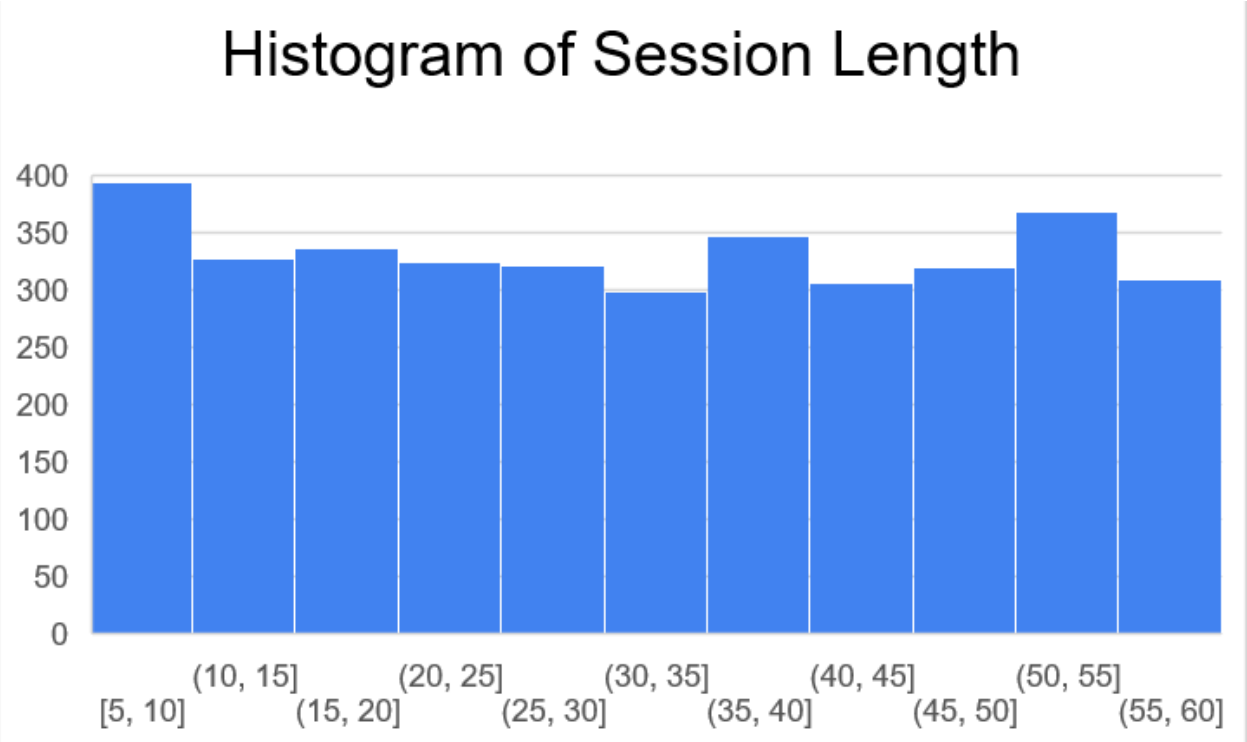


-Dữ liệu phản ánh tổng doanh thu (Total Revenue) theo quốc gia, với các giá trị ước lượng từ cột "Sum of Payment($)" trong sheet "Data". Các quốc gia được liệt kê bao gồm IN (Indonesia), JP (Nhật Bản), KR (Hàn Quốc), US (Mỹ), VN (Việt Nam)

-Biểu đồ "Revenue by Country" chỉ ra Mỹ là thị trường chủ lực với doanh thu cao nhất (khoảng 3500-4000 USD), trong khi các quốc gia như IN, JP đóng góp đồng đều (2500-3000 USD), và US thấp hơn kỳ vọng (2000-2500 USD).Các đề xuất hành động,

bao gồm cá nhân hóa chiến lược tiếp thị, điều chỉnh giá trị tại VN và KRsẽ giúp tối ưu hóa doanh thu và mở rộng thị phần.

f. Histogram of Session Length



Biểu đồ Histogram of Session Length cho thấy phân bố không đồng đều, với đỉnh cao ở khoảng [5, 10] (khoảng 400 phiên) và sự ổn định ở khoảng 10-25 phút, nhưng giảm dần ở các khoảng dài hơn (30-60 phút). Điều này gợi ý rằng người dùng có xu hướng chơi ngắn hoặc trung bình, với một nhóm nhỏ duy trì phiên dài. Các đề xuất hành động, bao gồm tối ưu hóa nội dung cho các khoảng thời gian khác nhau, điều chỉnh độ khó, và thêm động lực, sẽ giúp cải thiện trải nghiệm người dùng và tăng thời gian tương tác tổng thể.