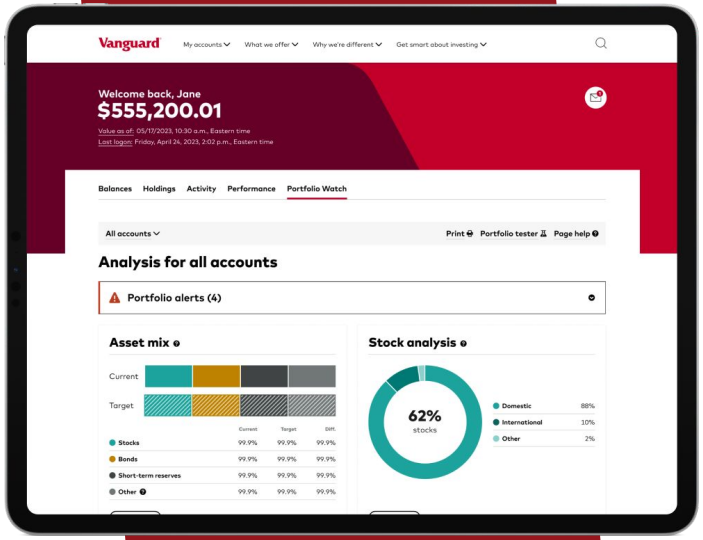
Vanguard®

NEW UI Analysis

CX Analyst Team

Aisyah Amatul Ghina | Sasha Crowe



INTRODUCTION

Vanguard is on a mission to offer more personalized financial advice, high-quality investments, retirement tools, and relevant market insights for their clients.

Main project pursuit:

“Did the **new UI lead** to **higher completion rates?**”

- Upgrade to a more modern & intuitive design
- Timely in-context prompts

3 month period (3/15/2017-6/20/2017)

DATA OVERVIEW

Client Profiles

Demographics, Tenure, Balance, etc.
Unique Client ID: 70,609

Digital Footprints

Visit ID, Process Steps, Timestamps
Unique Client ID: 120,157

Experiment Roster

Variation: Test, Control, NA (20,109)
Unique Client ID: 70,609

DATA WORKFLOW

Data Cleaning

Duplicates, merge and split datasets, etc.

Data Analysis

- Exploratory data analysis
- KPI calculation
- Hypothesis testing

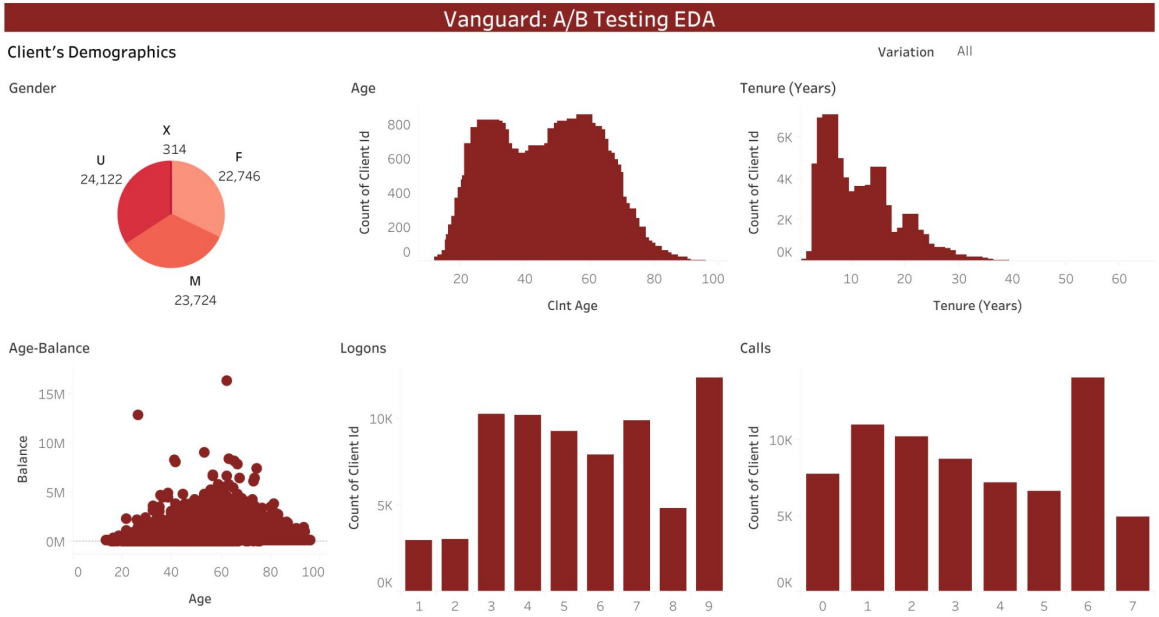
Data Visualization

Python and tableau

EXPLORATORY DATA ANALYSIS

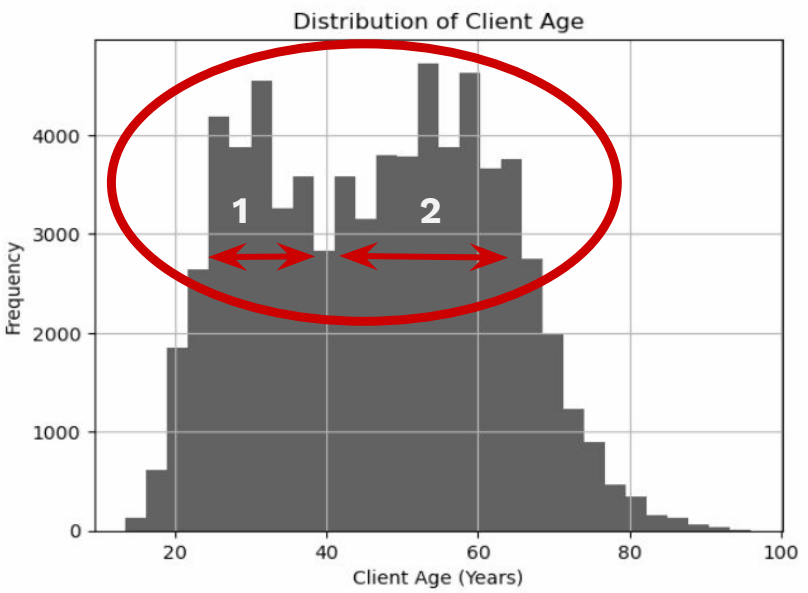
Vanguards: A/B Testing EDA

We explore each variables in client profile dataset.	We compare client's demographics information between test and control group to identify any potential biases.	Comparing the number of client visiting the platform between test and control group.



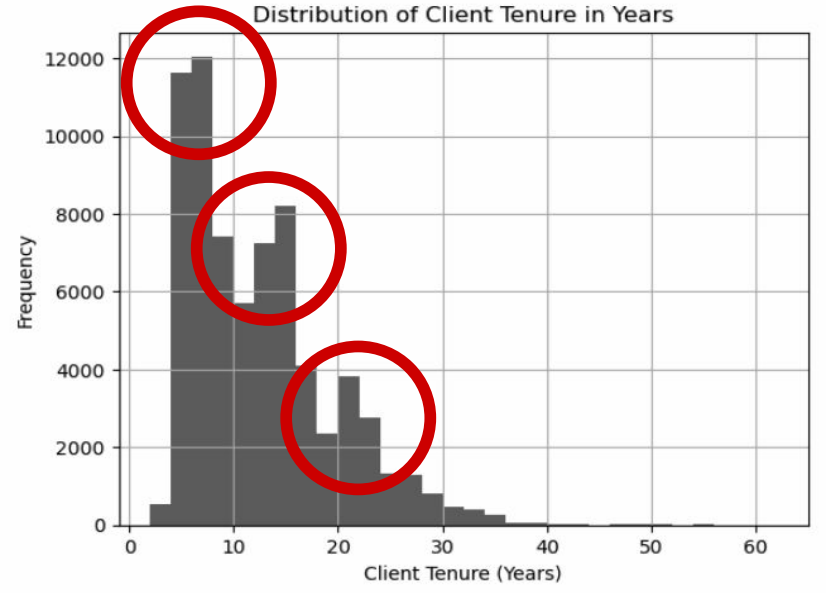
AGE & TENURE DISTRIBUTION

From youths to seniors...



Histogram shows **2** distinct peaks in age

New clients to long-standing clients...



Histogram shows **3** distinct peaks in tenure

CLIENT PROFILE CLUSTERS

Using **K-Means Clustering** we were able to identify 3 different distinct client profiles:

Cluster A
~ JOJO ~

29 yrs old
9 yr tenure
51,534.15 balance

Cluster B
~ VERONICA ~

48 yrs old
12 yr tenure
81,797.34 balance

Cluster C
~ KARL ~

65 yrs old
15 yr tenure
122,516.38 balance

Note: Gender was not included, as a third of the available gender information are unspecified, rendering it un-useful as a profile trait

KEY PERFORMANCE INDICATORS

Completion Rates

The proportion of users who reach the final ‘confirm’ step

A process is considered complete when the users follow the process step sequentially:

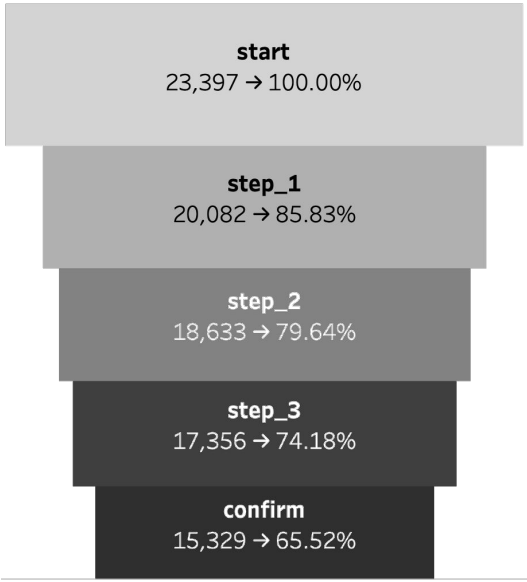
start > step 1 > step 2 > step 3 > confirm

Average Session Duration

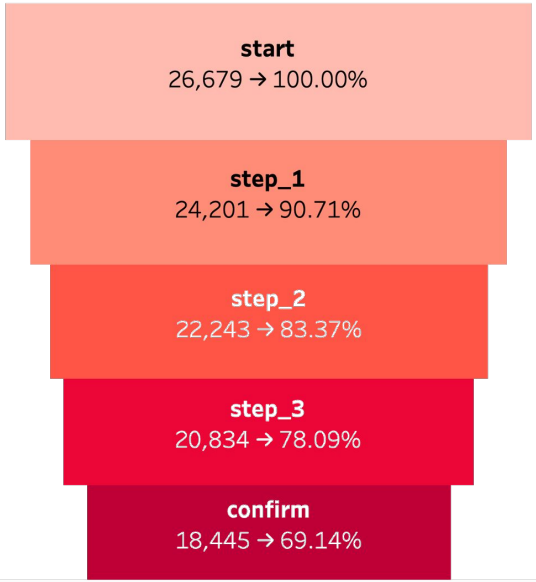
Error Rates

COMPLETION RATES (CR)

Control Group



Test Group



CR:	**Control:** 65.52%	**Test:** 69.14% ✅

Hypothesis Testing

H_0	CR_test <= CR_control
H_1	CR_test > CR_control

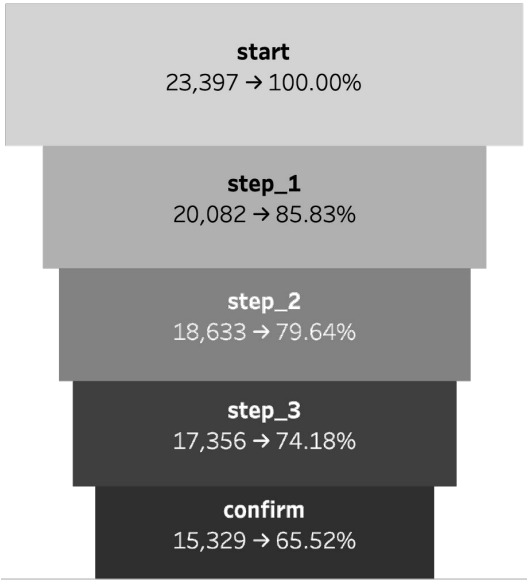
What are the completion rate (CR) results of the A/B test?

↑ 3.62 pp in test group CR

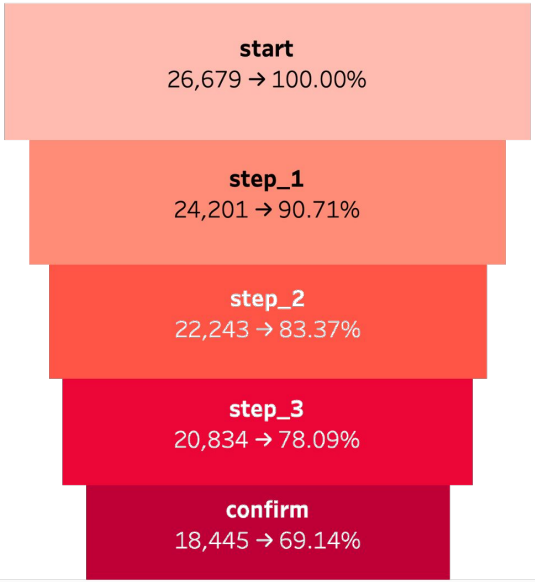
- z-stat = 8.62
- p-value = 0.00 **reject** H_0

COMPLETION RATES (CR)

Control Group



Test Group



CR: **Control:** 65.52% → ▲ < 5% ← **Test:** 69.14%

Hypothesis Testing

H_0	CR_test >= CR_control + 5%
H_1	CR_test < CR_control + 5%

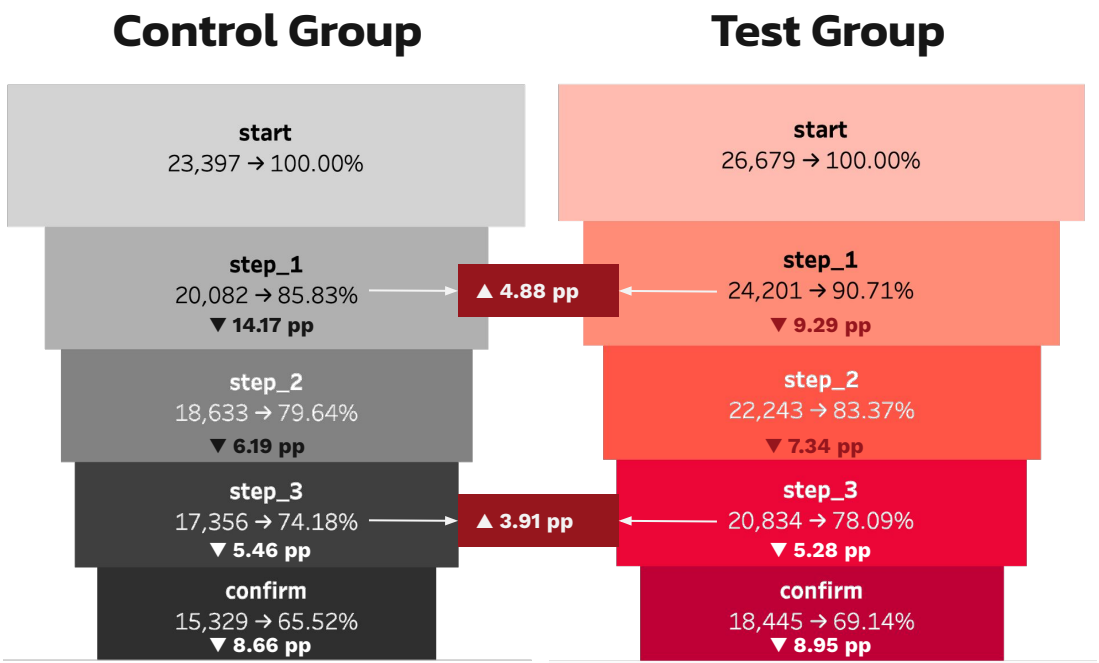
Cost-effectiveness Threshold

↑ 3.62 pp in test group CR

- z-stat = -4.88
- p-value = 0.00 **reject** H_0

With a ▲ of less than 5%, that is statistically reinforced, this A/B test **does not** meet the minimum improvement requirements

COMPLETION RATES (CR)



Hypothesis Testing

H_0	CR_test <= CR_control
H_1	CR_test > CR_control

Which steps performed better?

From Start - Step 1

- z-stat = 17.04
- p-value = 0.00 **reject** H_0

From Step 2 - Step 3

- z-stat = 2.11
- p-value = 0.02 **reject** H_0

KEY PERFORMANCE INDICATORS

Completion Rates

The proportion of users who reach the final ‘confirm’ step

Average Session Duration

The average duration users spend in the process

Keep in mind that despite there being 5 total steps, we are looking at the timing from one step to another → resulting in there being 4 time values

Error Rates

AVERAGE SESSION DURATION

Test Group

Start	Step 1	Step 2	Step 3	Confirm	Total
11.55	27.25	70.00	72.55		03:01

Control Group

Start	Step 1	Step 2	Step 3	Confirm	Total
21.43	24.05	71.65	90.65		03:28

Did the test group's avg. duration perform differently to the control group's? [3:01 vs 3:28]

- z-stat = -28.17
- p-value = 0.00 **reject** H_0

What about the individual steps?

	Start	Step-1	Step-2	Step-3
t-value	-79.47	15.19	-7.48	-25.09
p-value	< 0.00	< 0.00	< 0.00	< 0.00

Test

Start : Step1
Step2 : Step3
Step3 : Conf.

Hypothesis Testing

H_0	AD_test = AD_control
H_1	AD_test <> AD_control

AVERAGE SESSION DURATION

Test Group

Start	Step 1	Step 2	Step 3	Confirm	Total
11.55	27.25	70.00	72.55		03:01

Control Group

Start	Step 1	Step 2	Step 3	Confirm	Total
21.43	24.05	71.65	90.65		03:28

Did the test group's avg. duration perform differently to the control group's? [3:01 vs 3:28]

- z-stat = -28.17
- p-value < 0.00 **reject** H_0

What about the individual steps?

	Start	Step-1	Step-2	Step-3
t-value	-79.47	15.19	-7.48	-25.09
p-value	< 0.00 ✅	< 0.00 ✅	< 0.00 ✅	< 0.00 ✅

Test 📈

Start : Step1
Step2 : Step3
Step3 : Conf.

Hypothesis Testing

H_0	AD_test = AD_control
H_1	AD_test <> AD_control

KEY PERFORMANCE INDICATORS

Completion Rates

The proportion of users who reach the final ‘confirm’ step.

Average Session Duration

The average duration users spend on each step.

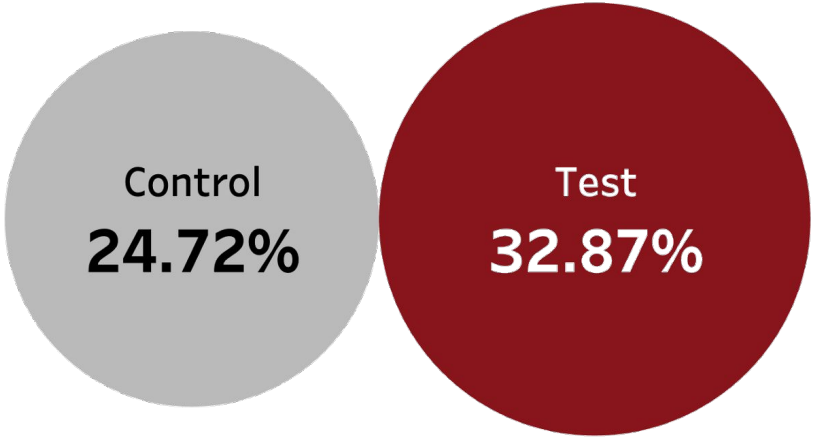
Error Rates

The proportion of users who go back to a previous step

If a user went back from confirm to start, it is not considered an error as we assume the session ended correctly and the user started a new session

ERROR RATES

Overall error rates in client %



Hypothesis Testing

H_0	ER_test = ER_control
H_1	ER_test <> ER_control

Does the test group have a sig. different error rate than the control group?

- z-stat = 20.12 & p-value = 0.00 → **reject** H_0

Result is not surprising, given the MVP nature of the experiment

ERROR RATES

Error occurrence by step per visit

	Control	Test
Step 1	2,491	6,404
Step 2	2,163	4,780
Step 3	4,247	4,744
Confirm	228	75

	Ctrl ER	Test ER	z-stat	p-value
Step 1	10.04 %	18.89 %	29.56	0.00
Step 2	9.23 %	16.49 %	24.39	0.00
Step 3	21.74 %	20.52 %	-3.06	0.00

H_0 No difference b/w steps | H_1 There's a difference

- Cnrl group has lower ER, apart from...
- **Step 3**; only lower ER for the test group 👀
- Could explain why S.3 had a shorter duration

Test CR outperformed Crtl CR, despite the high ER 👀

EXPERIMENT EVALUATION

	Control	Test	Winning Hypothesis	Better Performer
Completion Rates	65.52%	69.14%	CR_test > CR_control; CR_test < CR_control + 5%	**Test**
Average Session Duration	03:28	03:01	AD_test <> AD_control	**Test**
Error Rates	24.72%	32.87%	ER_test <> ER_control	**Control**

Experiment design observations:

- The sampling of the two groups was done well
  - Results did not appear skewed as each KPI result echoed each other
- The duration of the test was enough to reach a sturdy conclusion

Beneficial additional data:

- Difficult to interpret behaviour (different tabs? were some errors typo corrections?)
- More UI details in the project brief (longer/shorter duration is better?)

CONCLUSION

	CR	Duration	ER
Overall	**Test**	**Test**	**Ctrl**
Start to **Step 1**	Test sig. ↑	Test shorter	Ctrl
Step 1 to **Step 2**	Ctrl	Ctrl	Ctrl
Step 2 to **Step 3**	Test sig. ↑	Test shorter	Test ER ↓
Step 3 to **Confirm.**	Ctrl	Test shorter	Ctrl

→ Test outperformed Ctrl in ⅔ KPIs

→ Test outperformed Ctrl in all 3 KPIs ❗

~Insights Box~

While the test UI, in its current form, **did not** meet the cost-effective threshold, it was *very* close

➔ just shy by 1.38 pp

Look into clusters; **Cluster A** (Jojo) has compelling results

- Jojo (CA) had the lowest ER
- 72% CR → passing the C.E. threshold

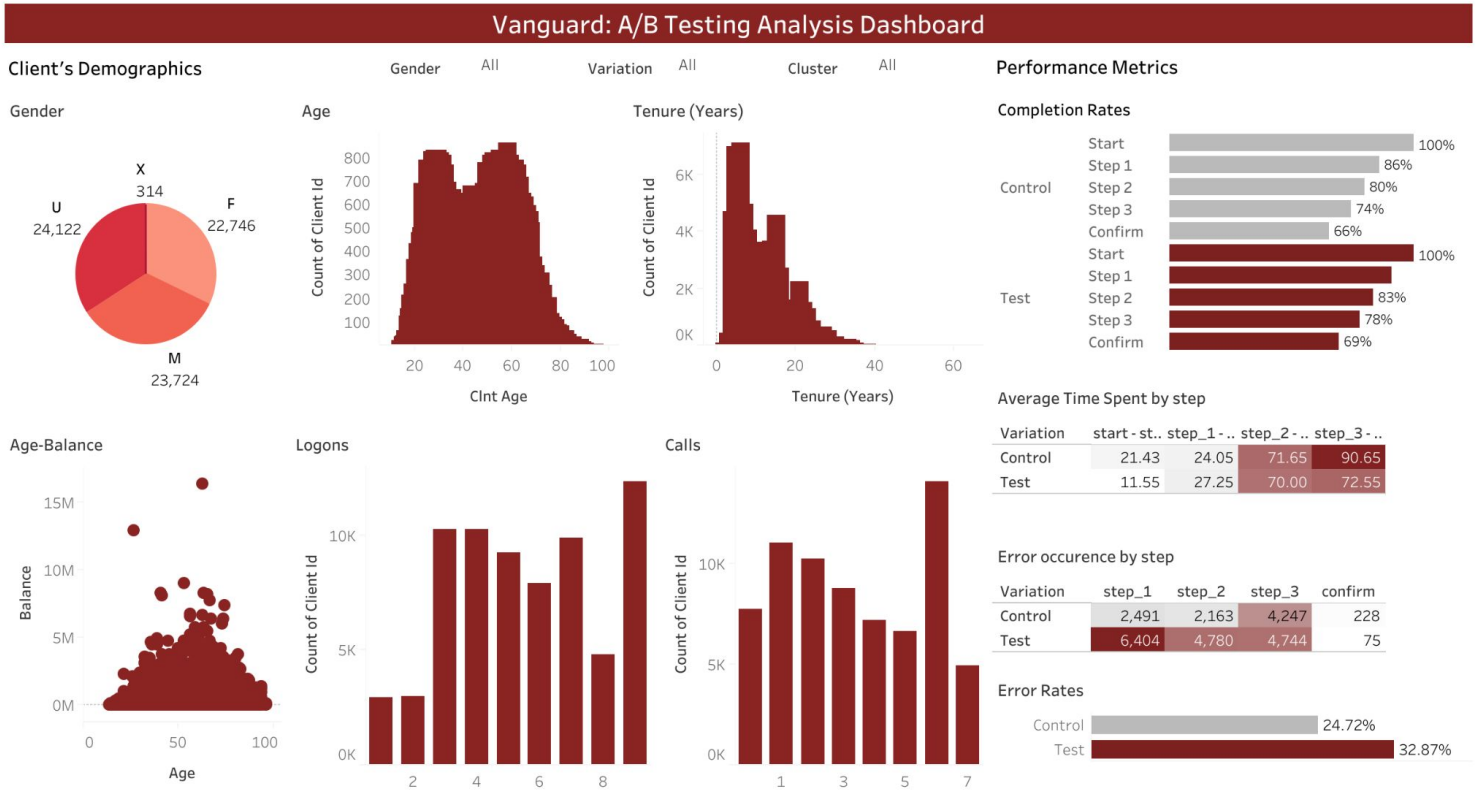
Our recommendations

1] Run another A/B test, using the insights from this A/B test

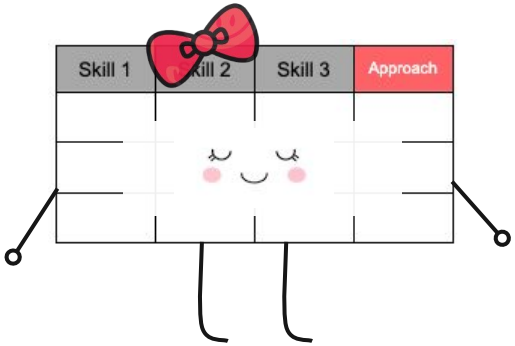
Start: Test → **S.1: Test** → **S.2 : Ctrl** → **S.3: Test** → **Con.: Test**

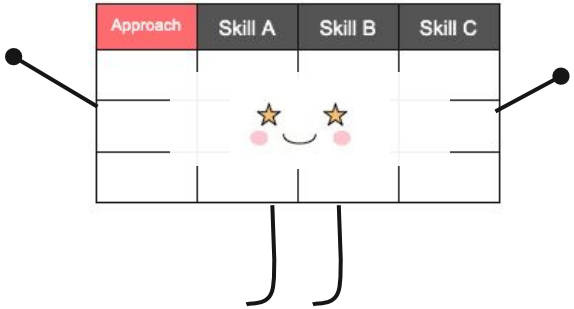
2] Reduce as many bugs /tech errors as feasibly possible for the test UI

TABLEAU VISUALIZATION



TEAMWORK & PROJECT MANAGEMENT





Two dataframes with a series of unique skills spot a column in common...
...giving both a reference to harmoniously merge on &...
...granting both access to a wider range of skills when combined

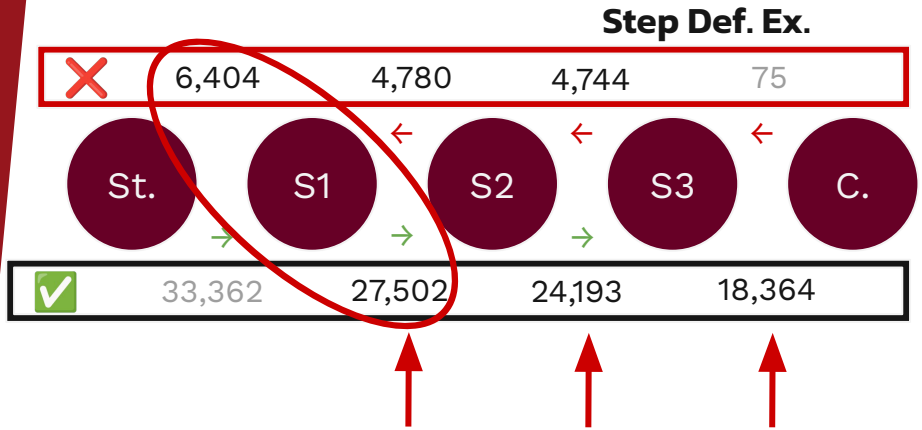
- Organized in a shared working document 📋
- Meet-up, discuss, go over completed tasks, agree on next steps, & 🔁
- Communicate when an issue arises 📢

THANKS

CX Analyst Team

Aisyah Amatul Ghina | Sasha Crowe

ERROR RATES



Formula: Error + Not Error = Total Observations

	Ctrl ER	Test ER	z-stat	p-value
Step 1	10.04 %	18.89 %	29.56	< 0.00
Step 2	9.23 %	16.49 %	24.39	< 0.00
Step 3	21.74 %	20.52 %	-3.06	< 0.00

Error occurrence by step per visit

	Control	Test
Step 1	2,491	6,404
Step 2	2,163	4,780
Step 3	4,247	4,744
Confirm	228	75

CHALLENGES

- Working with **large datasets in long format** to calculate performance metrics can be confusing and often involves some trial and error.
- Understanding and selecting the **appropriate statistical test** for each metric.
- Operating with infinite directions & dealing with different applicable meanings; making decisions when faced with uncertainties.

LEARNINGS

- Expanding knowledge in **Python methods** for data analysis, such as diff(), shift(), etc.
- Gain a clearer understanding of **how to perform hypothesis testing** and determine when to use a **z-test** or **t-test**, as well as whether to apply a **two-sided** or **one-sided** approach.
- It's important to pay attention to underlying assumptions.