



# Is the new UI better than the old one?

A/B testing

Beata, Pierre



# Plan of the presentation

- Introduction and statement of the problem
- Data overview and data cleaning
- EDA – Client's analysis
- Performance metrics:
  - Completion rate
  - Average time per each step
  - Error rate
- Hypothesis testing : is new interface better?
- Experiment evaluation : can we do better?
- General remarks
- Conclusions



# Introduction

## Who we are?

- Data analyst in the Customer Experience (CX) team at **Vanguard** (US-based investment management company)



## Challenge:

- Vanguard introduced **more intuitive and modern User Interface (UI)**, coupled with timely in-context prompts provided to users directly within the context of their current task or action.
- Would these changes encourage more clients to complete the process? Did the new UI lead to higher completion rates?



## Our goals:

- Analyse performance of ,new' and ,old' user interface
- Define relative KPIs
- Help to make business decisions

But first understand the nature and structure of your datasets.

# Data overview

- **Client Profiles** (df\_final\_demo): Demographics like age, gender, and account details of clients.

size: (70609, 9)

	client_id	clnt_tenure_yr	clnt_tenure_mnth	clnt_age	gendr	num_accts	bal	calls_6_mnth	logons_6_mnth
0	836976	6.0	73.0	60.5	U	2.0	45105.30	6.0	9.0
1	2304905	7.0	94.0	58.0	U	2.0	110860.30	6.0	9.0

- **Digital Footprints** (df\_final\_web\_data): A detailed trace of client interactions online.

size: (755405, 5)  
(df\_1 and df\_2)

	client_id	visitor_id	visit_id	process_step	date_time
0	9988021	580560515_7732621733	781255054_21935453173_531117	step_3	2017-04-17 15:27:07
1	9988021	580560515_7732621733	781255054_21935453173_531117	step_2	2017-04-17 15:26:51

- **Experiment Roster** (df\_final\_experiment\_clients): A list revealing which clients were part of the experiment

size: (70609, 2)

	client_id	Variation
0	9988021	Test
1	8320017	Test

# Data cleaning and merging

In **,client profiles'**:

- removing rows with *NaN* or *X* values for **,gendr'** column (14) rows, *NaN* value for **,clnt\_age'** column

→ merging with **,experiment rooster'** into **,clients'** df

- filling *NaN* values in **,Variation'** column with **,Unknown'** (20106)

Digital footprints:

- dropping duplicates (10764)
- performing analysis
- additional columns with: average step duration, error and progress counts, number of visits, numbers of each step

→ merging with **,Digital footprints'** with **,clients'** into **web\_data\_summary**

→ keeping only data for **,Control'** and **,Test'** groups

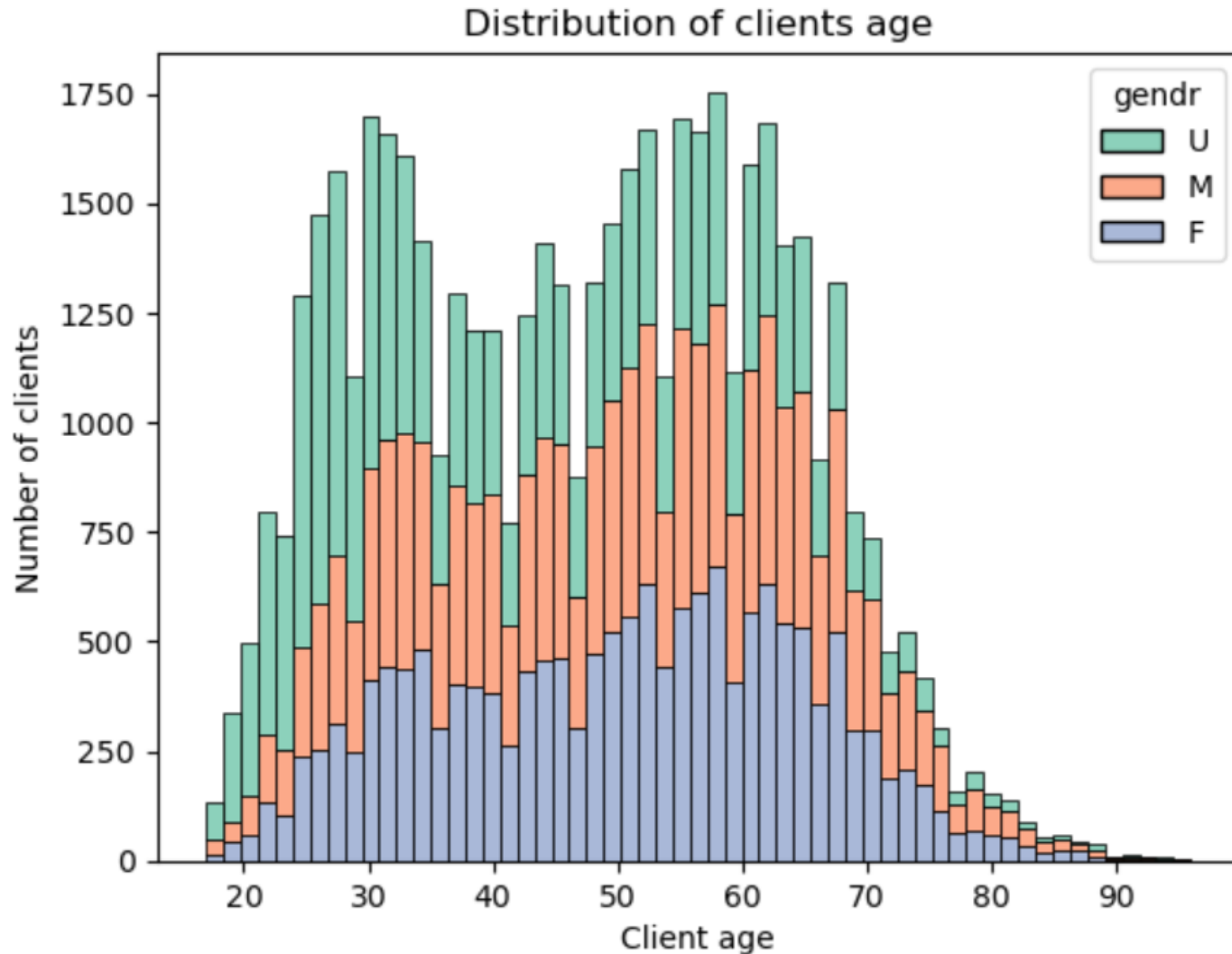
Final data frame size: **(50485, 25)**

Test 26959  
Control 23526

client_id	avg_step_duration	had_error	error_count	had_progress	progress_count	no_of_visits	is_confirmed	no_of_starts	no_of_step_1	no_of_step_2	no_of_step_3	no_of_confirms	step_duration_minutes
836976	0 days 00:03:46	False	0	True	4	2	True	5.0	1.0	1.0	1.0	3	3.766667
2304905	0 days 00:00:59	False	0	True	4	1	True	2.0	1.0	1.0	1.0	1	0.983333
1439522	0 days 00:00:39	True	1	True	3	2	False	2.0	1.0	1.0	1.0	0	0.650000

# EDA

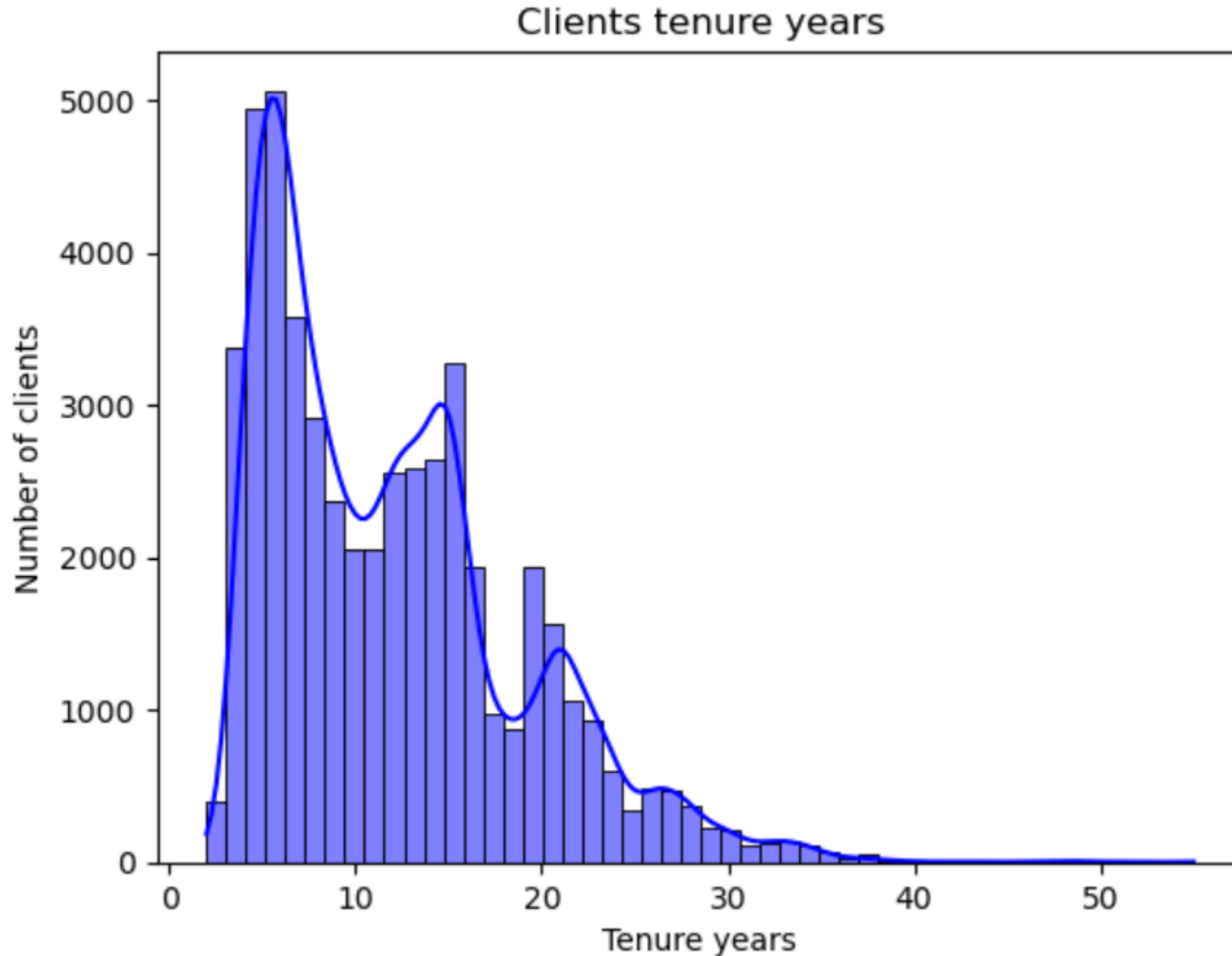
Who are the primary clients using online process? **Are the primary clients younger or older, new or long-standing?**



- Clients are evenly distributed between genders (34% for M and F, 32% for U).
- Average client age is 47 years.
- Only 25% of clients has less than 34 years.
- The oldest client is 96 😊

# EDA

Who are the primary clients using online process? Are the primary clients younger or older, new or long-standing?



- Average client's tenure years is 12.
- 75% of clients are long-standing with up to 16 tenure years.
- Only 25% are clients are there for less than 6 years.
- New clients with tenure years up to 2 years are minor group (only 61 persons).
- The longest tenure time is 55 years 😊

# Performance Metrics

## Test group 53%

Clients experienced the new, spruced-up digital interface



## Control group 47%

Clients interacted with traditional online process



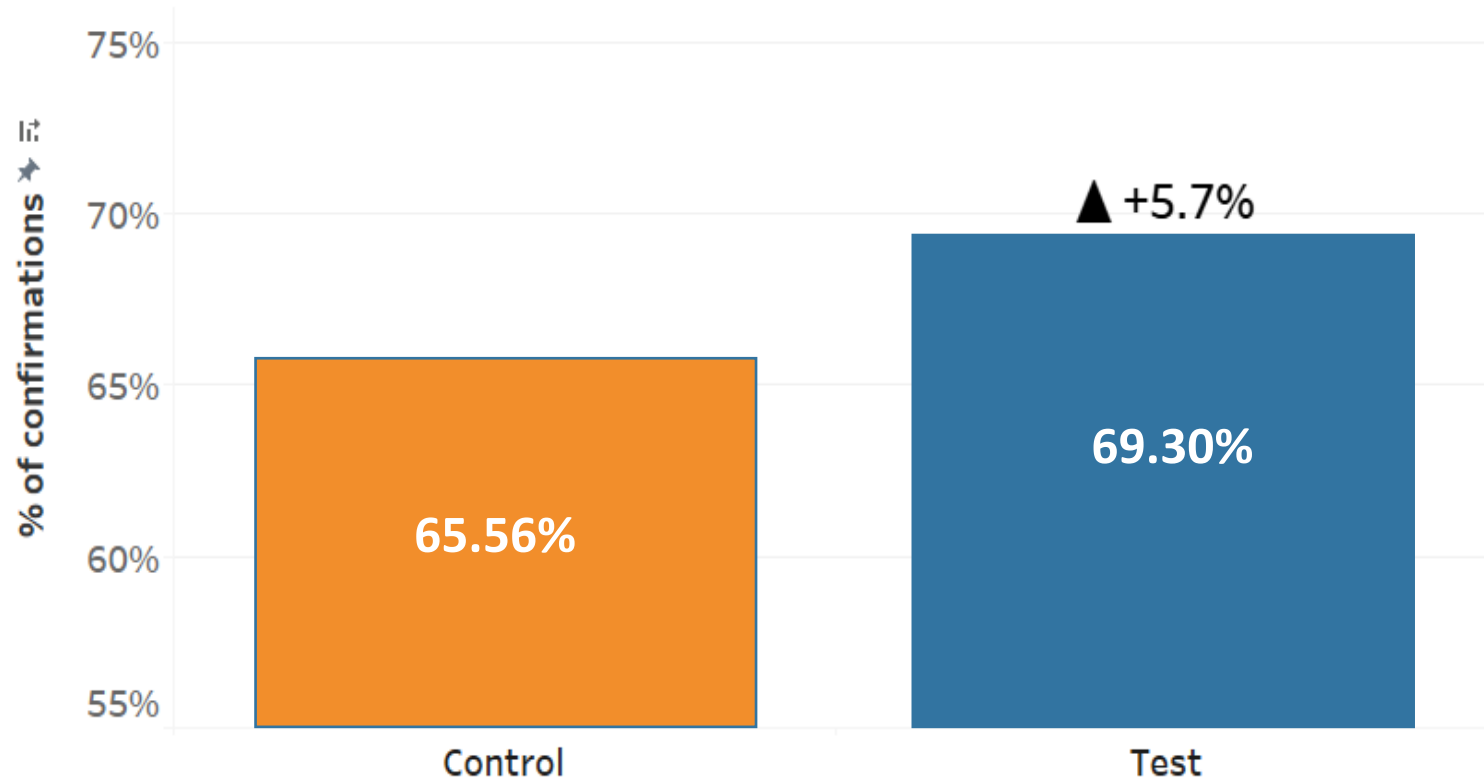
## Key Performances Indicators (KPIs):

- **Completion Rate:** The proportion of users who reach the final 'confirm' step.
- **Error Rates:** Moving from a later step to an earlier one.
- **Time Spent on Each Step:** The average duration users spend on each step.



# Completion Rate

What is proportion of clients who finished the process in Test and Control groups?

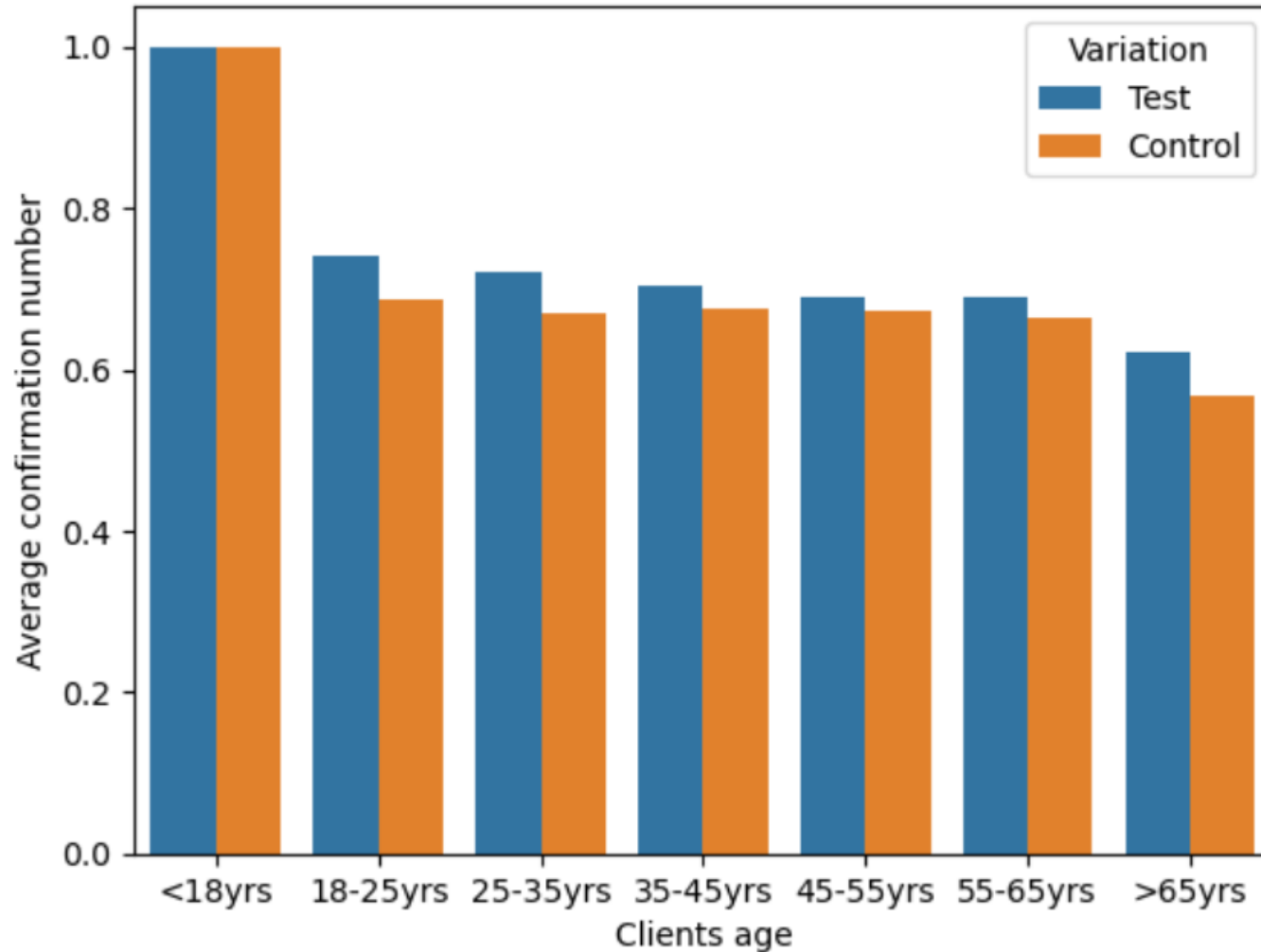


Great News !

Completion Rate increased in Test group by 5.7% with respect to Control group.

# Completion Rate

What is Completion rate in different age groups?

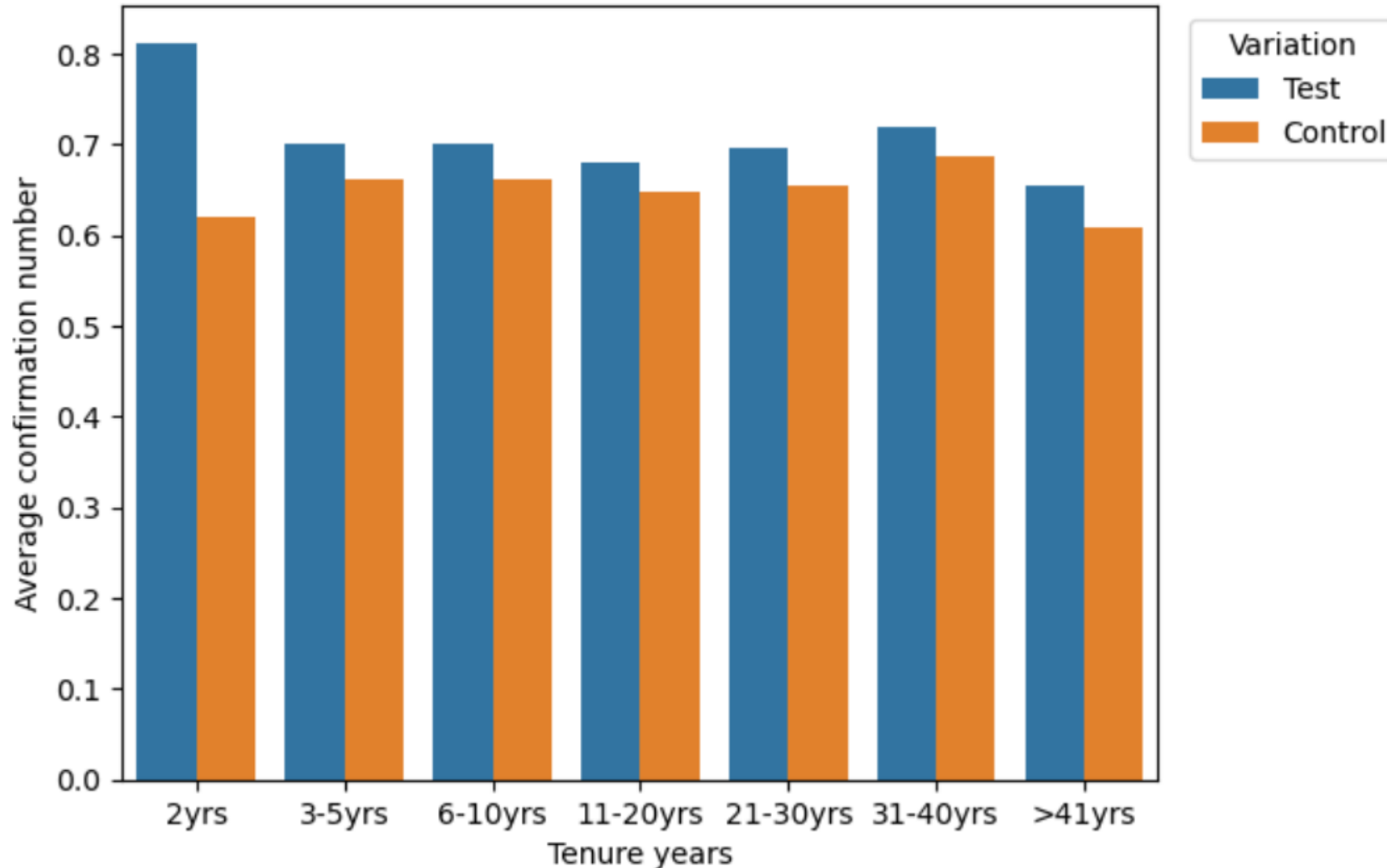


- Younger clients have the highest completion rate.
- In other age groups ,Test' group performs better.
- Completion rate decreases with age.

But be careful with data !  
The sizes of age groups are not the same.

# Completion Rate

What is Completion rate depending on client's tenure years?

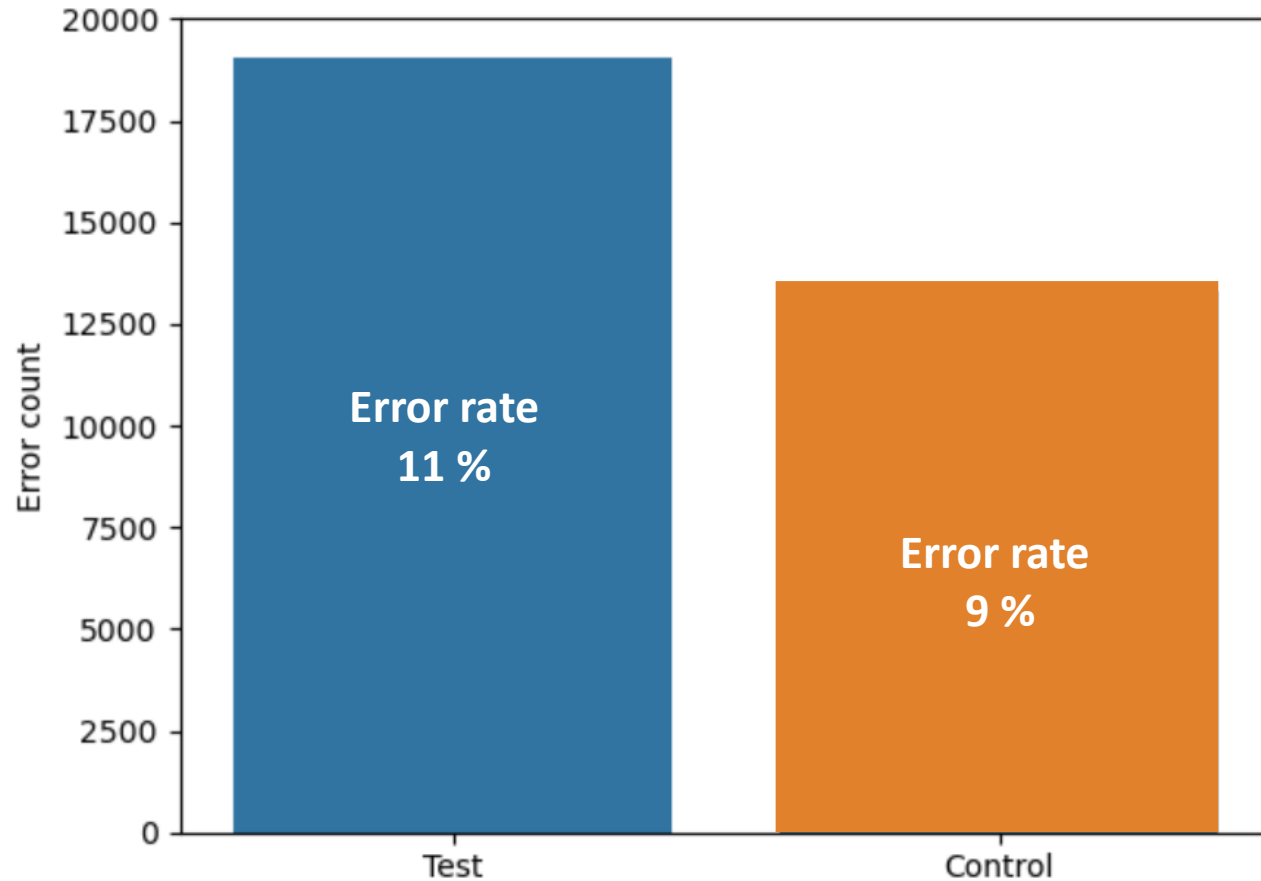


- New clients have the highest completion rate in the Test group and one of the lowest in Control group.

But again be careful with data !  
The sizes of age groups are not the same.

# Error Rates

What is the number of errors (going back to previous step) in Test and Control groups?



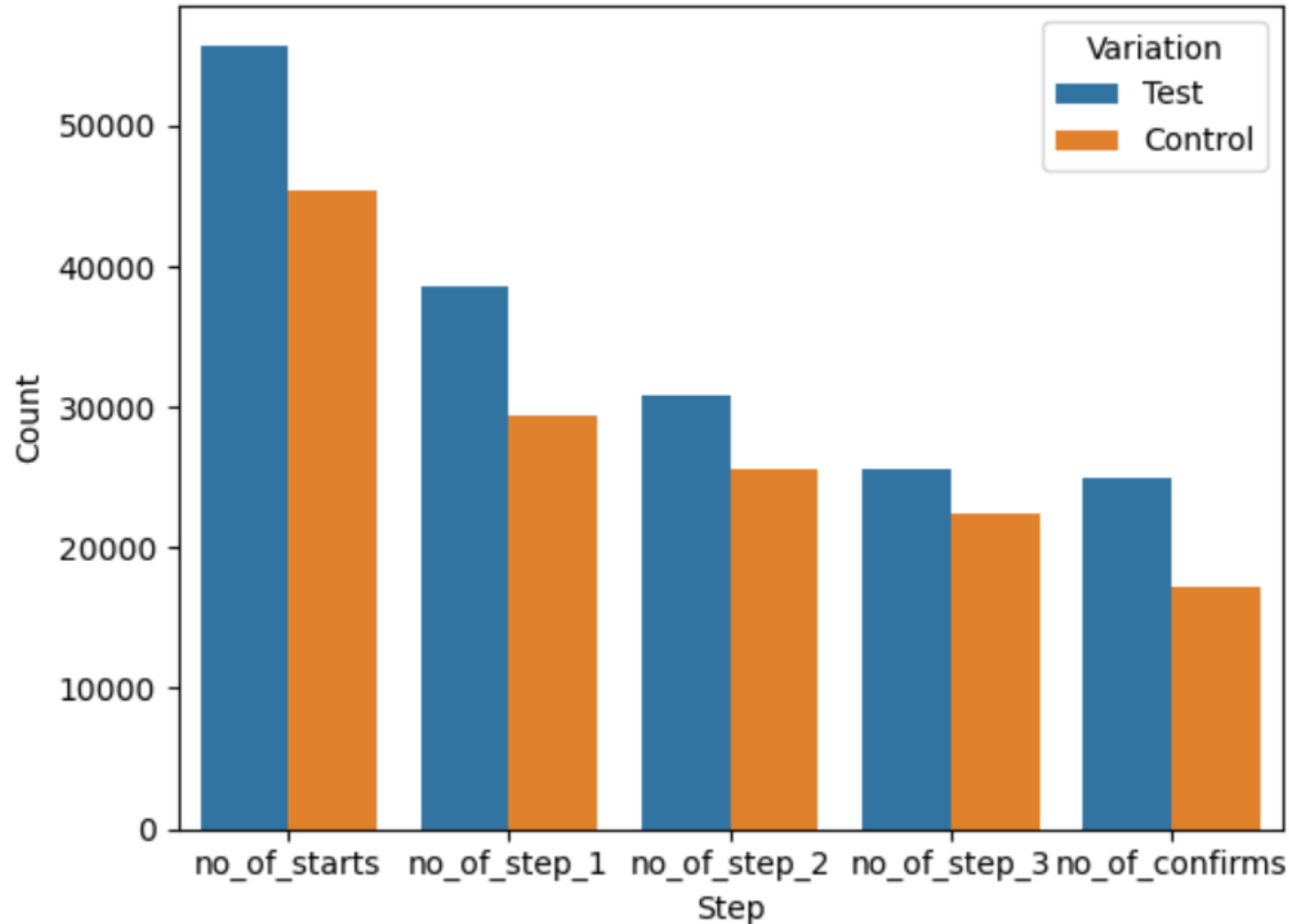
$$\text{Error rate} = \frac{\text{Number of 'backward' steps}}{\text{Total number of steps}}$$

Bad News ☹️

Error rates are higher in Test group!

# Error Rates

What is total number of each steps taken in Test and Control groups?



## Bad News ☹️

Clients were trying multiple times to start the process but without success...

Or were stuck on 'start' page...

Stacking on starting page indicates problems with application? Should it be treated as an error?

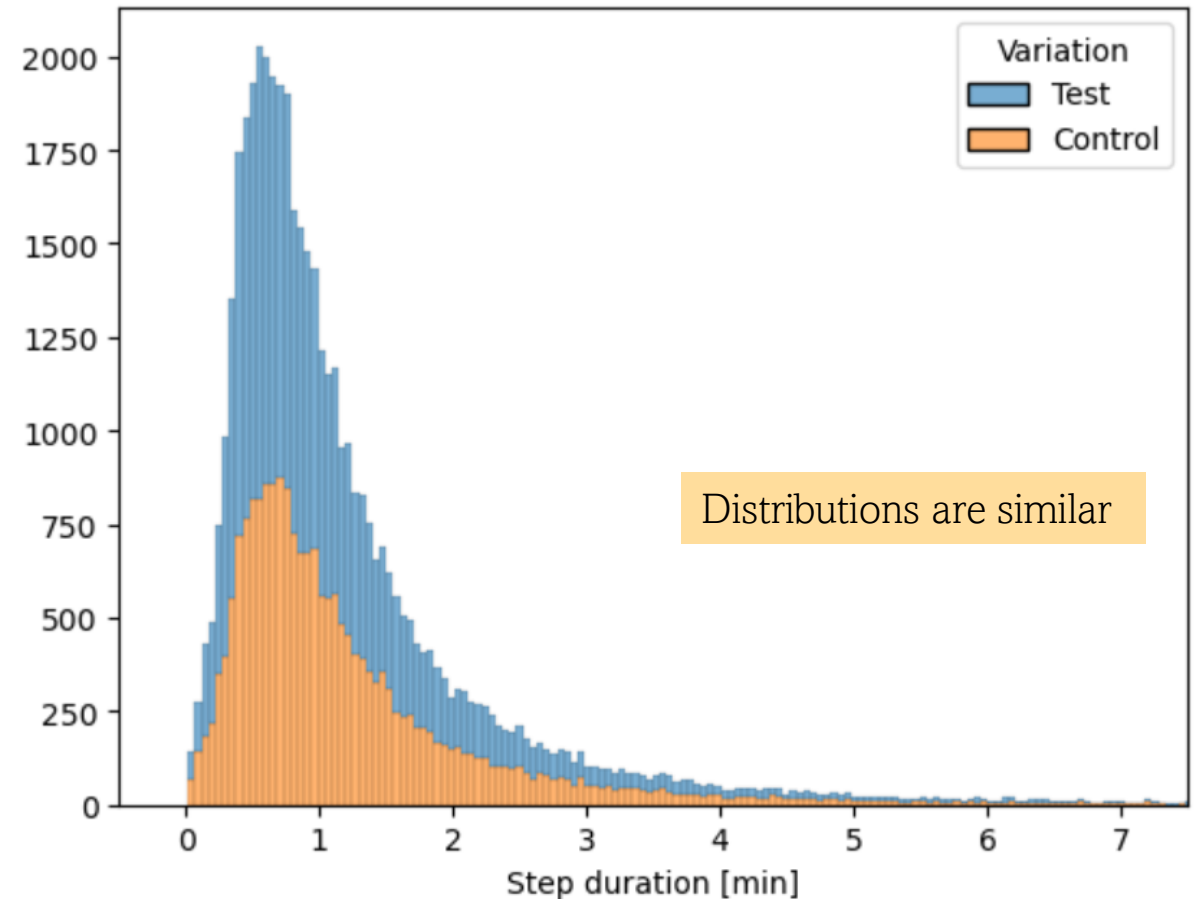


# Time Spent on each step

**Data cleaning** was required before analysing step duration data :

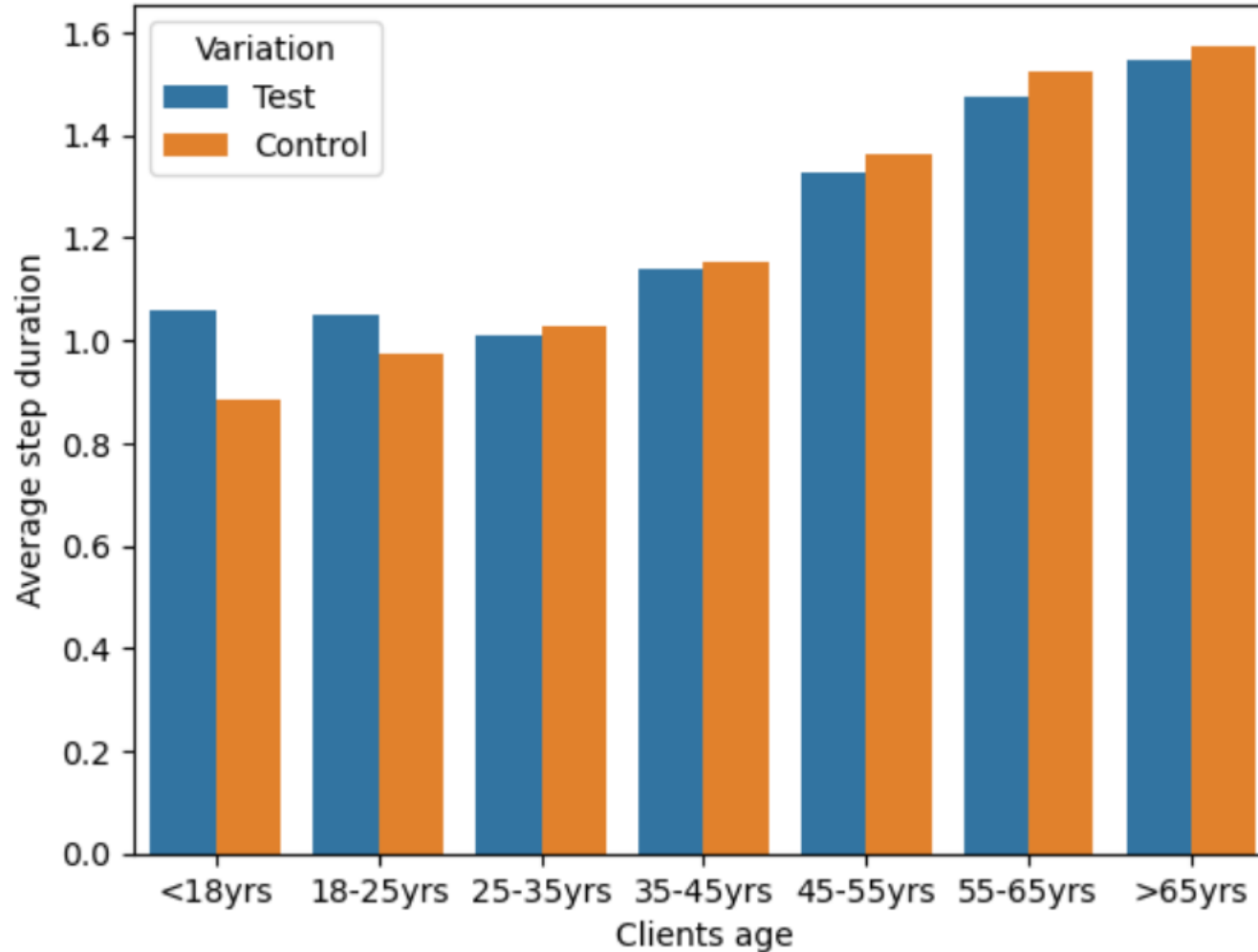
- remove rows with time equal to zero: 3937 records, from which 3916 had no progress
- remove outliers: only 28 client had time longer than 20 min, including 17 clients without any progress

Before			After		
	Control	Test		Control	Test
count	23526.000000	26959.000000		21096.000000	25424.000000
mean	1.185871	1.216098		1.300494	1.273674
std	1.515970	1.557726		1.290786	1.396329
min	0.000000	0.000000		0.016667	0.016667
25%	0.483333	0.504167		0.595833	0.554167
50%	0.847049	0.825000		0.937500	0.866667
75%	1.437500	1.404167		1.529167	1.458333
max	63.597222	65.883333		19.025000	19.916667



# Time Spent on each step

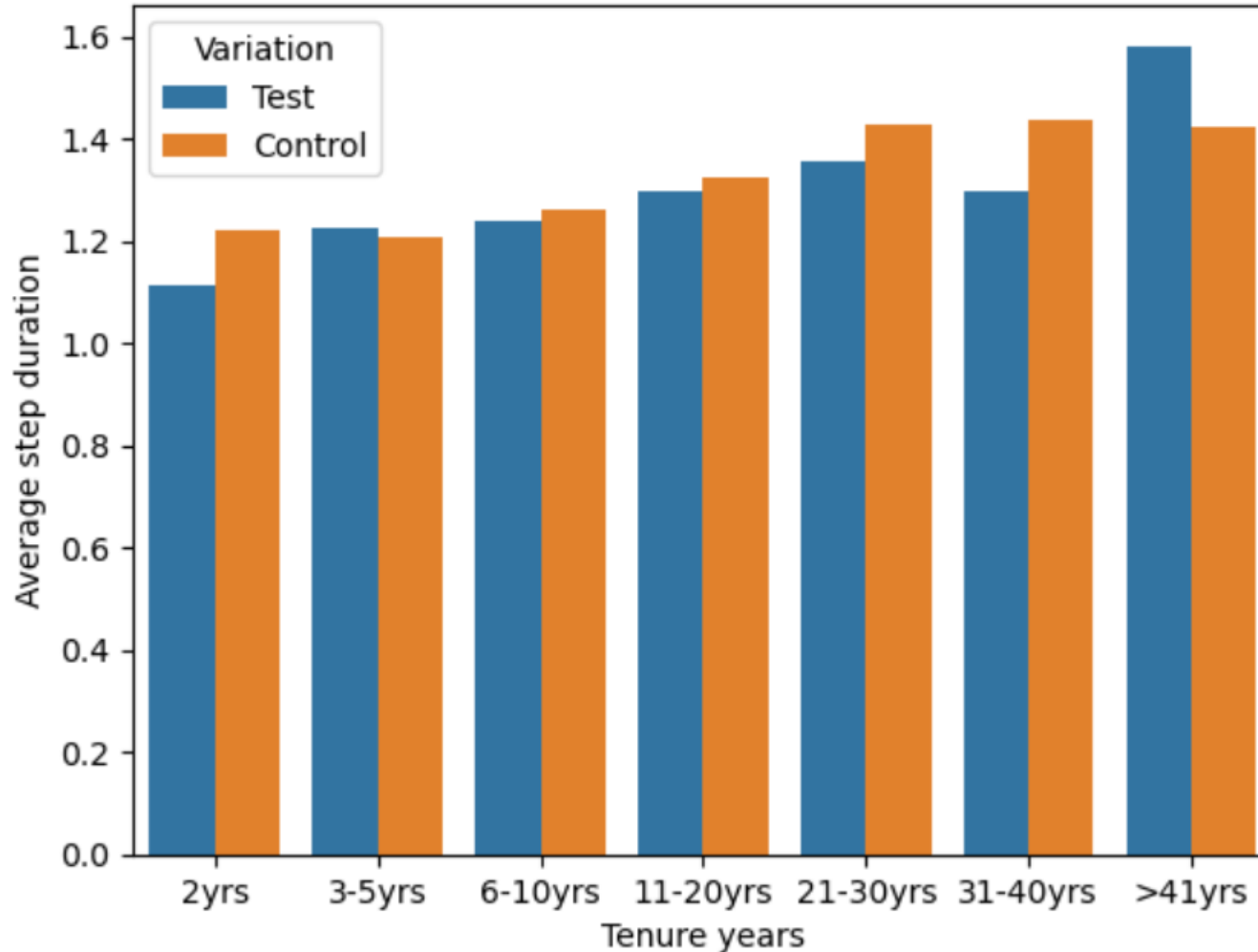
What is average step duration in different age groups?



- Average step duration increases with client's age.
- For clients under 25 years old average step duration is higher in Test group.
- For clients above 25 years average step duration is higher in Control group.

# Time Spent on each step

What is average step duration depending on client's tenure years?



- New clients with up to 2 tenure years have the smallest average step time.
- <2 tenure years Test group performs better than Control group.
- Average time increases with tenure years.

But again be careful with data !  
The sizes of age groups are not the same.

# Hypothesis Testing : is new interface better ?

**Null hypothesis:** The completion rate for the Test group (new design) is equal to the completion rate for the Control group (old design).

**Completion Rate:** Proportion z- test    significance level  $\alpha=0.05$

- P-Value :  $p = 0$
- We reject the null hypothesis (confirmation rates are the same) - and have high confidence the difference in completion rates between the groups are significantly different

**Completion Rate with a Cost-Effectiveness Threshold** (completion rate 5% better than the control ):

- P-Value :  $p = 0.055$
- We can not reject the null hypothesis (test confirmation rate is 5% higher) - not confident we will see a 5% lift after implementing the changes to UI.

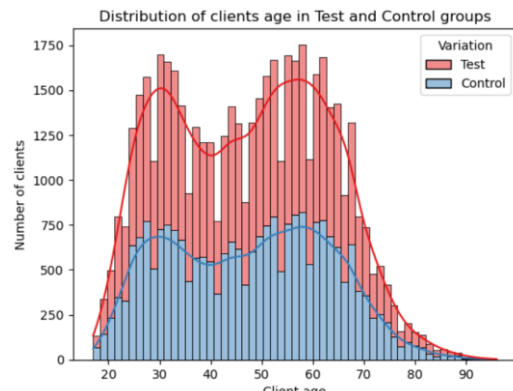
# Experiment evaluation

**Test group 53% (26,959)**

**Control group 47% (23,526 )**

Duration: from 3/15/2017 to 6/20/2017 - **4 months** (should be enough)

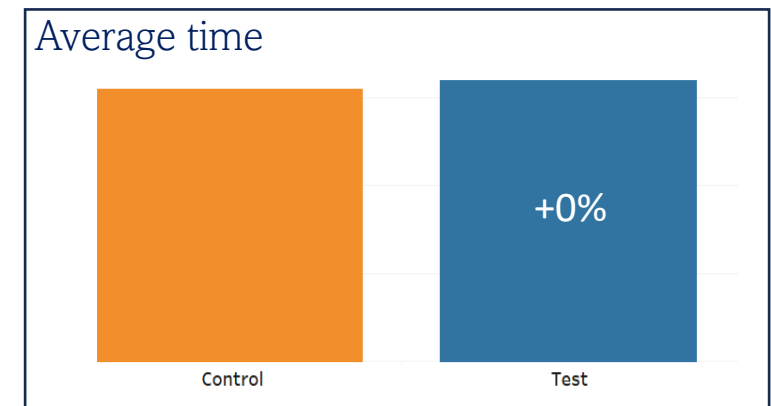
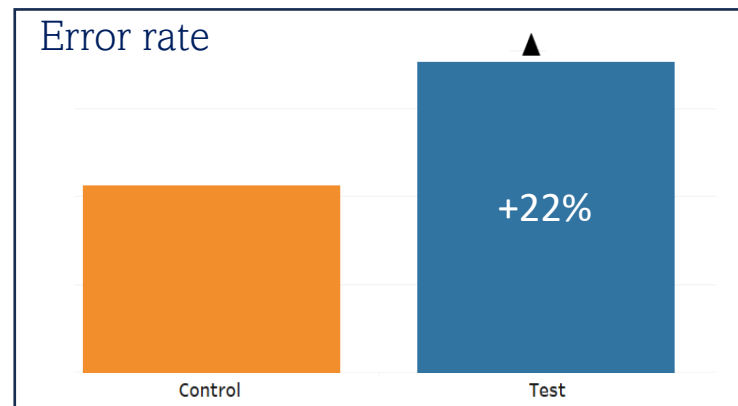
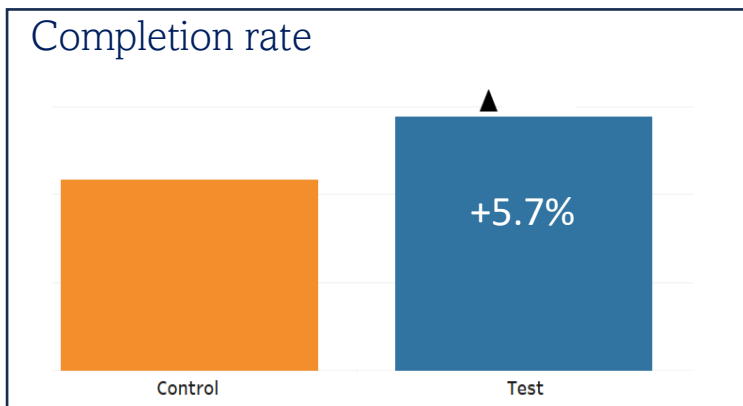
## Randomization :

- Test and Control group have **similar distributions of clients with respect to their age and tenure years**
- 
- But **number of clients in different age groups or tenure years is not equilibrated**. Young clients with short tenure years are in minority but seem to perform better in new application than the old one – worth checking !
  - Test and Control groups have almost the same number of clients but maybe this could be improved.
  - **How to define errors?** How to include clients stack on a particular step?
  - **Missing information : 20109 records** were removed because information about group was missing.



# Conclusions

- The **completion rate** for new interface increased by **5.7 %** with respect to an old design.
- But **error rate** also increased by **almost 22%** with respect to the old version.
- **Average time per step** for new and old design is **practically the same**.
- The analysis seems to indicate that new UI performs better for younger and recent tenure customers, but there is no sufficient data on these subpopulation to clearly define it. If this is an important target, additional test and/or analysis targeting them may be relevant.
- **Insufficient evidence to switch to new interface.**





THANK YOU !  
😊