MSDS-629 - 2023
Group Project

Online Control Experiments for Netflix Homepage Design

- with simulated experimental data

Group 16

Elena Liao

Kevin Kimmel

Patricia Ornelas Jauregui

Xinyi (Jessica) Wang

Table of Contents

Executive Summary

The goal of this project is to find the optimal Netflix homepage configuration to minimize the amount of time users spend browsing on average. An initial screening of the potential design factors showed that only content match score and preview length significantly influenced the response variable. Further experiments for the optimum levels led us to the optimal configuration of a 70% match score and 75-second teaser/trailer previews with the default 0.2 tile size ratio. Given this configuration, the expected value for average browsing time is 10.3486 minutes with a 95% confidence interval of [10.162, 10.535].

1. Introduction

In this project, we aim to find the optimal Netflix homepage configuration to minimize the amount of time users spend browsing on average. By analyzing data collected from each user's browsing time per session, we will be able to make inferences about the relationship between the user's homepage configuration and their browsing time.

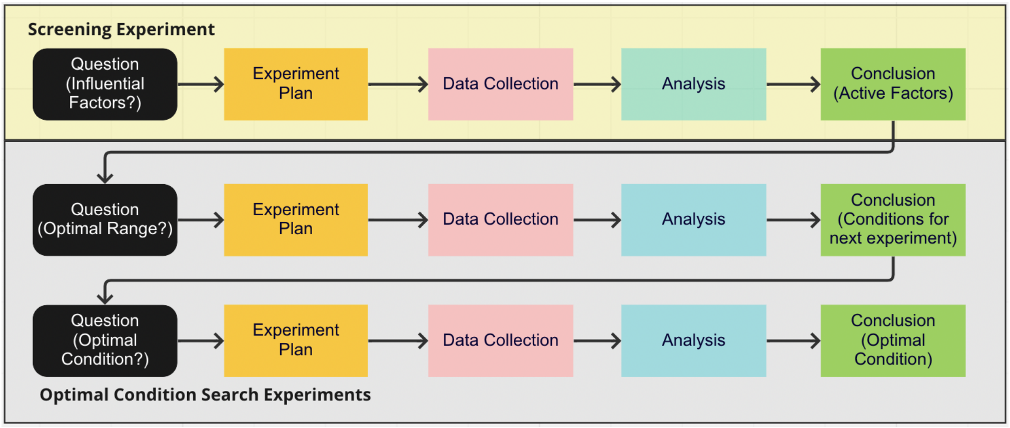
Our approach is to follow a three-phase experimental process. In phase 1, we plan to conduct an initial screening of our available factors (section 2.2) to determine the influential factors. This is a 2^4 factorial experiment where the high & low levels are far from each other. In phase 2, we move forward with only the influential design factors. We choose levels close in range to the level that gives a lower average browsing time in phase 1. In phase 3, we further narrow down our search by choosing levels for each design factor that are closer to the levels that produced shorter average browsing time. In the end, we select the condition that results in the minimum average browsing time from the conditions in phase 2 and phase 3. During this process, overall significance F-test, main and interaction effect plots, partial F- tests, and pairwise T-tests are applied to assist our analysis.

This report describes our experiments strategies and methods, and presents the data analysis results and conclusion.

2. The Experiments

As shown in the flowchart below, we integrated the Question, Plan, Data, Analysis, and Conclusion (QPDAC) approach. Without previous knowledge of the user browsing time on Netflix, we planned to perform a two-step experiment:

Step 1: Initial screening experiment to narrow down the influential design factors;
Step 2: Further experiments to analyze the effect of the influential design factors confirmed by Step 1 and search for the optimal condition for minimizing user browsing time on Netflix.


Screening Experiment
Question (Influential Factors?)
Experiment Plan
Data Collection
Analysis
Conclusion (Active Factors)
Question (Optimal Range?)
Experiment Plan
Data Collection
Analysis
Conclusion (Conditions for next experiment)
Question (Optimal Condition?)
Experiment Plan
Data Collection
Analysis
Conclusion (Optimal Condition)
Optimal Condition Search Experiments

2.1 Question

Experiment 1: the initial screening experiment primarily answers the question below:

- Among the potential factors, which are the ones influencing the response variable significantly?

Experiments 2 & 3: the optimal condition search experiments is to answer the following:

- Based on the influential factors we know, how do those factors influence the response variable? Which condition is optimal?

2.2 Plan

For all the experiments, the Metric of Interest (MOI) is the average browsing time in minutes, which is the time a user spends browsing Netflix. The Response Variable is the browsing time for a Netflix user.

For the screening experiment, all the potential design factors are considered:

- Tile Size: the ratio of a tile's height to the height of the screen
- Match score: the prediction score measures the user's enjoyment based on viewing history
- Preview length: the preview duration in seconds for a show or movie
- Preview type: the type of autoplay preview: teaser/trailer (TT) or actual content (AC)

After analyzing the screening experiment result, we decided to further explore the design factors of preview length and match score. The reason is that tile size was tested as a non-significant factor. Also, the type of teaser/trailer always delivers shorter browsing time in all circumstances, and it does not interact with other design factors significantly. Therefore, we set the preview type as TT for the ongoing experiments. Refer to the Analysis Section for more information (section 2.4).

The design factor levels considered for all experiments are presented in the table below:

Design Factor	Experiment 1		Experiment 2	Experiment 3
	Low-Value	**High-Value**		
Tile Size	0.1	0.5	0.2	0.2
Match Score	30	95	{50, 70, 90}	{65, 70, 75}
Preview Length	40	120	{75, 85, 95}	{60, 70, 80}
Preview Type	AC	TT	TT	TT

Experimental conditions were defined by combining the design factors and levels. Sixteen conditions for the screening experiment and nine conditions for each optimal search experiment were employed.

The experimental unit in this experiment is the Netflix user.

The experiment is run on a response surface simulator (Ref 1) for each experimental condition. For the sampling, the simulator randomly assigns 100 users to each condition.

2.3 Data

Based on the experiment plan, the simulator generates the observed browsing time for each experimental unit by mimicking the random assignment and observation.

2.4 Analysis

- **Overall Significance F-test**

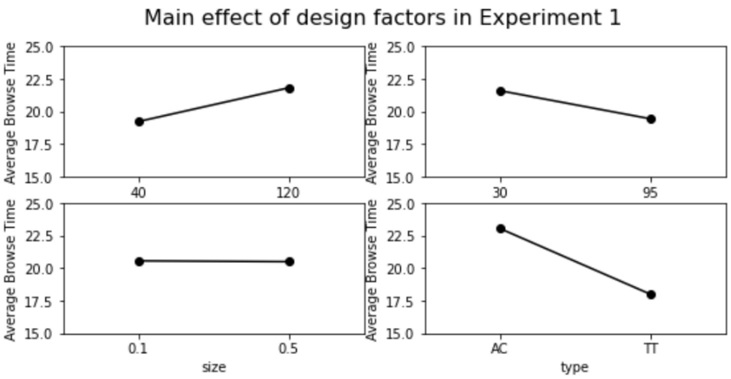
First, a ‘Gatekeeper’ F-test is performed to examine the overall significance. Linear regression models were established for these tests:

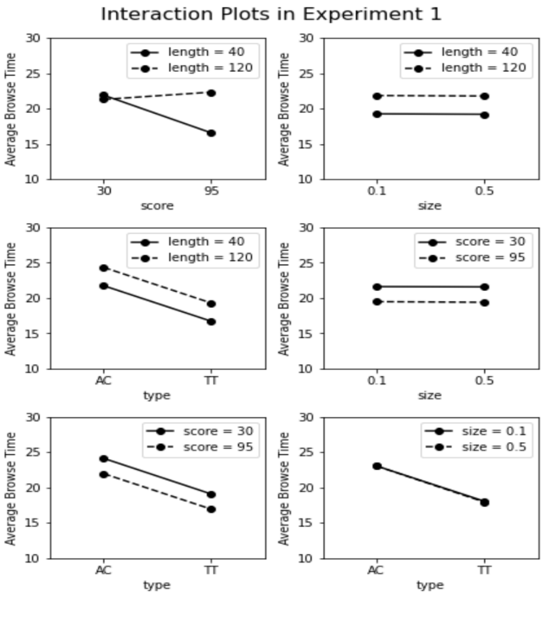
Experiment	Null Hypothesis	Alternate Hypothesis	Resulting p-value
1	$\mu_1 = \mu_2 = \dots = \mu_{16}$	$\mu_j \neq \mu_k$ $for\ some\ j \neq k$	0.00
2	$\mu_1 = \mu_2 = \dots = \mu_9$	$\mu_j \neq \mu_k$ $for\ some\ j \neq k$	7.93e-317
3	$\mu_1 = \mu_2 = \dots = \mu_9$	$\mu_j \neq \mu_k$ $for\ some\ j \neq k$	9.79e-117

In all three experiments, we reject the null hypothesis with the significance of 5%. This means that there is enough evidence to suggest that there is a difference among all conditions in each of our experiments in terms of average browsing time.

- **Main MOI Effect and interaction plots**

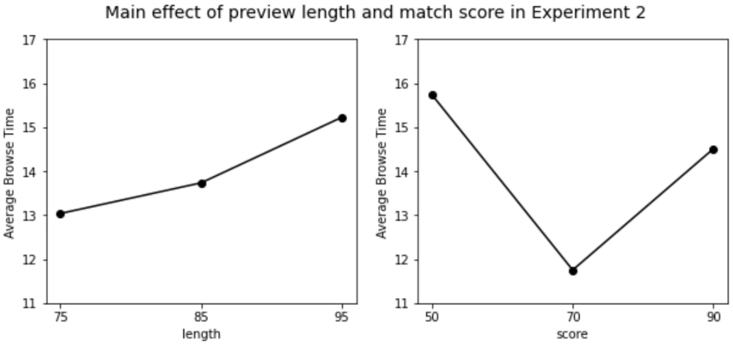
Experiment 1

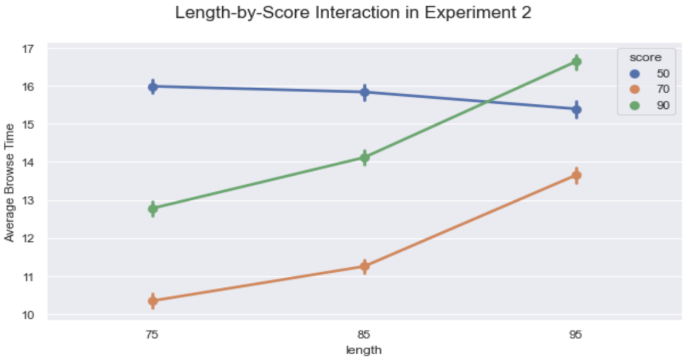




Based on the main effect and interaction plots above, the preview length, match score, and preview type all have significant main effects on the browsing time. The interaction plots indicate that only the preview length and match score have significant interaction effects with each other. A longer preview length increases the browse time when the match score moves from 30 to 95, while a shorter preview length decreases the browse time.

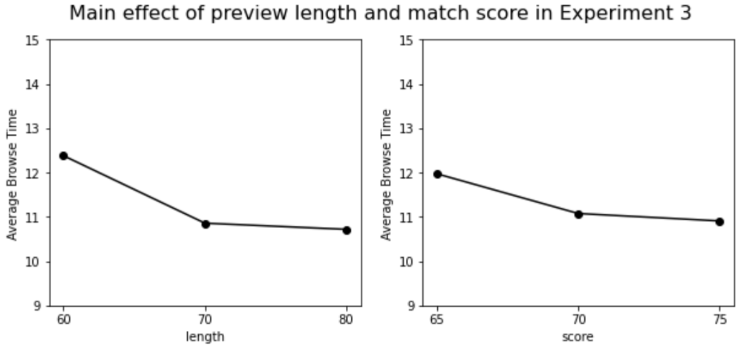
Experiment 2

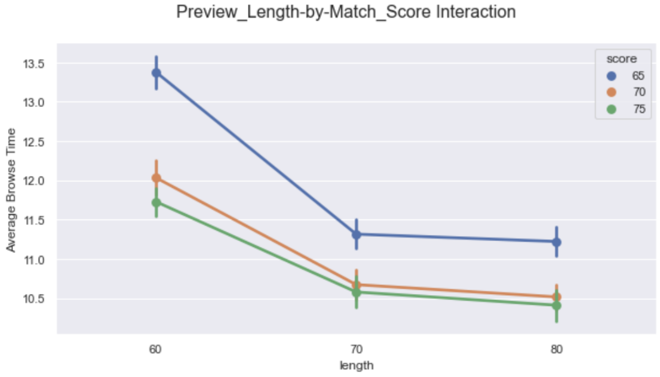




In the second experiment, we saw that the main effects of both factors appeared to be significant. The interaction plot further confirms that preview length and match score have significant interaction effects. Interestingly, we found out that a match score of 70 generally has the lowest average browsing time rather than a higher or lower match score.

Experiment 3





Based on the observation of the interaction plot, a match score of 75 generally has the lowest average browsing time.

- **Hypothesis Testing**

Hypothesis Tests were run to further confirm these observations. The test results are presented in the table below.

From these tests, we can conclude that the only significant interaction is between Match Score and Preview Length, and the only significant main effects are Match Score, Preview Length, and Preview Type. Using this information we can narrow our search in future experiments.

Test	Test Statistic	P-Value	Conclusion
4-Way Interactions	0.0899	0.764	Not significant
3-Way Interactions	2.367	0.0509	Not significant
2-Way Interactions	661.971	0.0	Significant
Match_Score:Tile_Size	0.3417	0.5589	Not Significant
Match_Score:Prev_Type	0.0368	0.8478	Not Significant
Tile_Size:Prev_Type	1.3088	0.2528	Not Significant
Match_Score:Prev_Length	3970.119	0.0	Significant
Tile_Size:Prev_Length	0.0099	0.9208	Not Significant
Prev_Type:Prev_Length	0.0118	0.9134	Not Significant
Tile_Size	0.43167	0.5112	Not Significant
Match_Score	522.112	3.345e-100	Significant
Prev_Length	752.663	4.64e-136	Significant
Prev_Type	2855.181	0.0	Significant

- **Pairwise Tests**

The following top 6 conditions with shortest average browsing time were considered for pairwise test:

- Condition 1: length = 75, score = 70, average browse time = 10.35
- Condition 2: length = 80, score = 75, average browse time = 10.41
- Condition 3: length = 80, score = 70, average browse time = 10.51
- Condition 4: length = 70, score = 75, average browse time = 10.58
- Condition 5: length = 70, score = 70, average browse time = 10.67
- Condition 6: length = 80, score = 65, average browse time = 11.22

The browsing time for above conditions have roughly normal distribution and equal variances (by F-test). Therefore, student T-tests were applied, and the corresponding results are shown below.

Test	H_0	H_A	p-value	Adjusted p-value (Holm)	Rejection
1	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	0.639	0.639	False
2	$\mu_1 = \mu_3$	$\mu_1 \neq \mu_3$	0.183	0.366	False
3	$\mu_1 = \mu_4$	$\mu_1 \neq \mu_4$	0.091	0.274	False
4	$\mu_1 = \mu_5$	$\mu_1 \neq \mu_5$	0.0156	0.062	False
5	$\mu_1 <= \mu_6$	$\mu_1 > 6$	3.4e-10	1.7e-9	True

The rejection results above indicate that the top 5 conditions are similar but significantly better than condition 6 in terms of minimizing the browsing time.

2.5 Conclusions

From the screening experiment, we discovered that the influential design factors are Match Score, Preview Length, and Preview Type. The only significant interaction is between Match Score and Preview Length.

For the optimal searching experiments, we focused on Match Score and Preview Length and confirmed our previous findings. The results indicate that a combination of 70 to 80 seconds for the Preview Length and a 70 to 75 Match Score delivers an optimal browsing time that is significantly better than other conditions. We selected the optimal condition of a 70% match score and 75-second teaser/trailer previews with the default 0.2 tile size ratio. Given this configuration, the expected value for average browsing time is 10.3486 minutes with a 95% confidence interval of [10.162, 10.535].

3. Conclusion

3.1 Results

In conclusion, our expected value for average browsing time with the optimal configuration of 70% match score, 75-second teaser/trailer previews, and 0.2 tile size ratio is 10.3486 minutes with 95% confidence interval [10.162, 10.535].

3.2 Experiment Limitations

The following aspects are not considered in this project:

Instead of selecting an optimal condition from the experimented ones, the response surface method can be an approach to finding the potential optimal point.

For efficiency purposes, we did not use a grid search approach to find the optimum design factor levels. This alternate approach would have been more computationally expensive, since we would have to collect data for more unique combinations of our design factors. Therefore, we may potentially miss a globally optimal solution.

Central Composite Design (CCD) can be applied to this project to find better candidate design factor levels in the future.

REFERENCES

1. https://nathaniel-t-stevens.shinyapps.io/Netflix_Simulator_v3/