Date: 12/04/2022
DSO 510: Professor Mohammed Alyakoob
Group members: Akanksha Arun, Christian Ingul, Juhil Ahir, Lavanya Desmukh, Paul Nguyen and Sagar Patel

Clicked on Ad Rate Analysis

Overview

We seek to understand how various factors, such as time spent on ads, the ad content, and the demographics of users, affect click-through rates. This can help businesses optimize their ad strategy to maximize clicks on ad rates/conversions while also minimizing expenditures.

Problem Statement

For most businesses, it is crucial to have the right marketing strategy in place. Strategies can come in many forms; some use a data-driven approach, and others rely on HIPPOs or a combination of both. Traditionally, successful marketing efforts generate higher revenue and, as a result, can help businesses grow and gain market share. However, marketing can also be a cost center as marketing budgets are often one of the most significant expenses businesses incur. Therefore, businesses should be attentive to where their marketing budgets are directed and what strategies to select.

Why is a data-driven approach important?

Using a data-driven approach which leverages historical data such as past performance while also performing A/B testing, a business can accurately allocate budget, highlight high-performing ads, reduce spending on ads with low click-through rates and narrow down an optimized ad strategy. The key is for a business to understand what variables contribute to higher click-through rates, and plan accordingly, which could lead to higher revenues and lower expenses and ultimately increase the bottom line.

Hypothesis Setting

H_0: Time spent on the website does not lead to better clicked on ad rates.
Ha: Time spent on the website does lead to better clicked on ad rates.

The Ideal Experiment

An ideal experiment for this scenario would be to perform A/B testing directed toward website marketing and having visitors randomly distributed into groups A and B. Group A would act as our control while group B would be our treatment variable for this experiment. Group A would represent clicked on ad rates, and we are looking to see how it changes by experimenting with the following B categories:

1. Impact of buzzwords: Analyze if any particular words contribute to higher clicked-on ad rates

2. Impact of ad content: Analyze how the average time spent on ads differs depending on the ad's content
3. Controlling for time spent on the website: Analyzing how two groups' time spent on the website affects the clicked-on ad rate

The end goal of the ideal experiment is to determine if there are optimal words, ad contents, or time users need to spend on the website to increase the clicked on ad rate. Ideally, the experiment should be carried out in phases, not testing all B's simultaneously, mitigating contamination and spillover effects. For this project, we are focused on analyzing the time spent on the website.

Data

We have utilized two Kaggle datasets and combined them into one master dataset:

Source 1: Link
Source 2: Link

Master dataset column and row's overview:
Column 1: "Daily Time Spent on Website": The amount of time spent on site (minutes).
Column 2: "Daily Internet Usage": The amount of time spent on the internet (minutes).
Column 3: "Ad Topic Line" : The topic line ad on the website.
Column 4: "Time Stamp": The time that the person entered the website.
Column 5: "Clicked on Ad": A binary value for whether or not the person clicked on the ad.
Column 6: "Gender": User gender.
Column 7: "Age": User age.
Column 8: "City": User city.
Column 9: "Country": User country.
Column 10: "Area Income": User average income by area.
Column 11: "Country Code": User country code.
Column 12: "Continent": User continent.
Rows: 50,000.

Data Cleaning and Manipulation

Actions taken:

1. We combined the two Kaggle datasets to add the column "Continents," which is based on the existing "Country" column, as we wanted to observe continent-specific data and categorize accordingly.
2. We created dummy variables for the categorical variables gender, continents, and most popular ad topic lines.
3. Columns "Age" and "Time of the Day" were divided into buckets allowing for increased flexibility with the model and making the data easier to interpret.
4. Further, we added a column for ad length based on the "Ad-topic line" column accounting for the number of words to control for the effects of shorter or longer topic lines excluding hyphens.

Visualizations

Clicked on ad rate per country heat map (Appendix A, slide 11):

The heat map shows the countries with the highest clicked on ad rate. Furthermore, these countries are filtered by the "clicked on ad" sum being greater than 100. By adding this filter, we exclude countries with high clicked-on ad rates and low "click on ad" from the heat map compared to the other countries. Therefore, in the heat map, the darker the color, the higher the clicked on ad rate for that country. If segregated by continent, the countries that are the most active or have higher clicked on ad rates are as follows. In North America, the country with the highest rate is the United States. For South America, we have Chile and French Guiana. In Africa, we have Burkina Faso and Zimbabwe. For Europe, we have Lithuania and Portugal. Lastly, in Asia, we have the United Arab Emirates.

Clicked on ad demographics (Appendix B, slide 12):
Furthermore, we wanted to take it further by analyzing what an average clicker looks like for those countries. As seen in Appendix B, we have included a summarized table group by gender. In this table, we can observe that our average clickers are in their thirties, with average area income being around fifty-five thousand and average daily time spent on site being around one hour. These averages are consistent across both genders. In other words, people who click on an ad have relatively similar characteristics.

Clicked on ad time of the day (Appendix C, slide 13):
Our Appendix C is a compact table representing the different clicked on ad rates made at different times of the day. As seen in the table, the highest clicked on ad rate happens to be during the evening time (5 pm to 9 pm), which implies that users have more available time during the evening, once they are back from their work, to navigate through the website and click on ads.

Time spent on the website before clicking on an ad (Appendix D, slide 14):
Appendix D showcases the relationship between time spent on a website and clicking on the ad. More time spent on the website will lead to a higher clicked on ad rate. However, that is not the case. As shown in the visualization, the clicked on rate peaks between 51 to 70 minutes, where the rate is 53%, and drops to 46% after 70 minutes. This tells us that more time spent on the website does not correlate with a high click on ad rate. This information can be helpful in the marketing department so they can strategize their approach and make the most out of it.

Regression model 1 outputs (Appendix E, slide 15):
We want to draw attention to the negative coefficient for time spent on a website. The coefficient tells us that the longer a person spends on the website, the less likely they are to click on the ad; in fact, the probability of clicking on the ad decreases by -0.0012. However, because it is a linear model, it fails to capture non-linearity, which is the increase from 50 minutes to 70 minutes, and the downfall in clicked on rate from 70 minutes onward. The coefficient is negative because the decrease in the probability of clicks is steeper than the increase in the probability of clicks.

When comparing time spent on site and the outcome variable, an R-Squared of 0.256 is okay. Variation in click rate is hard to predict; many things outside our model affect whether they click. We care about daily time spent on site and controlling for things that might be correlated with time spent on site and the outcome variable.

Clicked on ad per month variation vs time spent on the website variation (Appendix F, slide 17):
When looking at the average time spent on websites monthly, we observed that the time is relatively constant. We believed this was interesting and decided to add a "click on rate" line to see if the rate would differ from the average time spent. Instead, we noticed that the clicked on rate per month varied from 31% to over 75%, which led us to believe another factor affected the click-on ad rate.

The most popular ad titles and the clicked on rate (Appendix G, slide 18).
Here we can see the top 6 ad titles based on the sum of clicked on an ad and the clicked on ad rates. This is interesting because it makes up 26% of ad clicks. This is major considering that we have 676 unique titles; 6 ads make up 26% of all clicked on ads.

Regression model 2 outputs (Appendix H):
We decided to run another regression with the 6 most popular ad topic lines included. Here we can see that the impact of the "Most popular topics" is 0.2312, which means that these ad topics increase the probability of clicking on the ad.

Limitations:

1. Non-linearity is not captured by linear models, which is reflected in the negative time spent on the website coefficient.
2. We have a discrete variable as our response variable, which fits poorly in a linear model; a logistic regression or a more flexible model might have yielded a better prediction.
3. There are a large number of factors that can affect the user's decision to click on an ad which makes it harder to predict click-through rate with the data we have currently. This leads to a smaller R-squared value for our model.

Conclusion and takeaway:
From our above findings, we can statistically reject our null hypothesis (Time spent on a website does NOT lead to better clicked on ad rates). However, we believe there is a stronger correlation between click on ad rates and the ad topic line. For example, with a coefficient of 0.2312, our 6 most popular titles of 676 unique titles make up over 26% of the total clicks. Variations in clicked on ad rates are hard to predict, and many factors outside the model might affect a user's decision to click on the ad. Other factors that lead to a smaller R-Squared would include but are not limited to: content, graphics, sound effects, website space, and frequency.

Explanation of time spent on the website's effect on clicked on ad rates:
After completing the presentation, Professor Alyakoob asked a question regarding the time spent on a website's effect on clicked on ad rates. The group agrees that intuitively it seems peculiar that the sweet spot for clicked on ad rates is between 51-70 minutes. This, as most clicked on ad rates are maximized by catching the website visitor's attention early. However, in our case, given the nature of the website ad topic line and content being specific to the software, we find the results to be more logical under the following assumptions:

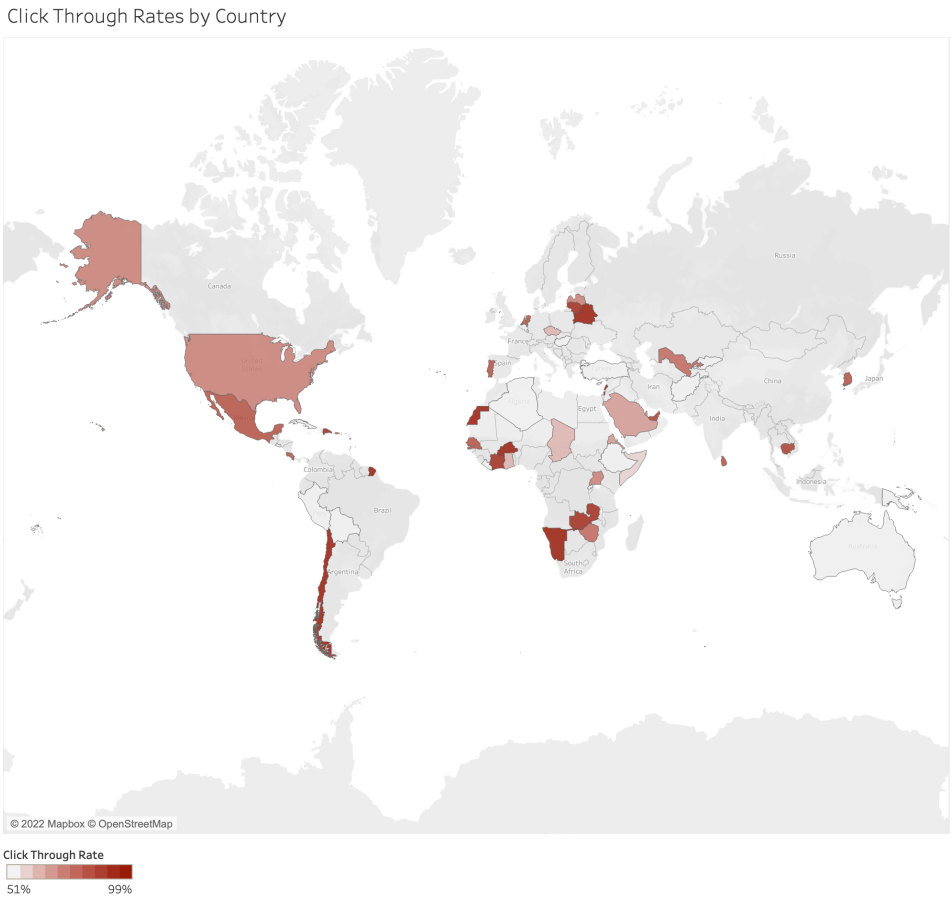
1. Consumer preferences and characteristics stay consistent over time.
2. The nature of the website content is more sophisticated than other types of ads and would require more careful reading increasing the time spent on the website.
3. While a visitor is browsing the website, its administrators analyze the content viewed and require up to 50-70 minutes to recommend an ad. Further, this would suggest that if the visitor's attention is not caught within 50-70 minutes, the clicked on ad rate should dwindle when exceeding the timeframe, which our visualizations and data showcase.

Next steps:
To improve the R-squared value, we would require to gather supplemental data on the factors outside the model. This would include but not be limited to the above-mentioned variables: content, graphics, sound effects, website space, and frequency.

Appendices

- Appendix A - Clicked on ad rate per country heat map
- Appendix B - Clicked on ad demographics
- Appendix C - Clicked on ad time of the day
- Appendix D - Time spent on the website before clicking on an ad
- Appendix E - Regression model 1 outputs
- Appendix F - Clicked on ad per month variation vs time spent on the website variation
- Appendix G - The most popular ad titles and the clicked on rate
- Appendix H - Regression model 2 outputs

Appendix A: Clicked on ad rate per country heat map.



Appendix B: Clicked on ad demographics.

Clicker Demographics

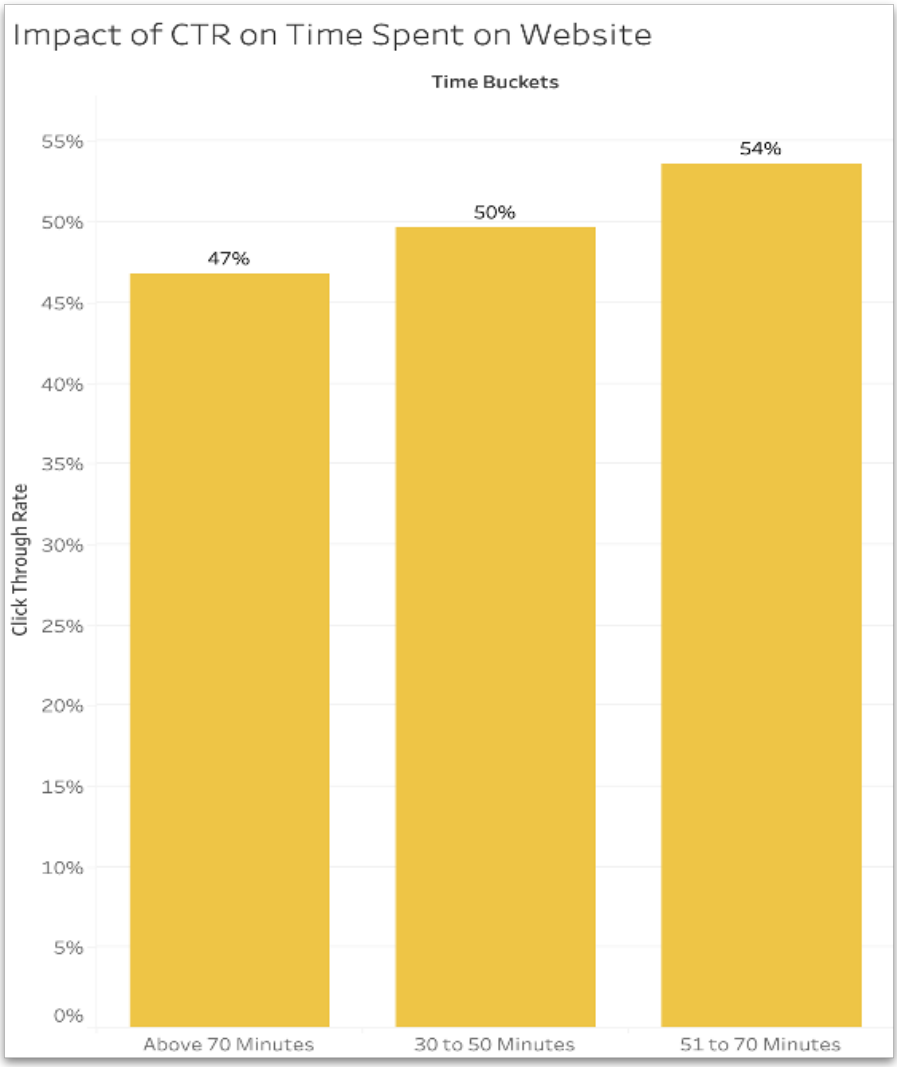
Gender	Continent	Avg. Age	Avg. Area Income	Avg. Daily Time Spent on Site
Female	Africa	38	$51,916.96	64
	Americas	37	$54,778.64	64
	Asia	37	$54,297.27	63
	Australia	33	$54,506.44	65
	Europe	38	$54,133.92	64
Male	Africa	38	$53,222.71	63
	Americas	36	$53,642.27	64
	Asia	38	$53,113.35	63
	Australia	31	$53,859.93	65
	Europe	39	$55,673.78	62

Appendix C: Clicked on ad time of the day.

Time of Day Analysis

Time of Day..	Click Through Rate	Total Ads
Morning	46%	15,145
Afternoon	38%	10,105
Evening	61%	12,280
Night	53%	12,470

Appendix D: Time spent on the website before clicking on an ad.



Appendix E: Regression model 1 outputs.

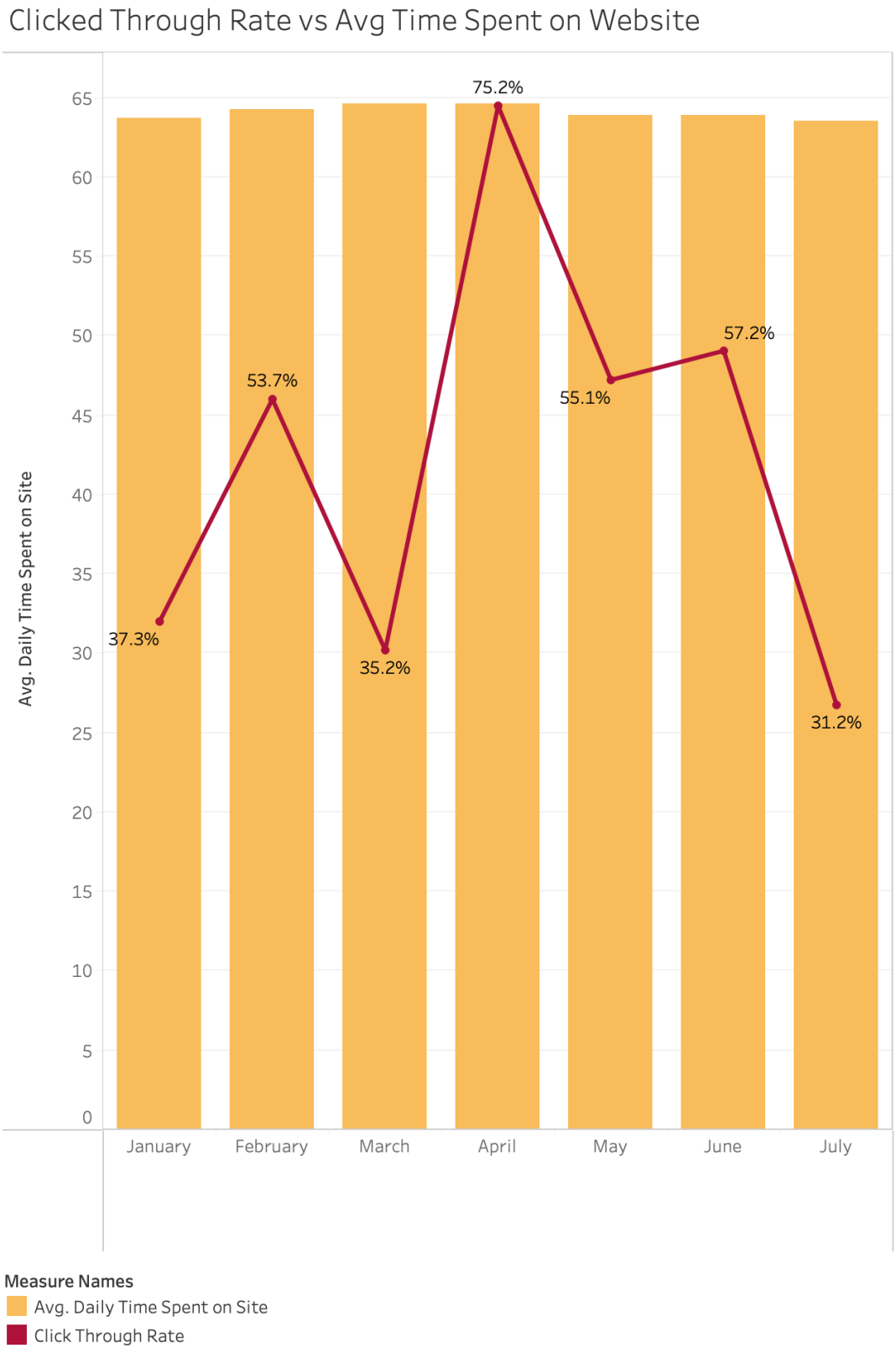
OLS Regression Results

Dep. Variable:	Clicked on Ad New	R-squared:	0.256
Model:	OLS	Adj. R-squared:	0.255
Method:	Least Squares	F-statistic:	1430.
Date:	Sun, 04 Dec 2022	Prob (F-statistic):	0.00
Time:	20:48:51	Log-Likelihood:	-28911.
No. Observations:	50000	AIC:	5.785e+04
Df Residuals:	49987	BIC:	5.796e+04
Df Model:	12		
Covariance Type:	nonrobust		

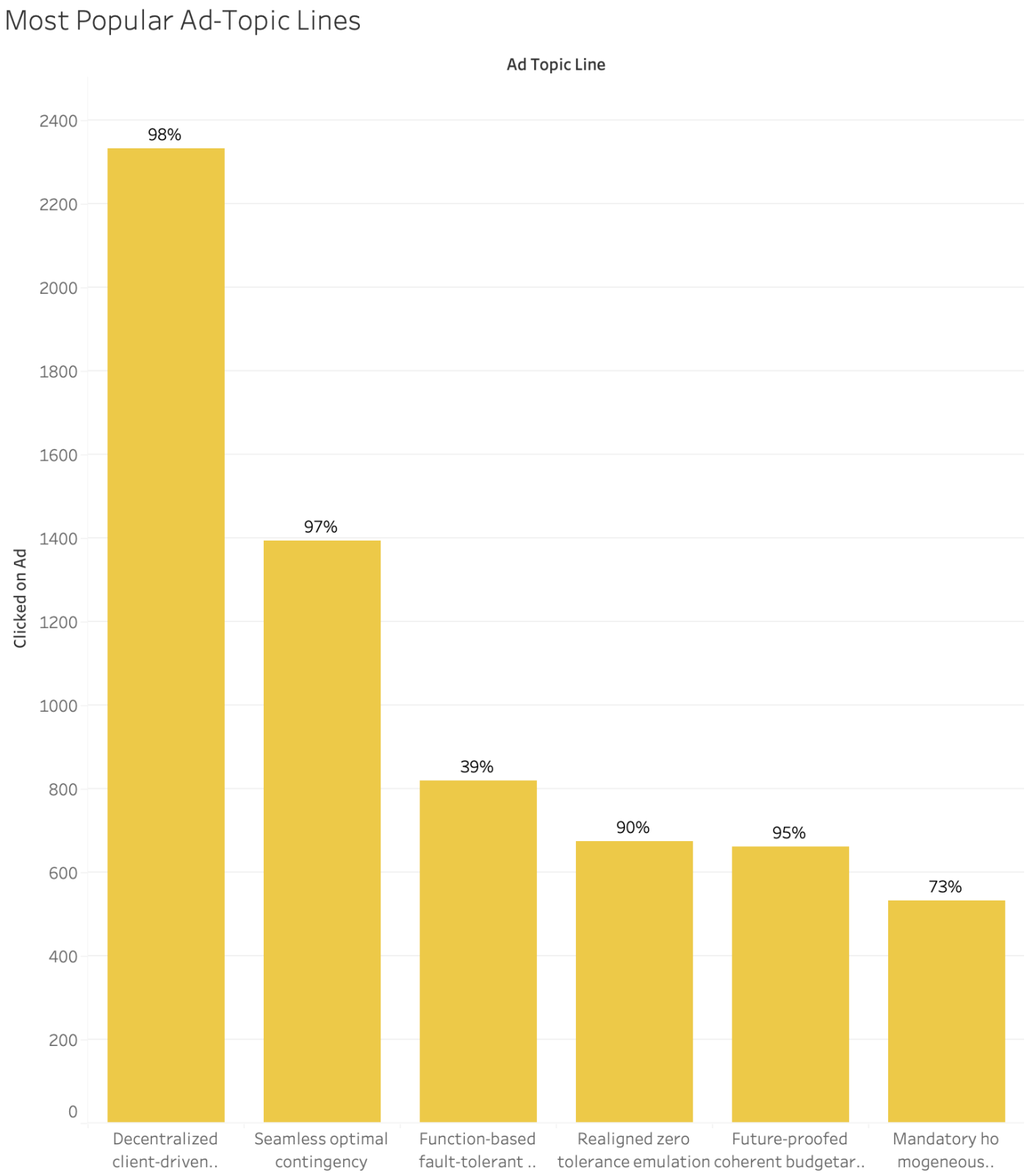
	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.1339	0.021	-6.296	0.000	-0.176	-0.092
Gender_dummies	0.0036	0.004	0.927	0.354	-0.004	0.011
Daily Time Spent on Site	-0.0012	0.000	-9.460	0.000	-0.001	-0.001
Age	0.0245	0.000	109.723	0.000	0.024	0.025
Area Income	-5.258e-07	1.7e-07	-3.097	0.002	-8.59e-07	-1.93e-07
Daily Internet Usage	-0.0016	4.69e-05	-33.109	0.000	-0.002	-0.001
Ad length	0.0067	0.002	2.894	0.004	0.002	0.011
Africa	0.1010	0.012	8.160	0.000	0.077	0.125
Americas	0.0036	0.012	0.287	0.774	-0.021	0.028
Antartica	-0.2914	0.069	-4.238	0.000	-0.426	-0.157
Asia	0.0560	0.013	4.471	0.000	0.031	0.081
Australia	-0.0552	0.013	-4.250	0.000	-0.081	-0.030
Europe	0.0522	0.013	4.165	0.000	0.028	0.077
Quarter Year	0.0382	0.003	12.522	0.000	0.032	0.044

Omnibus:	9157.627	Durbin-Watson:	1.184
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2044.751
Skew:	0.134	Prob(JB):	0.00
Kurtosis:	2.046	Cond. No.	2.35e+20

Appendix F: Clicked on ad per month variation vs time spent on the website variation



Appendix G: The most popular ad titles and the clicked on rate



Appendix H: Regression model 2 outputs.

OLS Regression Results

Dep. Variable:	Clicked on Ad New	R-squared:	0.282
Model:	OLS	Adj. R-squared:	0.282
Method:	Least Squares	F-statistic:	1785.
Date:	Sun, 04 Dec 2022	Prob (F-statistic):	0.00
Time:	20:48:51	Log-Likelihood:	-28007.
No. Observations:	50000	AIC:	5.604e+04
Df Residuals:	49988	BIC:	5.614e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0917	0.019	-4.777	0.000	-0.129	-0.054
most_popular_titles	0.2315	0.005	43.017	0.000	0.221	0.242
Daily Time Spent on Site	-0.0014	0.000	-10.643	0.000	-0.002	-0.001
Age	0.0225	0.000	99.974	0.000	0.022	0.023
Area Income	5.896e-07	1.69e-07	3.494	0.000	2.59e-07	9.2e-07
Daily Internet Usage	-0.0017	4.61e-05	-36.449	0.000	-0.002	-0.002
Africa	0.0986	0.012	8.149	0.000	0.075	0.122
Americas	0.0132	0.012	1.097	0.273	-0.010	0.037
Antartica	-0.2532	0.067	-3.751	0.000	-0.385	-0.121
Asia	0.0477	0.012	3.909	0.000	0.024	0.072
Australia	-0.0455	0.013	-3.586	0.000	-0.070	-0.021
Europe	0.0475	0.012	3.877	0.000	0.023	0.071
Quarter Year	0.0344	0.003	11.498	0.000	0.029	0.040

Omnibus:	10354.361	Durbin-Watson:	1.228
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2283.427
Skew:	0.191	Prob(JB):	0.00
Kurtosis:	2.025	Cond. No.	3.05e+20