STAT – S670: Exploratory Data Analysis

Final Project: Advertisement Clicking

Authors: Saranjeet Singh Saluja, Sricharraan Ramaswamy, Athulya Anand

1. Introduction:

1.1. Need of the Project:

- In today's era, just as the internet has become a part of our everyday life, advertisers have also been using this platform to market and sell their products.
- Marketing Analysis plays a significant part in Advertising and involves a component of marketing that identifies the customer needs and determines the best to meet those needs.
- Click-through rate is a significant component in advertisement marketing that advertisers focus on.
- CTR allows you to better understand your customers by revealing what works and what doesn't while trying to reach your target demographic.
- A low CTR could indicate that you're targeting the wrong audience or that you're not using persuasive language to persuade them to click.

Through this project, we predominantly focus on bringing out a relationship between the daily time spent on site and the probability of a consumer clicking on ads while using the internet daily. Likewise, we focus on the following problem statements mentioned below. The graphs we analyze in this project function in the following way:

- How does internet usage or daily time spent on advertisements help understand if the consumer will click on the ad?
- What is the relationship between the Daily time spent on site and the Daily usage of the Internet?
- Identify the parameters like age, gender, etc. that affects the probability of clicking on ads.

1.2 Data Description

In this project, we have used the **Advertising.csv** dataset for analysis. This dataset consists of data on 1000 users indicating whether a particular internet consumer clicked on an Advertisement on a company website. Some of the important variables in the dataset are as follows:

- **Daily Time Spent on Site**: consumer time on-site in minutes
- **Age**: Customer age in years
- **Area Income**: Avg. Income of geographical area of consumer
- **Daily Internet Usage**: Avg. minutes a day consumer is on the internet
- **Ad Topic Line**: Headline of the advertisement
- **City**: City of consumer
- **Male**: 0 or 1 indicated whether the consumer was Male or not (0 – Female, 1 – Male)
- **Country**: Country of consumer
- **Timestamp**: Time at which consumer clicked on Ad or closed window
- **Clicked on Ad**: 0 or 1 indicated clicking on Ad (0 – NO, 1 – YES)

In this exploratory data analysis project, we have our target variable as *Clicked on Ad*. Other important variables considered include *Data Time Spent on Site, Daily Internet Usage, Age, and Male.*

2. Relationship between target variable Clicked on Ad and Daily Time Spent on Site or Daily Internet Usage:

2.1. Likelihood of Clicking on Ad:

As per Fig. 1, we can see the distribution of daily time spent on site for clicking on the ad i.e.: clicking on the ad is represented in red and not clicking on the ad is represented in blue. We can clearly see that the distribution of daily time spent on site for clicking on the ad are very different from the distribution for not clicking on the ad which is left-skewed.

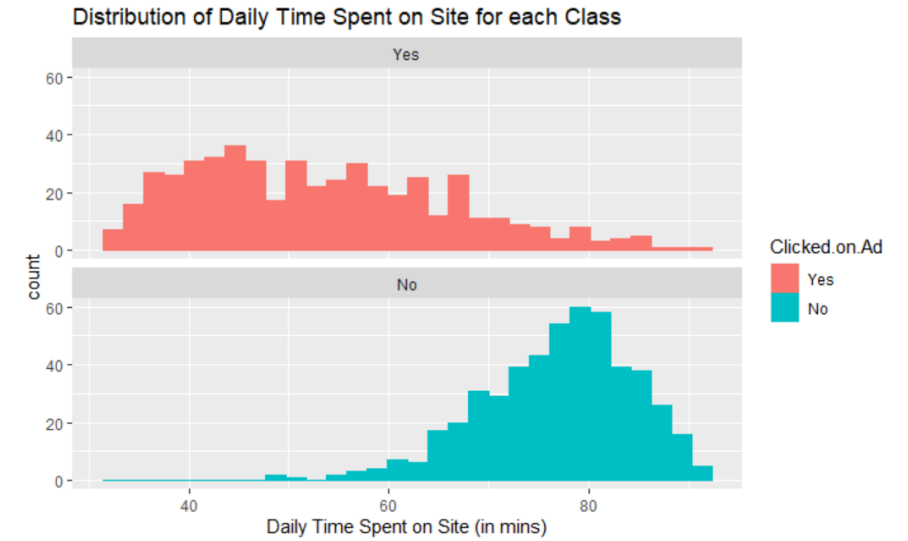

Fig 1. Distribution of Daily Time Spent on Site for whether the consumer clicks on an Ad

As per Fig. 2, we can see the distribution of daily internet usage for clicking on the ad i.e.: clicking on the ad is represented in red and not clicking on the ad is represented in blue. We can clearly see that the distribution of daily internet usage for clicking on the ad is right-skewed and are very different from the distribution for not clicking on the ad.

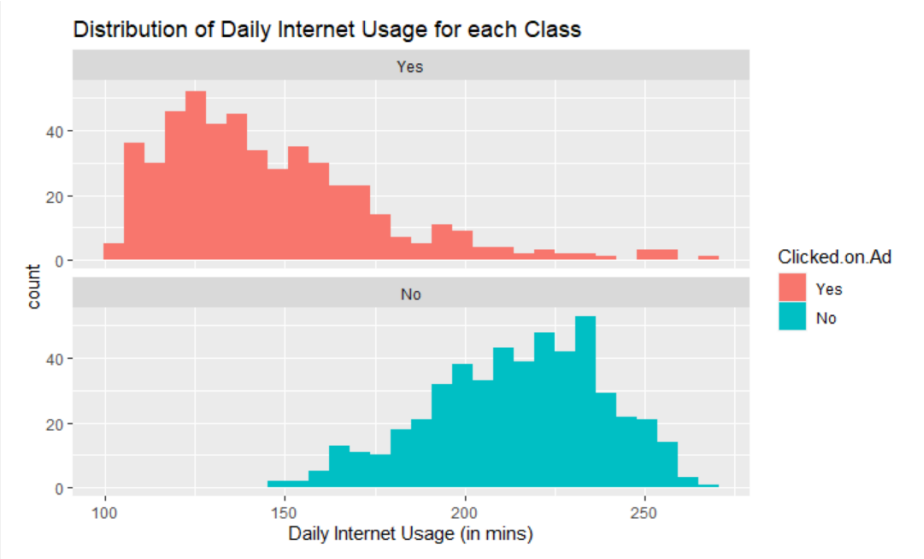

Fig 2. Distribution of Daily Internet Usage for whether the consumer clicks on an Ad

2.2. Daily time spent on site Vs Daily Internet Usage:

Fig. 3 is a scatterplot plotted mainly to depict and understand the relationship between Daily time spent on the site vs the daily internet usage.

- As you can see the users who click on ads are represented by red data points and those who do not are depicted by the blue data points.
- This scatter plot indicates a clear distinction between those who click on ads and those who do not.
- It can be inferred that users who spend a relatively low time on sites daily are more likely to click on ads as opposed to those who spend lesser time on the internet and are less likely to click on ads.
- We can also witness a few outliers on this plot when the daily time spent on site is around 60mins.

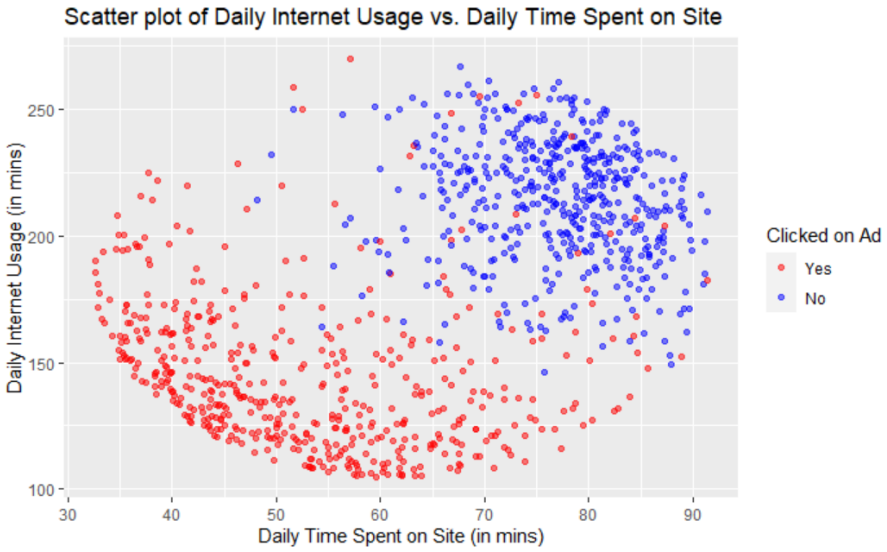

Fig 3. Scatter Plot: Daily time spent on site Vs Daily internet usage

- Before plotting this scatter plot, so we expected the red data points to be in place of the blue data points and vice versa.
- Thus, from Fig 3., we hypothesize that consumers like young teens mostly know exactly what they want to browse or which ad they want to click on by spending less time on sites. On the contrary, the consumers of the middle or higher age group may or may not know if the ad is of their interest, thereby spending more time on sites.
- Further analysis has been done in later parts regarding the same.

Furthermore, we plot a density plot to confirm the same, as they provide a better distinction when comparing two distributions.

Fig 4., shows a clear distinction between those who click on ads and those who do not. We can observe that people who spend more time on sites and who use the internet for a longer duration do not tend to click on ads.

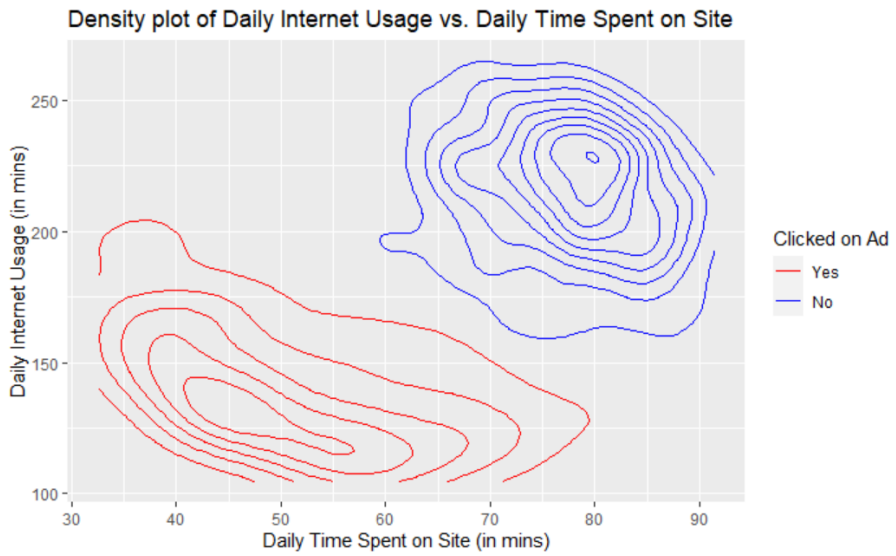

Fig 4. Density Plot: Daily Time Spent on Site Vs Daily Internet Usage

2.3. Modelling to predict the probability of clicking an ad:

2.3.1. Logistic Regression:

Firstly, we aim to build on simple models and then move on to complex models. In Fig. 5., we start predicting the probability of clicking on ads given the daily time spent on site. The probability of clicking on Ads decreases as the Daily time spent on the site increases.

It is difficult to create a viable model based on a single variable, so we will investigate additional variables to include in our model.

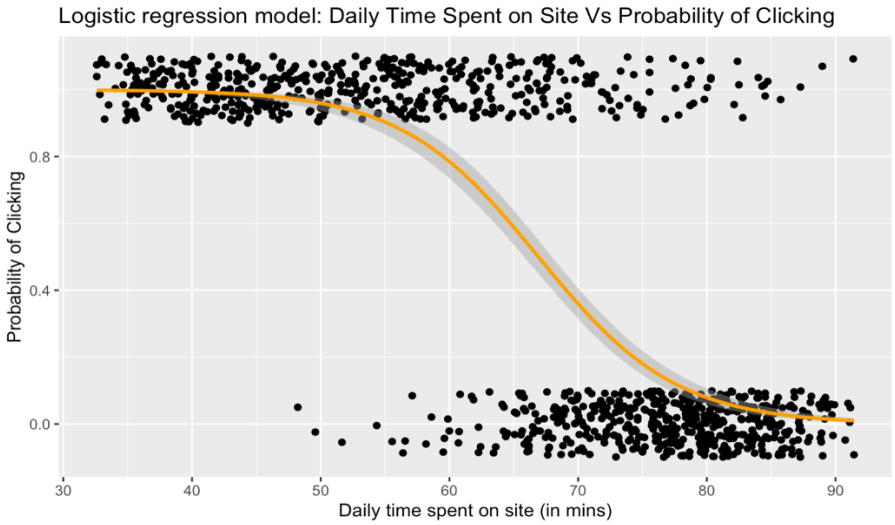

Fig 5. Logistic Regression Model

2.3.2. Checking Daily Time Spent on Site per quantile:

To incorporate Daily Time Spent on Site within our model, we take the interaction between Daily Time Spent on Site and Daily Internet Usage. To further understand the interaction between the variables, in Fig. 6, we aim to check the change in the probability of clicking on ads, given each interval of Daily Internet Usage. The four quantiles (in mins) are represented by:

- 1 - [105,139],
- 2 - (139,183],
- 3 - (183,219] &
- 4 - (219,270],

represented in red, green, blue, and purple respectively.

- *Quantile 1*- The probability of clicking on the ad is 1 irrespective of the Daily Time Spent on Site.
- *Quantile 2*- The probability of clicking on the ad is close to 1 till Daily Time on Site is up to 50 mins and drops as the Daily Time on Site increases.
- *Quantiles 3 & 4* – The probability of clicking on the ad converges to 0 for these quantiles as the Daily Time Spent on Site increases.

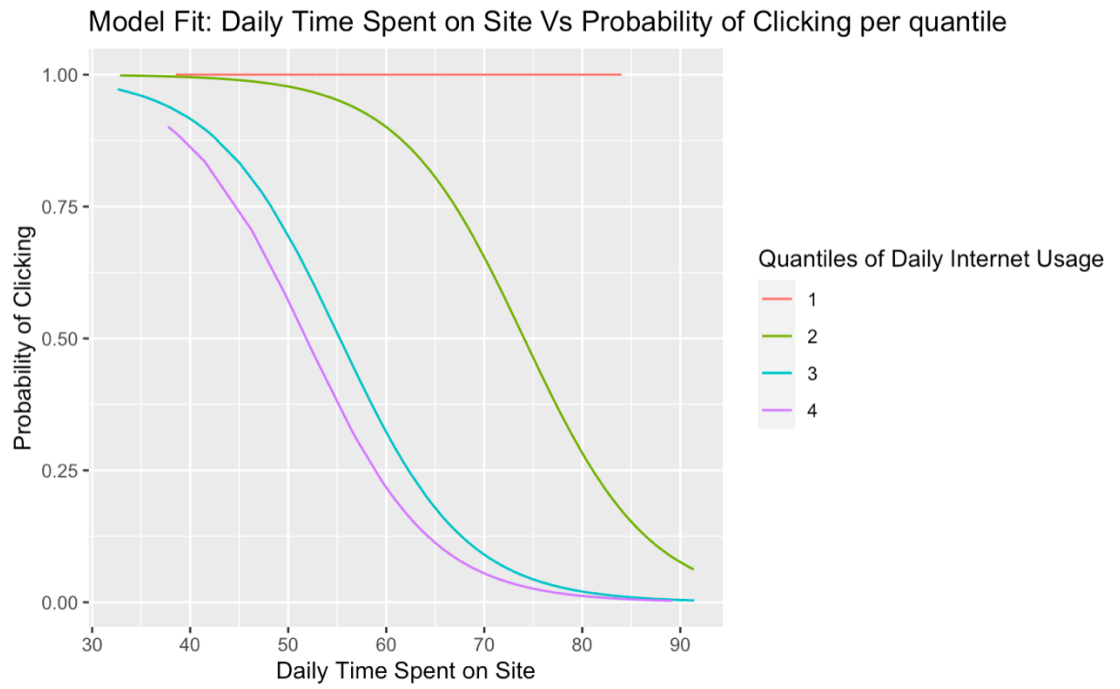

Fig 6. Analysis of daily time spent on site per quantile

3. Checking how other factors affect the probability of clicking ads:

3.1. Age Vs Daily Time Spent on Site and Daily Internet usage:

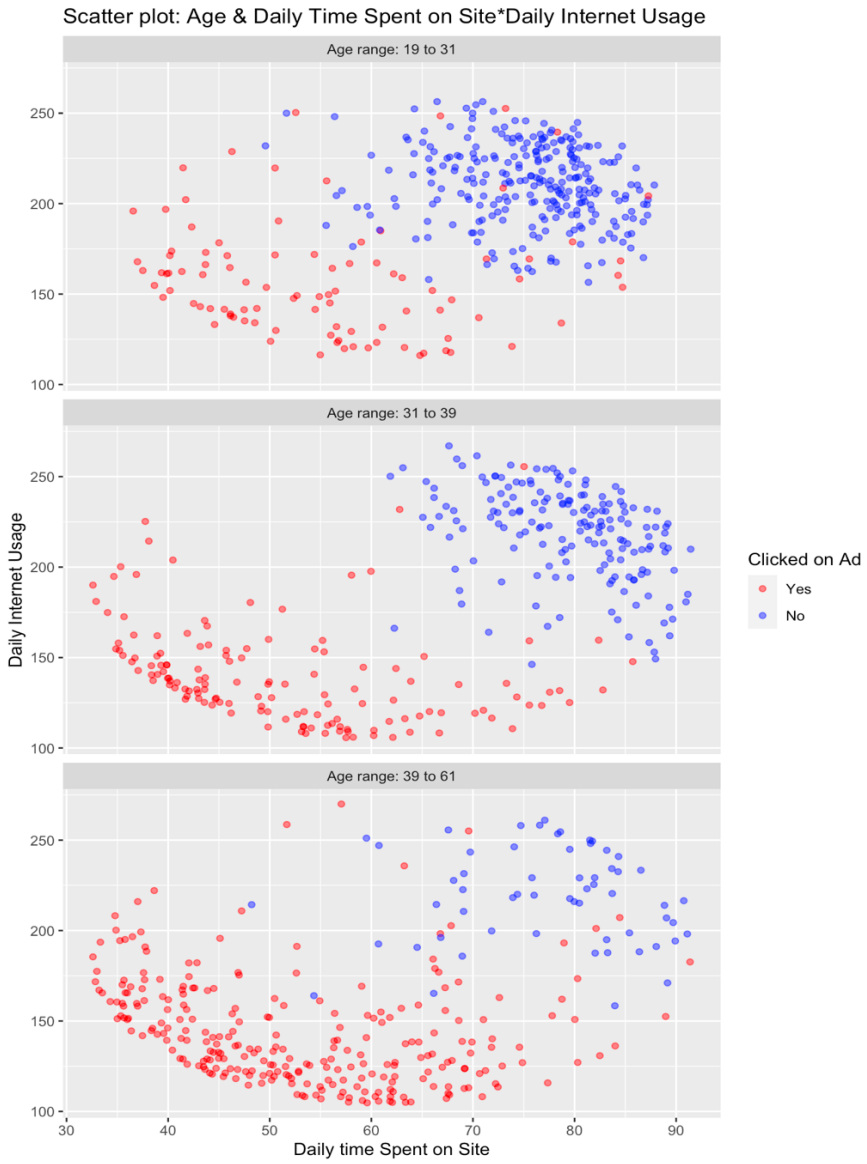

Fig 7: Scatter Plot: Age Vs Daily time spent on site and Daily internet usage

Fig. 7, depicts a scatter plot of Daily Time Spent on Site vs the Daily Internet Usage of consumers, faceted by Age. Our aim is to study the probability of clicking on the ad depending on the age group of consumers and their respective time on the internet.

Users who click on ads are represented by red data points, while those who do not are represented by blue data points. We can conclude that consumers aged 19 to 31 use the internet and spend a lot of time on websites, but they don't click on ads very often. Although there are few consumers over the age of mid-twenties who click on ads, the population is not dense. As a result, we can conclude that this age group of consumers knows exactly what they want to browse and which ad to click on based on their needs.

Consumers between the ages of 31 and 39 make good use of the internet and spend a reasonable amount of time on websites. There is a clear distinction between those who click and those who do not click on advertisements. However, because the number of "Yes" and "No" responses is similar, it is difficult to determine whether they clicked on ads based on their needs or due to a lack of knowledge about that product.

Lastly, consumers in the age group 39-61 clearly tell us that even though they use the internet less and spent less time on sites, they end up clicking on ads. Using this, we conclude that the older consumers may or may not have much knowledge of online advertising leading them to click on ads.

Recalling our hypothesis made in Section 2.2 we can confirm how an age group range of consumers (E.g.: 19 – 31-year-old consumers) clicks on ads. Thus, from the above analysis, we conclude that our hypothesis made was correct.

3.2. Modelling of Age Vs Daily Time Spent on Site and Daily Internet usage

Building on our previous model we attempt to incorporate the 'Age' variable in the model too. Considering the interaction between Daily Time Spent on Site and Internet Quantiles (Quantiles of Daily Internet Usage) we include the Age variable as an addition to this interaction. Fig. 8, shows the fitted model discussed.

```
click.age.interact.logit = glm(Clicked.on.Ad ~ Daily.Time.Spent.on.Site:InternetCat + Age , family = "binomial", data = ads)
summary(click.age.interact.logit)
```

Fig 8. Fitted Model

Coefficients:

	Estimate	Std. Error	z-value	p-value
Intercept	6.47842	1.11812	5.794	6.87e-09
Age	0.10983	0.01974	5.563	2.65e-08
Daily.Time.Spent.on.Site:InternetCat1	0.18742	18.46996	0.010	0.992
Daily.Time.Spent.on.Site:InternetCat2	-0.13633	0.01474	-9.251	< 2e-16
Daily.Time.Spent.on.Site:InternetCat3	-0.17672	0.01631	-10.832	< 2e-16
Daily.Time.Spent.on.Site:InternetCat4	-0.18917	0.01690	-11.193	< 2e-16

Table 1. Coefficients and p-values of Fitted Model

Table 1 shows the model coefficients and p-values for every term. Not only does the model have a low AIC score of 262.66 as compared to other models, but it is also intuitively logical too. The Age variable has a positive coefficient which is reasonable because as we observed in Fig. 8, older customers tend to click on the ad more often. Similarly, the interaction terms show a negative coefficient which is also justified based on our previous observations where the probability of clicking on the ad is higher for a low Time Spent on Site and Low Internet Usage. Overall, this model is a good fit for our dataset

3.3. Check if gender affects Clicked on Ad:

Since Male is a categorical independent variable and Clicked on Ad is a categorical dependent variable, the approximate relationship can be visualized through bar graphs or mosaic plots. This time, we will explore the relationship between the two variables through a mosaic plot in Fig. 9.

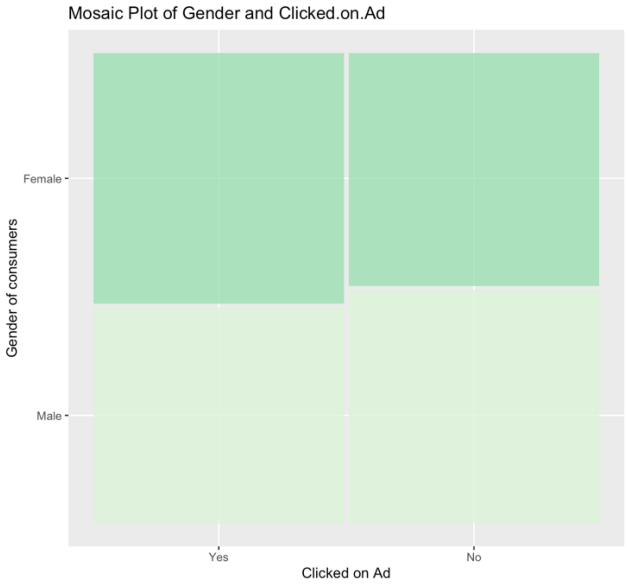

Fig 9. Mosaic Plot: Gender Vs Clicked on Ad

```{r}
chisq.test(ads$Male, ads$Clicked.on.Ad)
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  ads$Male and ads$Clicked.on.Ad
X-squared = 1.2979, df = 1, p-value = 0.2546
```

Fig 10. Pearson's Chi-squared test

For the advertising data, we want to know if clicking on the ad is gender independent: if it isn't, we might suspect that there is some sort of gender-based variation happening in ad clicking, i.e., both genders are responding to ad clicking in a different way.

Fig. 9, displays a mosaic plot of the advertising data. Here we're looking at clicked on the ad and gender. The area of each tile shows us how many consumers were in each of the four categories: either clicked on ad-yes/male, clicked on ad-yes/female, did not click on ad/male, or did not click on ad/female.

We carry out a chi-square test to statistically check and prove whether clicking on the ad is independent of gender. From Fig. 10, ***The chi-square test yielded a value of 0.2546 with a p-value greater than 0.05***. Therefore, it can be judged that the percentage of clicks on advertisements is the same for both men and women, in other words, the Male variable and the Clicked-on Ad variables are independent of each other. We conclude that gender has no effect on clicking on ads and hence, we do not include it in our model building process.

4. Limitations and Future Scope:

Some of the limitations that we faced in this project are as stated below:

- The dataset was categorized country-wise and city-wise. These two factors were difficult for us to consider in our analysis as we would have had to categorize each city under a country and each country under a continent for better analysis, making it very tedious. Thus, we could not understand how consumers react to ads in different parts of the world.
- The above limitation made us not consider the Income as it is most likely dependent on the country and/or city.
- The advertisement category was not available. Only the headline of the advertisement is available from the dataset. This further limits our analysis, as we cannot check what type of ads are more likely to be clicked by consumers

Although, we had to tackle the above limitations we aim to find an advanced way in considering them in our analysis in the future.

5. Conclusion:

Through this project, we have gained insights on factors that potentially affect the clicks on advertisements. As per our analysis, the factors which affect clicking on advertisements are Daily Time Spent on the Site, Daily Internet Usage, and the Age of the consumer. Thus, in conclusion we state that these interrelated factors could be significant to advertisers in studying the Click-Through-Rate, essentially understanding the needs of their target demographic.