PREDICTING ADVERTISEMENT CLICKS

GAGANA UDAY KUMAR

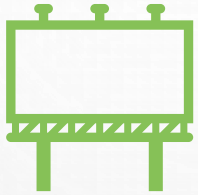
WOV796

BACKGROUND

Online Advertising Landscape

Online ads are essential for businesses to reach potential customers.

Not all users click on ads, prompting the need for targeted advertising strategies.



Challenges in Online Advertising

Businesses face challenges in effectively targeting their audience and maximizing ad engagement.

Maximizing ad engagement and click-through rates is crucial for advertising success.

Project Objective

Develop a predictive model for ad click-through rates using data from a marketing agency.

The goal is to predict if a user will click on an online advertisement, helping businesses target their ads more effectively.

DESCRIPTION OF THE DATA

- Overview of dataset
  - The dataset “advertisement” was obtained from kaggle
  - The dataset contains information from a marketing agency about user interactions with online ads.
- Key variables
  - Daily time spent on site: average time spent by users on the website.
  - Age: age of the user interacting with the ad.
  - Area income: average income of the area where the user resides.
  - Daily internet usage: amount of time spent by the user on the internet daily.
  - Ad topic line: title or description of the ad.
  - City: city of residence of the user.
  - Male: binary variable indicating gender (1 for male, 0 for female).
  - Country: country of residence of the user.
  - Timestamp: date and time when the ad interaction occurred.
  - Clicked on ad: binary variable indicating if the user clicked on the ad (1 for clicked, 0 for not clicked).

PROPOSED ANALYSIS

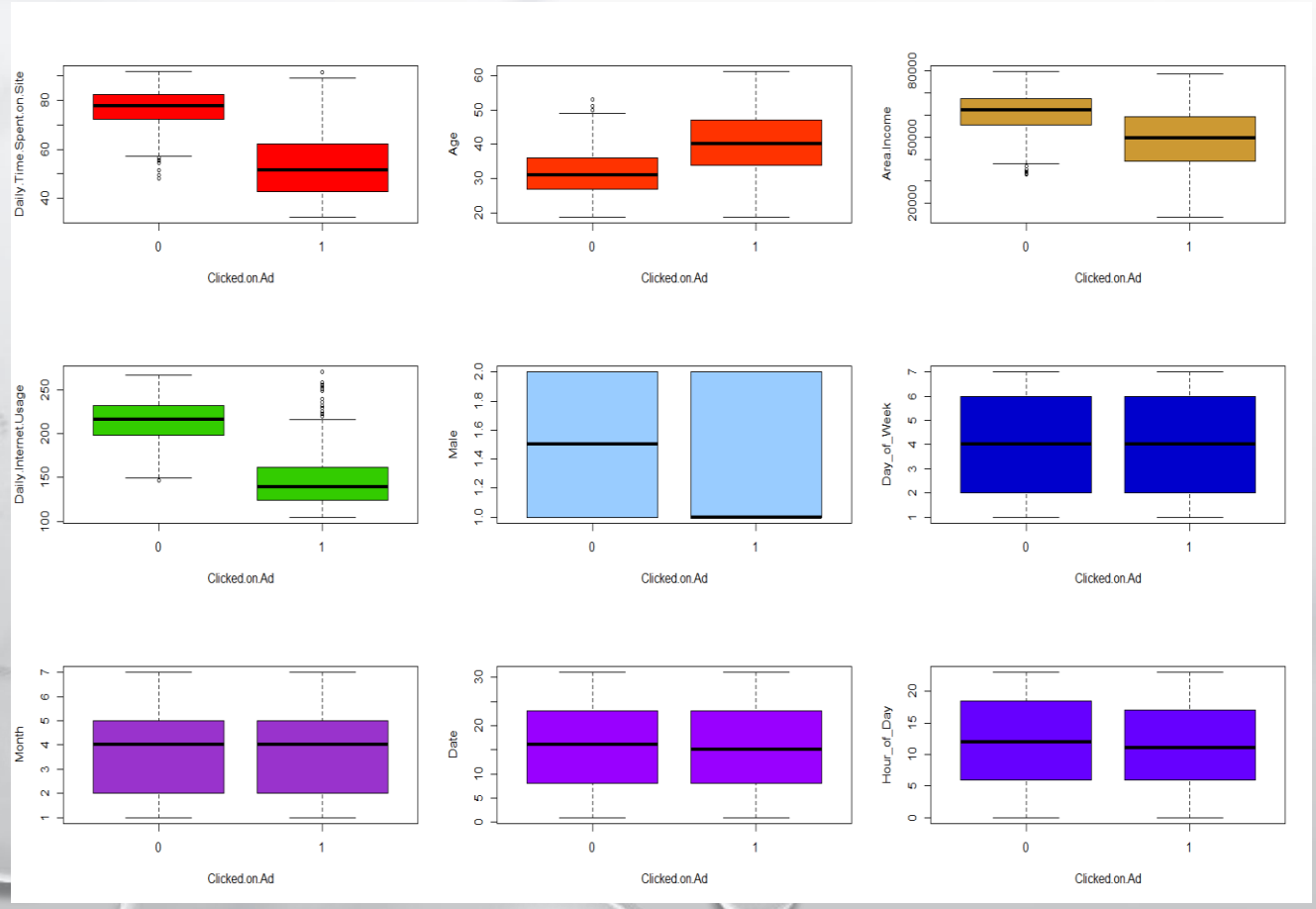
Here are the few analysis that I will be focusing on:

1. Build a model with good accuracy to predict which user would click on the advertisement.
2. I would like to see if there is any correlation with time spent on the site and 'clicked on ad'.
3. What is the average time spent on the site daily and average area income of users who click on the ad.
4. I would like to investigate if mean income changes with click on the ad.
5. I will check if total daily internet use and time spent by users on the site are related to each other in some way.
6. What are the general characteristics of the users who click on the ad.
7. Which day, which month, which date do users usually click on the ad.
8. If we divide the time into morning, afternoon and night, which part of the day do users click on ad?

DATA CLEANING

- No empty cells
- Converted the data type of clicked.On.Ad and male to factor
- Removed character variables ad.Topic.Line, city, country
- Created day_of_week, month, date, hour_of_day from timestamp variable
- Created levels for variables day_of_week and month
- Deleted the variable timestamp
- Data was split into 80-20 percentage

EXPLORATORY ANALYSIS



	Daily.Time.Spent.on.Site	Age	Area.Income	Daily.Internet.Usage	Male	Clicked.on.Ad
Daily.Time.Spent.on.Site	1.00	-0.33	0.31	0.52		-0.75
Age	-0.33	1.00	-0.18	-0.37		0.49
Area.Income	0.31	-0.18	1.00	0.34		-0.48
Daily.Internet.Usage	0.52	-0.37	0.34	1.00		-0.79
Male					1.00	
Clicked.on.Ad	-0.75	0.49	-0.48	-0.79		1.00

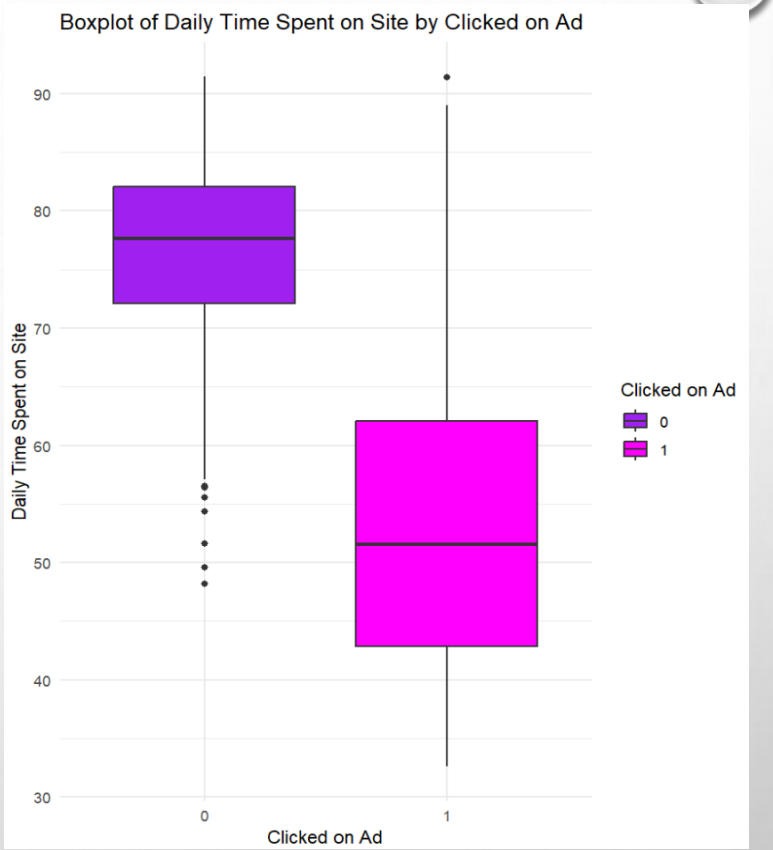
1 0.8 0.6 0.4 0.2 0 -0.2 -0.4 -0.6 -0.8 -1

DATA ANALYSIS

Correlation Between Time Spent On The Site And 'Clicked On Ad'

WILCOXON RANK SUM TEST :

- Reject the null hypothesis.
- There is a statistically significant difference in daily time spent on site between the two groups (Clicked on Ad = 0 and Clicked on Ad = 1).




```
        Wilcoxon rank sum test with continuity correction

data:  ad$Daily.Time.Spent.on.Site[ad$Clicked.on.Ad == 0] and ad$Daily.Time.Spent.on.Site[ad$Clicked.on.Ad == 1]
W = 232513, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Average Time Spent On The Site And Area Income Of Users Who Click On The Ad

TIME SPENT ON THE SITE

Min - 32.6

Max - 91.37

Average - 53.15

AREA INCOME

Min - 13996.5

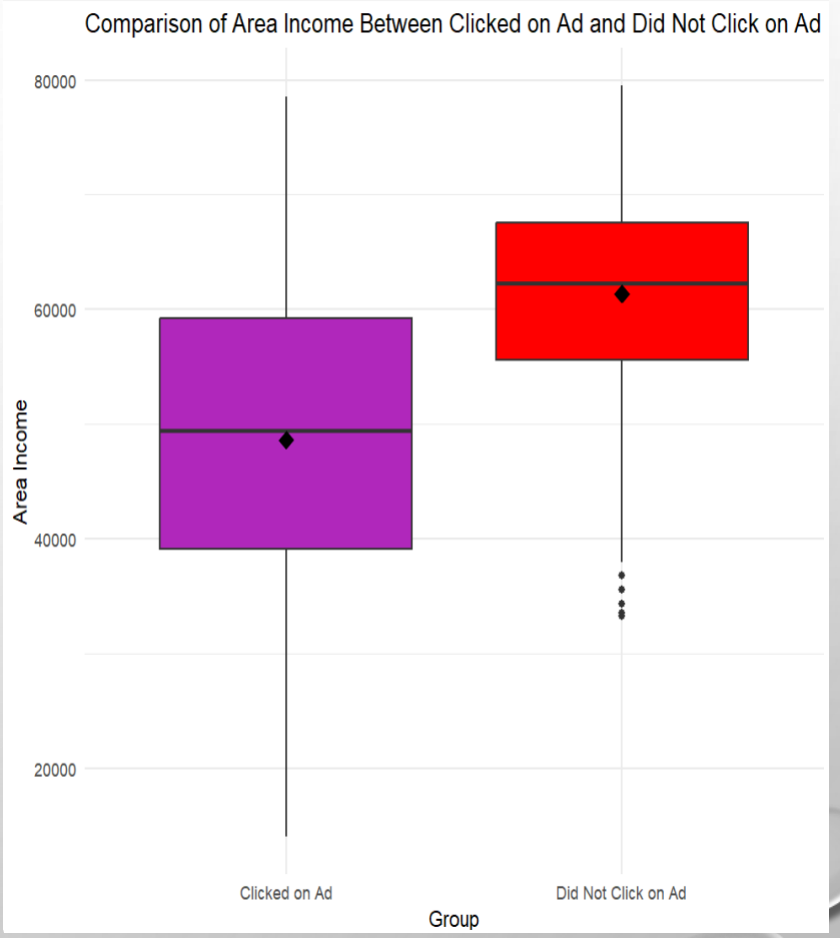
Max - 78520.99

Average - 48614.41

Investigating If Mean Income Changes With Ad Clicks

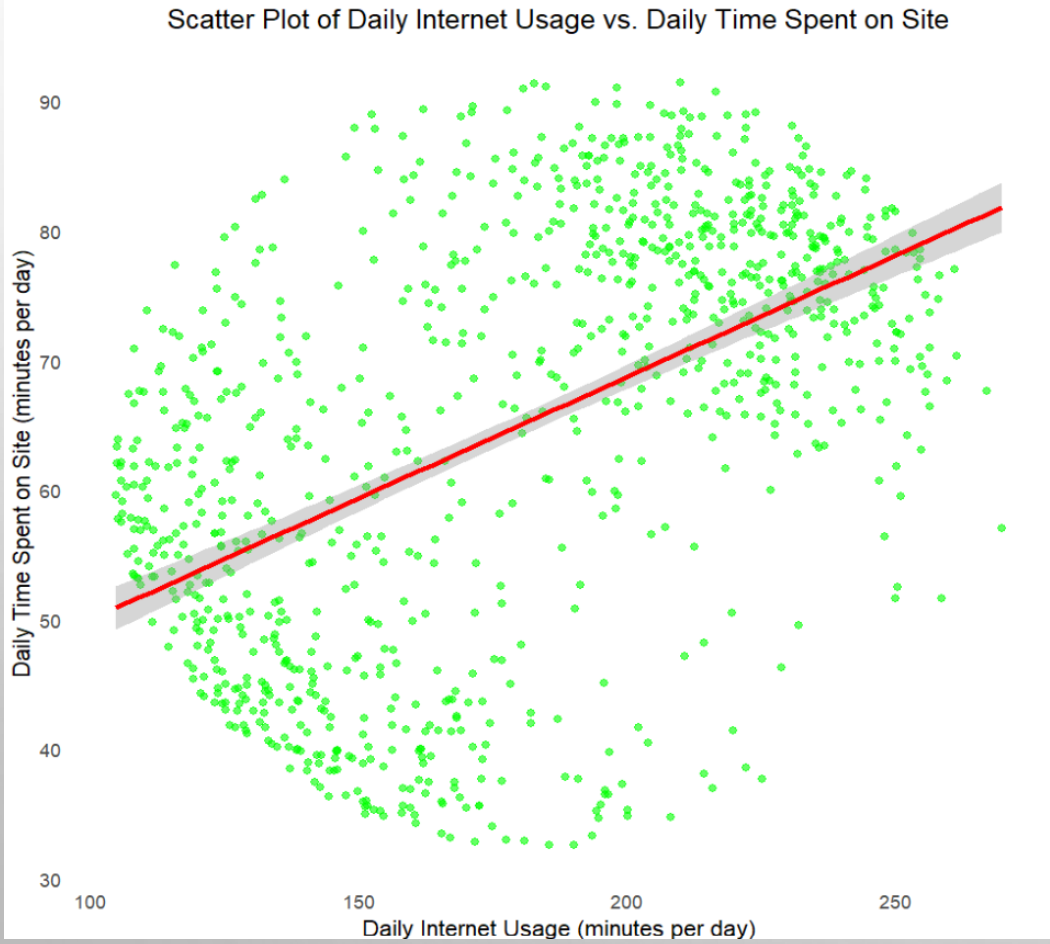
WILCOXON RANK SUM TEST :

- Highly significant difference in mean area income between individuals who clicked on the ad and those who did not.
- The mean income is lower for users who clicked on the ad compared to those who did not.




```
        Wilcoxon rank sum test with continuity correction

data:  ad$Area.Income[ad$Clicked.on.Ad == 0] and ad$Area.Income[ad$Clicked.on.Ad == 1]
W = 192438, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

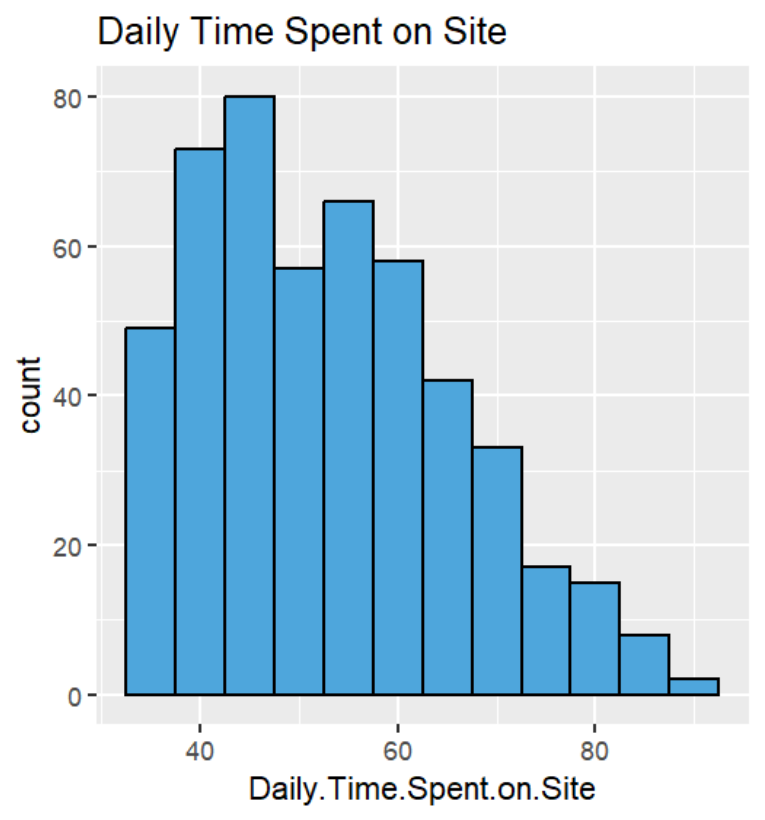
Exploring The Relationship Between Total Daily Internet Use And Time Spent On The Site

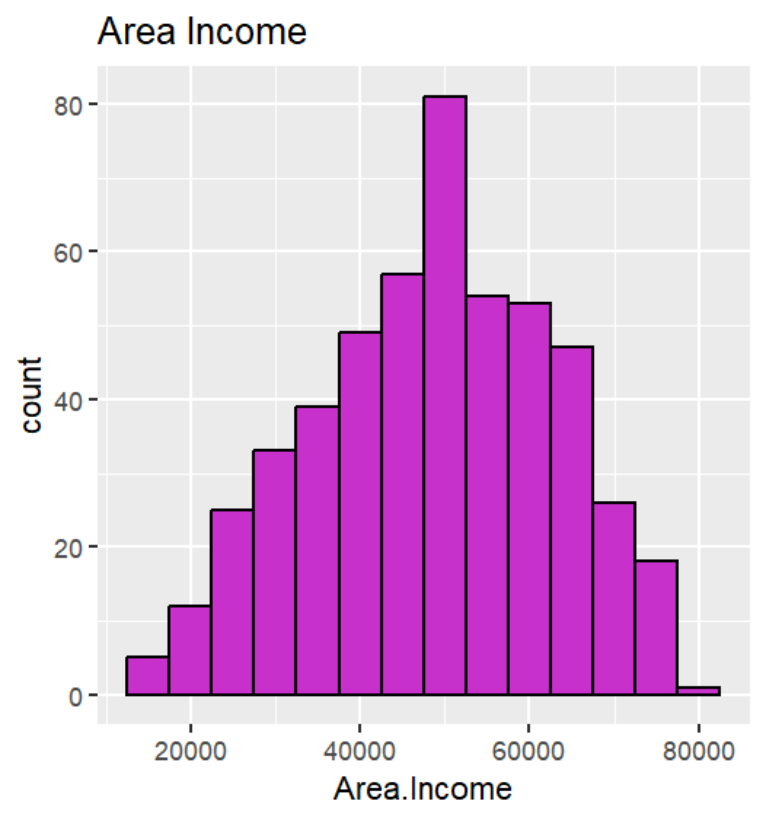


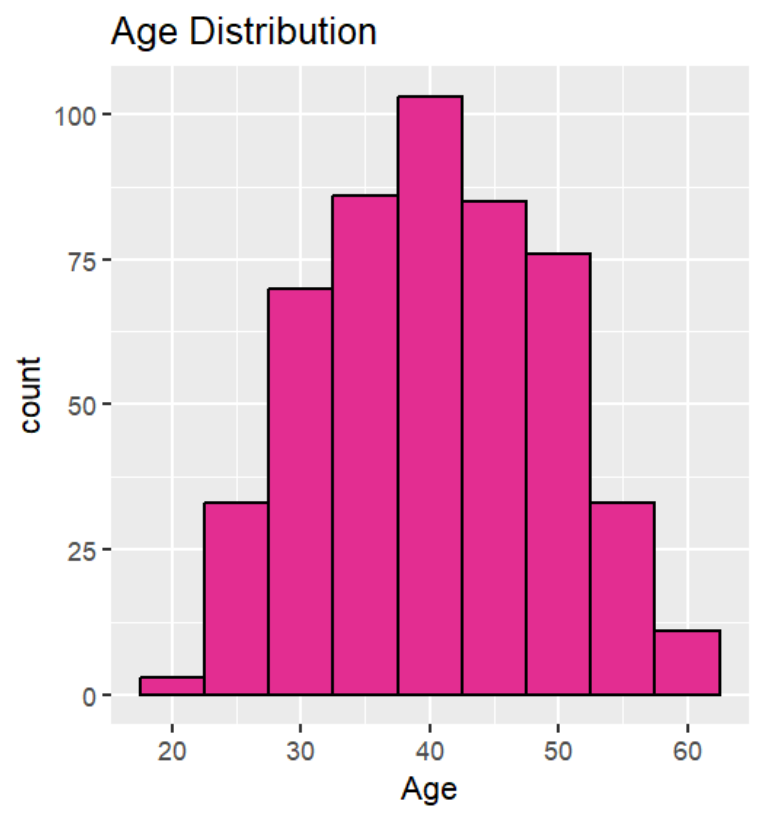
- The correlation coefficient of 0.52 suggests a moderate positive linear relationship between daily internet usage and daily time spent on site
- The relationship is not strong enough

```
> print(correlation)
[1] 0.5186585
```

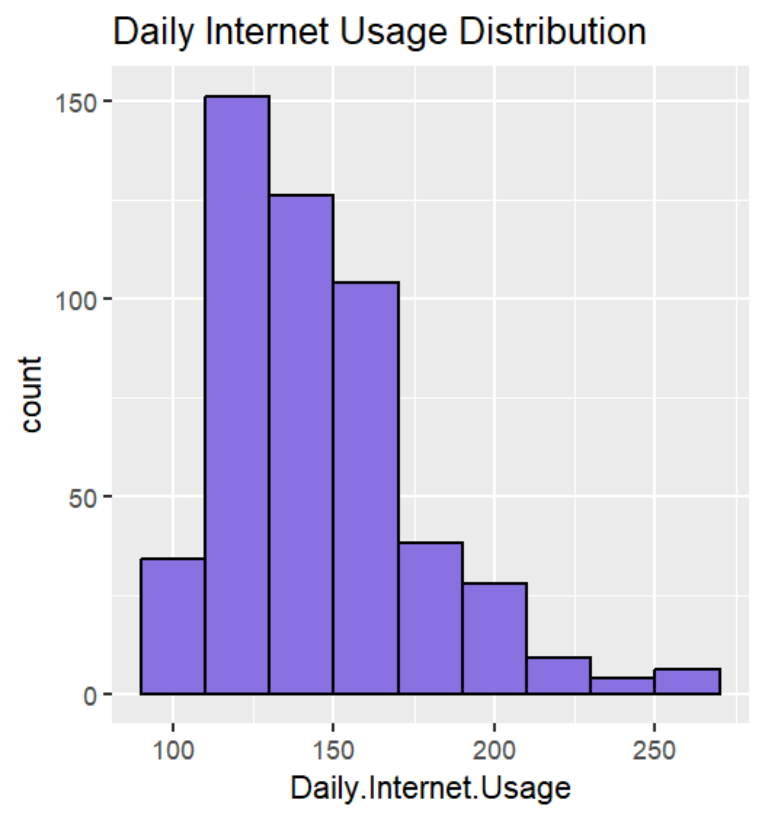
General Characteristics Of Users Who Click On The Ad

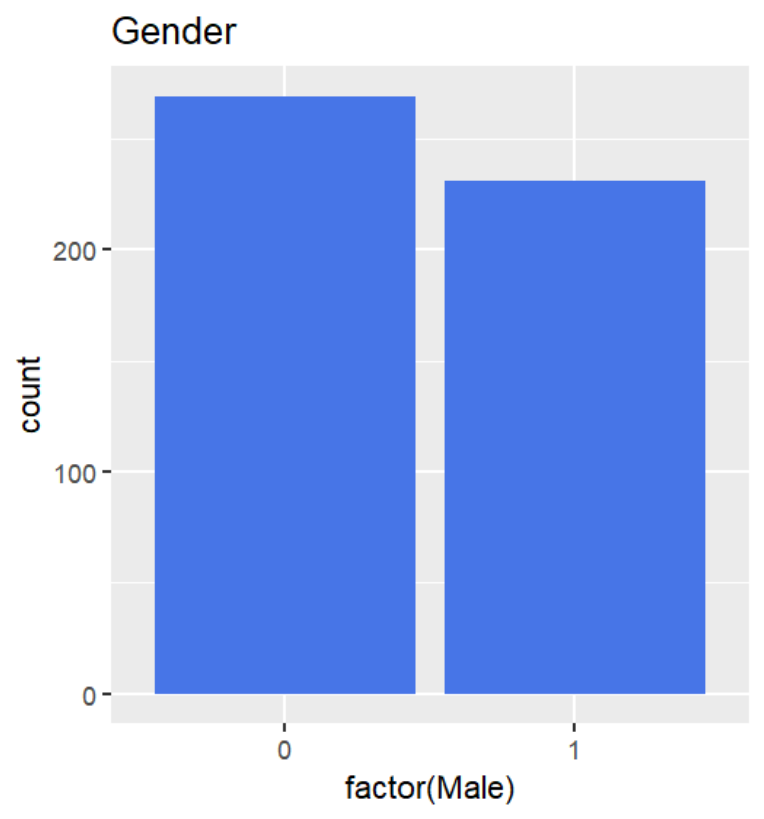




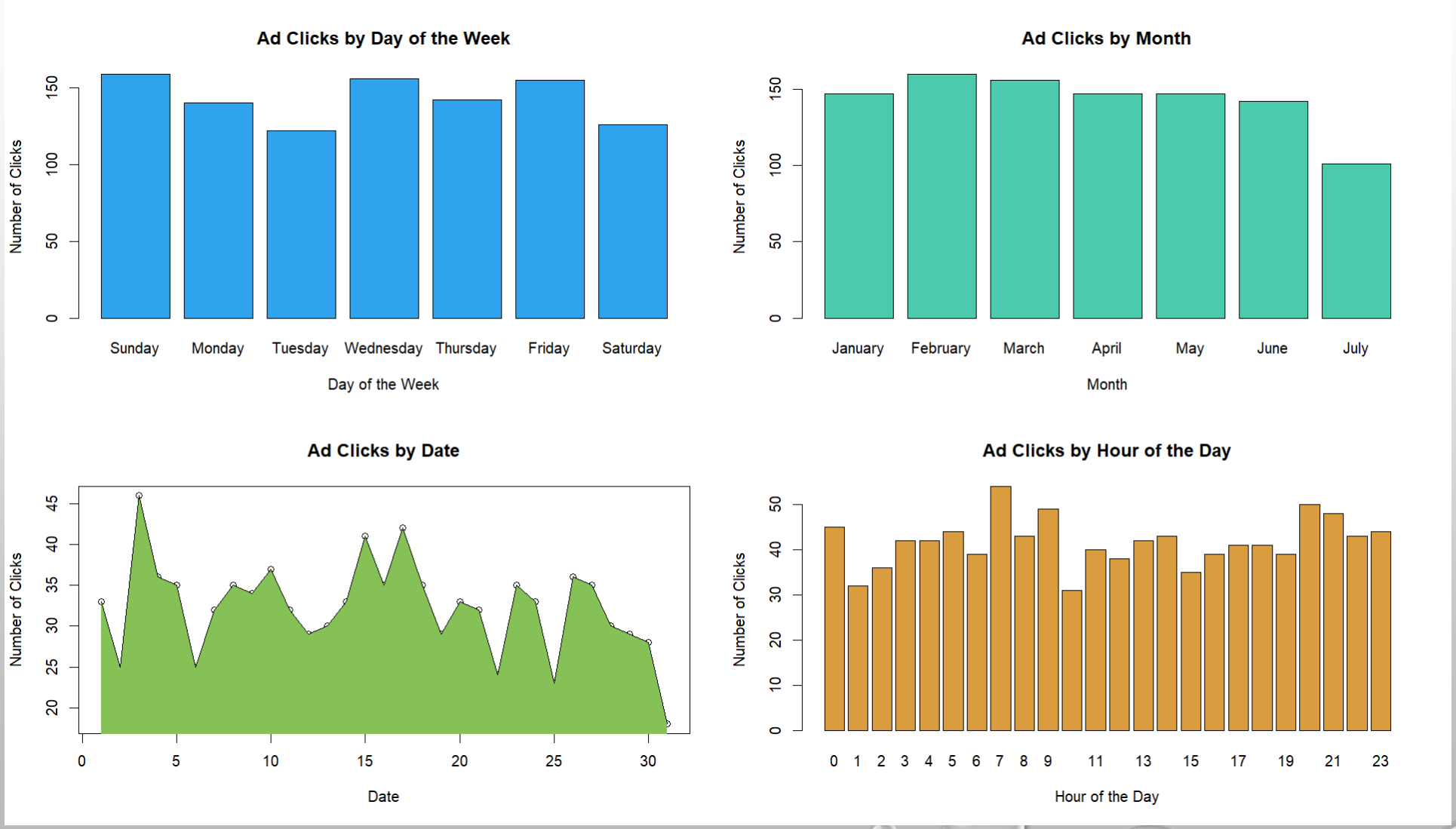


General Characteristics Of Users Who Click On The Ad



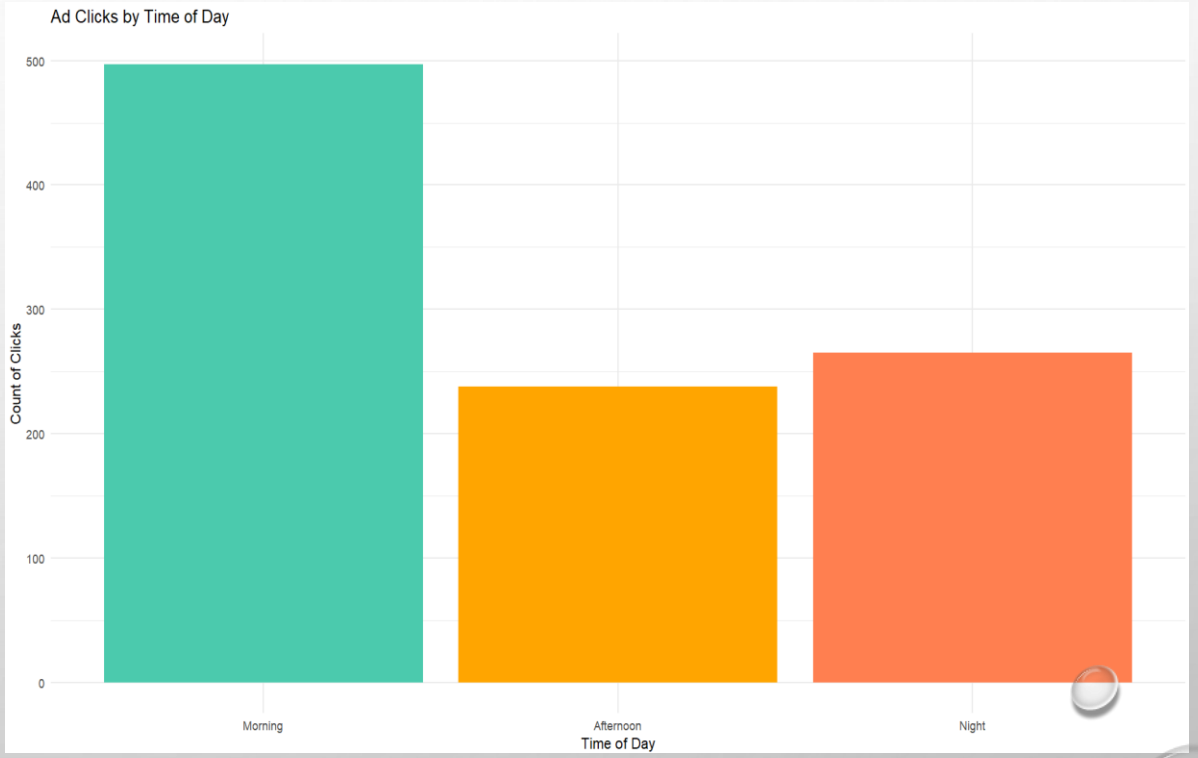


Timing Of Ad Clicks



Part Of The Day Users Click On Ad

- Categorized the time into morning, afternoon and night
- do not have sufficient evidence to reject the null hypothesis
- No significant association between the time of day (morning, afternoon, night) and whether a user clicks on an ad or not




```
> print(chi_sq_result)

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 2.3062, df = 2, p-value = 0.3157
```

Build A Model With Good Accuracy To Predict Which User Would Click On The Advertisement

LOGISTIC REGRESSION

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 102   3
         1   1  94

               Accuracy : 0.98
                 95% CI : (0.9496, 0.9945)
    No Information Rate : 0.515
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9599

 Mcnemar's Test P-Value : 0.6171

            Sensitivity : 0.9691
            Specificity : 0.9903
         Pos Pred Value : 0.9895
         Neg Pred Value : 0.9714
             Prevalence : 0.4850
         Detection Rate : 0.4700
   Detection Prevalence : 0.4750
      Balanced Accuracy : 0.9797

       'Positive' Class : 1
```

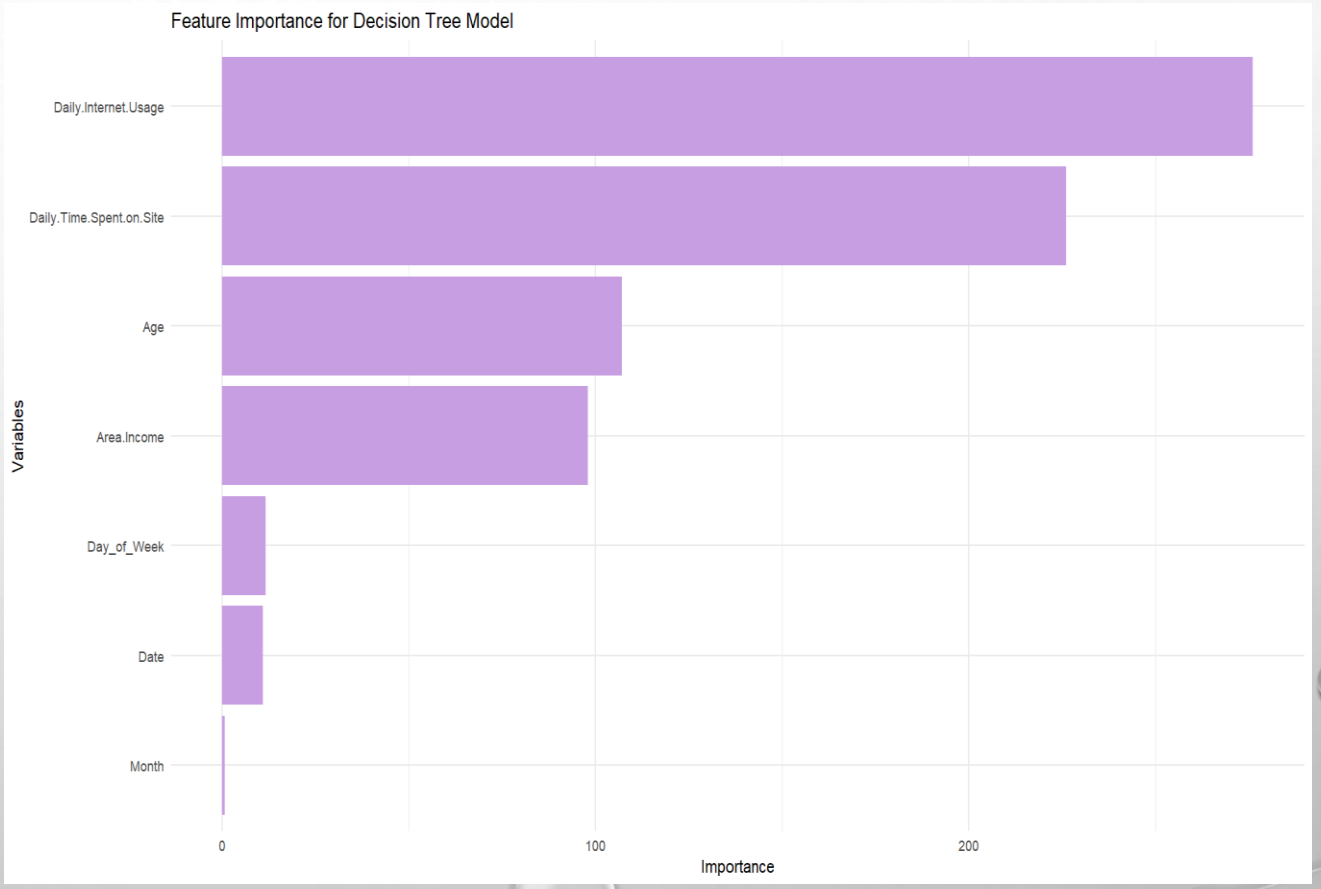
SVM

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 103   5
         1   0  92

               Accuracy : 0.975
                 95% CI : (0.9426, 0.9918)
    No Information Rate : 0.515
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9499

 Mcnemar's Test P-Value : 0.07364

            Sensitivity : 1.0000
            Specificity : 0.9485
         Pos Pred Value : 0.9537
         Neg Pred Value : 1.0000
             Prevalence : 0.5150
         Detection Rate : 0.5150
   Detection Prevalence : 0.5400
      Balanced Accuracy : 0.9742

       'Positive' Class : 0
```

Build A Model With Good Accuracy To Predict Which User Would Click On The Advertisement

DECISION TREE

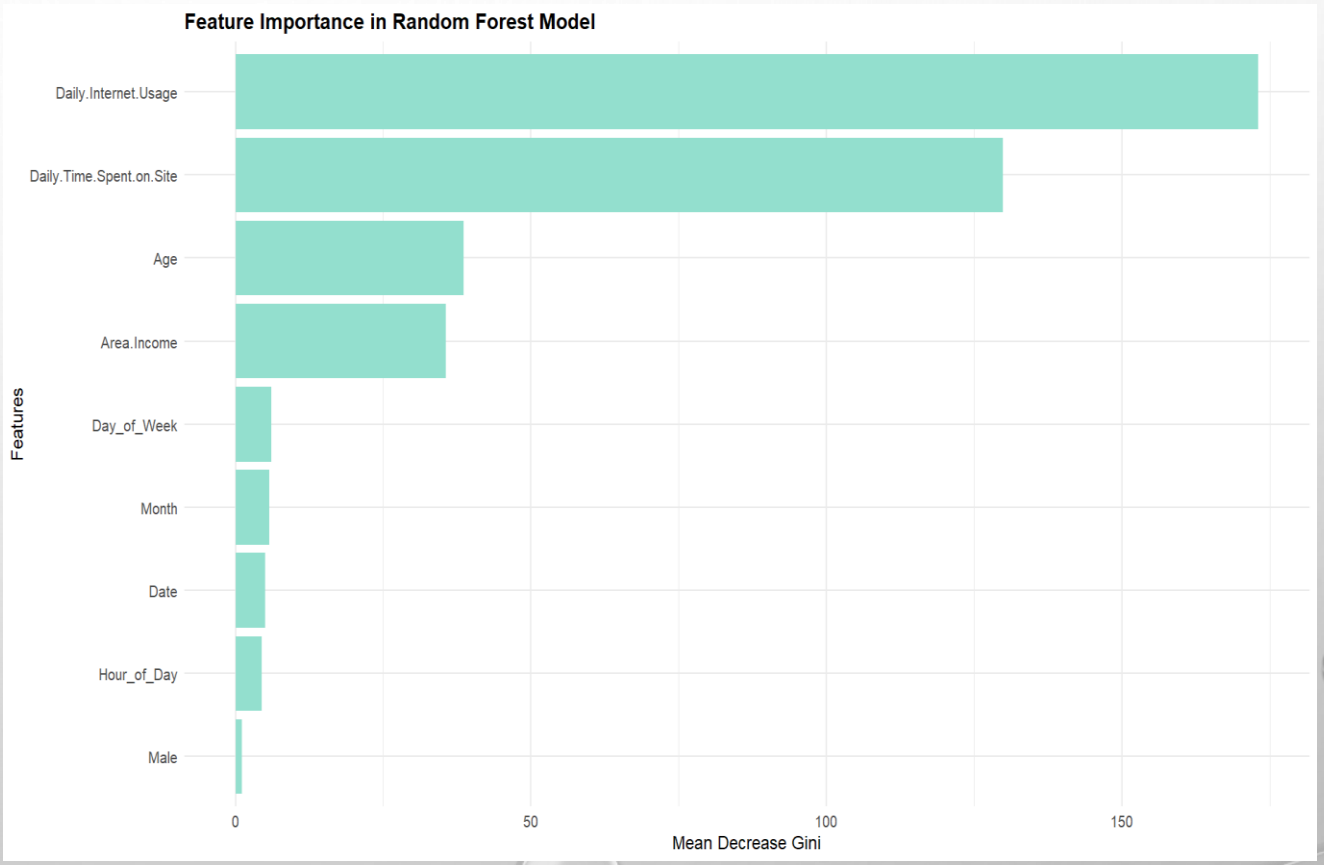
```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 95  8
         1  8 89

               Accuracy : 0.92
                 95% CI : (0.8733, 0.9536)
    No Information Rate : 0.515
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8399

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9223
            Specificity : 0.9175
         Pos Pred Value : 0.9223
         Neg Pred Value : 0.9175
             Prevalence : 0.5150
         Detection Rate : 0.4750
   Detection Prevalence : 0.5150
      Balanced Accuracy : 0.9199

       'Positive' Class : 0
```



Build A Model With Good Accuracy To Predict Which User Would Click On The Advertisement

RANDOM FOREST

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 101   4
         1   2  93

               Accuracy : 0.97
                 95% CI : (0.9358, 0.9889)
    No Information Rate : 0.515
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9399

 Mcnemar's Test P-Value : 0.6831

            Sensitivity : 0.9806
            Specificity : 0.9588
         Pos Pred Value : 0.9619
         Neg Pred Value : 0.9789
             Prevalence : 0.5150
         Detection Rate : 0.5050
   Detection Prevalence : 0.5250
      Balanced Accuracy : 0.9697

       'Positive' Class : 0
```



CONCLUSION

Influence of Daily Time Spent: The amount of daily time users spend on the site significantly impacts whether they click on an ad.

Mean Income and Ad Clicks: There is a significant difference in area income between users who click on ads and those who do not.

Timing of Ad Clicks: Users tend to click on ads most frequently in the morning and evening hours.

Model Performance: Logistic regression, decision tree, and random forest models all achieve high accuracy in predicting ad clicks, with Logistic regression performing particularly well.

Feature Importance: Daily internet usage and daily time spent on the site are key features in predicting ad clicks, as highlighted by decision tree and random forest models.

Future Directions: Further investigation into user behavior and its impact on ad interaction can lead to more targeted marketing strategies.