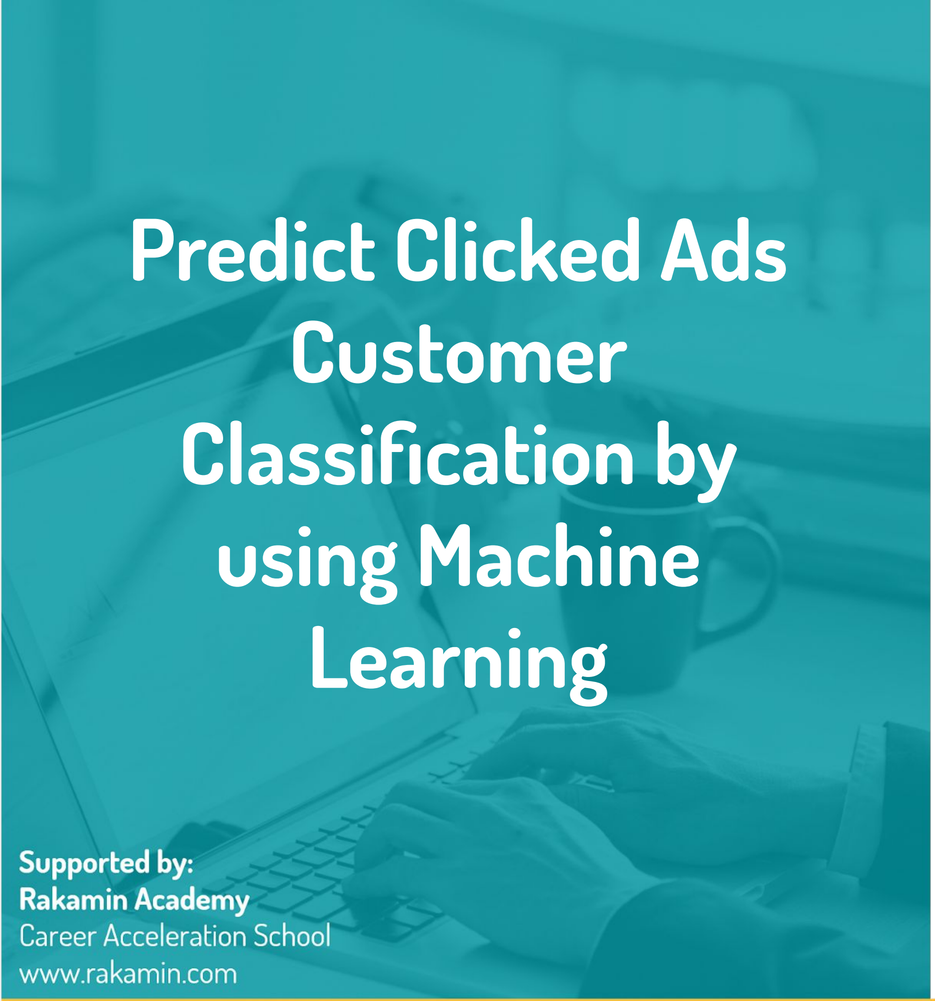


Predict Clicked Ads Customer Classification by using Machine Learning

Created by:
Tsabitah Karimah
tsabitahkarimah@gmail.com
LinkedIn

“I am a dietetics graduate with a newfound interest for data science, seeking to leverage analytical skills and healthcare knowledge in a data scientist role. Completed an intensive 5-month data science course, gaining proficiency in key programming languages and machine learning techniques. Eager to apply newly acquired skills in data manipulation, visualization, and predictive modeling to extract meaningful insights from complex datasets. Combines technical acumen with strong problem-solving abilities to drive data-informed decision-making in a data scientist role.”

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Overview

“An Indonesian company wants to evaluate the effectiveness of an advertisement they've aired. This is crucial for the company to understand the extent to which their advertisement has reached its target audience and successfully attracted customers.

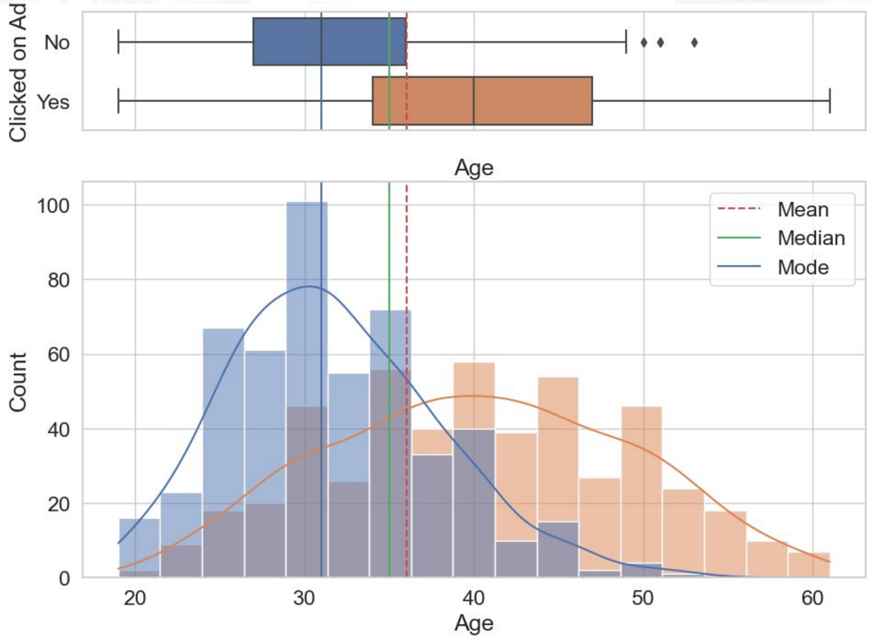
By processing historical advertisement data and uncovering insights and patterns, the company can better determine its target market. The focus of this case is to create a machine learning classification model that can accurately identify the right target customers.”

Exploration Data Analysis (EDA)

- At a glance, the data has 1000 rows and 11 columns
- There are 4 columns with missing values;
    - ‘Daily Time Spent on Site’
    - ‘Daily Internet Usage’
    - ‘Area Income’
    - ‘Male’
- All data values matched with the data type except for timestamp
    - Hence, the data type was converted from object to timestamp
- All the features then grouped based on its data type.

For further details, please refer to this Jupyter Notebook

Univariate analysis - Age

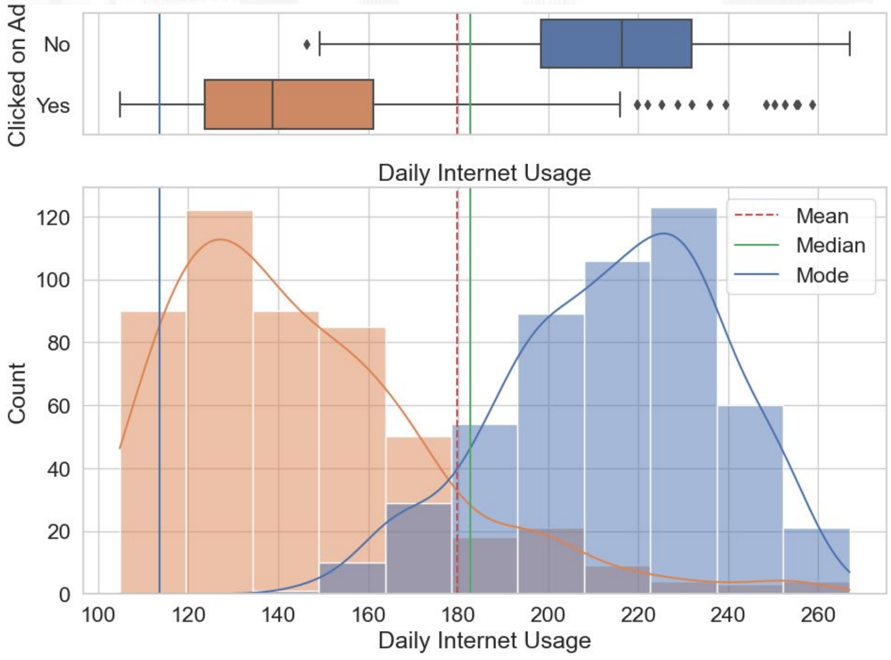


- Customer's age is ranging from 19-61 years old
- The data is slightly positive skewed as the mean (36 years old) is higher than the median (35 years old)
- The median age of those who clicked on the ad appears to be slightly lower than the median age of those who didn't.
- The histogram suggests that the age distribution for those who didn't clicked on the ad might be slightly skewed towards younger ages compared to those who did.
- The box plot shows a number of outliers in the "No" group, indicating some individuals in this group are significantly older or younger than the majority.

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

Univariate analysis - Daily Internet Usage

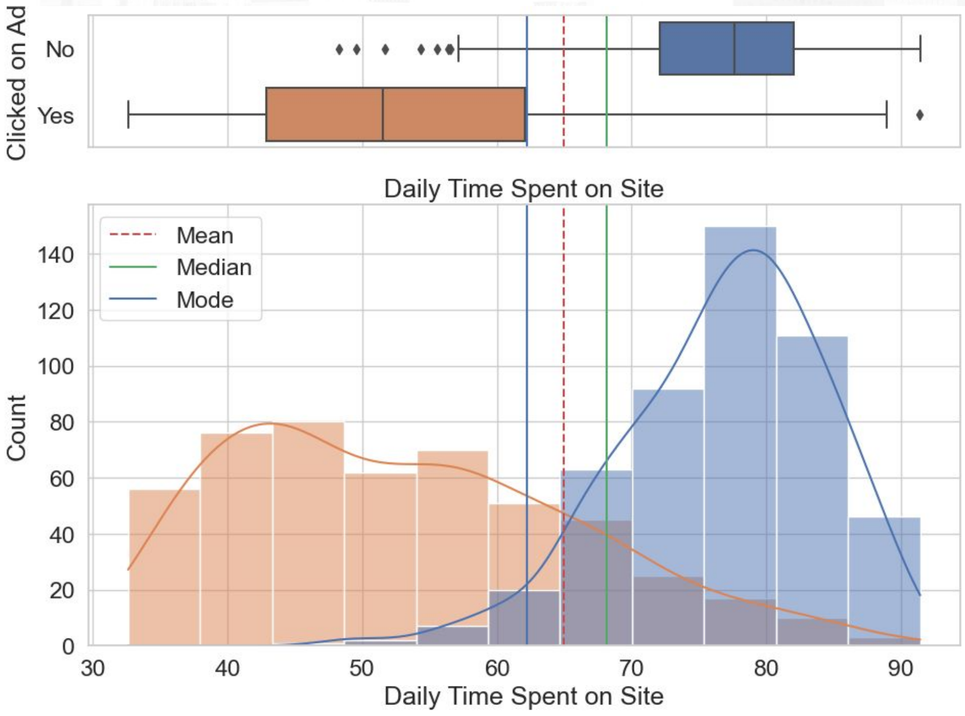


- Daily internet usage among customers is ranging from 105-267 gb(assumed)
- The data distribution is slightly negative skewed as shown by the median (183) higher than the mean (180)
- The median daily internet usage of those who clicked on the ad appears to be slightly higher than the median usage of those who didn't.
- The box plot shows a few outliers in both groups, indicating some individuals have significantly higher or lower daily internet usage.
- The histogram suggests that the distribution of daily internet usage for those who didn't clicked on the ad might be slightly skewed towards higher usage compared to those who did.
- The majority of respondents in both groups have a daily internet usage between 150 and 250 minutes.

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

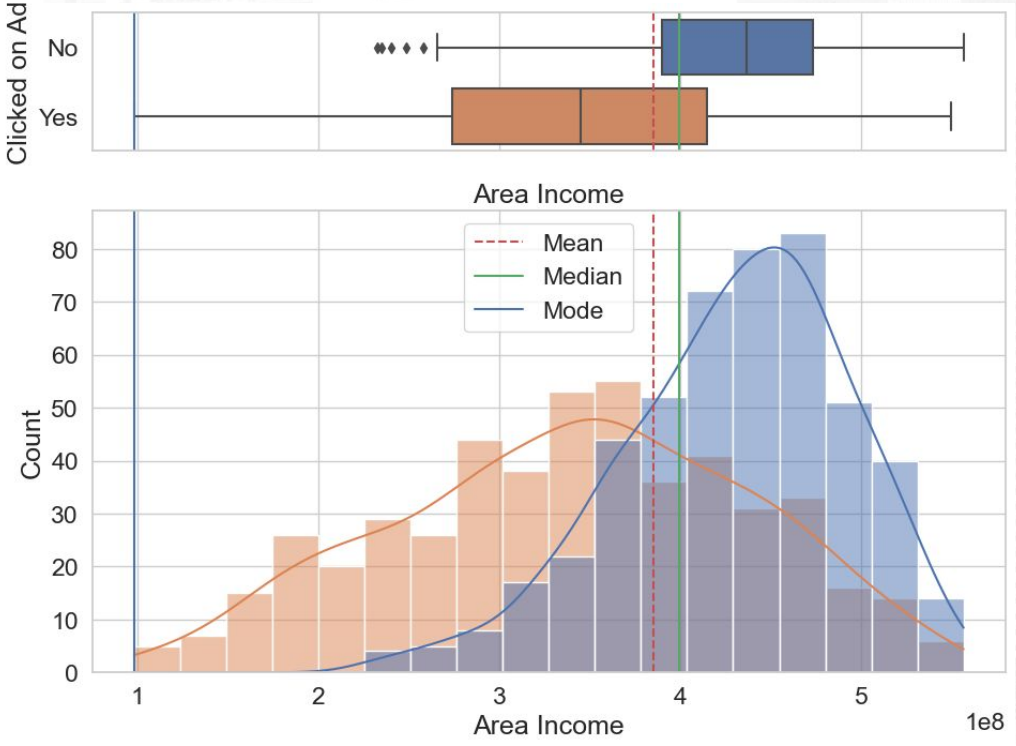
Univariate analysis - Daily Time Spent on Site



- Daily time spent on site ranging from approximately 33 mins to 91 mins
- The data distribution is slightly negative skewed as shown by the median (68 mins) that is higher than the mean (65 mins)
- The box plot shows numbers of outliers of those who didn’t clicked on the ad.
- The median daily time spent on the site for those who clicked on the ad appears to be slightly higher than the median for those who didn't.
- The histogram suggests that the distribution of daily time spent on the site for those who did not clicked on the ad might be slightly skewed towards higher usage compared to those who did.

For further details, please refer to this [Jupyter Notebook]

Univariate analysis - Area Income

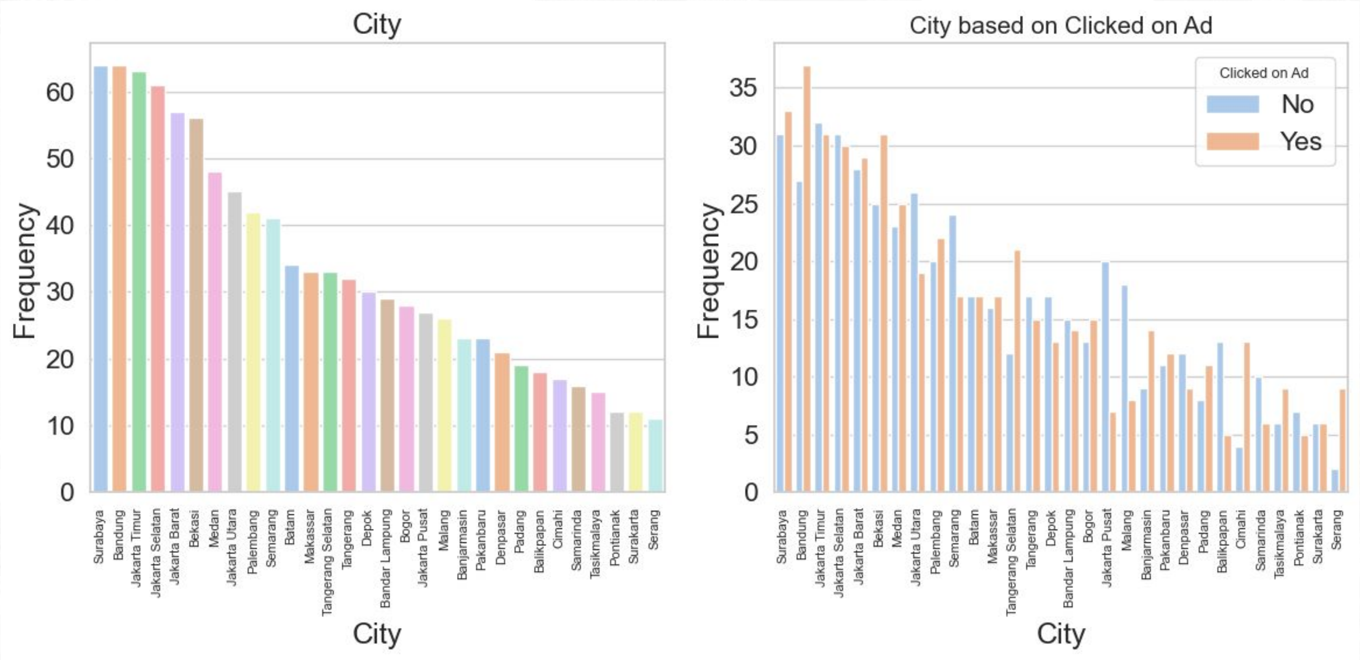


Area Income: in the range of 97.975.500 - 556.393.600 million, median (399.068.320) > mean (384.864.670), slightly negative distribution

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

Univariate analysis - City



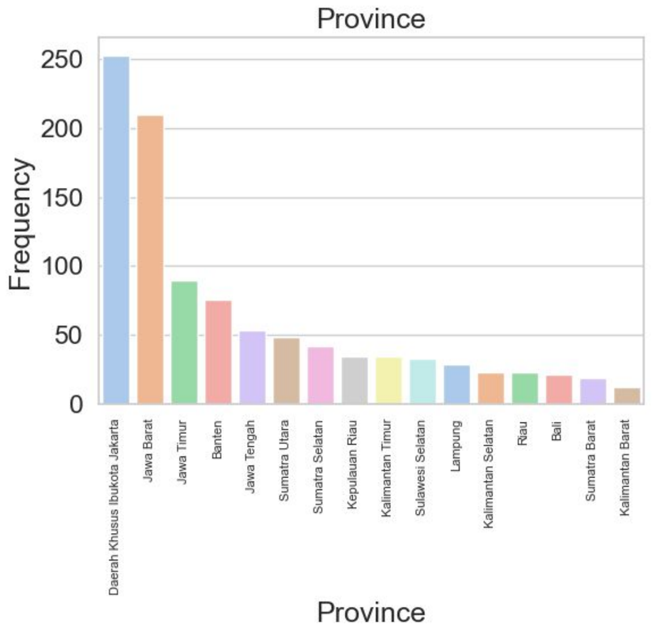
- The customers are located in a total of 30 cities with Surabaya and Bandung as the city where most customers are located.
- Customers who clicked on ad mostly located in Bandung, Surabaya, and Bekasi.

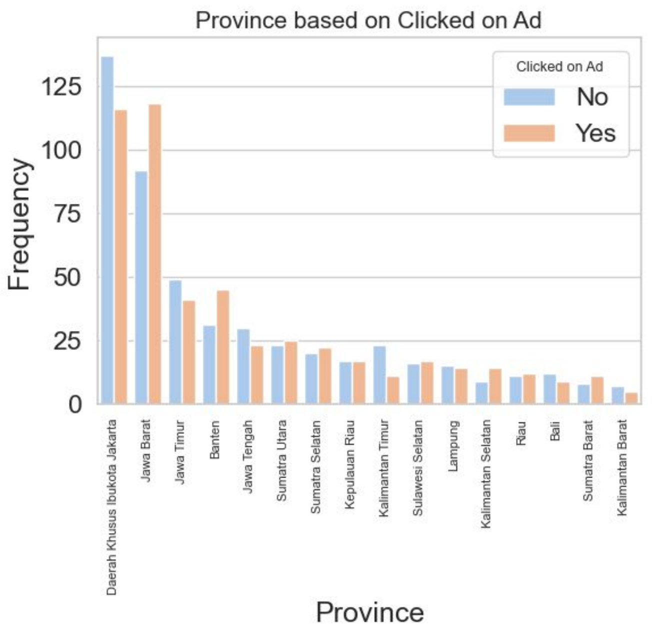
For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

Rakamin Academy

Univariate analysis - Province



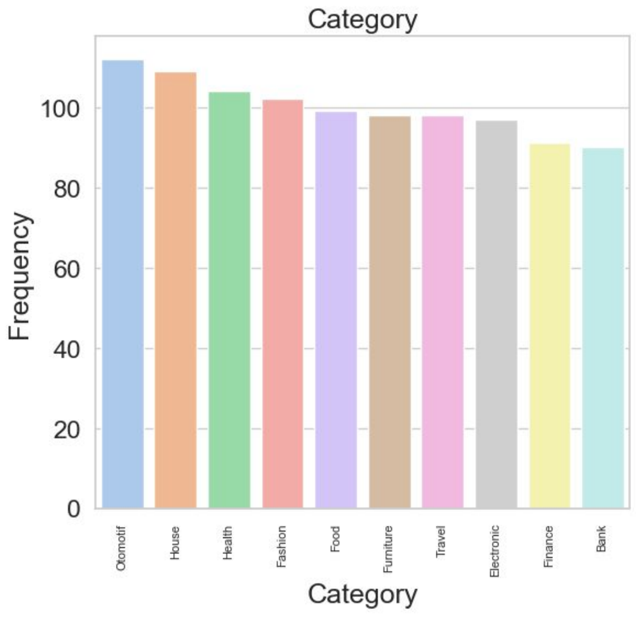


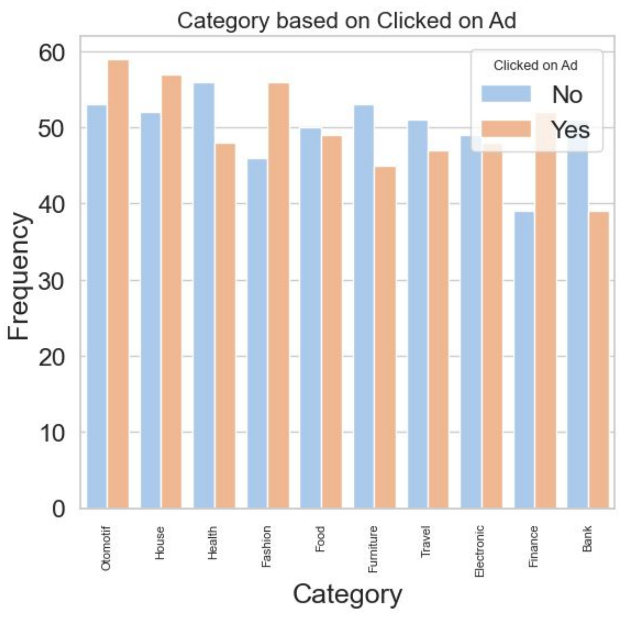
- Customers are located in 16 province with most customers who clicked on the ad located in DKI Jakarta and Jawa Barat.

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

Univariate analysis - Category





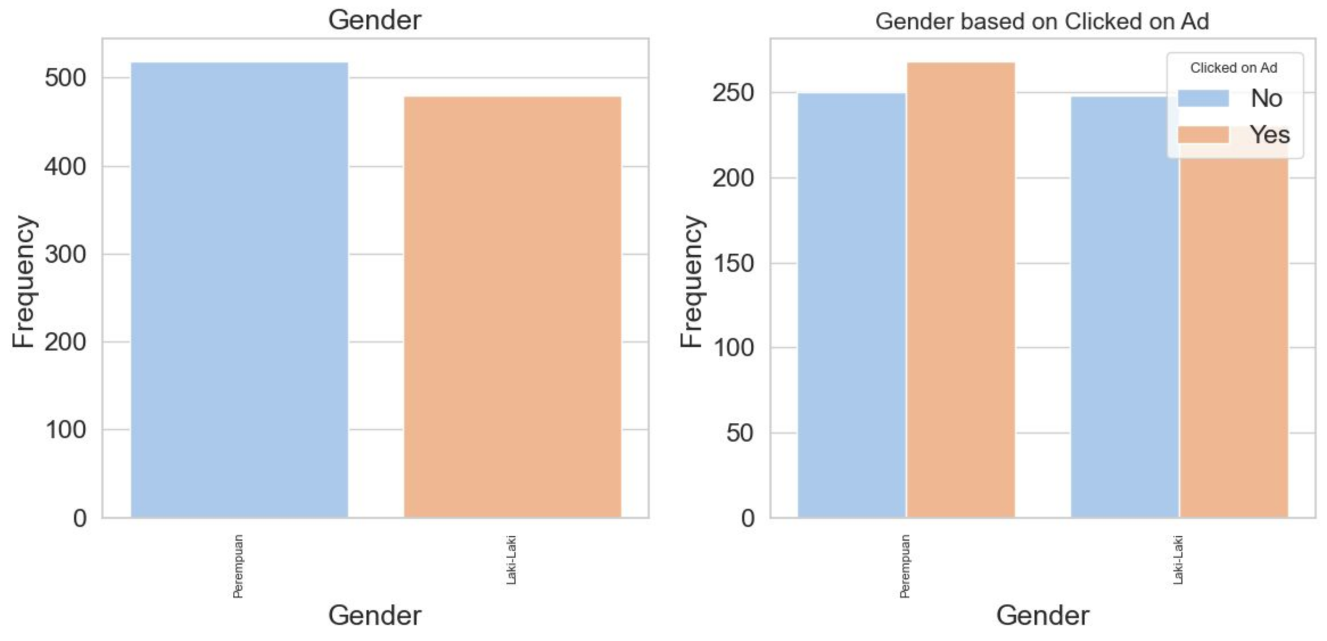
- The category most selected by the customers is Otomotif, followed by House, Health and Fashion.
- The least selected categories are Electronic, Finance, and Bank.
- Customers who clicked on the ad are mostly from category Otomotif, House, and Fashion.

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

Rakamin Academy

Univariate analysis - Male

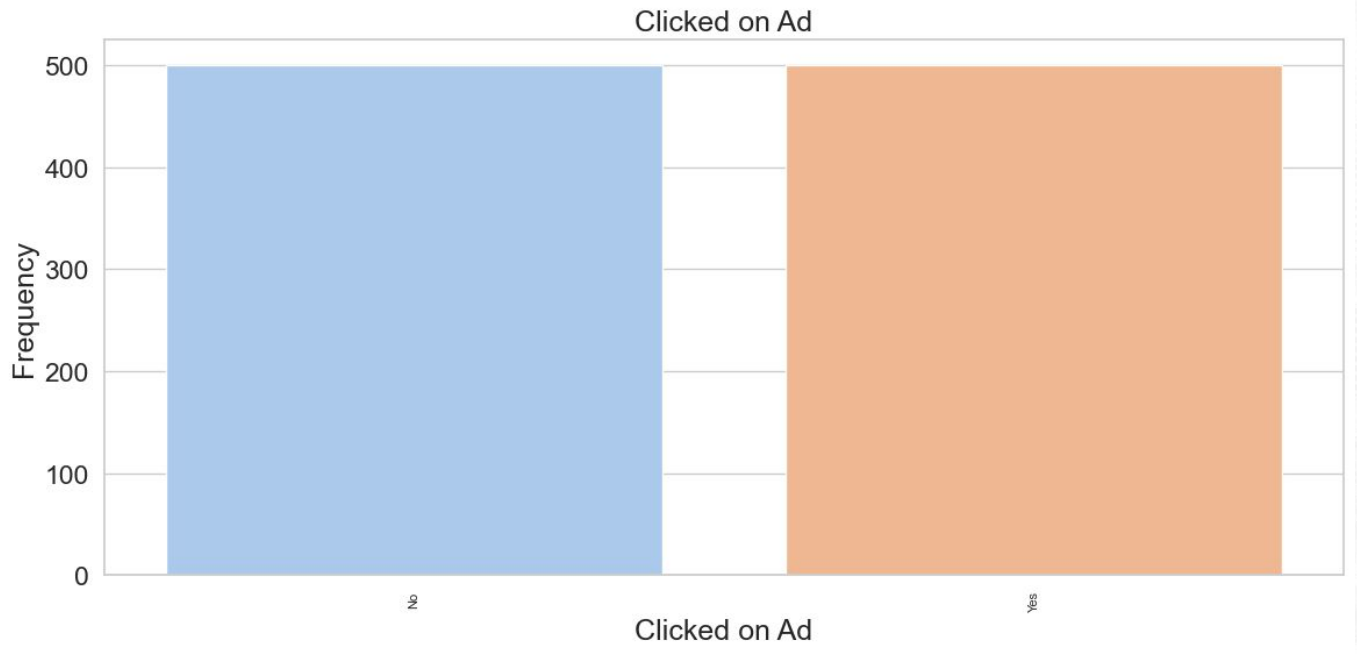


- Most of the customers are female and so are those who clicked on the ad

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

Univariate analysis - Clicked on Ad



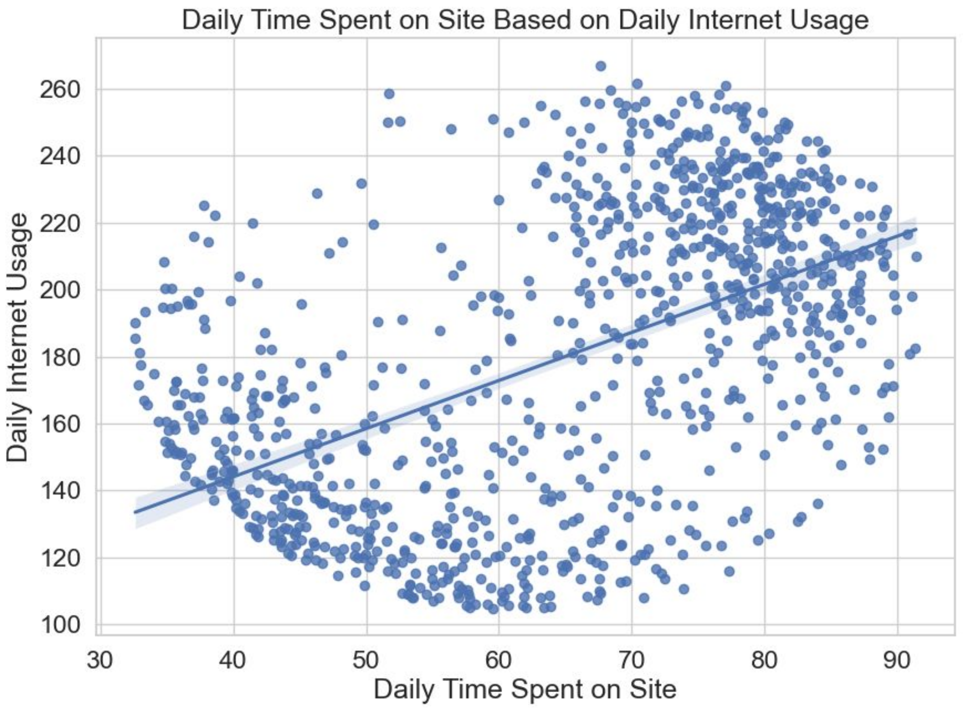
- The number of customers who clicked and didn’t click the ad are equal

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

Rakamin Academy

Bivariate analysis



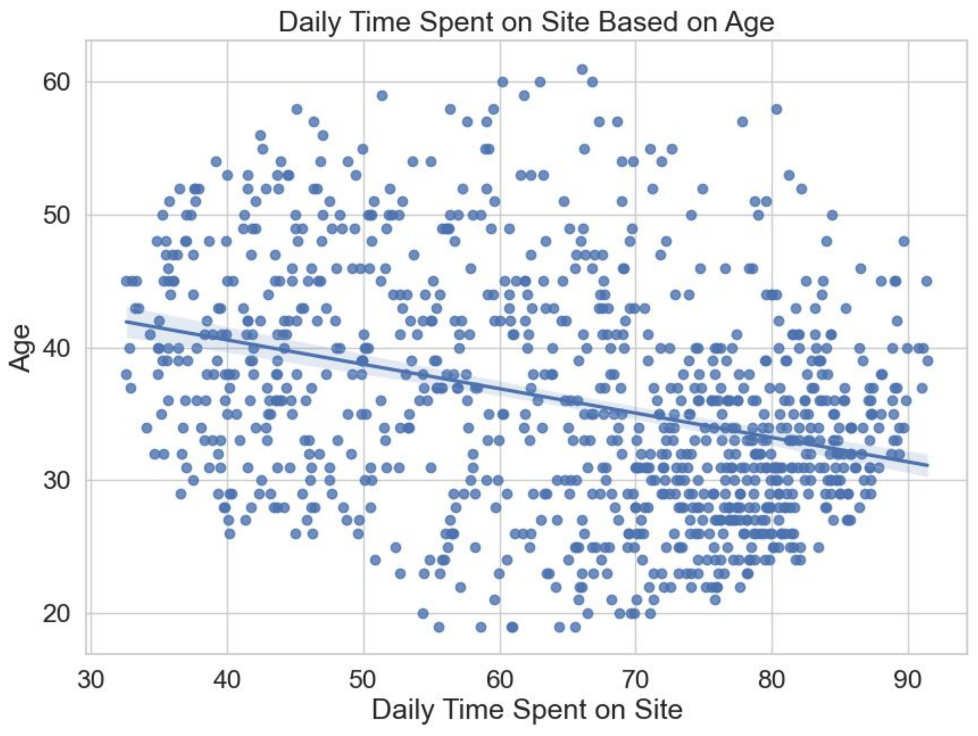
- The scatter plot shows positive relationship between "Daily Time Spent on Site" and "Daily Internet Usage"
- This implies that as the customer spent more time on the site, the daily internet usage also increases.

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

Rakamin Academy

Bivariate analysis



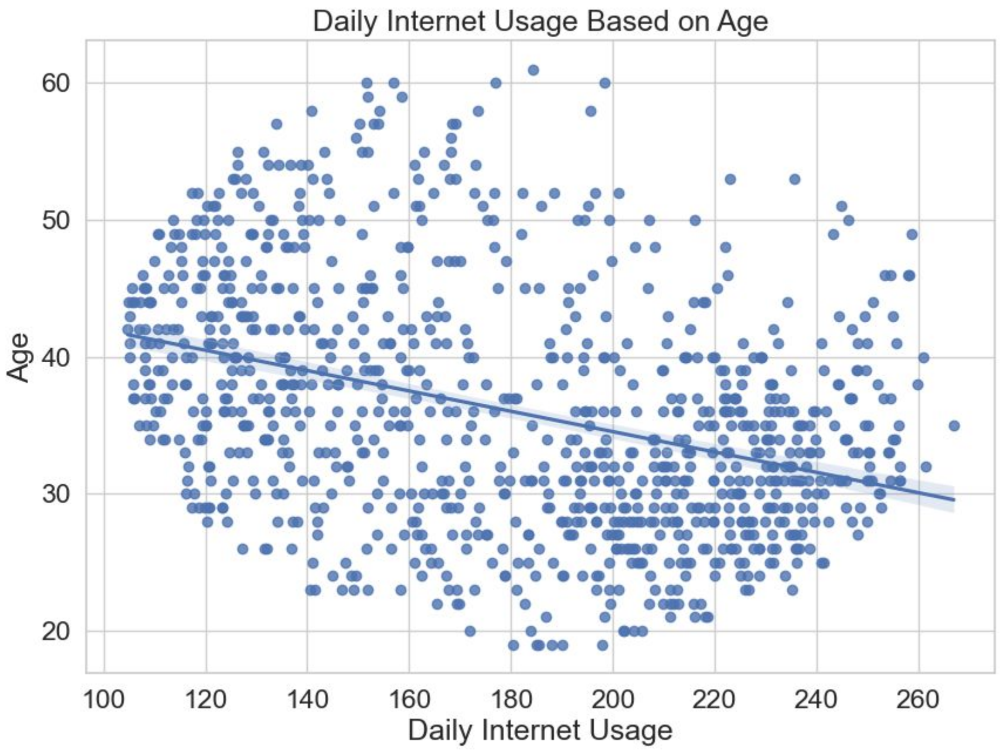
- The scatter plot shows negative relationship between "Daily Time Spent on Site" and "Age"
- This implies that the older the customer, the daily time spent on site is lesser

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

Rakamin Academy

Bivariate analysis

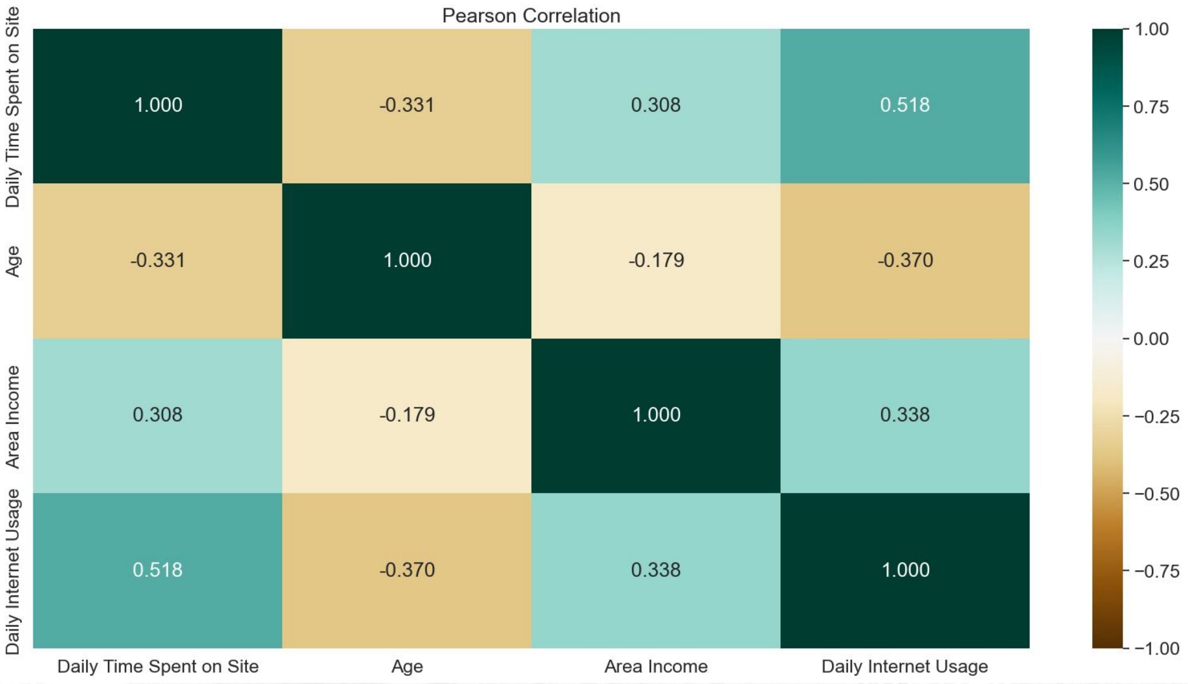


- The scatter plot shows positive relationship between "Daily Internet Usage" and “Age”
- This implies that as the age declines, the lesser daily internet usage

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

Multivariate analysis



- Strong positive correlation can be seen between “Daily Time Spent on Site” and “Daily Internet Usage” (r=0.518).
- Moderate positive correlation can be seen between “Daily Time Spent on Site” and “Area Income” (r=0.308). Customers with higher income tend to spend more time on site.
- Weak negative correlation can be seen between “Daily Time Spent on Site” and “Age” (r=-0.370)
- There is no significant association between “Age” and “Area Income” (r=-0.179).

For further details, please refer to this Jupyter Notebook