Linear Models Project Report

Symbiosis Statistical Institute

Submitted By:
Devanshu Gohil (22060641021)
Ashish Joshi (220060641028)
Pragati Dixit (220060641042)
Siddhant Desai (220060641014)

Customer Behavior Analysis

Table of Content

Predicting customer ad-clicks

- **Introduction**:

  Customer ad click prediction refers to the process of using machine learning algorithms to predict the probability of a customer clicking on an advertisement & involves analyzing various factors such as customer behavior, demographics, and other relevant data to predict the likelihood of a customer clicking on a particular ad.

  The goal of customer ad click prediction is to help advertisers optimize their ad campaigns by identifying the most effective ads and targeting strategies. By predicting which ads are most likely to be clicked on, advertisers can allocate their resources more efficiently. It is commonly used in digital advertising, such as display ads, social media ads, and search engine ads.

- **Business Problem Statement**

  Prediction of customer ad-clicks is dependent on various independent factors like the average engagement of the website where the ad is been displayed (i.e., Daily time spent on site), Time of the day and Day of the week, Relevance and Ad format these are some of the critical factors that can help us to draw insights.

  So in this particular project, by applying different models we are trying to analyze how can we achieve the overall goal of increasing the ad reach effectively?

- **Glimpse of Data**

  We have used two datasets for analysis

  Extracted the advertisement (1000 entries)

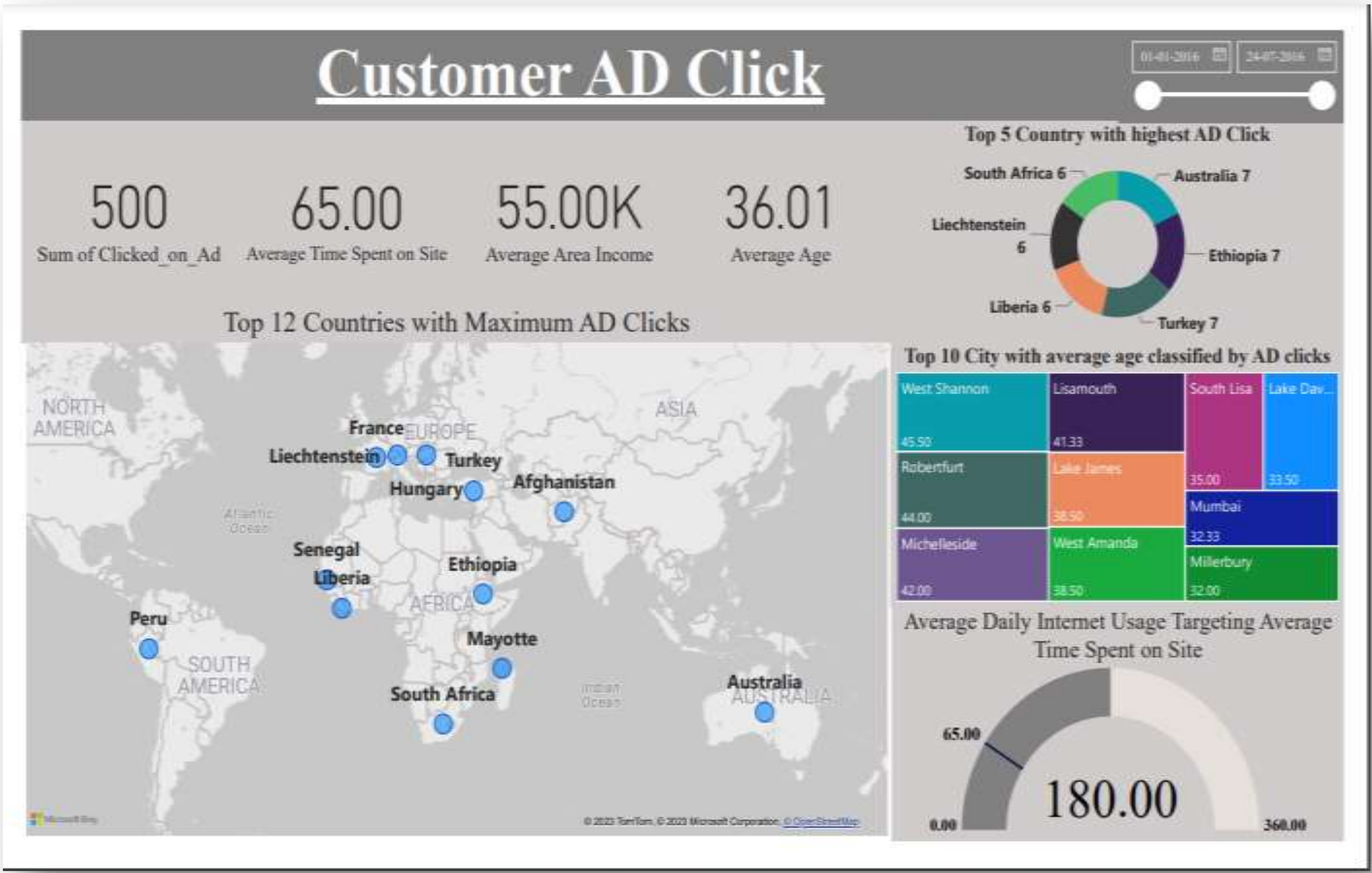
	A	B	C	D	E	F	G	H	I	J
1	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Male	Country	Timestamp	Clicked_on_Ad
2	68.95	35	61833.9	256.09	Cloned 5thgeneration orchestration	Wrightburgh	0	Tunisia	27-03-2016 00:53	0
3	80.23	31	68441.85	193.77	Monitored national standardization	West Jodi	1	Nauru	04-04-2016 01:39	0
4	69.47	26	59785.94	236.5	Organic bottom-line service-desk	Davidton	0	San Marin	13-03-2016 20:35	0
5	74.15	29	54806.18	245.89	Triple-buffered reciprocal time-fran	West Terrifurt	1	Italy	10-01-2016 02:31	0
6	68.37	35	73889.99	225.58	Robust logistical utilization	South Manuel	0	Iceland	03-06-2016 03:36	0
7	59.99	23	59761.56	226.74	Sharable client-driven software	Jamieberg	1	Norway	19-05-2016 14:30	0
8	88.91	33	53852.85	208.36	Enhanced dedicated support	Brandonstad	0	Myanmar	28-01-2016 20:59	0
9	66	48	24593.33	131.76	Reactive local challenge	Port Jefferybury	1	Australia	07-03-2016 01:40	1
10	74.53	30	68862	221.51	Configurable coherent function	West Colin	1	Grenada	18-04-2016 09:33	0
11	69.88	20	55642.32	183.82	Mandatory homogeneous architect	Ramirezton	1	Ghana	11-07-2016 01:42	0
12	47.64	49	45632.51	122.02	Centralized neutral neural-net	West Brandontc	0	Qatar	16-03-2016 20:19	1
13	83.07	37	62491.01	230.87	Team-oriented grid-enabled Local A	East Theresashi	1	Burundi	08-05-2016 08:10	0
14	69.57	48	51636.92	113.12	Centralized content-based focus gr	West Katiefurt	1	Egypt	03-06-2016 01:14	1
15	79.52	24	51739.63	214.23	Synergistic fresh-thinking array	North Tara	0	Bosnia anc	20-04-2016 21:49	0
16	42.95	33	30976	143.56	Grass-roots coherent extranet	West William	0	Barbados	24-03-2016 09:31	1
17	63.45	23	52182.23	140.64	Persistent demand-driven interface	New Travistowr	1	Spain	09-03-2016 03:41	1
18	55.39	37	23936.86	129.41	Customizable multi-tasking website	West Dylanberg	0	Palestiniar	30-01-2016 19:20	1
19	82.03	41	71511.08	187.53	Intuitive dynamic attitude	Pruittmouth	0	Afghanista	02-05-2016 07:00	0
20	54.7	36	31087.54	118.39	Grass-roots solution-oriented congl	Jessicastad	1	British Indi	13-02-2016 07:53	1

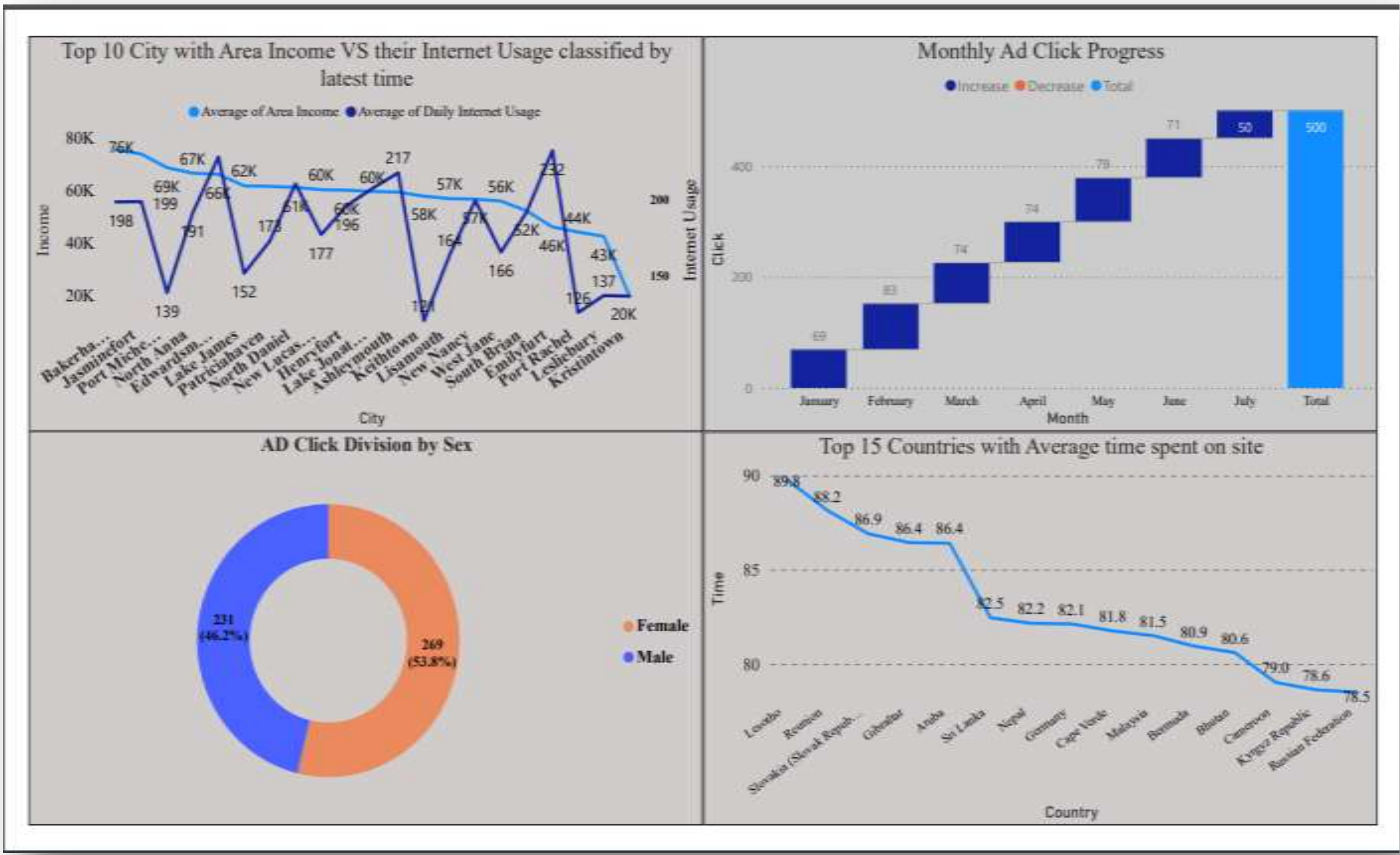
Column Name	Interpretation
Daily Time Spent on Site	Daily Time Spent on Site refers to the average amount of time that a user spends on a website in a day.
Age	Age of Customer
Area Income	Area Income refers to the average amount of income earned by individuals living in a particular geographic region
Daily Internet Usage	Daily Internet Usage refers to the amount of time that an individual spends using the internet on a daily basis, measured in minutes
Ad Topic Line	Ad Topic Line refers to the headline or title of an advertisement that is intended to capture the viewer's attention and encourage them to read further or take action.
City	City of Customer
Male	Gender of Person Given in Binary (0 and 1)
Country	Country of Customer
Clicked On Ad	Binary Values(0 and 1) describing whether customer has clicked on AD or not

- **Motive of Analysis**

  The motive of customer ad click analysis is to understand the behavior of customers who click on ads and to identify patterns and trends that can inform marketing strategies. This analysis can help businesses optimize their AD campaigns and improve the effectiveness of their marketing efforts. Ultimately, the goal is to increase the conversion rate and drive more sales. This analysis is done by Logistic Regression model. The analysis could identify which factors have the most significant impact on AD clicks and help analyst make data-driven decisions to improve their ad targeting campaign.

- **Data Visualizations using Power BI**





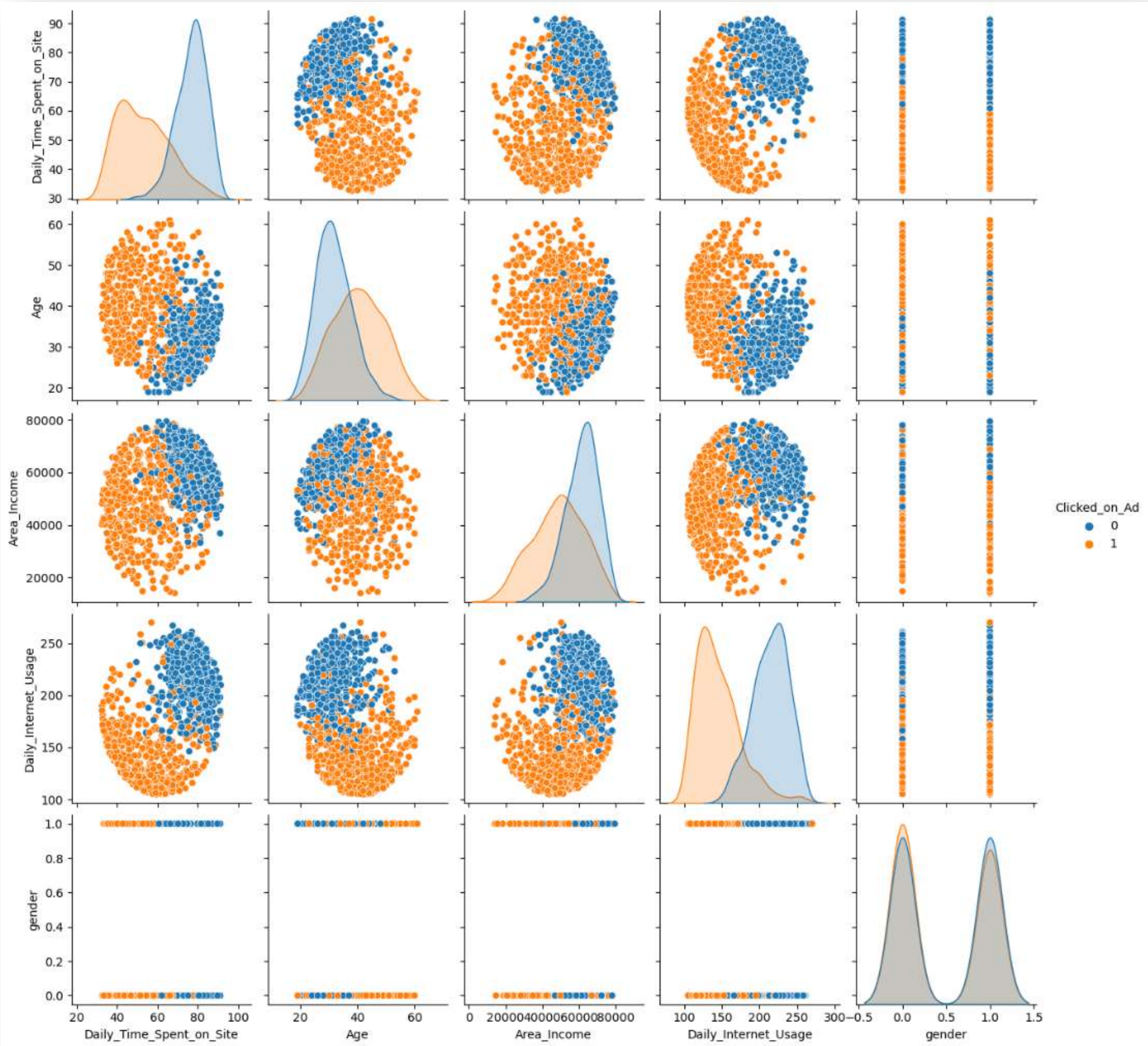
- **Exploratory Data Analysis – EDA**

  We have used function data.describe( ) to get the summary statistics of the numerical columns of a dataset. There is total 1000 observations, whereas average Daily time spent on site is 65 which is good as it's more than half of maximum value. Average age of people clicking on ads is 36 yrs.

	Daily_Time_Spent_on_Site	Age	Area_Income	Daily_Internet_Usage	gender	Clicked_on_Ad
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000
mean	65.000200	36.009000	55000.000080	180.000100	0.481000	0.50000
std	15.853615	8.785562	13414.634022	43.902339	0.499889	0.50025
min	32.600000	19.000000	13996.500000	104.780000	0.000000	0.00000
25%	51.360000	29.000000	47031.802500	138.830000	0.000000	0.00000
50%	68.215000	35.000000	57012.300000	183.130000	0.000000	0.50000
75%	78.547500	42.000000	65470.635000	218.792500	1.000000	1.00000
max	91.430000	61.000000	79484.800000	269.960000	1.000000	1.00000

1) Pair Plot

A pair plot is a type of data visualization that displays the pairwise relationships between multiple variables in a dataset. It is used to identify patterns and correlations between variables, as well as to observe the distribution of each variable individually. The plot consists of scatterplots for each pair of variables and histograms along the diagonal.



1) From above pair plot we can see that all age groups have clicked on ad but by spending less time.

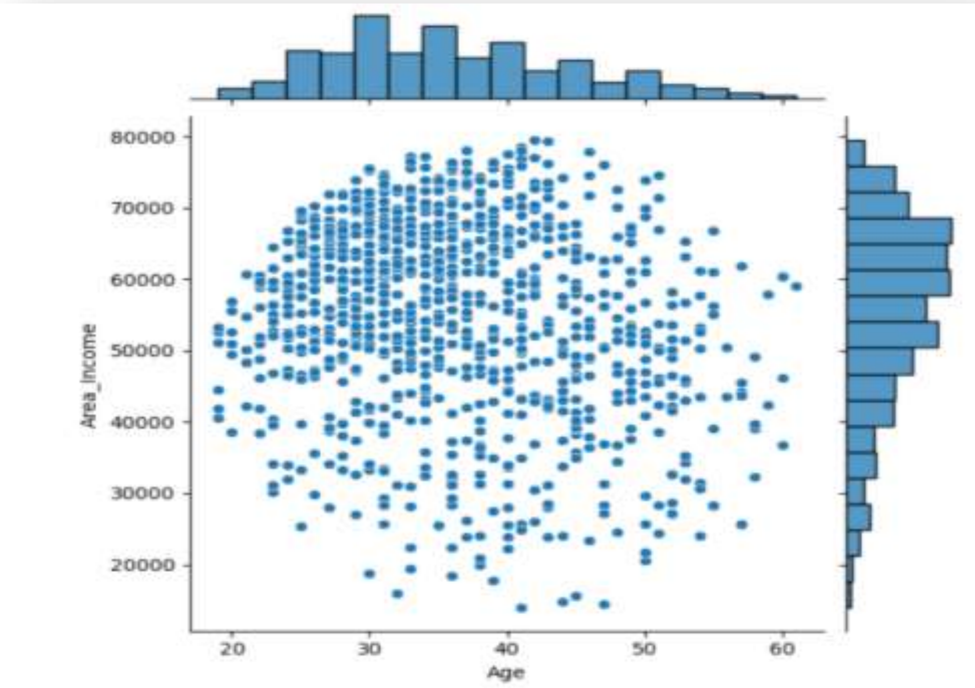
2) Similarly we can check that low income people with less internet usage have clicked on ad much larger as compare to others.

3)by looking at diagonal plots we can conclude that there is different skewness positive and negative depending on variable in context of ad click

2) Joint Plot

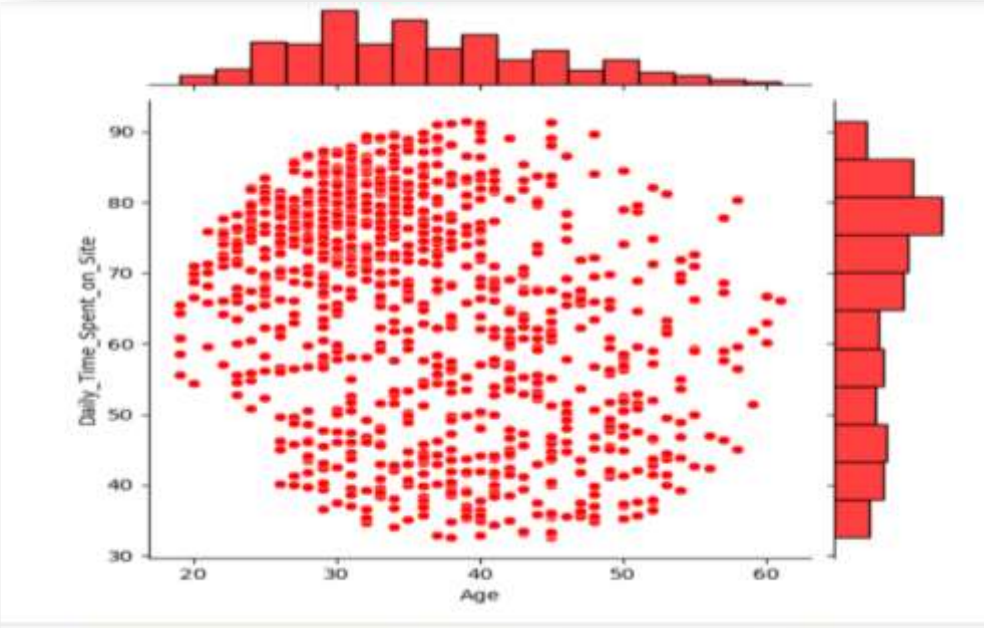
A joint plot is a type of data visualization that displays the relationship between two variables using both a scatterplot and a histogram. It is used to identify patterns and correlations between variables, as well as to observe the distribution of each variable individually. The plot consists of a scatterplot in the center and histograms along the x and y axes.

a) Joint plot for Area Income vs Age



A by looking at above plot we come to know that there are many youngsters living in high area income.

b) Joint plot for Daily Time Spent on site vs Age

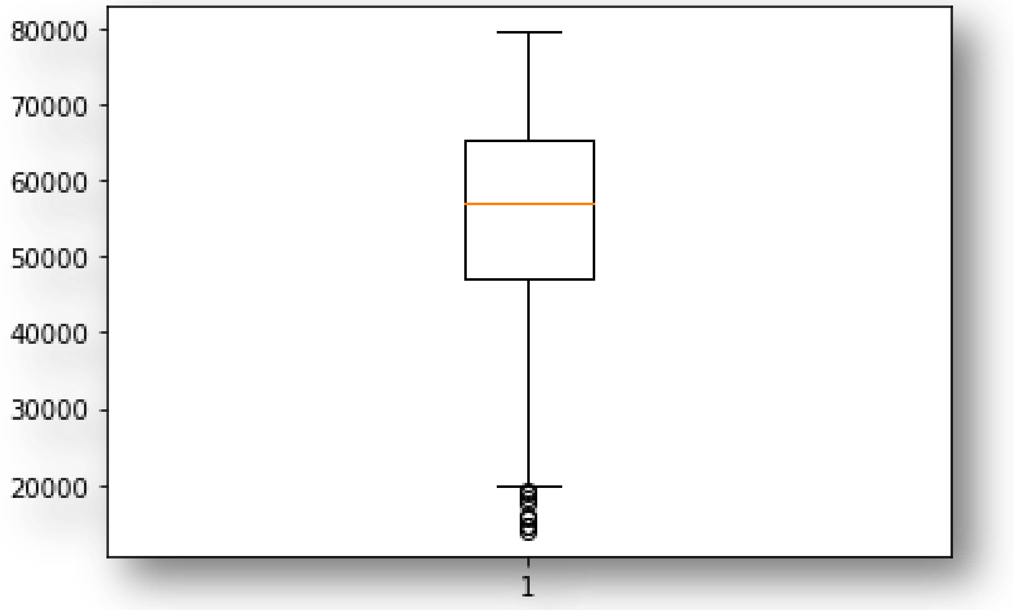


By above plot we can conclude youngsters spend much more time on the site as compared to old age people

3) Box Plot

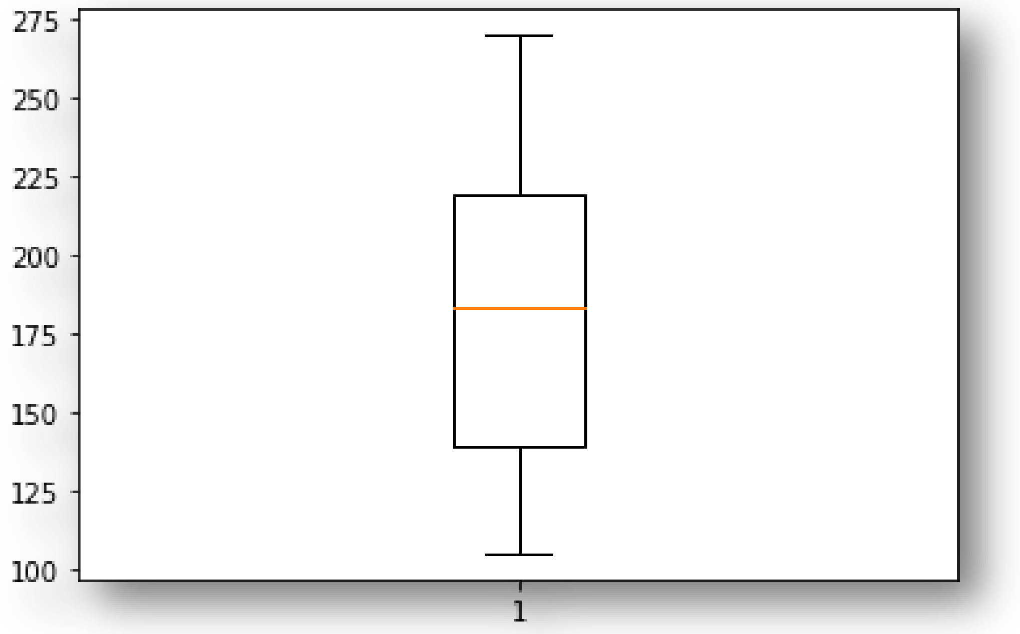
A box plot, also known as a box and whisker plot, is a graphical representation of the distribution of a dataset through its quartiles. It is used to display the range and variability of the data, as well as to identify outliers. The box in the plot represents the middle 50% of the data, with the median represented by a line in the box. The whiskers extend from the box to show the range of the data, and any outliers beyond the whiskers are shown as individual points.

a) Box plot for Area Income



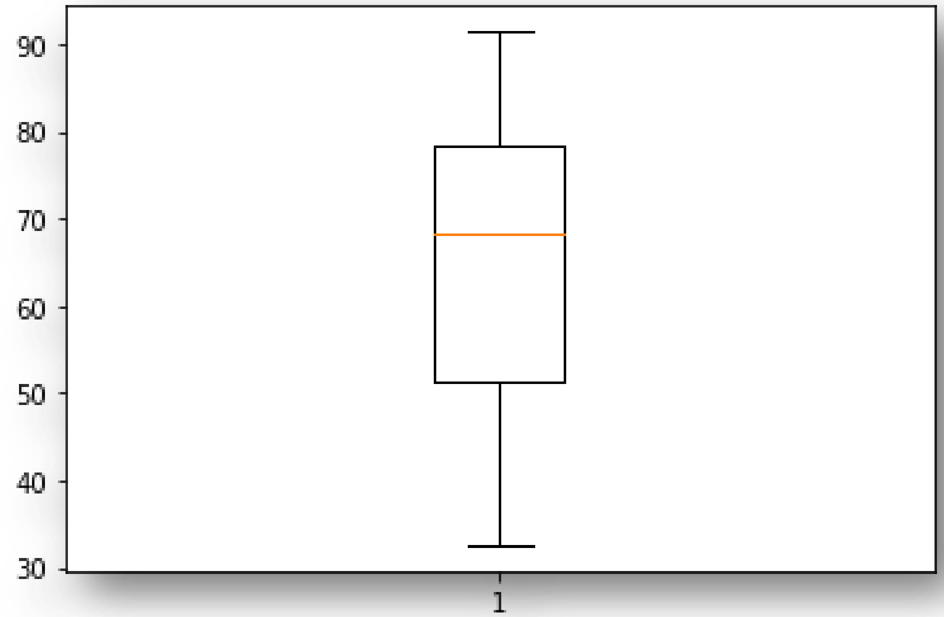
We can see that here there are some outliers as there are points below 1st Quartile.

b) Box plot for Daily Internet Usage



We can see that here there are no outliers as there are no points below 1st Quartile and above 3rd Quartile.

c) Box plot for Daily Time Spent on Site



We can see that here there are no outliers as there are no points below 1st Quartile and above 3rd Quartile.

4) Correlation Matrix

Function used here is data.corr( ) which give us the correlation coefficient between the all the possible combinations of variables in our dataset. Higher value (i.e. close to ±1) of correlation coefficient means higher the strength of the linear relationship between the variables.

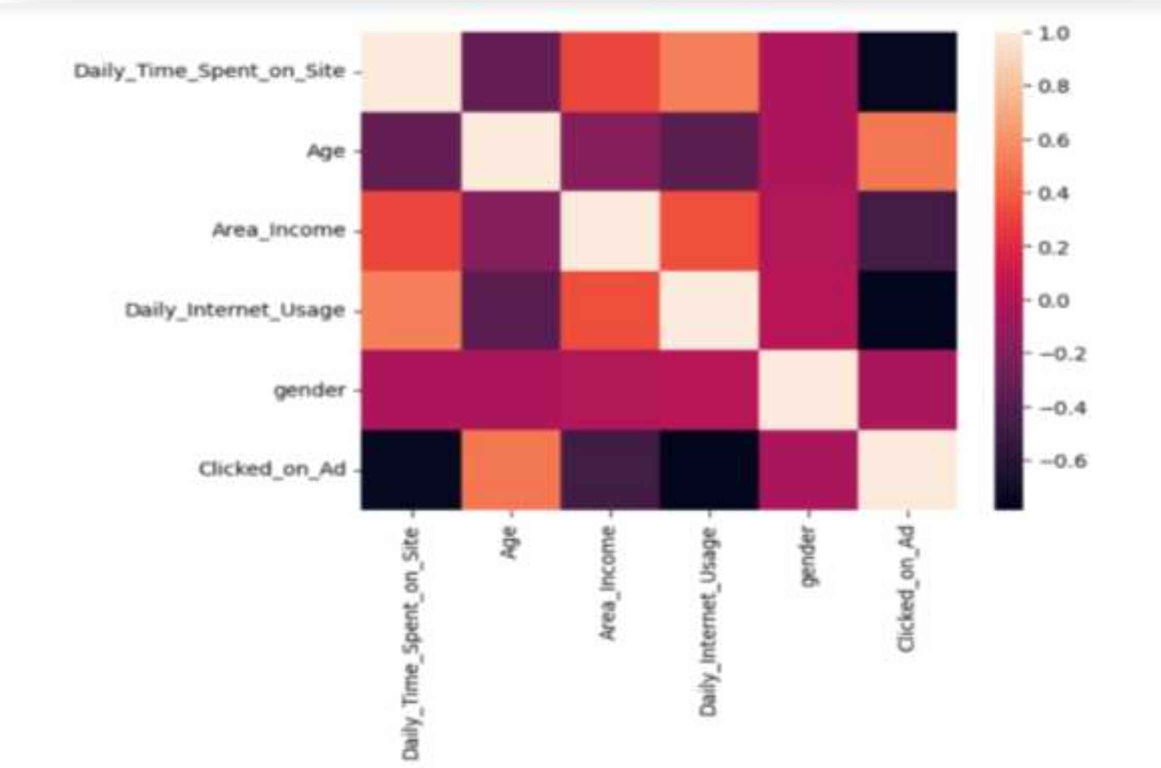
	Daily_Time_Spent_on_Site	Age	Area_Income	Daily_Internet_Usage	gender	Clicked_on_Ad
Daily_Time_Spent_on_Site	1.000000	-0.331513	0.310954	0.518658	-0.018951	-0.748117
Age	-0.331513	1.000000	-0.182605	-0.367209	-0.021044	0.492531
Area_Income	0.310954	-0.182605	1.000000	0.337496	0.001322	-0.476255
Daily_Internet_Usage	0.518658	-0.367209	0.337496	1.000000	0.028012	-0.786539
gender	-0.018951	-0.021044	0.001322	0.028012	1.000000	-0.038027
Clicked_on_Ad	-0.748117	0.492531	-0.476255	-0.786539	-0.038027	1.000000

Here we can see that for

- Daily Time Spent on Site have negative Correlation (-0.7481) with Clicked on Ad and have partial positive correlation (0.5186) with Daily Internet Usage.
- Age has partial positive correlation (0.4925) with Clicked on Ad.
- Area Income has partial negative correlation (-0.4762) with Clicked on Ad.
- Daily Internet Usage has negative correlation (-0.7865) with Clicked on Ad.

5) Correlation Heat map

A heat map is a graphical representation of data that uses a color-coding scheme to represent different values in a matrix. In a heat map, each cell in the matrix is assigned a color based on its value, with darker colors indicating higher values and lighter colors indicating lower values.



- ## Methodology (Logistic Regression)

For analysis of above data, and to meet the requirement of our problem statement we used two models in this:
We used Logistic Regression model (also known as logit model) for Dataset

Logistic regression, also known as the logit model, is a statistical method used to analyze and model the relationship between a binary (two-class) categorical response variable and one or more predictor variables. This logistic function is represented by the following formulas:

Log$\left(\frac{p}{1-p}\right)$= $\boldsymbol{\beta'}$**x**

Where **p=probability of success; odds =**$\frac{p}{1-p}$;

Log-odds= log$\left(\frac{p}{1-p}\right)$

Solving for μ,gives the logistic function:

$$\boldsymbol{\mu = \frac{1}{1 + e^{-\beta' x}}}$$

Here $\boldsymbol{\mu}$ = ***click-on ad*** which takes values 0 or 1

Whereas X_1= **Daily Time Spent on Site**

X_2=**Age**

X_3=**Area Income**

X_4=**Daily Internet Usage**

X_5= **Gender**

- ## Modeling & Results

Result obtained after apply the logit model to our variables of interest:
Model 1:
Clicked on Ad ~ Daily Time Spent on Site + Age + Area Income +Daily Internet Usage+ gender

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:           Clicked_on_Ad   No. Observations:                 1000
Model:                             GLM   Df Residuals:                      994
Model Family:                 Binomial   Df Model:                            5
Link Function:                   Logit   Scale:                          1.0000
Method:                           IRLS   Log-Likelihood:                -90.904
Date:                 Sat, 29 Apr 2023   Deviance:                       181.81
Time:                         17:26:15   Pearson chi2:                     806.
No. Iterations:                      9   Pseudo R-squ. (CS):             0.7002
Covariance Type:             nonrobust
==========================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept                 27.3606      2.736      9.999      0.000      21.997      32.724
Daily_Time_Spent_on_Site  -0.1927      0.021     -9.286      0.000      -0.233      -0.152
Age                        0.1709      0.026      6.607      0.000       0.120       0.222
Area_Income               -0.0001   1.88e-05     -7.245      0.000      -0.000   -9.93e-05
Daily_Internet_Usage      -0.0635      0.007     -9.390      0.000      -0.077      -0.050
gender                    -0.4217      0.404     -1.043      0.297      -1.214       0.371
==========================================================================================
```

- Covariance type refers to the method used to estimate the variance-covariance matrix of the parameter estimates in the GLM regression model. We use this matrix to estimate the standard errors of the coefficients and to test the statistical significance of the predictor variables. Covariance type non-robust assumes that the errors of the model are homoscedastic and have a normal distribution.
- Family= binomial family states us that the outputs are binary which is required for logistic regression.
- Df Residuals = 994, it means that the model was fit using 1000 observations, and 6 parameters were estimated (5 coefficients and 1 intercept). Therefore, the Df Residuals is calculated as the difference between the total number of observations (1000) and the number of parameters estimated (6), which is equal to 994.

- In a logistic regression model, the log-likelihood is a measure of the goodness of fit of the model to the data. Specifically, it is the logarithm of the likelihood function, which represents the probability of observing the data given the model parameters, a log-likelihood value of -90.904 means that the model is a good fit to the data. The log-likelihood value is negative, which is expected because the likelihood function is always less than or equal to 1.
- The intercept is 27.3606, which represents the log odds of the response variable (Clicked on Ad) when all predictor variables are equal to zero.
- In the given logistic regression results, the Pseudo R-squared value of 0.7002 indicates that the model explains 70.02% of the variability in the dependent variable, which is a relatively good fit.

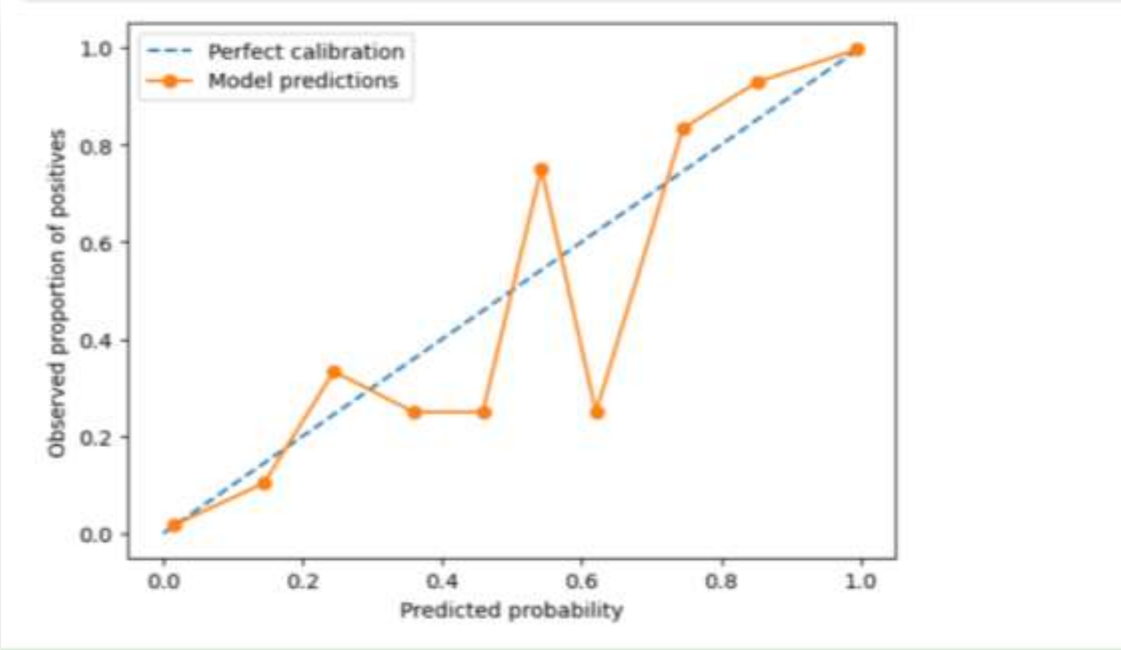
- **Confusion Matrix**

```
[[491   9]
 [ 19 481]]
```

In the given scenario, 491 cases were accurately identified as negative and confirmed as true negatives, while 9 cases were identified as positive but were actually negative, which were false positives. Moreover, 19 cases that were actually positive were predicted as negative, which were false negatives. Finally, 481 cases were correctly predicted as positive and were actually positive, and these were considered true positives.

- **Calibration curve**

A calibration curve is a plot that helps to evaluate the performance of a classification model by comparing the predicted probabilities to the observed proportions. The plot typically shows the predicted probabilities on the x-axis and the observed proportions on the y-axis.



Ideally, the points on the calibration curve should lie close to the diagonal line, which indicates perfect calibration. As we can see this is a poorly calibrated model, it has a curve that deviates from the diagonal line, indicating poor agreement between predicted probabilities and observed proportions.

Predicting Consumer's Ad Click in a Facebook Ad Campaign

- ## Introduction

  Customer ad click prediction refers to the process of using machine learning algorithms to predict the probability of a customer clicking on an advertisement & involves analyzing various factors such as customer behavior, demographics, and other relevant data to predict the likelihood of a customer clicking on a particular ad.

  The goal of customer ad click prediction is to help advertisers optimize their ad campaigns by identifying the most effective ads and targeting strategies. By predicting which ads are most likely to be clicked on, advertisers can allocate their resources more efficiently. It is commonly used in digital advertising, such as display ads, social media ads, and search engine ads.

- ## Business Problem Statement

  The dataset provided contains information about Facebook posts made by a cosmetics brand. The task is to analyse the data and provide insights on factors affecting the engagement of posts on the brand's Facebook page.

- ## Glimpse of Data

  The data used for the analysis is the number of clicks received by various Facebook ads during a specific campaign. The following table gives us a clear idea about the columns in the dataset.

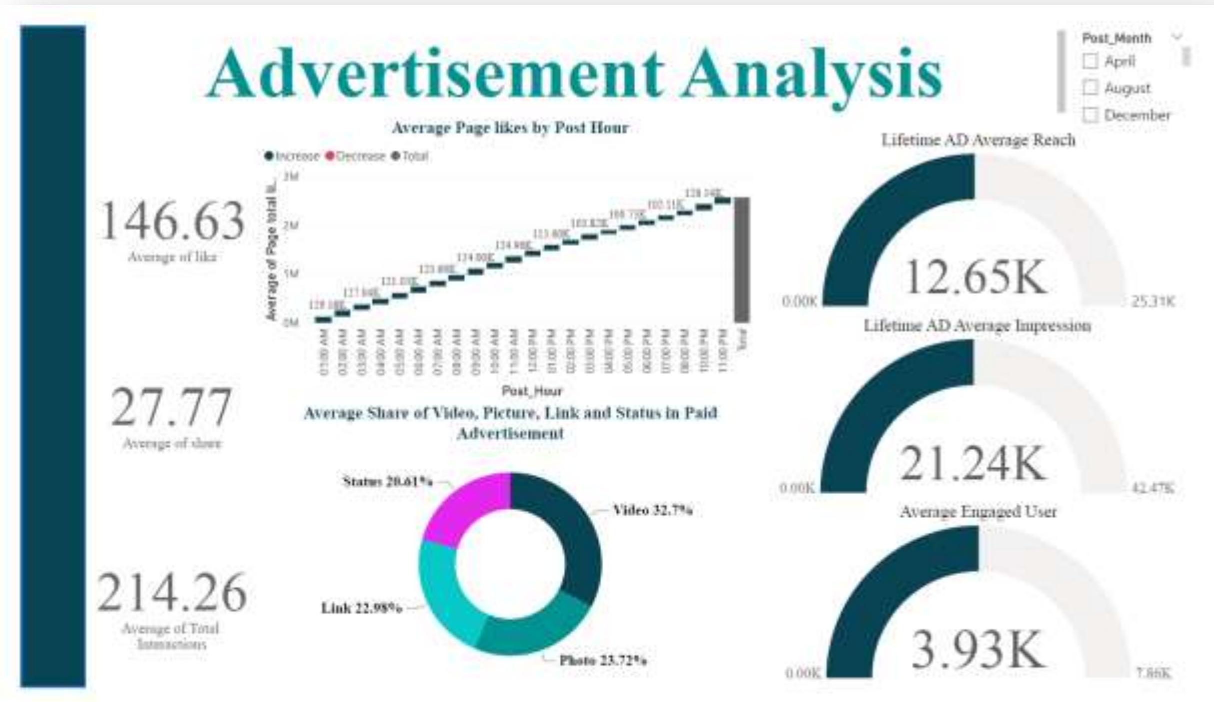
Column Name	Interpretation
Page total likes	The number of likes the Facebook page had at the time the post was made
Advertisement Type	The type of post that was made (photo, status, link, etc.)
Category	The category of the Facebook page
Post Month	The month in which the post was made
Post Weekday	The day of the week on which the post was made
Post Hour	The hour of the day at which the post was made
Paid	Whether the post was a paid promotion or not
Lifetime ad Total Reach	The number of unique Facebook users who saw the post
Lifetime ad Total Impressions	The total number of times the post was displayed to Facebook users
Lifetime Engaged Users	The number of Facebook users who clicked on the post, including likes, comments, and shares
Lifetime Post Consumers	The number of unique Facebook users who clicked on the post
Lifetime ad Consumptions	The total number of clicks on the post, including clicks on links and photos
Lifetime ad Impressions by people who have liked your Page	The total number of times the post was displayed to Facebook users who have liked the page
Lifetime ad reach by people who like your Page	The number of unique Facebook users who have liked the page and who saw the post
Lifetime People who have liked your Page and engaged with your post	The number of Facebook users who have liked the page and who engaged with the post
Comments	The number of comments on the post
Likes	The number of likes on the post
Share	The number of shares of the post
Total Interactions	The total number of likes, comments and shares on the post

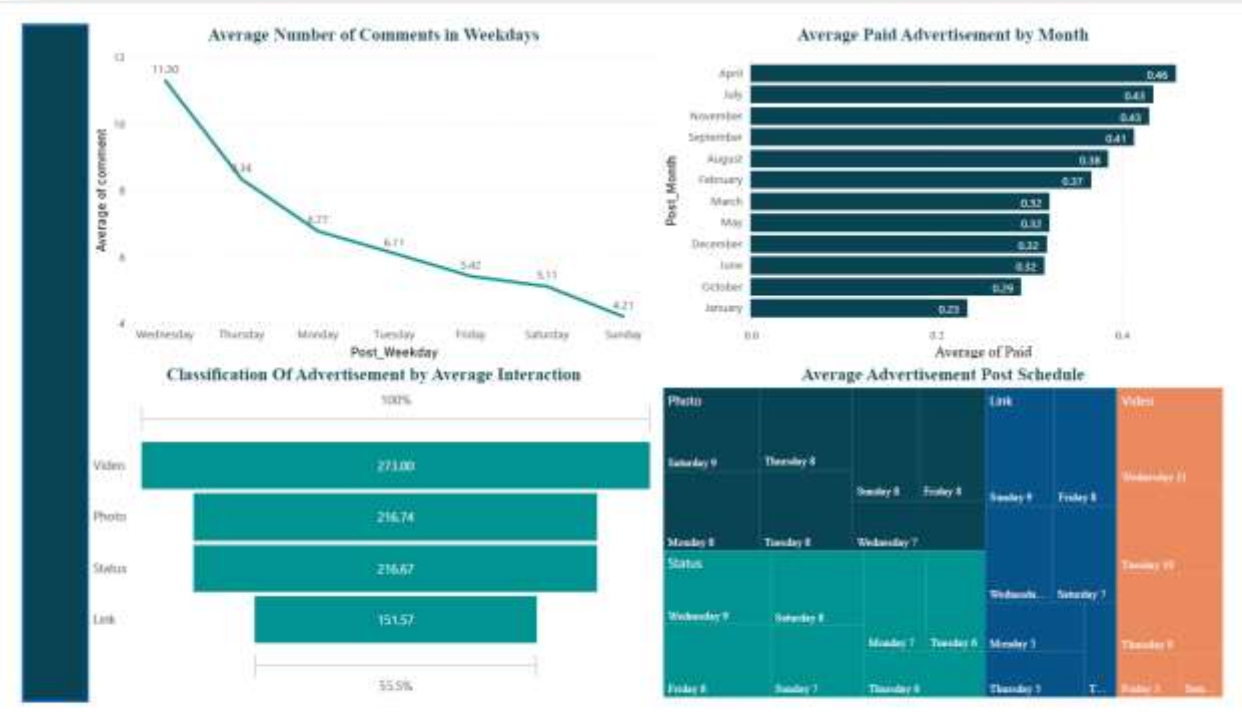
Page t...	Advert...	Categ...	Post M...	Post W...	Post H...	Paid	Lifetim...	Lifetime ad Total Impressi...	Lifetime Engaged Us...	Lifetime Post Consum...	Lifetim...	Lifetim...	Lifetim...	Lifetim...	comm...	like	share	Total (...)actions
139441	Photo	2	12	4	3	0	2752	5091	178	109	159	3078	1640	119	4	79	17	100
139441	Status	2	12	3	10	0	10460	19057	1457	1361	1674	11710	6112	1108	5	130	29	164
139441	Photo	3	12	3	3	0	2413	4373	177	113	154	2812	1503	132	0	66	14	80
139441	Photo	2	12	2	10	1	50128	87991	2211	790	1119	61027	32048	1386	58	1572	147	1777
139441	Photo	2	12	2	3	0	7244	13594	671	410	580	6228	3200	396	19	325	49	393
139441	Status	2	12	1	9	0	10472	20849	1191	1073	1389	16034	7852	1016	1	152	33	186
139441	Photo	3	12	1	3	1	11692	19479	481	265	364	15432	9328	379	3	249	27	279
139441	Photo	3	12	7	9	1	13720	24137	537	232	305	19728	11056	422	0	325	14	339
139441	Status	2	12	7	3	0	11844	22538	1530	1407	1692	15220	7912	1250	0	161	31	192
139441	Photo	3	12	6	10	0	4694	8668	280	183	250	4309	2324	199	3	113	26	142
139441	Status	2	12	5	10	0	21744	42334	4258	4100	4540	37849	18952	3798	0	233	19	252
139441	Photo	2	12	5	10	0	3112	5590	208	127	145	3887	2174	165	0	88	18	106
139441	Photo	2	12	5	10	0	2847	5133	193	115	133	3779	2071	152	0	90	14	104
139441	Photo	2	12	5	3	0	2549	4896	249	134	168	3631	1917	183	5	137	10	152
138414	Photo	2	12	4	5	1	22784	39941	887	337	417	34415	19511	684	2	577	20	599
138414	Status	2	12	3	10	0	10060	19680	1264	1209	1425	17272	8548	1162	4	86	18	108
138414	Photo	3	12	3	3	0	1722	2981	165	123	148	1868	1050	123	2	40	12	54
138414	Photo	1	12	2	12	1	53264	111785	1706	1103	1655	92512	39776	1307	15	678	20	713

- **Motive of Analysis**

  The main motive of analyzing the data is to understand the relationship between the independent variables (comments, likes, share, etc.) and the dependent variable (total interactions) in order to optimize Facebook ad campaigns for maximum clicks and return on investment (in terms of developing maximum impressions from the camp). This analysis is done by **Simple Linear Regression, Multiple Linear Regression and Lasso Regression** model. The analysis could identify which factors have the most significant impact on ad clicks and help analyst make data-driven decisions to improve their ad targeting and messaging campaign.

- **Data Visualizations using Power BI**





- # Exploratory Data Analysis – EDA

Using the **pd.read()** function, the dataset is initially uploaded to Python. The following output is generated by the **data.info()** function

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 746 entries, 0 to 745
Data columns (total 19 columns):
 #   Column                                                    Non-Null Count  Dtype
---  ------                                                    --------------  -----
 0   Page total likes                                          746 non-null    int64
 1   Advertisement Type                                        746 non-null    object
 2   Category                                                  746 non-null    int64
 3   Post Month                                                746 non-null    int64
 4   Post Weekday                                              746 non-null    int64
 5   Post Hour                                                 746 non-null    int64
 6   Paid                                                      746 non-null    int64
 7   Lifetime ad Total Reach                                   746 non-null    int64
 8   Lifetime ad Total Impressions                             746 non-null    int64
 9   Lifetime Engaged Users                                    746 non-null    int64
 10  Lifetime Post Consumers                                   746 non-null    int64
 11  Lifetime ad Consumptions                                  746 non-null    int64
 12  Lifetime ad Impressions by people who have liked your Page  746 non-null    int64
 13  Lifetime ad reach by people who like your Page            746 non-null    int64
 14  Lifetime People who have liked your Page and engaged with your post  746 non-null    int64
 15  comment                                                   746 non-null    int64
 16  like                                                      745 non-null    float64
 17  share                                                     745 non-null    float64
 18  Total Interactions                                        746 non-null    int64
dtypes: float64(2), int64(16), object(1)
memory usage: 110.9+ KB
```

From the output, we can interpret that there are 746 observations and 19 columns in the dataset. Except for the ***like*** and ***share*** columns, which contain one missing value, all columns have 746 observations.

1. Mean

It is a measure of central tendency that represents the average value of a set of numbers.

2. Median

It is a measure of central tendency that represents the middle value of a dataset when it is ordered from smallest to largest

3. Quartiles

Quartiles are a way to measure the spread and distribution of a dataset, and are often used in conjunction with box plots and other graphical representations of data. Quartiles are values that divide a dataset into four equal parts: -

Q1: The median of the lower half of the data

Q2: The median of the entire dataset

Q3: The median of the upper half of the data

4. Standard Deviation

It is a measure of the amount of variation or dispersion in a set of data.

Mean, median, quartiles and standard deviation is calculated by the data.describe() function. The following output is achieved

	Page total likes	Advertisement Type	Category	Post Month
count	746.000000	746.000000	746.000000	746.000000
mean	122447.950445	1.273458	1.930338	6.947721
std	16024.317033	0.723725	0.837903	3.326718
min	81370.000000	1.000000	1.000000	1.000000
25%	109670.000000	1.000000	1.000000	4.000000
50%	129600.000000	1.000000	2.000000	7.000000
75%	136642.000000	1.000000	3.000000	10.000000
max	139441.000000	4.000000	3.000000	12.000000

	Post Weekday	Post Hour	Paid	Lifetime ad Total Reach
count	746.000000	746.000000	746.000000	746.000000
mean	4.139410	7.809464	0.359249	12654.644772
std	2.057398	3.987223	0.480102	21119.033389
min	1.000000	1.000000	0.000000	238.000000
25%	2.000000	4.000000	0.000000	3486.000000
50%	4.000000	9.000000	0.000000	5830.000000
75%	6.000000	11.000000	1.000000	11700.250000
max	7.000000	23.000000	1.000000	180480.000000

	Lifetime ad Total Impressions	Lifetime Engaged Users
count	7.460000e+02	746.000000
mean	2.123657e+04	3931.018767
std	6.398916e+04	5389.489821
min	3.000000e+00	9.000000
25%	4.210250e+03	526.500000
50%	6.882000e+03	1007.000000
75%	1.267800e+04	5429.500000
max	1.110282e+06	19953.000000

5. Correlation Coefficient

Correlation is a statistical measure that describes the degree of association between two or more variables. It indicates the strength and direction of the linear relationship between two variables. A correlation coefficient is a value that ranges from -1 to +1, where -1 indicates perfect negative correlation, +1 indicates perfect positive correlation and 0 indicates no correlation.

	Total Interactions
Page total likes	0.031496
Advertisement Type	-0.041541
Category	0.100925
Post Month	0.021096
Post Weekday	-0.062144
Post Hour	-0.030622
Paid	0.102889
Lifetime ad Total Reach	0.459485
Lifetime ad Total Impressions	0.327865
Lifetime Engaged Users	0.103001
Lifetime Post Consumers	0.319876
Lifetime ad Consumptions	0.215403
Lifetime ad Impressions by people who have like...	0.244976
Lifetime ad reach by people who like your Page	0.542708
Lifetime People who have liked your Page and en...	0.458384
comment	0.841937
like	0.963758
share	0.882572

1) As we can see 'comment' has very high correlation with our dependent variable 'total interaction.

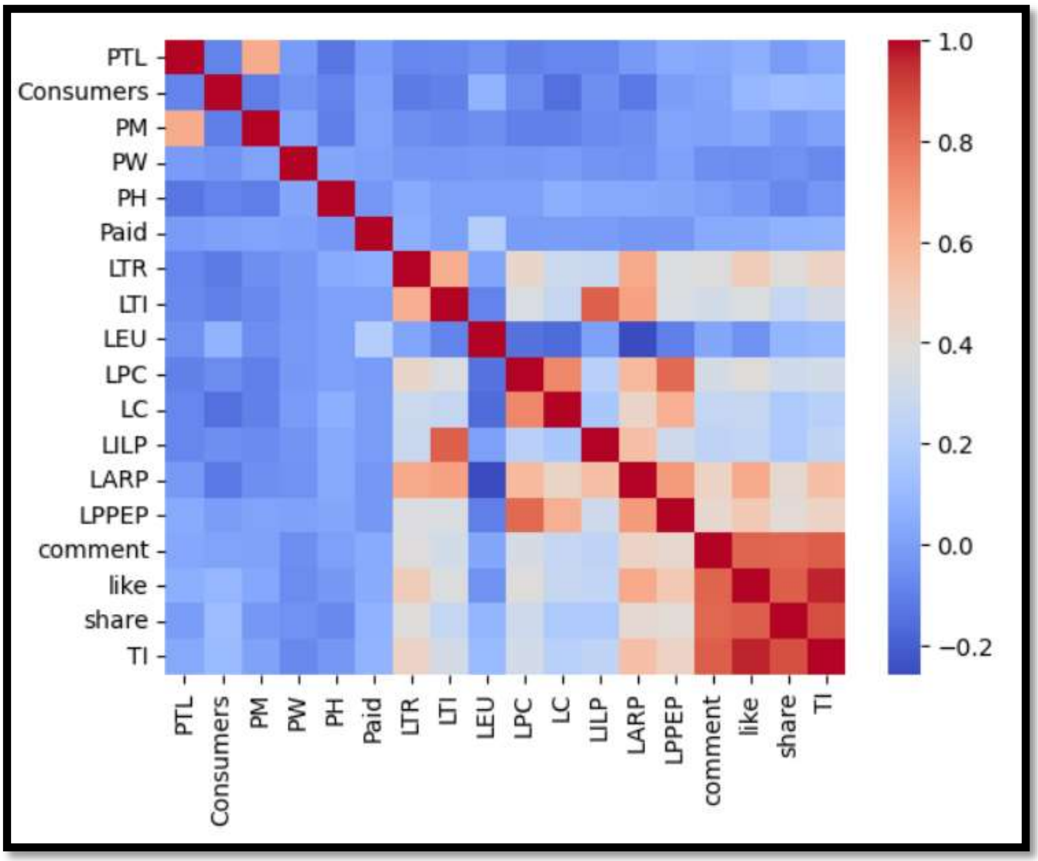
2) Similarly like and share also shows very high correlation with our dependent variable.

So now we can state that these variable explains a lot about our dependent variable.

3) Similarly other variables like lifetime people who have like your page and got engaged and lifetime ad reach by people also shows a good amount of relation with dependent variable.

6. Correlation Heatmap

It is a graphical representation of the correlation matrix, which shows the correlation coefficients between pairs of variables in a dataset. It is a useful tool for visualizing the strength and direction of the relationships between variables in a dataset.



7. Filling Missing Observations

Filling missing observations in data is important for several reasons

a. **To Prevent Bias**

If a significant number of observations are missing, it can result in biased results, especially if the missing data is not random. Filling in the missing data can help to reduce bias and provide more accurate results.

b. **To increase statistical power**

We can increase the statistical power of our analysis and increase the chances of detecting significant results.

c. **To improve accuracy**

If we have incomplete data, it can affect the accuracy of our analysis. Filling in the missing data can help us to get a more accurate estimate of the true value of a variable.

d. **To maintain sample size**

If we have a large amount of missing data, we may need to exclude certain observations from our analysis, which can reduce our sample size and decrease the reliability of our results. Filling in the missing data can help us to maintain a larger sample size and improve the reliability of our results.

➢ **Output**

```
Page total likes                                                    0
Advertisement Type                                                  0
Category                                                            0
Post Month                                                          0
Post Weekday                                                        0
Post Hour                                                           0
Paid                                                                0
Lifetime ad Total Reach                                             0
Lifetime ad Total Impressions                                       0
Lifetime Engaged Users                                              0
Lifetime Post Consumers                                             0
Lifetime ad Consumptions                                            0
Lifetime ad Impressions by people who have liked your Page          0
Lifetime ad reach by people who like your Page                      0
Lifetime People who have liked your Page and engaged with your post 0
comment                                                             0
like                                                                0
share                                                               0
Total Interactions                                                  0
dtype: int64
```

- **Methodology & Results**

1. Simple Linear Regression

Simple linear regression is a statistical method that models the relationship between two continuous variables by fitting a linear equation to the observed data. The goal of simple linear regression is to determine whether there is a significant relationship between the two variables, and to predict the value of one variable based on the value of the other. The equation for simple linear regression is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Where:

y is the dependent variable (response variable)

x is the independent variable (explanatory variable)

β0 is the intercept (the value of y when x is 0)

β1 is the regression coefficient (the change in y for a unit change in x)

ε is the error term (the difference between the predicted value of y and the actual value of y)

Once we have estimated the values of β0 and β1, we can use the regression equation to predict the value of y for any value of x. We can also use the regression equation to test whether there is a significant relationship between the two variables, by calculating the p-value associated with the regression coefficient β1. If the p-value is less than a predetermined significance level (usually 0.05), we can conclude that there is a significant relationship between the two variables.

Here, in this dataset y is Total Interactions. We have fitted two simple linear regression models. The independent variables are Like and Share columns.

The model fitted for the Share column is

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:     Total Interactions   R-squared (uncentered):                   0.849
Model:                            OLS   Adj. R-squared (uncentered):              0.849
Method:                 Least Squares   F-statistic:                              4191.
Date:                Fri, 28 Apr 2023   Prob (F-statistic):                   4.00e-308
Time:                        10:32:10   Log-Likelihood:                         -4790.1
No. Observations:                 746   AIC:                                      9582.
Df Residuals:                     745   BIC:                                      9587.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
share          7.7433      0.120     64.737      0.000       7.508       7.978
==============================================================================
Omnibus:                      401.718   Durbin-Watson:                   1.920
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5502.432
Skew:                           2.105   Prob(JB):                         0.00
Kurtosis:                      15.621   Cond. No.                         1.00
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Following results can be interpreted from the above output: -

i) The R-squared value of 0.849 indicates that 84.9% of the variation in Total Interactions can be explained by the share variable. The adjusted R-squared value is also 0.849, which indicates that there is no penalty for adding the independent variable share to the model.

ii) The coefficient for the share variable is 7.7433, which indicates that for every one unit increase in share, the Total Interactions is expected to increase by 7.7433 units. The standard error for this coefficient is 0.120, which indicates the precision of this estimate.

iii) The t-value for the share variable is 64.737, with a p-value of 0.000. This indicates that the share variable is highly statistically significant and is likely to have a true effect on the Total Interactions.

iv) The output provides additional information about the goodness of fit of the model, including the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values. These values can be used to compare different models and select the best one based on their fit and complexity.

v) High values of Jarque-bera statistics states that there is normality in our data. And 1.9 value of Durbin-watson shows that there is no autocorrelation in our data.

Similarly, we can interpret by fitting the Like column: -

```
                             OLS Regression Results
=======================================================================================
Dep. Variable:     Total Interactions   R-squared (uncentered):                  0.938
Model:                            OLS   Adj. R-squared (uncentered):             0.938
Method:                 Least Squares   F-statistic:                         1.124e+04
Date:                Fri, 28 Apr 2023   Prob (F-statistic):                       0.00
Time:                        10:28:45   Log-Likelihood:                        -4459.1
No. Observations:                 746   AIC:                                     8920.
Df Residuals:                     745   BIC:                                     8925.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
like           1.2101      0.011    106.039      0.000       1.188       1.233
==============================================================================
Omnibus:                      163.528   Durbin-Watson:                   0.871
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              301.069
Skew:                           1.293   Prob(JB):                     4.20e-66
Kurtosis:                       4.731   Cond. No.                         1.00
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

i) The R-squared value is 0.938, indicating that 93.8% of the variation in Total Interactions can be explained by the variation in likes.

ii) The coefficient of the independent variable (like) is 1.2101, which means that for every one-unit increase in likes, Total Interactions is expected to increase by 1.2101 units.

iii) The p-value for the coefficient is less than 0.05, indicating that the relationship between the two variables is statistically significant.

iv) The standard error for the coefficient is 0.011. This suggests that the estimate is precise.

v) Here also there is no autocorrelation and presence of normality in our data by looking at Durbin Watson and Jarque-bera values

2. Multiple Linear Regression

Multiple linear regression is a statistical method used to examine the relationship between a dependent variable and multiple independent variables. It assumes a linear relationship between the dependent variable and the independent variables, and the goal is to identify which independent variables are most strongly associated with the dependent variable.

The formula for multiple linear regression can be expressed as:

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon}$$

Where:

Y is the dependent variable (response variable)

X1 X2... Xn are the independent variables (explanatory variables)

β0 is the intercept (constant)

β1, β2, ..., βn are the regression coefficients for each independent variable

ε is the error term

The coefficients β1, β2, ..., βn represent the change in Y for a one-unit change in each respective independent variable while holding all other variables constant. The regression model estimates the values of the coefficients based on the data, and the goal is to find the values of β0, β1, β2, ..., βn that provide the best fit to the data.

```
==============================================================================
Dep. Variable:     Total Interactions   R-squared (uncentered):           0.970
Model:                            OLS   Adj. R-squared (uncentered):      0.969
Method:                 Least Squares   F-statistic:                      1368.
Date:                Sun, 30 Apr 2023   Prob (F-statistic):                0.00
Time:                        13:28:01   Log-Likelihood:                 -4196.8
No. Observations:                 746   AIC:                              8428.
Df Residuals:                     729   BIC:                              8506.
Df Model:                          17
Covariance Type:            nonrobust
===========================================================================================
========
                                                                        coef    std err          t      P>|t|      [0.025
0.975]
-------------------------------------------------------------------------------------------
--------
Advertisement Type                                                    5.0660      3.342      1.516      0.130      -1.496
11.628
Category                                                              5.9837      2.622      2.282      0.023       0.835
11.132
Post Month                                                           -1.1542      0.678     -1.702      0.089      -2.486
0.177
Post Weekday                                                          0.6063      1.135      0.534      0.594      -1.623
2.835
Post Hour                                                             0.6829      0.576      1.186      0.236      -0.447
1.813
Paid                                                                  5.8247      5.287      1.102      0.271      -4.555
16.204
Lifetime ad Total Reach                                               0.0016      0.000      7.028      0.000       0.001
0.002
Lifetime ad Total Impressions                                        -0.0006      0.000     -4.995      0.000      -0.001
-0.000
Lifetime Engaged Users                                                0.0055      0.001      9.925      0.000       0.004
0.007
Lifetime Post Consumers                                              -0.0182      0.007     -2.576      0.010      -0.032
-0.004
Lifetime ad Consumptions                                             -0.0016      0.002     -0.749      0.454      -0.006
0.003
Lifetime ad Impressions by people who have liked your Page            0.0007      0.000      5.531      0.000       0.000
0.001
Lifetime ad reach by people who like your Page                       -0.0039      0.001     -4.717      0.000      -0.006
-0.002
Lifetime People who have liked your Page and engaged with your post   0.0227      0.010      2.192      0.029       0.002
0.043
comment                                                               1.4279      0.293      4.878      0.000       0.853
2.003
like                                                                  0.9627      0.025     39.279      0.000       0.915
1.011
share                                                                 0.9844      0.152      6.474      0.000       0.686
1.283
==============================================================================
Omnibus:                      140.937   Durbin-Watson:                    2.115
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               557.416
Skew:                           0.830   Prob(JB):                     9.09e-122
Kurtosis:                       6.896   Cond. No.                      1.78e+05
==============================================================================
```

i) R-squared value is 0.970, indicating that the model explains 97% of the variance in the dependent variable.
ii) Standard Errors assume that the covariance matrix of the errors is correctly specified.
iii) The condition number is large, 1.78e+05. This might indicate that there are strong multicollinearity or other numerical problems.
If we have strong multicollinearity in this model then we lasso regression model
iv) but values of Durbin-watson and jarque-bera clearly states that there is no autocorrelation and normality assumption is satisfied.

3. Lasso Regression Model

Lasso regression is a linear regression technique that uses L1 regularization to shrink the coefficients of the input features towards zero, effectively performing feature selection and preventing overfitting. The L1 regularization penalty is defined as the absolute value of the sum of the coefficients. The lasso regression model can be written as:

$y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta nxn + \varepsilon$

where y is the dependent variable, x1, x2, ..., xn are the independent variables, β0, β1, β2, ..., βn are the regression coefficients, and ε is the error term.

The lasso regression model seeks to minimize the following objective function:

min $||y - X*\beta||2 + ||\lambda*\beta||$

where,

X is the matrix of input features

β is the vector of regression coefficients

|| || denotes the L1 norm

λ is the regularization parameter that controls the strength of the regularization penalty.

```
Selected features: ['like', 'comment', 'Lifetime People who have liked your Page and engaged with your post', 'Advertisement Ty
pe', 'Post Month', 'Lifetime ad Total Impressions', 'Lifetime Engaged Users', 'Lifetime ad Total Reach']
MSE: 5244.865643012645
R^2: 0.947595192087289
```

➢ **Variance Inflating Factor (VIF):**

Variance Inflation Factor (VIF) is a measure of multicollinearity in a linear regression model. It measures the degree to which the variance of the estimated regression coefficients is increased due to the presence of correlated predictor variables. The VIF for a given predictor variable is calculated as:

$$\mathbf{VIF = 1 / (1 - R^2)}$$

Where, R2 is the coefficient of determination obtained by regressing the predictor variable on all other predictor variables. The VIF value ranges from 1 upwards, with a value of 1 indicating no multicollinearity (i.e., no correlation between the predictor variable and the other predictor variables), and higher values indicating increasing levels of multicollinearity.

In general, a VIF value greater than 5 or 10 is considered to indicate problematic levels of multicollinearity, although the specific threshold may depend on the context and goals of the analysis.

```
        vif                                           features
0  5.057131                                               like
1  3.709574                                            comment
2  2.690483  Lifetime People who have liked your Page and e...
3  3.304676                                 Advertisement Type
4  3.103155                                         Post Month
5  1.903999                     Lifetime ad Total Impressions
6  1.489002                            Lifetime Engaged Users
7  2.607751                            Lifetime ad Total Reach
```

From the above output we can conclude: -

When we include like variable, we obtain VIF value 5.05, which is slightly higher than 5 but explains a lot about the model, i.e., it enhances accuracy to 94.7%, therefore we will consider it as our ideal match for our response variable. If we include share variable, the accuracy reduces to 85% while retaining all VIF values.

Now, we will predict Total Interactions column by data.predict() function

Post kday	Post Hour	Paid	Lifetime ad Total Reach	Lifetime ad Total Impressions	Lifetime Engaged Users	Lifetime Post Consumers	Lifetime ad Consumptions	Lifetime ad Impressions by people who have liked your Page	Lifetime ad reach by people who like your Page	Lifetime People who have liked your Page and engaged with your post	comment	like	share	Total Interactions	predicted
2	13	0	44464	66824	1052	930	1571	22904	14080	559	4	154.0	30.0	188	203.087123
2	12	0	2718	4698	566	528	663	3601	1992	306	0	50.0	10.0	60	80.528694
5	4	0	9703	5379	2664	439	155	12667	592	380	3	63.0	22.0	211	123.018587
5	3	1	11608	15323	985	705	940	8419	5840	594	4	330.0	29.0	363	375.829333
4	1	0	5568	10282	746	545	867	5696	3162	537	13	319.0	55.0	387	378.052669
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7	10	1	3934	6330	512	437	599	5010	3082	384	3	113.0	17.0	133	146.962250
6	6	0	2812	4954	536	485	672	3382	1853	323	4	79.0	16.0	99	118.986474
7	11	0	3558	5396	621	568	775	3708	2392	403	0	78.0	16.0	94	106.982541
1	7	1	6327	5921	7657	330	54	1589	586	489	1	100.0	42.0	313	182.857872
4	3	1	7968	13023	206	158	223	6734	3492	138	4	57.0	10.0	71	102.519722

Conclusion:

1) As we can see our model is best fit with lasso regression which gives the accuracy of 94.7% and removes a multicollinearity issue from our model.

2) We have also checked our model for heteroscedasticity where we got the results as
Lagrange multiplier statistic = 174.282972828887
p-value = 3.136268375303688e-34
Which shows that the variances are not constant.

3) To tackle the issue of heteroscedasticity and make our model an ideal best fit we should use weighted least square estimation.

4) Finally we have got a model with 94.7% accuracy , no autocorrelation , no outliers and model which satisfies normality condition of residuals so we have a good fit mode.