

# Conversion Rate Optimization for XYZ Company's Advertising Campaign

**NAME - Kartik Dhiman**

**[Follow me!](#)**



**Introduction:** This report presents the findings of an analysis of a dataset related to an advertising campaign aimed at promoting a product. The dataset includes information such as ad\_id, campaign\_id, age, gender, interest, impressions, clicks, spent, total\_conversion, and approved\_conversion. The objective of this analysis was to gain insights from the dataset and build a predictive model to predict whether an ad click will lead to a conversion or not.

---

**Dataset:** The provided dataset contains information about each ad, including: [Link](#)

```
df.head()
```

	ad_id	xyz_campaign_id	fb_campaign_id	age	gender	interest	Impressions	Clicks	Spent	Total_Conversion	Approved_Conversion
0	708746	916	103916	30-34	M	15	7350	1	1.43	2	1
1	708749	916	103917	30-34	M	16	17861	2	1.82	2	0
2	708771	916	103920	30-34	M	20	693	0	0.00	1	0
3	708815	916	103928	30-34	M	28	4259	1	1.25	1	0
4	708818	916	103928	30-34	M	28	4133	1	1.29	1	1

This data is about a company's social media ad campaign. It contains different factors considered for social media campaigning. We have a total of 1143 rows and 11 columns.

- ad\_id: a unique ID for each ad.
- xyz\_campaign\_id: an ID associated with each ad campaign of XYZ company.
- fb\_campaign\_id: an ID associated with how Facebook tracks each campaign.
- age: age of the person to whom the ad is shown.
- gender: gender of the person to whom the ad is shown.
- interest: a code specifying the category to which the person's interest belongs (interests are as mentioned in the person's Facebook public profile).
- Impressions: the number of times the ad was shown.
- Clicks: number of clicks on for that ad.
- Spent: Amount paid by XYZ company to Facebook, to show that ad.
- Total\_conversion: Total number of people who inquired about the product after seeing the ad.
- Approved\_conversion: Total number of people who bought the product after seeing the ad.

## Descriptive Statistics:

```
df.describe()
```

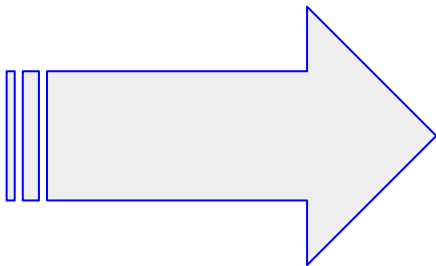
	ad_id	xyz_campaign_id	fb_campaign_id	interest	Impressions	Clicks	Spent	Total_Conversion	Approved_Conversion
count	1.143000e+03	1143.000000	1143.000000	1143.000000	1.143000e+03	1143.000000	1143.000000	1143.000000	1143.000000
mean	9.872611e+05	1067.382327	133783.989501	32.766404	1.867321e+05	33.390201	51.360656	2.855643	0.944007
std	1.939928e+05	121.629393	20500.308622	26.952131	3.127622e+05	56.892438	86.908418	4.483593	1.737708
min	7.087460e+05	916.000000	103916.000000	2.000000	8.700000e+01	0.000000	0.000000	0.000000	0.000000
25%	7.776325e+05	936.000000	115716.000000	16.000000	6.503500e+03	1.000000	1.480000	1.000000	0.000000
50%	1.121185e+06	1178.000000	144549.000000	25.000000	5.150900e+04	8.000000	12.370000	1.000000	1.000000
75%	1.121804e+06	1178.000000	144657.500000	31.000000	2.217690e+05	37.500000	60.025000	3.000000	1.000000
max	1.314415e+06	1178.000000	179982.000000	114.000000	3.052003e+06	421.000000	639.949998	60.000000	21.000000

We can understand from this following table that average of 2.85 people who enquired about each ad and we have maximum 21 people who buys a product after enquiring and have a maximum 60 people who enquiring about the ad. The maximum clicks we receive in an ad is 421

- The dataset includes 1,143 rows of data.
- The mean ad\_id is 987,261, which suggests that there are many unique ad\_ids in the dataset.
- The mean xyz\_campaign\_id is 1,067, which indicates that there are multiple campaigns in the dataset.
- The mean fb\_campaign\_id is 133,784, which suggests that there are many unique fb\_campaign\_ids in the dataset.
- The mean interest is 32.77, which indicates that the average interest level of the target audience is moderate.
- The mean number of impressions is 186,732, which suggests that the ads have been viewed a large number of times.
- The mean number of clicks is 33.39, which indicates that the click-through rate is relatively low.
- The mean amount spent is \$51.36, which suggests that the cost per click is relatively low.
- The mean total conversion is 2.86, which indicates that, on average, each click leads to almost three conversions.
- The mean approved conversion is 0.94, which suggests that, on average, less than one conversion is approved for each click.
- The standard deviation of the variables is relatively high, which suggests that there is a wide range of values for each variable.
- The minimum value for each variable indicates the lower limit of the range of values, while the maximum value indicates the upper limit of the range of values.
- The quartile values provide information on the distribution of the variables, with the median

# Data Cleaning:

We have 0 Missing values and 0 Duplicates as you can see in these Screenshots Below.



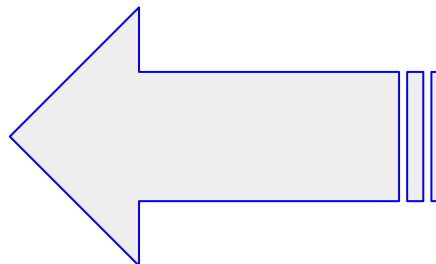
```
print(df.isnull().sum())
```

ad_id	0
xyz_campaign_id	0
fb_campaign_id	0
age	0
gender	0
interest	0
Impressions	0
Clicks	0
Spent	0
Total_Conversion	0
Approved_Conversion	0
dtype: int64	

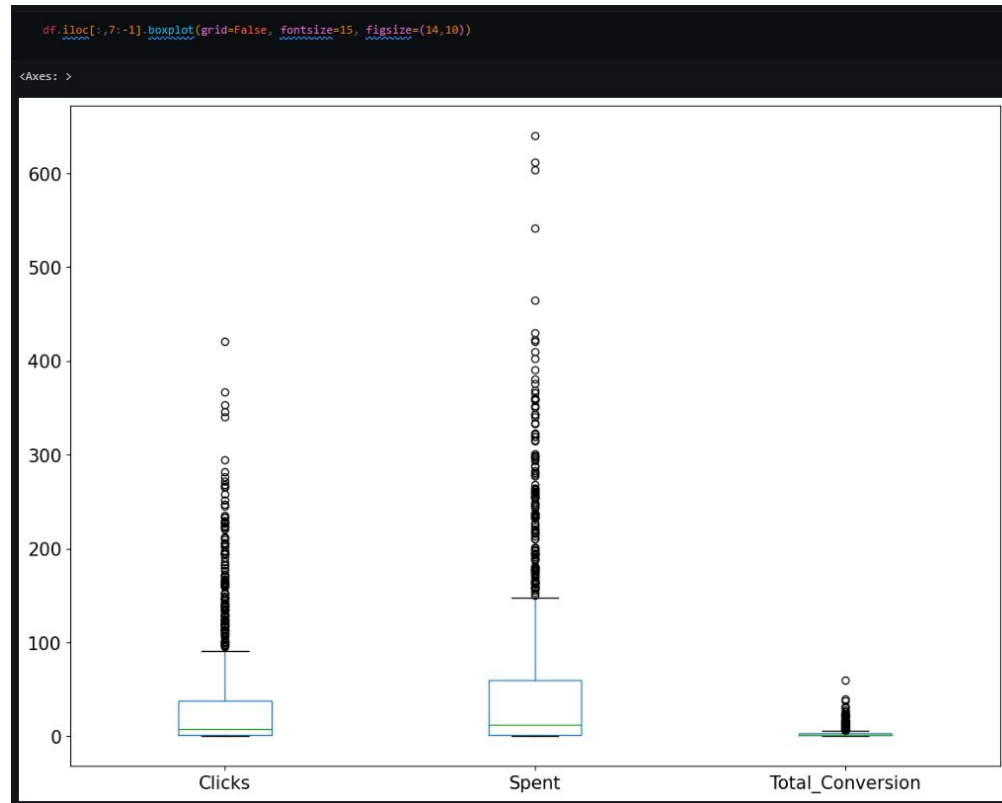
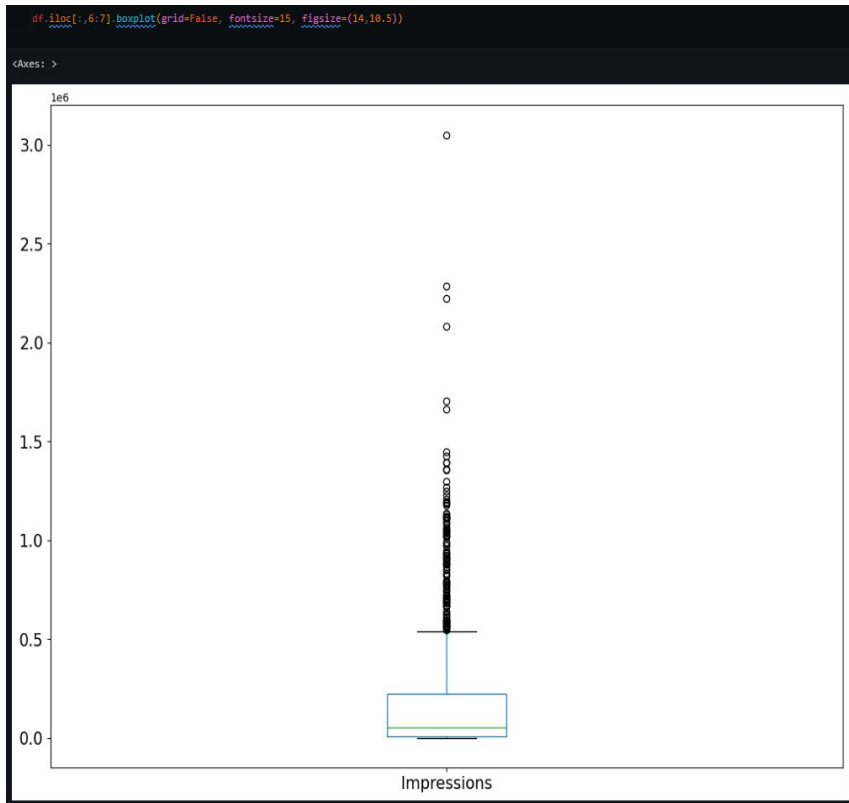
Hurray! Here we can see that there is no missing values in our data.

```
#Checking for duplicates values  
df.duplicated().sum()
```

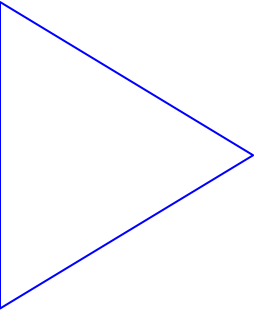
0



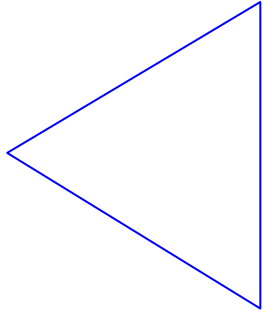
# We have some Outliers Present in Spent, Impression And Clicks



I dropped two features xyz\_campaign\_id And FB\_Ads\_id and also saved clean data in cleaned.csv file you can this in under the data folder.



```
Cleaned data 🍌  
  
data.drop(['fb_campaign_id', 'xyz_campaign_id'], axis = 1, inplace = True)  
  
data.to_csv("Data\cleaned\Clean_Data.csv")
```

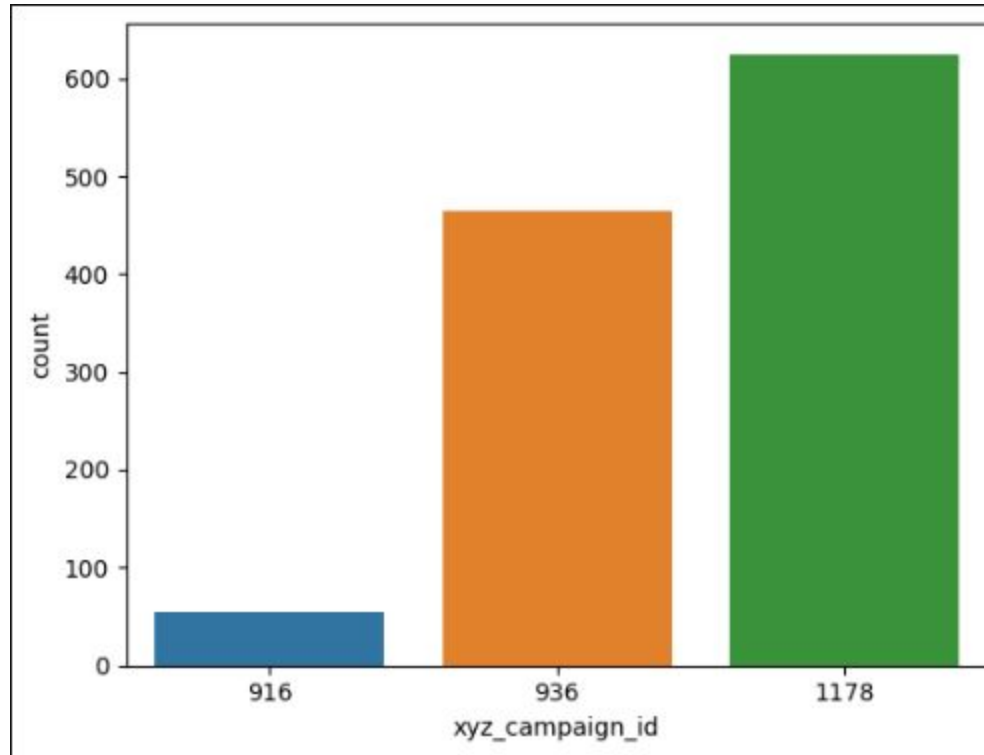


Also I done Label Encoding on Gender Column:

```
from sklearn.preprocessing import LabelEncoder  
lb = LabelEncoder()  
df['gender'] = lb.fit_transform(df['gender'])  
  
df['gender'].value_counts()  
  
1    592  
0    551  
Name: gender, dtype: int64
```

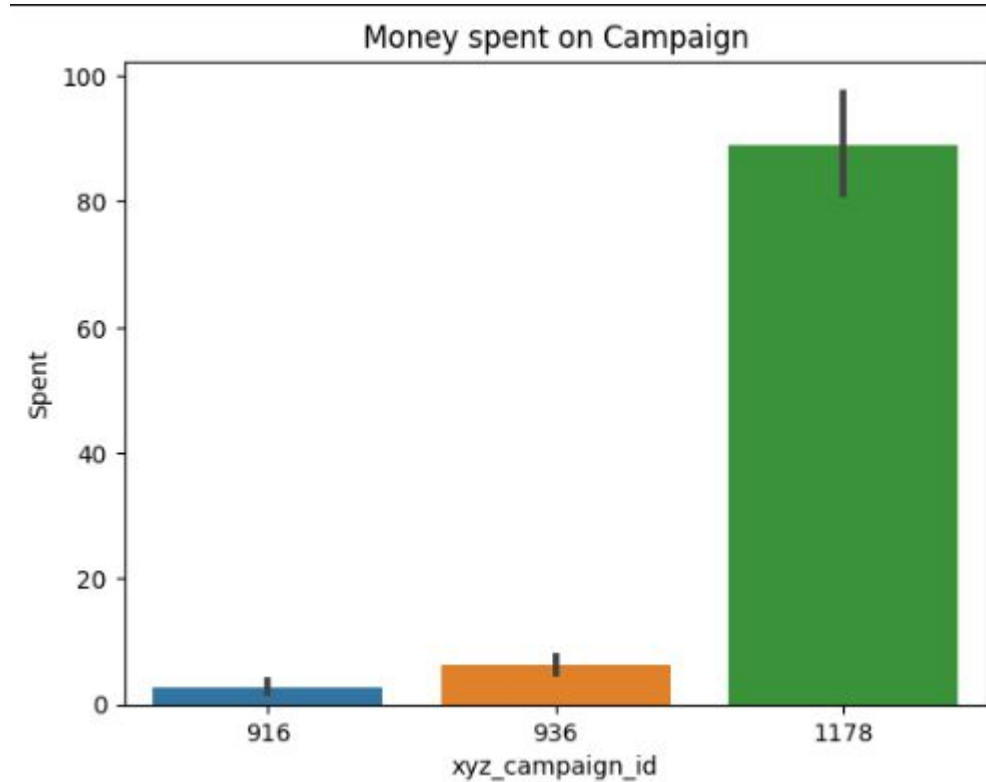


# Exploratory Data Analysis: 😊

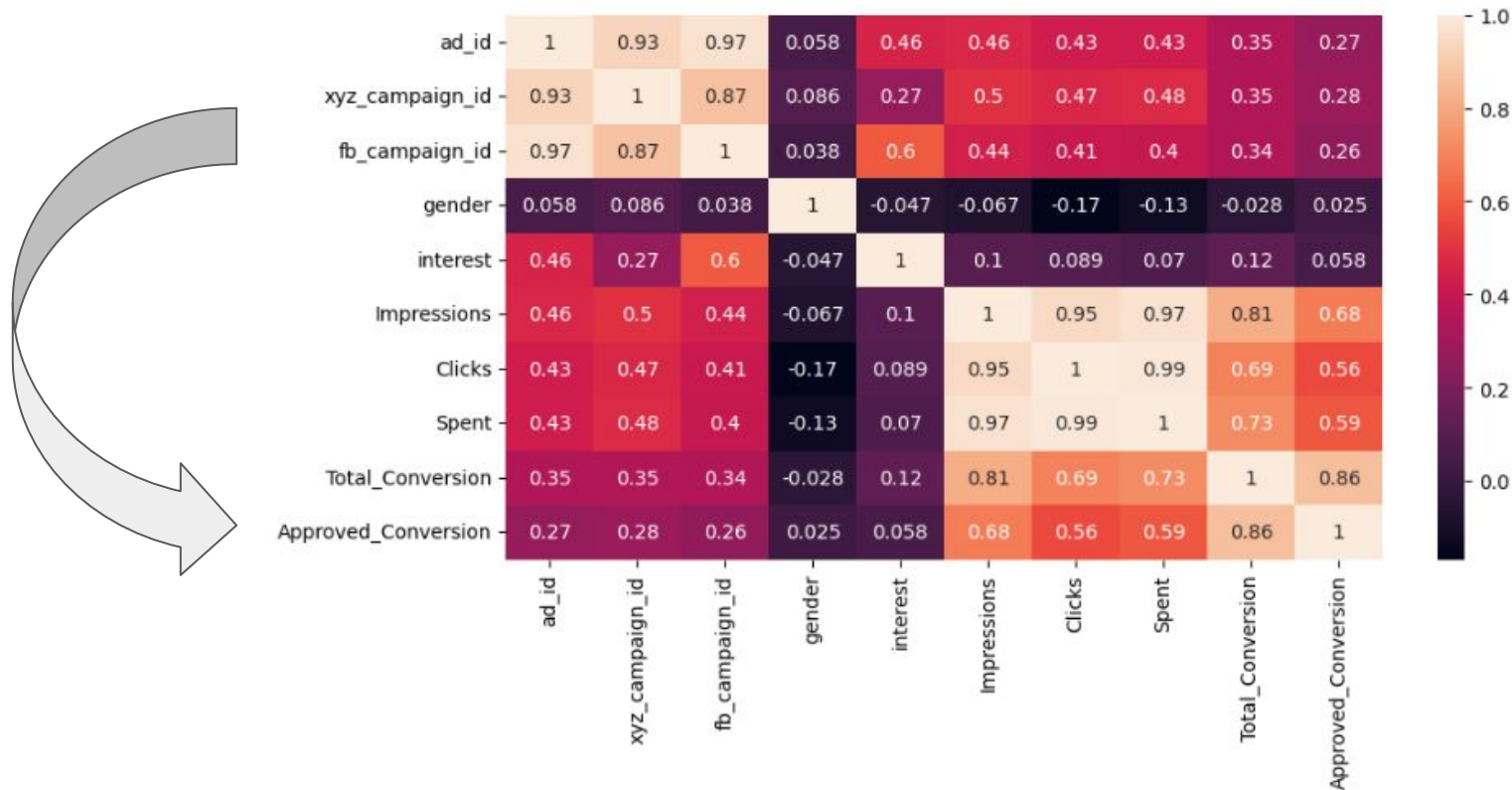


- There are Three Campaign Id's Present.

Money spent on campaign with campaign id 1178 is almost 4 times of the other two campaign.

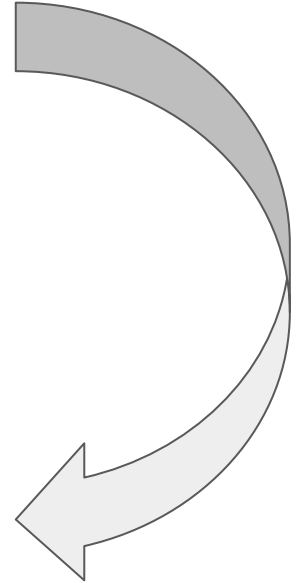
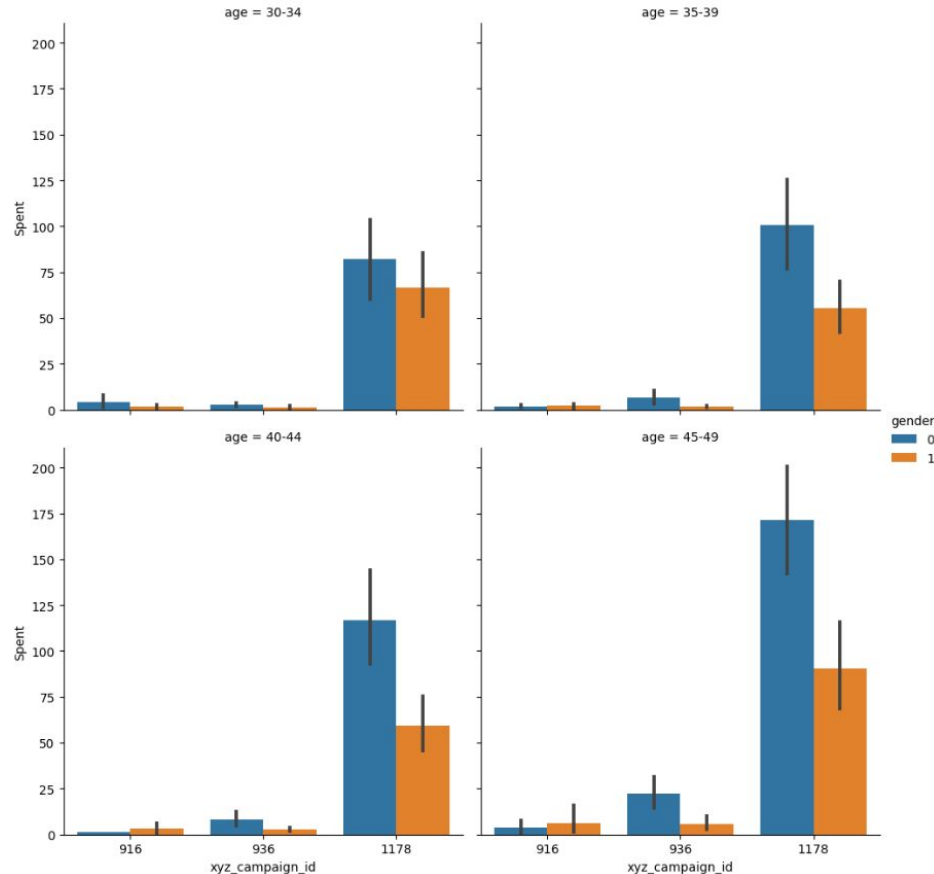


# Once Check Correlation Matrices:



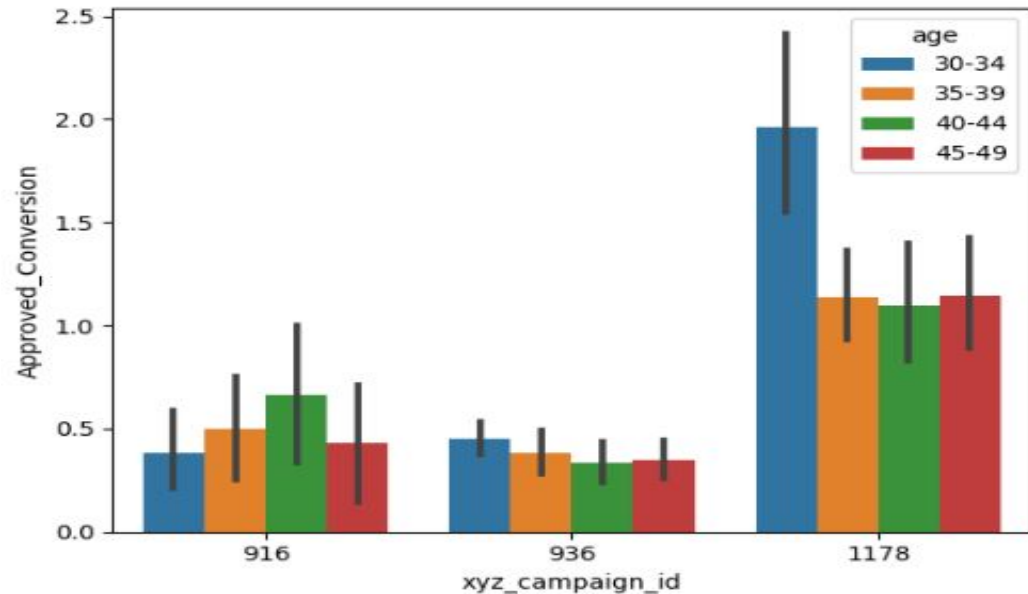
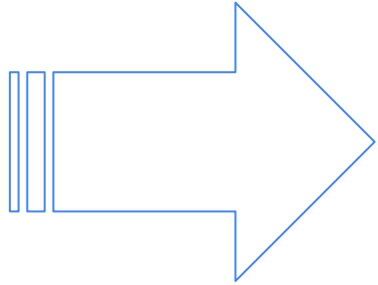
Here it's clear, "Impressions" and "Total\_Conversion" are more correlated with "Approved\_Conversion", "Clicks" and "Spent".

The graph show that age group 45-49 has clicked on ad maximum number of time as compared to other age group people. Females of the group has clicked more time on ads in comparison with males.

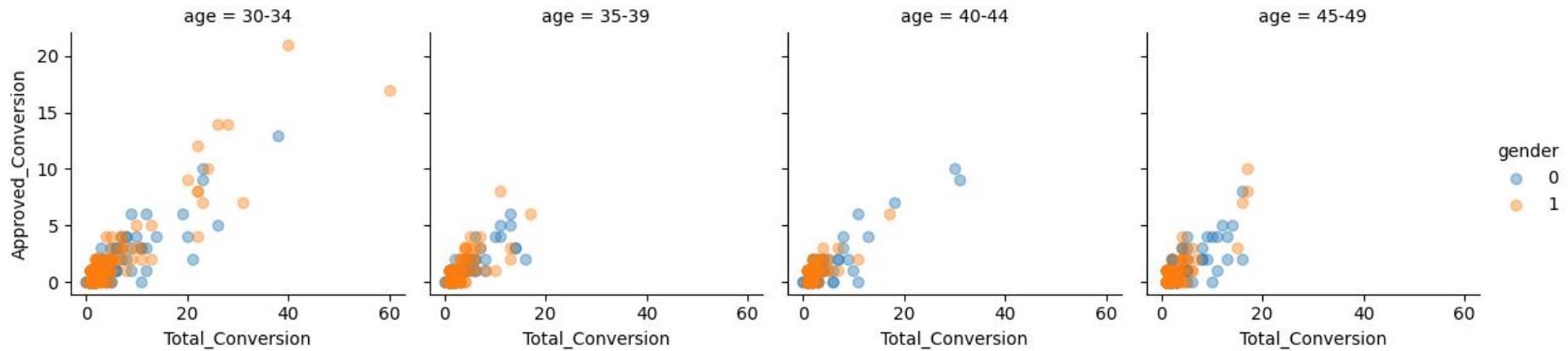


As we have seen earlier that in campaign id 1178 maximum click was being done with people of age group 45-49. But in the above graph we can see that maximum approved conversion are being made by people of age group 30-34, which has made less number of click as compared to all the other age group.

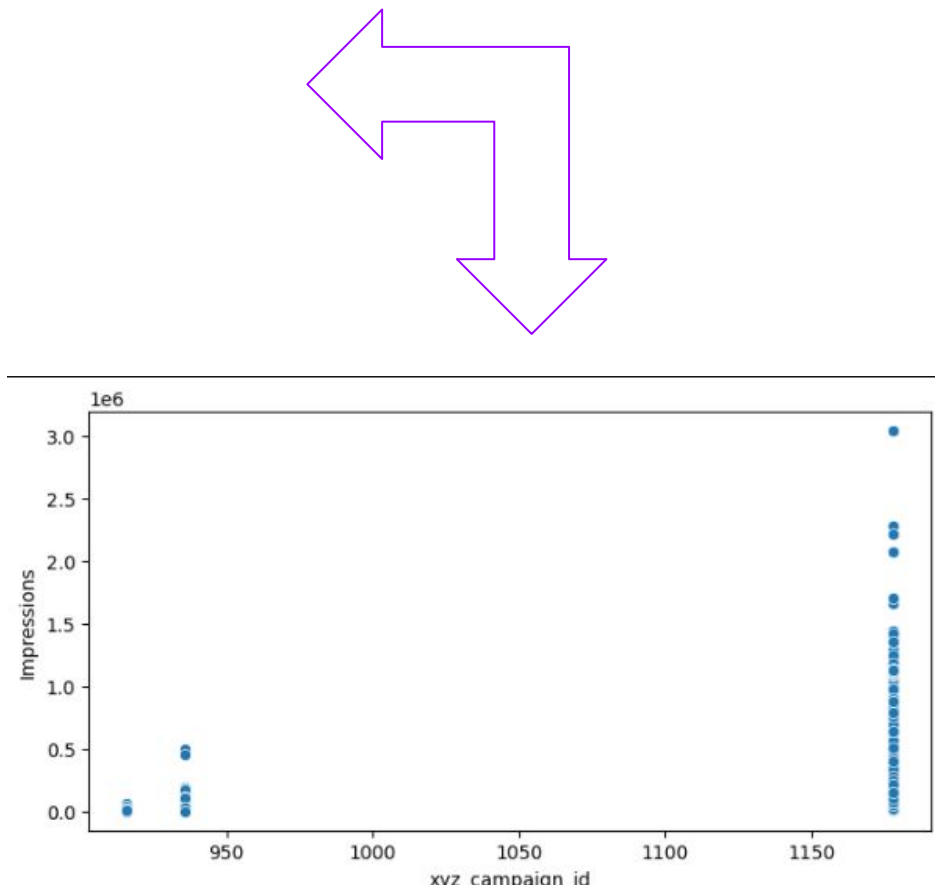
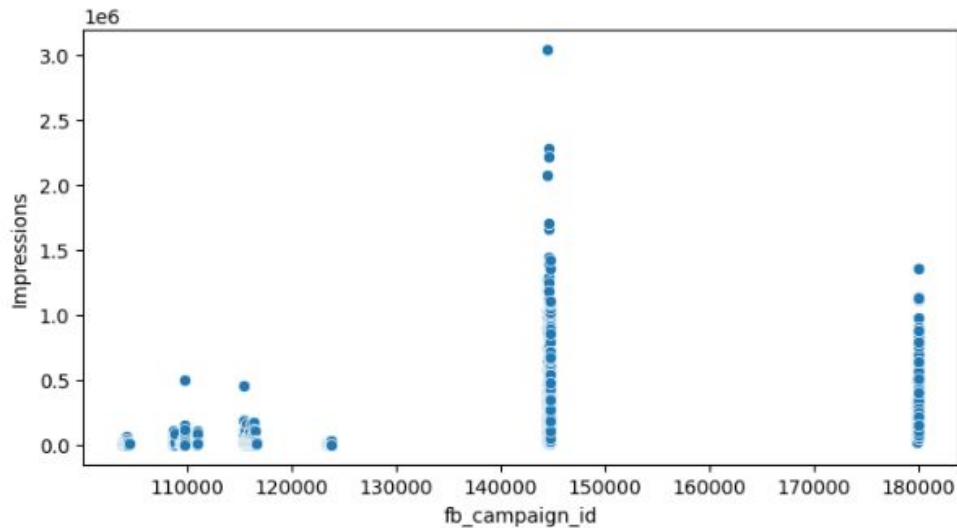
So, ads with campaign id 1178 can target people of age group 30-34, whereas 936 can target mostly on age group 30-34 as well. Campaign id 916 can target people of age group 40-44 as they purchase more number of time.



This graph clearly depicts that men and women in the age group of 30-34 have bought the product after enquiring about it. The age group of 30-34 enquired about the product and bought the product more as compared to the rest of the age groups



The ads by xyz companies were displayed relatively fewer times than facebook ads.  
Hence fewer clicks than facebook ads.



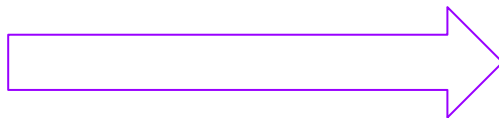
## Dummy Variable: I Created a Dummy Variable of Age.

```
#Converting categorical variables into numerical  
df = pd.get_dummies(df, columns=['age'])
```

	ad_id	xyz_campaign_id	fb_campaign_id	gender	interest	Impressions	Clicks	Spent	Total_Conversion	Approved_Conversion	age_30-34	age_35-39	age_40-44	age_45-49
0	708746.0	916.0	103916.0	1.0	15.0	7350.0	1.0	1.43	2.0	1.0	1.0	0.0	0.0	0.0
1	708749.0	916.0	103917.0	1.0	16.0	17861.0	2.0	1.82	2.0	0.0	1.0	0.0	0.0	0.0
2	708771.0	916.0	103920.0	1.0	20.0	693.0	0.0	0.00	1.0	0.0	1.0	0.0	0.0	0.0
3	708815.0	916.0	103928.0	1.0	28.0	4259.0	1.0	1.25	1.0	0.0	1.0	0.0	0.0	0.0
4	708818.0	916.0	103928.0	1.0	28.0	4133.0	1.0	1.29	1.0	1.0	1.0	0.0	0.0	0.0

---

## Feature Engineering:

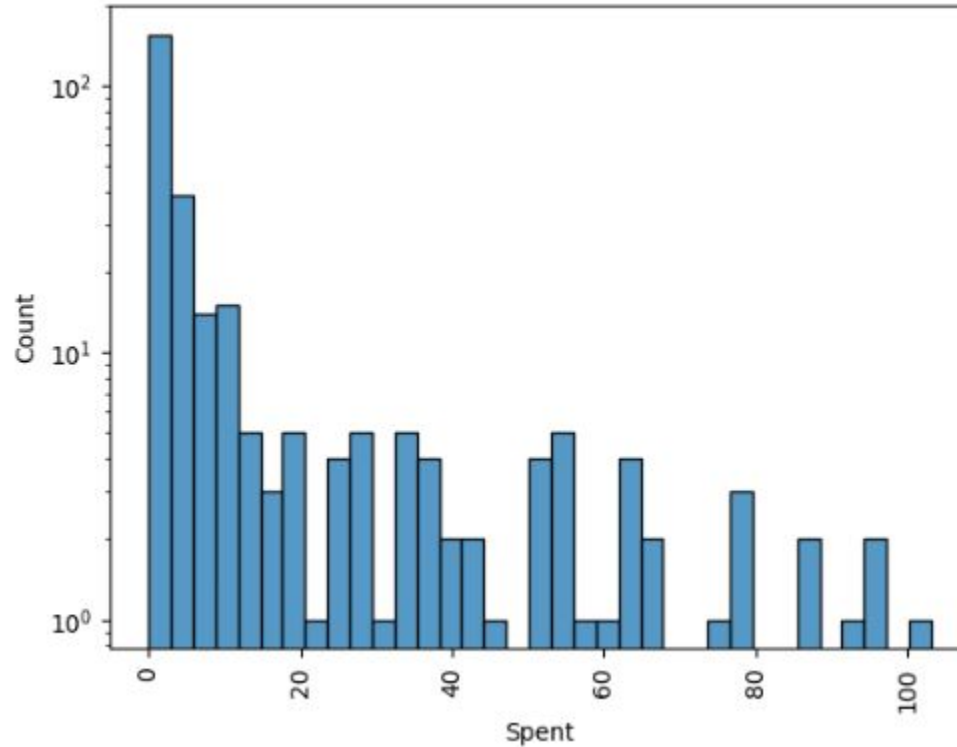




it's time for some feature engineering and let's introduce some additional features with the definitions

- **Click-through-rate (CTR)** - This is the percentage of how many of our impressions became clicks. A high CTR is often seen as a sign of good creative being presented to a relevant audience. A low click through rate is suggestive of less-than-engaging adverts (design and / or messaging) and / or presentation of adverts to an inappropriate audience. What is seen as a good CTR will depend on the type of advert (website banner, Google Shopping ad, search network text ad etc.) and can vary across sectors, but 2% would be a reasonable benchmark.
- **Conversion Rate (CR)** - This is the percentage of clicks that result in a 'conversion'. What a conversion is will be determined by the objectives of the campaign. It could be a sale, someone completing a contact form on a landing page, downloading an e-book, watching a video, or simply spending more than a particular amount of time or viewing over a target number of pages on a website.
- **Cost Per Click (CPC)** - how much (on average) did each click cost. While it can often be seen as desirable to reduce the cost per click, the CPC needs to be considered along with other variables. For example, a campaign with an average CPC of £0.5 and a CR of 5% is likely achieving more with its budget than one with a CPC of £0.2 and a CR of 1% (assuming the conversion value is the same).
- **Cost Per Conversion** - Another simple metric, this figure is often more relevant than the CPC, as it combines the CPC and CR metrics, giving us an easy way to quickly get a feel for campaign effectiveness.

# Removing Outliers Using Interquartile Method:



Checking after removing outliers, back then it has 600+ now it is 100, so we can say we have removed outliers.

# Feature Importance:

---

In the Logistic Regression model, each feature is assigned a weight (positive or negative) indicating its impact on the predicted probability of Conversion\_Rate. For example, the feature "Conversion\_Rate <= 0.00" has a negative weight of -0.654, which means that when the value of this feature is true (i.e. Conversion\_Rate is less than or equal to 0.00), it decreases the predicted probability of Conversion\_Rate. On the other hand, the feature "Impressions <= 3116.50" has a positive weight of 0.123, which means that when the value of this feature is true (i.e. Impressions is less than or equal to 3116.50), it increases the predicted probability of Conversion\_Rate.

In the Decision Tree model, each feature is assigned a score indicating its importance in predicting the target variable. The feature with the highest score has the most significant impact on the predicted probability of Conversion\_Rate. For example, the feature "Conversion\_Rate <= 0.00" has the highest score of -0.786, which means that it has the most significant impact on the predicted probability of Conversion\_Rate.

# Summary:

---

This report presents an analysis of a dataset related to an advertising campaign aimed at promoting a product. The dataset includes information such as `ad_id`, `campaign_id`, `age`, `gender`, `interest`, `impressions`, `clicks`, `spent`, `total_conversion`, and `approved_conversion`. The objective of this analysis was to gain insights from the dataset and build a predictive model to predict whether an ad click will lead to a conversion or not.

The analysis revealed that the age group of 30-34 has the most approved conversions, while women have clicked on ads more than men. However, in the age group of 30-34, men have enquired about the product and bought it more than women. The decision tree and logistic regression models were used to predict whether an ad click will lead to a conversion or not, and both models performed well with high accuracy, precision, recall, and F1 score.

Based on the insights gained from the analysis, it is recommended that the advertising campaign should target people of the age group of 30-34, as they have the most approved conversions. Additionally, the advertising campaign should also target women, as they have clicked on ads more than men. However, in the age group of 30-34, men have enquired about the product and bought it more than women. Therefore, the advertising campaign should be tailored to both men and women in this age group. The decision tree model may be more suitable in situations where the false positive rate is important to control, but both models can be used for prediction.

# Methodology:

---

Exploratory data analysis was performed to gain insights into the data. The age group vs clicks graph showed that the age group of 45-49 has clicked on ads the most, while the gender vs clicks graph showed that women have clicked on ads more than men. The age group vs approved conversions graph showed that the age group of 30-34 has the most approved conversions. The graph of age group vs total\_conversion and age group vs approved\_conversion showed that the age group of 30-34 enquired about the product and bought it more than the other age groups.

The decision tree and logistic regression models were used to predict whether an ad click will lead to a conversion or not. The dataset was divided into training and testing sets. The logistic regression model had an overall accuracy of 0.9827586206896551, with precision, recall, and F1 score all equal to 0.97. The decision tree model had an overall accuracy of 0.9655172413793104, with precision, recall, and F1 score all equal to 0.97. However, the ROC AUC score of the decision tree model was slightly higher at 0.977485380116959, compared to the logistic regression model's ROC AUC score of 0.9307692307692309.

Based on the insights gained from the analysis, it is recommended that the advertising campaign should target people of the age group of 30-34, as they have the most approved conversions. Additionally, the advertising campaign should also target women, as they have clicked on ads more than men. However, in the age group of 30-34, men have enquired about the product and bought it more than women. Therefore, the advertising campaign should be tailored to both.

- **User targeting:** It's important to ensure that your advertising campaign is targeting the right audience. Review your targeting options and make sure that your ads are being shown to the right people. Consider refining your audience by factors such as demographics, interests, and behaviors.
- **Ad design:** The ad design should be visually appealing and compelling enough to grab the user's attention. Consider testing different ad creatives, including images and videos, to see which ones perform best. Ensure that your ad copy is concise and clear, and that it includes a clear call-to-action.
- **Budget allocation:** Analyze the performance of your current ads and allocate more budget to those that are performing well. Consider reducing the budget for underperforming ads or pausing them altogether.
- **Ad placement strategies:** Test different ad placements and ad formats to see which ones perform best. Consider running ads on different platforms and targeting different devices.
- **Landing page optimization:** Ensure that your landing pages are optimized for conversions. This includes having clear and concise messaging, a strong call-to-action, and a seamless user experience. Consider testing different landing pages to see which ones perform best.
- **Continual monitoring and optimization:** Continuously monitor the performance of your ads and make adjustments as needed. Test different targeting options, ad formats, and ad creatives to see what works best for your audience.