SD03Q06 – Kaalishwar.R – 1832027

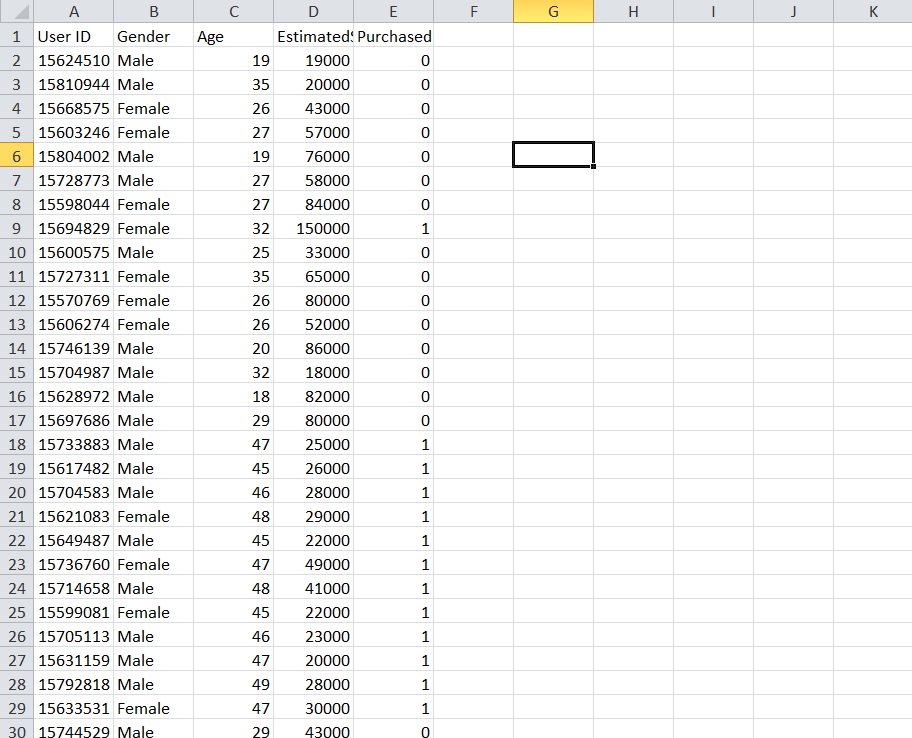
Social\_Networks\_Ad

Problem statement:

The problem revolves around the usage of social networking and the ad that is being produced by the sponsors and buying the products. We have got a dataset to describe the person buying product based on the advertisement.

Dataset:

The dataset contains 400 rows of information.



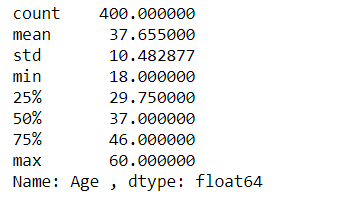
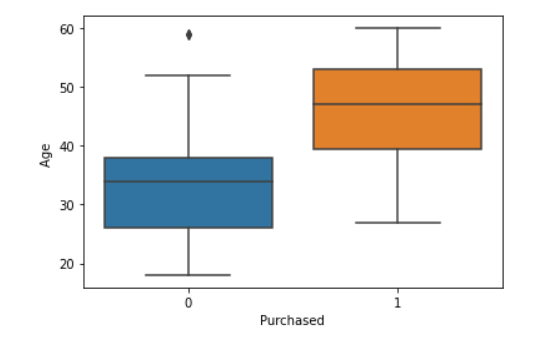
Data pre-processing:

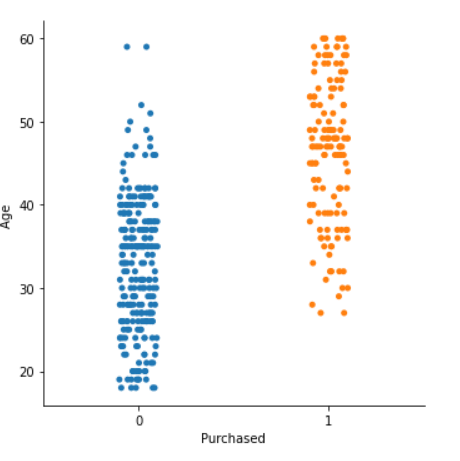
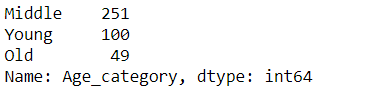
The data has been pre-processed and some of the missing values have been replaced by mean or median values based on the distribution of the data in a particular attribute. The data has been modified that it doesn’t affect the flow of the data.

Social\_Network\_Ads:

The exploratory data analysis (EDA) has been done.

The box plots have been made to know the outliers for the attributes and to see whether the outliers affect the distribution. The below shows that EDA for age attribute.



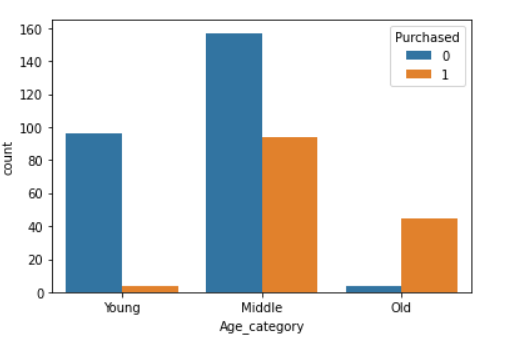


From the above plot we can see the distribution of purchased with age attribute. The latter shows the count of the people purchased based on the age.

The summary statistics about the age attribute, we can categorize the customers into three age categories:

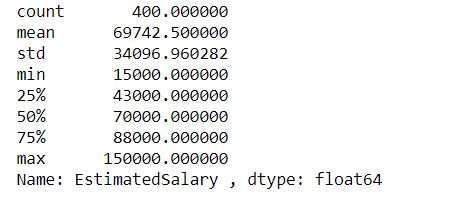
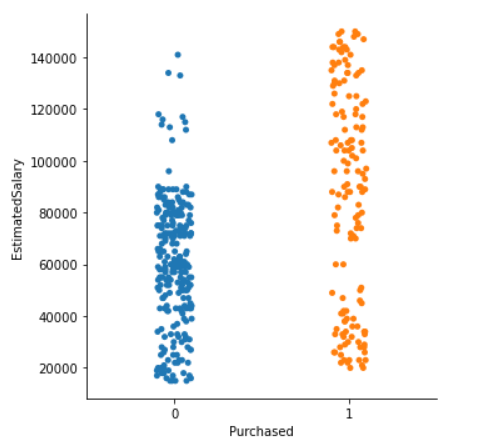
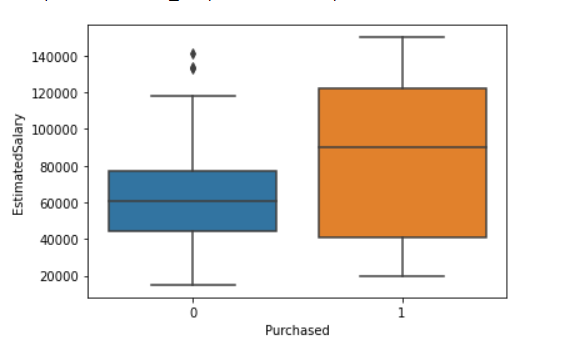
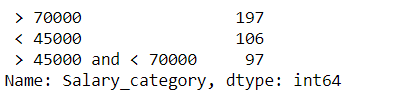
* less than 30 years as "Young"
* 30 to 50 years as "MIDDLE"
* Greater than 50 years as "OLD"

We can now plot the count of customers in each category who have purchased our product.



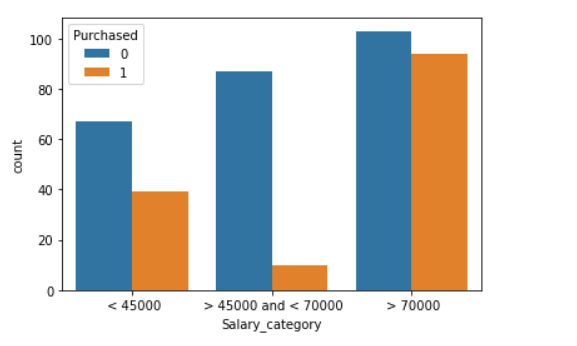
The above plot shows the purchased based on the age category and from the plot the participation of middle age people have been promising using social networks.

For the estimated salary the EDA, graphs and plots have been made and the count being made based on the category.

From the above box plots we can see that 50% of the customers who purchased from the ads earn between 40000 to 100000.So that we can differentiate them based on three classes.

* <45000
* >45000 and <70000
* >70000



The above plots shows that the most people are from the >70000 category. The purchase percentage of customers to whom the ads are shown is very high in the salary range >70000 than the customers with less salary. But it is a bit strange to see customers with salary greater than 45000 purchase less than customers with salary less than 45000.

KNN Model:

The KNN model is one of the simplest model easy to understand based on the supervised learning technique. The KNN algorithm can be used for both regression and classification but it is more suitable for classification. It is a non-parametric algorithm, since it does not make any assumption on underlying data.

KNN Algorithm:

* Select the number K of the neighbours
* Calculate the Euclidean distance of K number of neighbours. The Euclidean distance shows that the distance between the data points and the distribution of the data points.
* Take the K nearest neighbours as per the calculated Euclidean distance.
* Among these k neighbours, count the number of the data points in each category.

Assign the new data points to that category for which the number of the neighbour is maximum.

The model has been made. The data is being pre – processed and it is suitable for the model. The data is being separated as test set and train set. The model is being trained by the train set and it is being tested by the test set. The accuracy of the model is 92.5%.



Conclusion:

The model built on KNN algorithm shows the accuracy of 92.5% which gives a good and excellent and predicted output.