PROJECT2

Data: - bank_data.csv

Introduction:

The dataset available comprises of details about bank clients and their subscription status for a recently introduced deposit account. The purpose of this analysis is to extract valuable information regarding the factors that influence customers' decisions to subscribe, as well as to identify patterns within the data that can be beneficial for the bank's marketing and customer acquisition strategies. By utilizing a range of analytical techniques and tools taught in the course, we will thoroughly examine the dataset, perform descriptive statistics, and create informative plots and charts to derive significant insights and conclusions. Ultimately, our aim is to provide actionable recommendations based on these findings in order to enhance the bank's marketing efforts and improve customer engagement.

Descriptive Statistics:

Insights about the descriptive analysis of the data after removing null values

Age Distribution:

The average age of bank clients is approximately 38.66 years, with a minimum age of 21 years and a maximum age of 94 years.

The majority of clients fall between the age range of 31 and 45 years, as indicated by the 25th percentile (31 years) and the 75th percentile (45 years).

Job Distribution:

The most prevalent occupation among clients is administration ("admin."), with a total of 419 occurrences.

There are a total of 11 distinct job categories, highlighting the diversity in the occupations of bank clients.

Marital Status Distribution:

The majority of clients are married, with a total of 848 occurrences, followed by singles and divorced/widowed clients.

There are three unique marital status categories: married, single, and unknown.

Education Level:

The most common education level among clients is a university degree, with a total of 528 occurrences.

There are a total of seven distinct education categories, ranging from basic education to university degrees.

Contact Method:

The preferred method of contact for most clients is cellular communication, with a total of 973 occurrences, compared to telephone communication.

There are two unique contact communication types: cellular and telephone.

Month of Contact:

The most frequent month for the last contact with clients is May, with a total of 475 occurrences, followed by other months.

There are a total of 10 unique months in which clients were last contacted.

Day of Week of Contact:

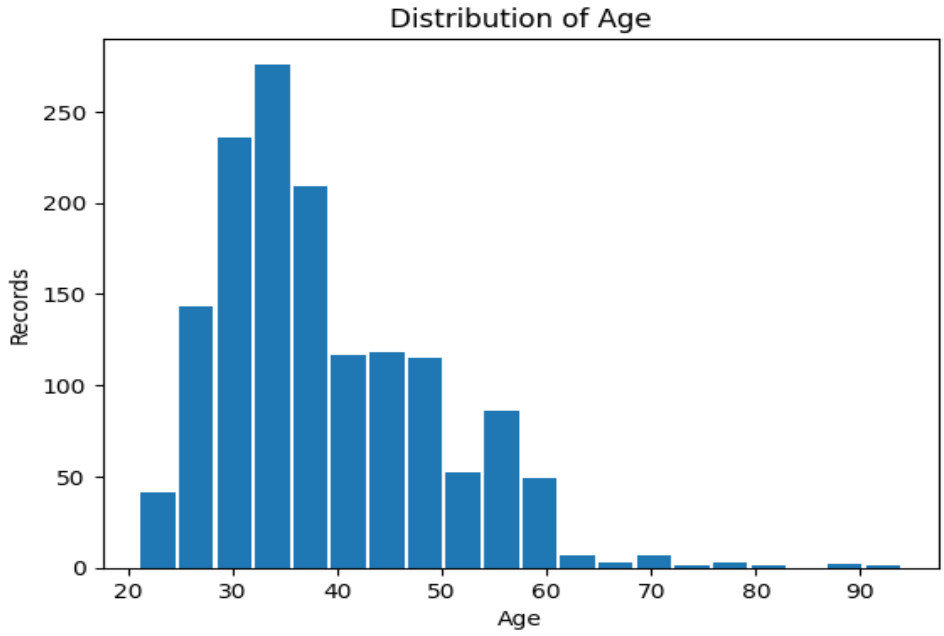
Thursday (Thu) appears to be the most common day of the week for the last contact, with a total of 323 occurrences.

There are five unique days of the week: Monday (Mon) to Friday (Fri).

Subscription Status:

The majority of clients did not subscribe to a new deposit account, with a total of 1285 occurrences of "no" compared to "yes".

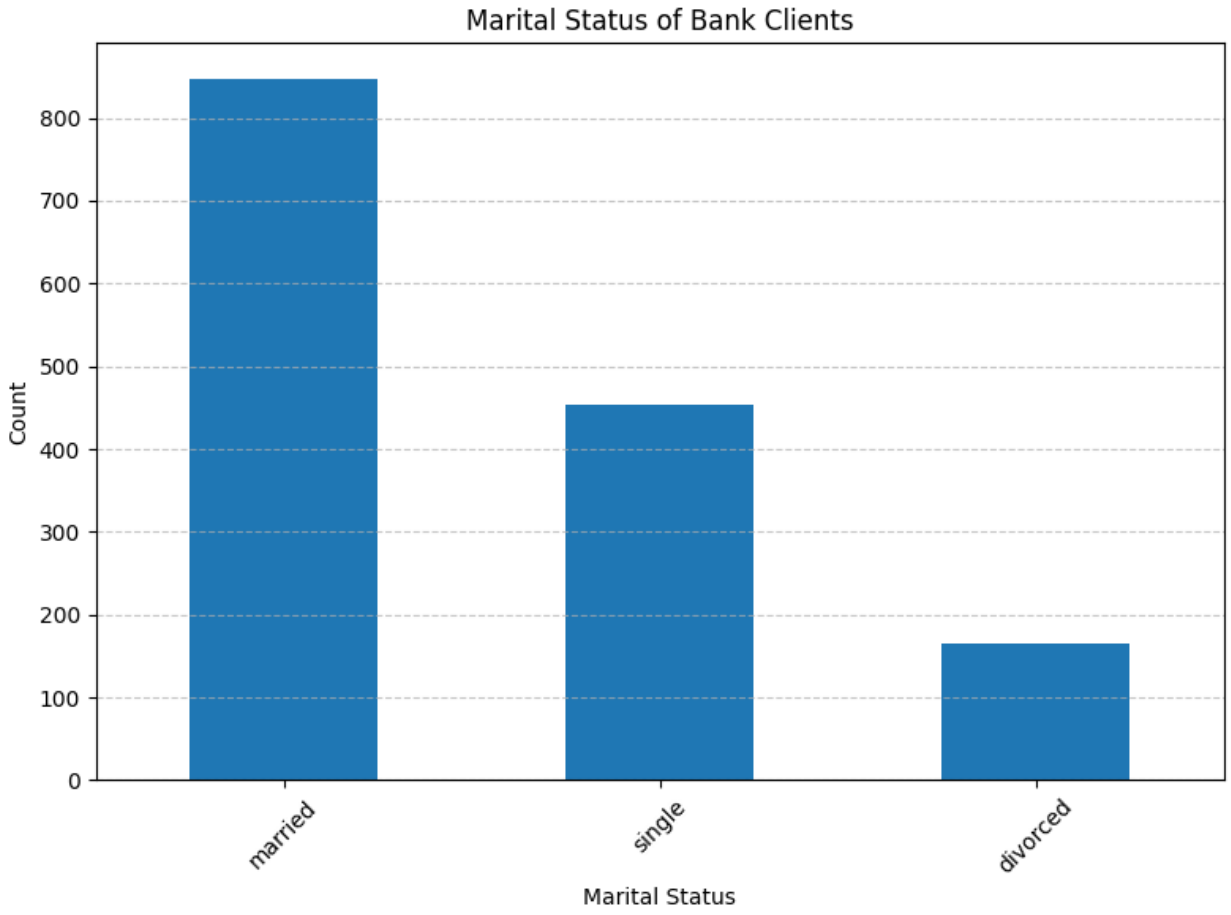
Histogram Plot



The age group with the most records is between 30 and 40 years old. There are around 220 records in this group.

The number of records decreases as age increases. There are very few records of people over 65 years old.

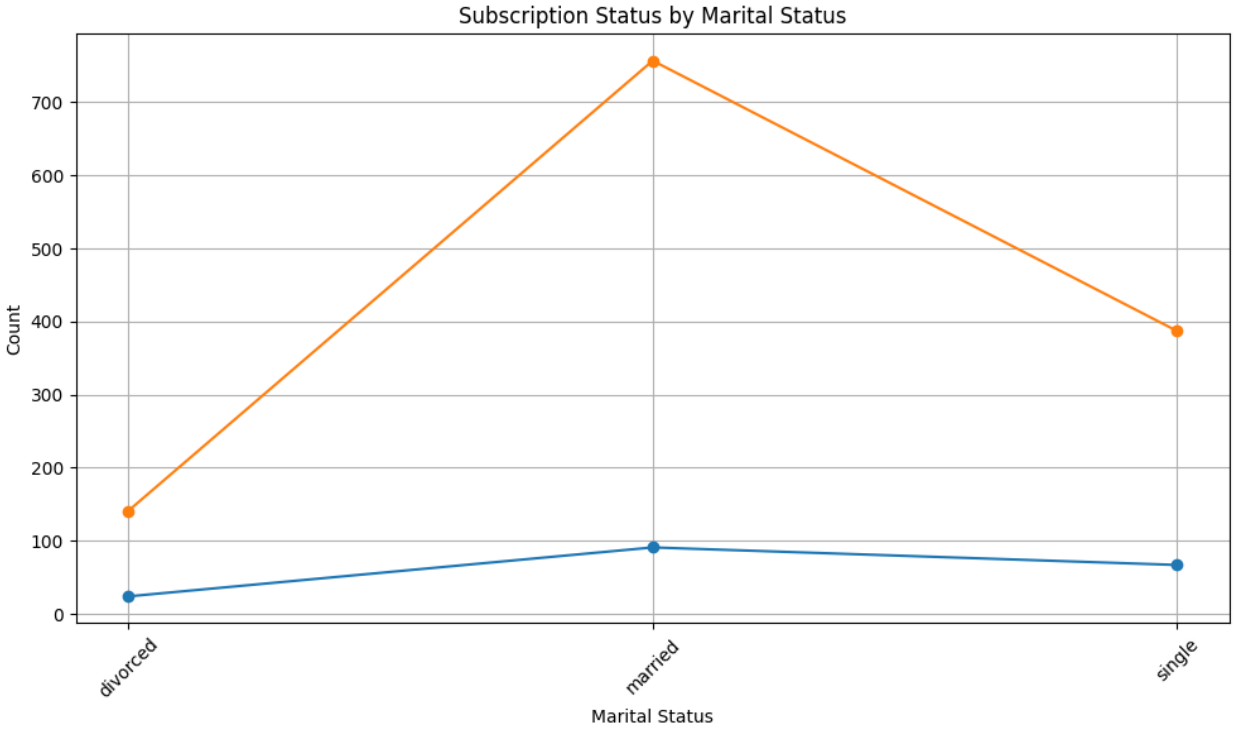
Bar Graph



Married clients constitute the majority of the bank's clientele, accounting for approximately 60% of the total.

Single clients make up around 20%, while divorced clients represent approximately 10% of the client base.

Line Plot

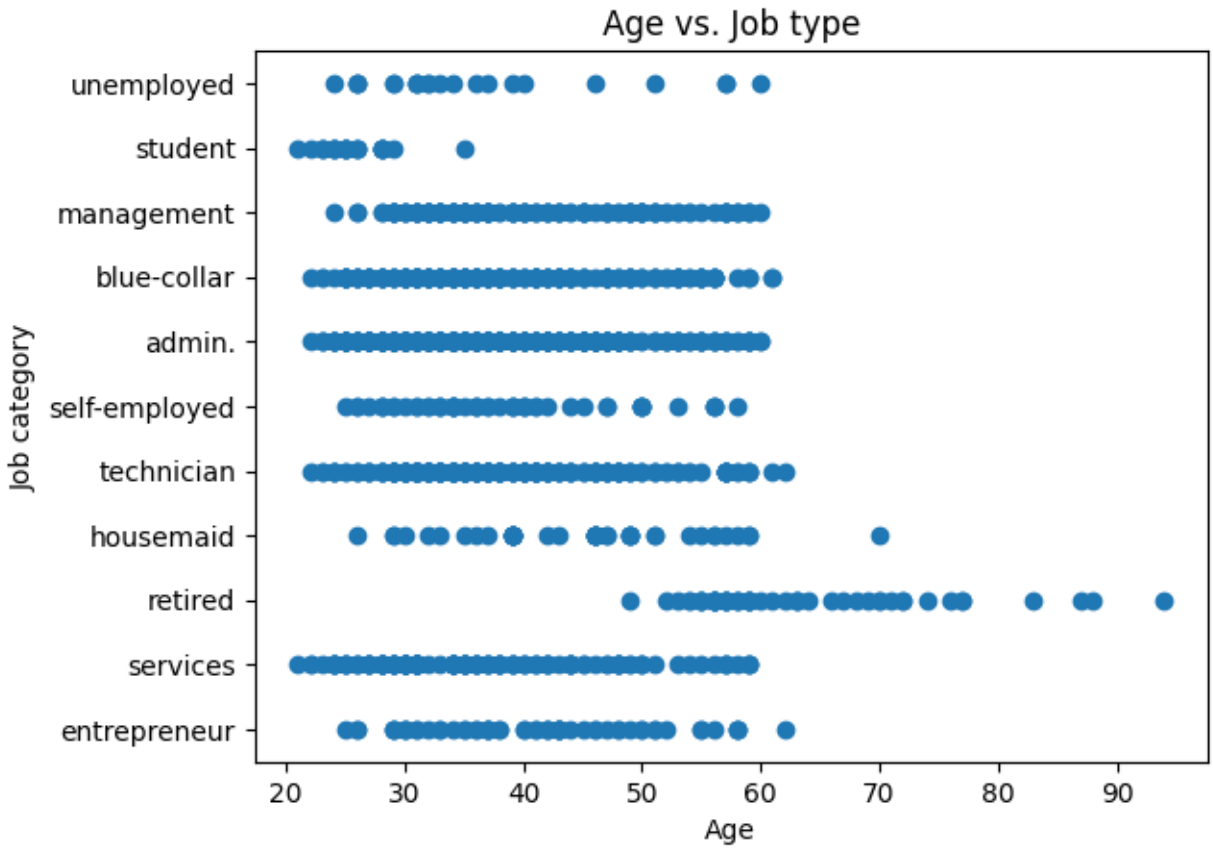


The graph shows the number of subscriptions by marital status. It appears the number of subscriptions is increasing over time for all three marital statuses: married, single, and divorced.

Married group have more subscriptions than any other group.

The rate of increase seems to be greatest for divorced customers.

Scatter Plot



Younger workers (under 30) tend to be concentrated in service jobs, student positions, and technician roles.

Meanwhile, older workers (over 50) are more likely to be found in management positions, retired, or self-employed.

Linear Regression

The R-squared value of the model is extremely low at 0.011, indicating that only around 1.1% of the variance in the dependent variable (subscribed) is explained by the independent variables included in the model.

Significance of Predictor Variables:

Among the categorical variables representing education levels, basic.6y, basic.9y, and high.school are statistically significant at conventional levels ($p < 0.05$). However, illiterate, professional.course, and university.degree do not seem to be statistically significant.

The age variable also does not appear to be statistically significant, as its coefficient is very close to zero and its p-value is high ($p > 0.05$).

Interpretation of Coefficients:

The intercept represents the estimated value of subscribed when all other predictors are zero, which is 0.173 in this case.

The coefficients for the education categories represent the estimated change in subscribed compared to the baseline category when all other predictors are held constant.

For instance, individuals with a basic.6y education level are estimated to have a subscribed value approximately 0.138 lower than those in the baseline category, holding all other variables constant.

Similarly, the coefficient for age is almost zero, indicating that age has almost no effect on the subscribed variable.

Predicted Value:

Based on the coefficients provided, the predicted value of subscribed is approximately 0.088.

Logistic Regression:

Precision: The precision is calculated to be approximately 89.23%, indicating that around 89.23% of the positive predictions made by the model are accurate.

Recall (Sensitivity): The recall is approximately 53.54%, showing that the model correctly identifies about 53.54% of the actual positive cases.

- The model correctly predicted 688 instances of clients not subscribing to the deposit account.
- However, it incorrectly predicted 83 instances as subscribing when they did not.
- The model missed predicting 597 instances of clients who actually subscribed to the deposit account.
- It correctly predicted 99 instances of clients subscribing to the deposit account.

For Subscribed:

H0: There is no significant difference between the count for subscribed for married compared to single.

H1: There is a significant difference between the count for the subscribed for married compared to single.

Conduct t-test

T-test results for difference between the count for the subscribed for married compared to single.

The p-value for number of teens is lesser than 0.05, we reject the null hypothesis and conclude that there is significant difference between the count for subscribed, for married compared to single.

Conclusion:

In Conclusion, the examination of the bank dataset uncovered various crucial findings concerning the factors impacting customers' decisions to subscribe and their demographic characteristics. Despite the limited predictive accuracy of linear regression and logistic regression models, the analysis revealed notable variations in subscription rates based on marital status. Specifically, married individuals exhibited higher subscription rates compared to single customers. Moreover, the descriptive statistics offered valuable insights into the age distribution, job roles, education levels, and preferred contact methods among bank clientele. These results emphasize the significance of targeted marketing approaches customized for diverse demographic groups to boost subscription rates and enhance customer engagement. Furthermore, delving deeper into the efficacy of marketing campaigns and customer communication channels could yield further insights for optimizing subscription outcomes and overall business performance.