BANK MARKETING (CAMPAIGN)

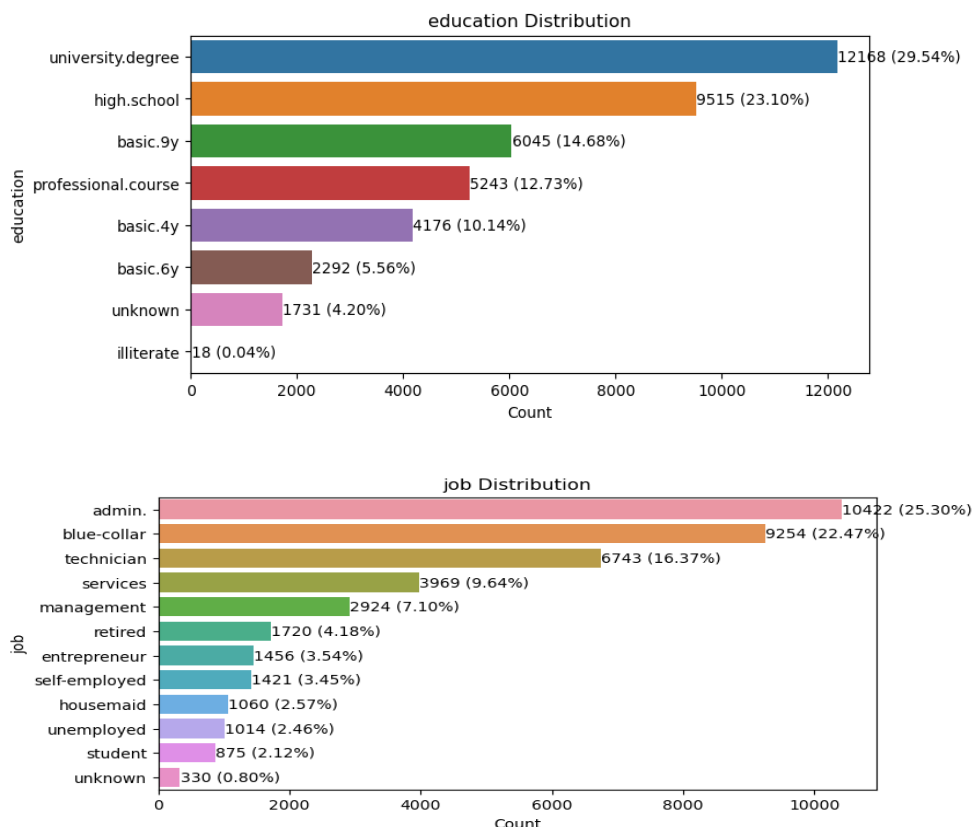## WEEK 9

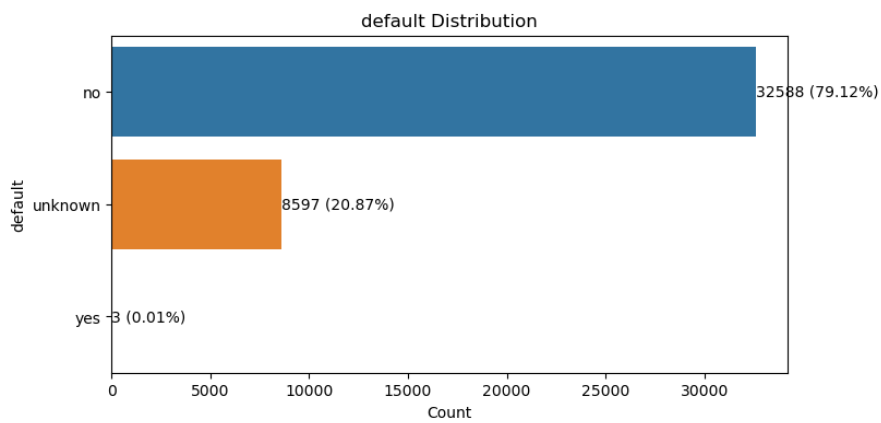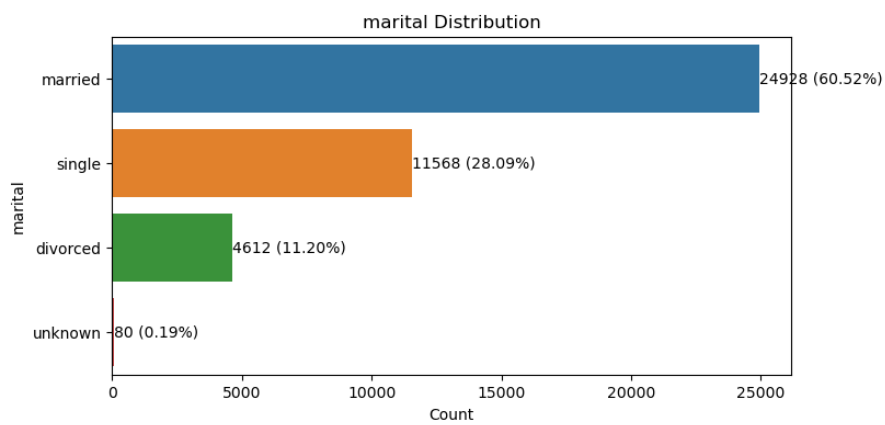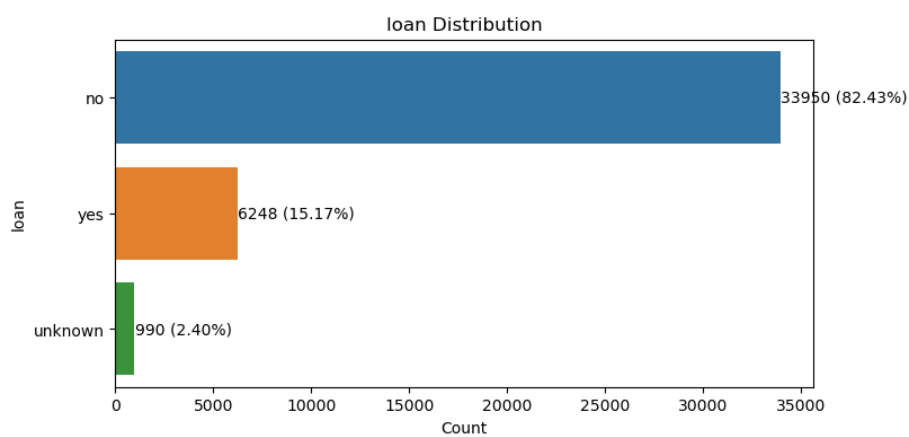| Group Name	Tentacool
Name	Murat Kıran
Email	murattkiran@gmail.com
Country	Türkiye
Specialization	Data Science

1. Problem description:

ABC Bank is preparing to launch a term deposit product and seeks to develop a predictive model to determine if customers will purchase the product based on their past interactions with the bank or other financial institutions.

2. Data Cleansing and Transformation:

There are unknown values for many variables in the dataset. Variables with unknown values include 'education,' 'job,' 'housing,' 'loan,' 'default,' and 'marital'.

education Distribution

job Distribution

housing Distribution

	Count
yes	21576 (52.38%)
no	18622 (45.21%)
unknown	990 (2.40%)

loan Distribution

	Count
no	33950 (82.43%)
yes	6248 (15.17%)
unknown	990 (2.40%)

marital Distribution

	Count
married	24928 (60.52%)
single	11568 (28.09%)
divorced	4612 (11.20%)
unknown	80 (0.19%)

default Distribution

	Count
no	32588 (79.12%)
unknown	8597 (20.87%)
yes	3 (0.01%)
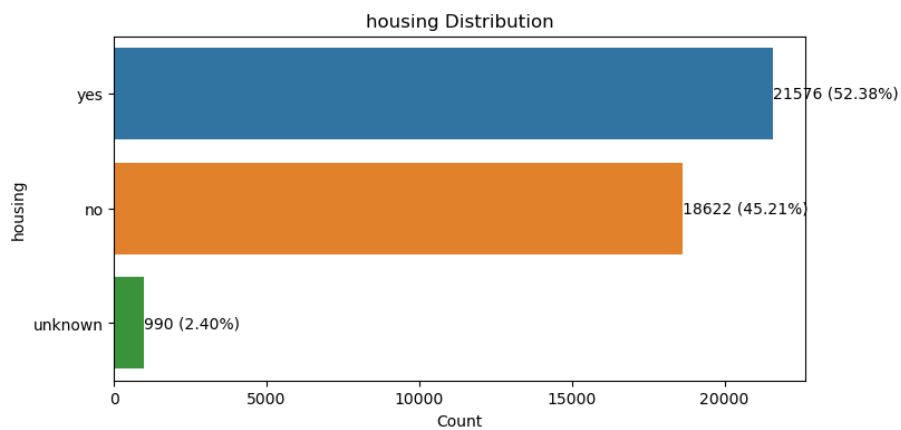
Unknown values in the dataset were not deleted; instead, sensible imputations were carried out, associating education, housing, and loan variables with job. This approach is grounded in the real-world scenario where job is correlated with education, housing, and loan status.

The quantity of 'unknown' values within the 'marital' variable is exceedingly low. Given the potential that customers might choose not to disclose their marital status, no action was taken regarding the 'unknown' values in this instance.

The unknown values in the 'default' variable are considered instances where information has intentionally not been disclosed. It is plausible that customers may choose not to reveal this specific information to the banking representative. Therefore, the unknown value in 'default' essentially represents a distinct category.

In the dataset, outliers were identified in the variables "campaign," "pdays," and "previous." The "campaign" variable represents the number of contacts performed during the current campaign for a specific client, including the last contact, while "previous" denotes the number of contacts made before the current campaign for the same client.

After examining the "campaign" and "previous" variables and taking real-world situations into account, no measures were taken to suppress the outliers in these values.

It is essential to note that outliers in these variables may naturally occur in the context of marketing campaigns, reflecting instances where a client is contacted multiple times during a campaign or has a history of previous contacts. Consequently, these outliers were retained in the dataset without modification, as they align with the inherent characteristics of the data and contribute valuable information to the analysis.

The 'pdays' variable was eliminated from the dataset and replaced with a new categorical variable, 'pdays_category,' featuring the following categories: 'not.contacted,' 'recent.contacted,' 'moderate.contacted,' and 'distant.contacted.