Modelling Visualisation and Analysis

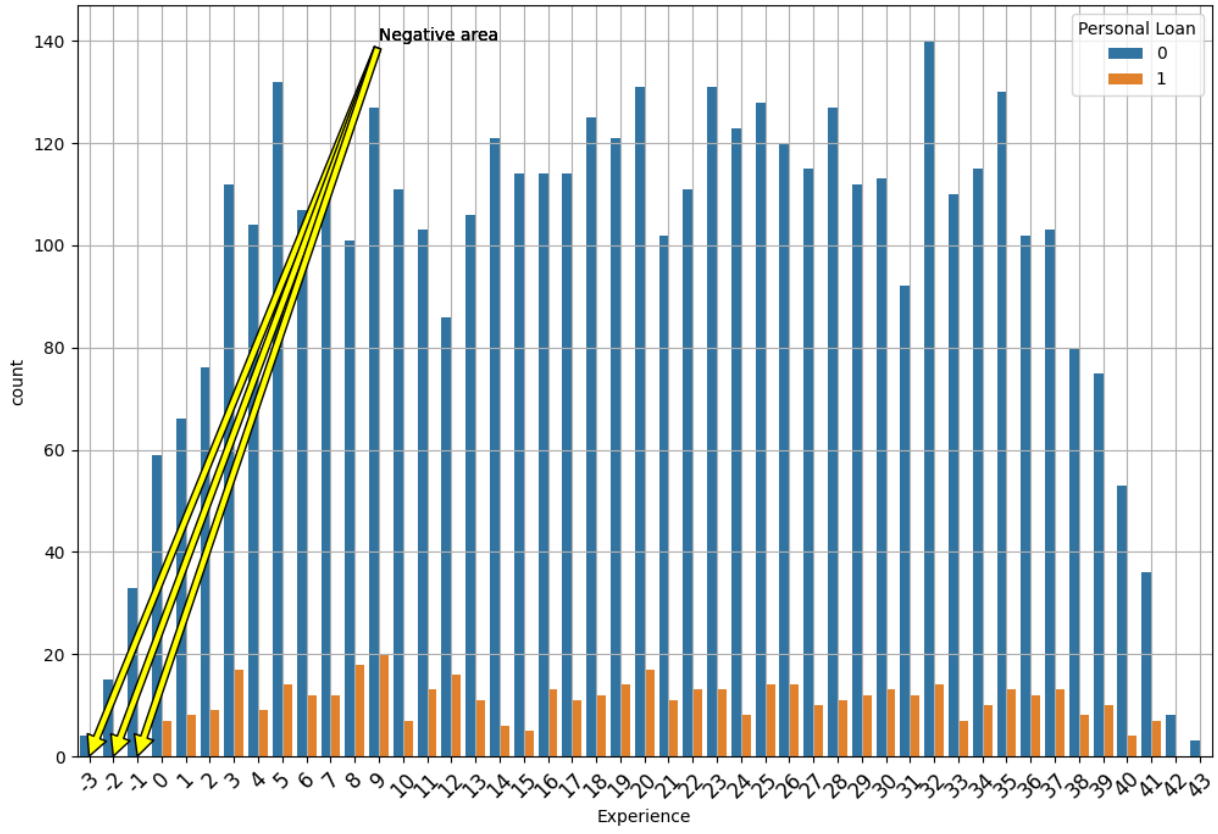
This section provides a concise analysis of the visualisations generated during the modelling process, covering data preparation and exploration phases. It's important to note that the focus of this project lies in evaluating and optimising models. Therefore, the analysis primarily aims to elucidate model-centric insights and discoveries made throughout this journey, rather than emphasising business functions and objectives.

1. Data Preparation

1.1 Anomalous value in ‘Experience’

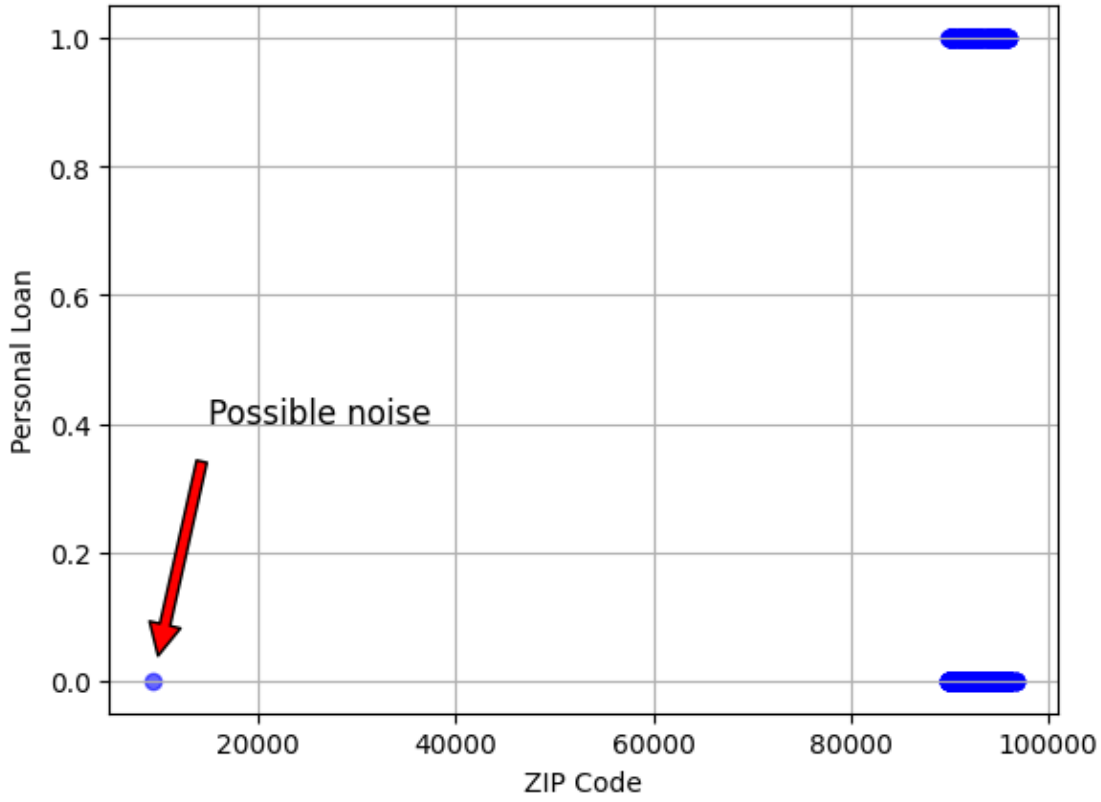
From the distribution, it's evident that some experience values fall within the negative range, specifically between -3 and -1, which is illogical in this context. Assuming data entry errors, we convert these negative numbers to positive using absolute conversion.

Note: The numbers in the ‘Experience’ column represent years of professional experience.



1.2 Noise filtering in ‘ZIP Code’

Upon examination of the ZIP Code data, it becomes apparent that the majority of the entries locates in the range of 90000 to 97000. However, the scatter plot highlights potential outliers that may stem from input errors. To ensure the integrity of our modelling process, we opt to remove these outliers to mitigate the impact of unusual data.



1.3 ZIP Code conversion

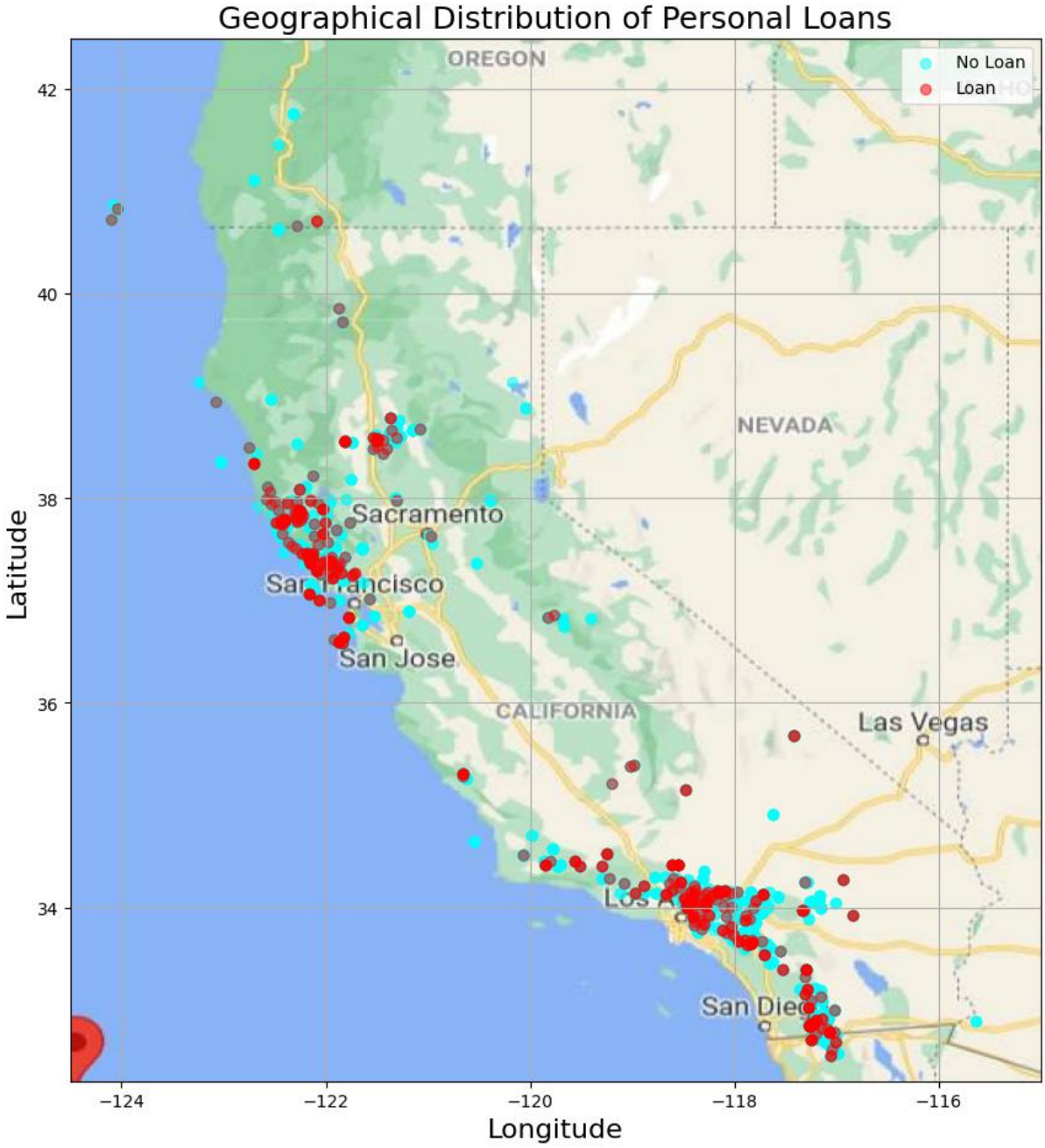
In the following step, we use the zip codes to allocate places, latitude, and longitude for each ZIP code entry. But, the incomplete generation requires manual input to fill in the gaps as some places are not automatically identified by the system. Hence, I searched on a postcode website via Google to fill in the missing information based on the search result. It turns out the ZIP code '96651' situates in Slovakia rather than the USA. Due to the lack of technical insight into the dataset, I decided to leave it (Slovakia ZIP code) unchanged to maintain input effectiveness and accuracy.

```
[ ]  df[df['Place']=='not found']['ZIP Code'].value_counts()
```

```
ZIP Code
92717    22
96651     6
92634     5
Name: count, dtype: int64
```

- 92717: Irvine (USA)
- 92634: Fullerton (USA)
- 96651: Rudno nad Hronom (Slovakia)

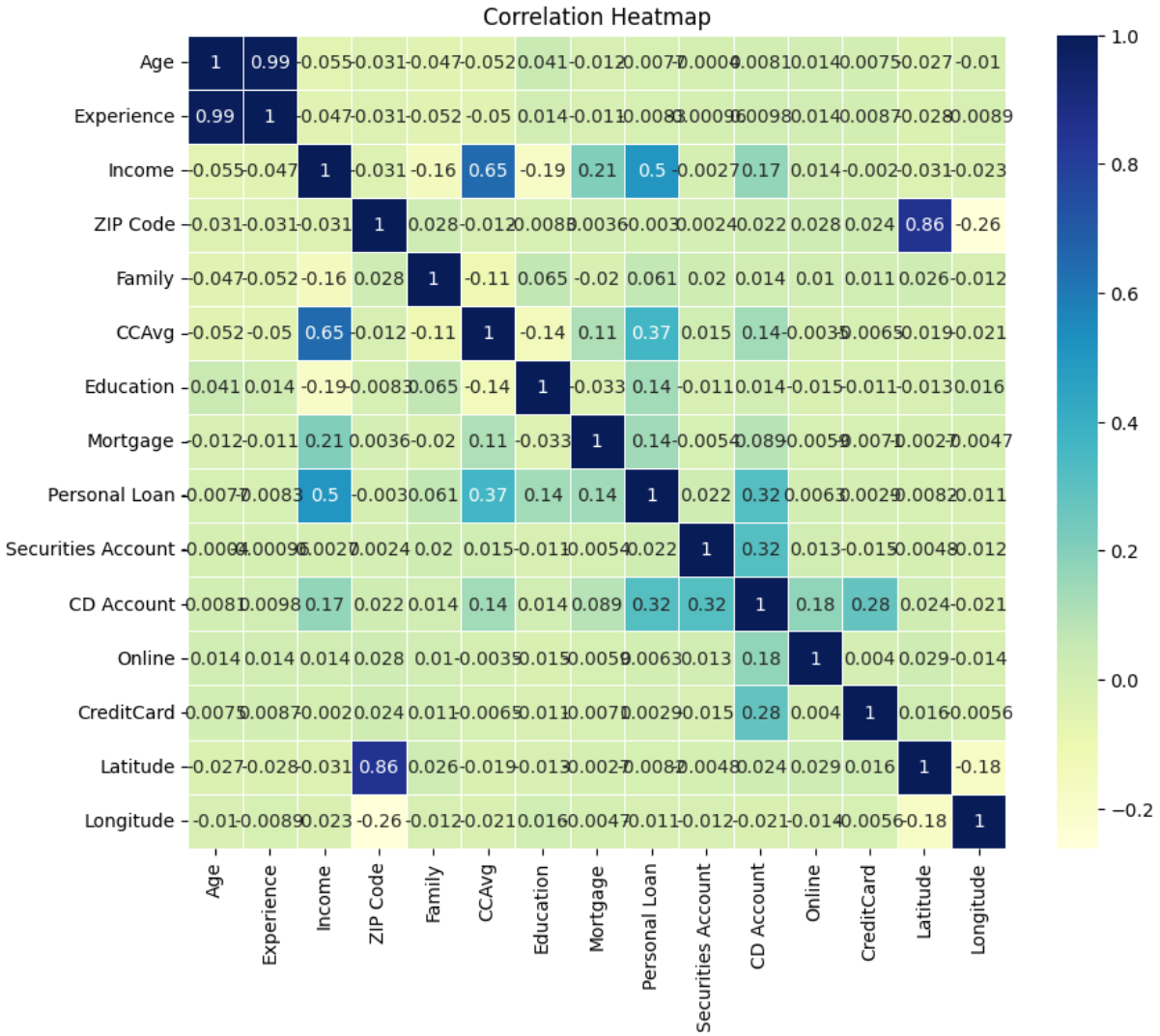
Through the converted data, we are able to present a coordinate map showing the geographical distribution of individuals based on their exact location and personal loan status. The map highlights clustering around western coastal regions in California, primarily San Francisco Bay areas, Los Angeles, and San Diego. Also, a concentration of individuals accepting personal loan offers is observed within these regions, this suggests a likelihood of residing in affluent areas.



2. Data Exploration

2.1 Correlation heatmap

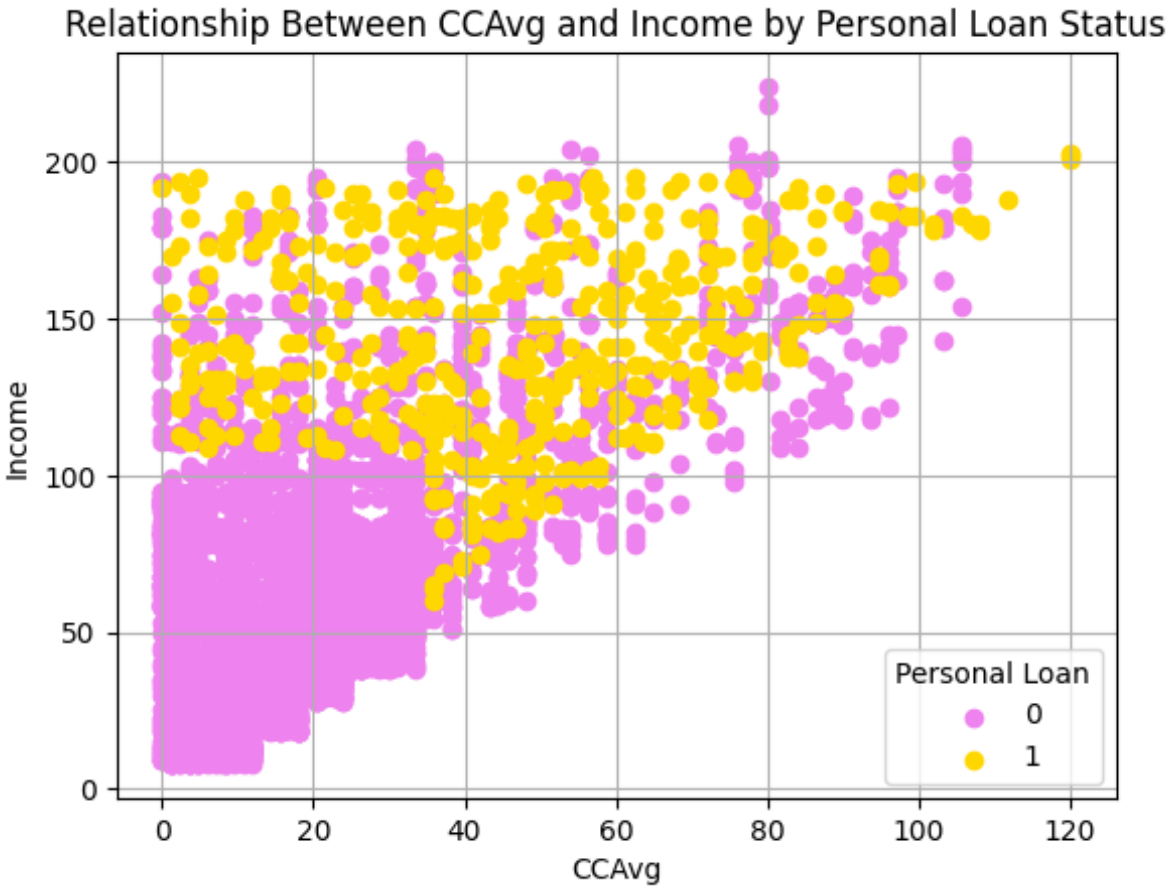
The heatmap reveals a positive correlation between age and professional experience, which means older individuals tend to have more work experience. There's a moderate correlation (0.65) between credit card spending (CCAvg) and income, giving the potential to explore the relationship between personal loan acceptance and higher income or credit card usage.



2.2 Relationship evaluation

The scatter plot demonstrates a strong positive triangular relationship among credit card spending, income, and personal loan acceptance. Yellow dots, which signify individuals who have accepted personal loans, depict that areas beyond an income threshold of 100 and credit card spending exceeding 35 are more likely to accept loan offers. This insight recommends that financial institutions prioritise marketing

efforts towards customers with incomes above 100, while also considering those with credit card spending exceeding 35 as potential targets.



In summary, understanding these connections enables client institutions to tailor loan offerings and marketing strategies more effectively. By analysing these relationships, institutions can refine risk assessment models, upskill sales staff, and enhance the accuracy of loan approvals, ultimately leading to increased customer satisfaction.