1. Reading Data

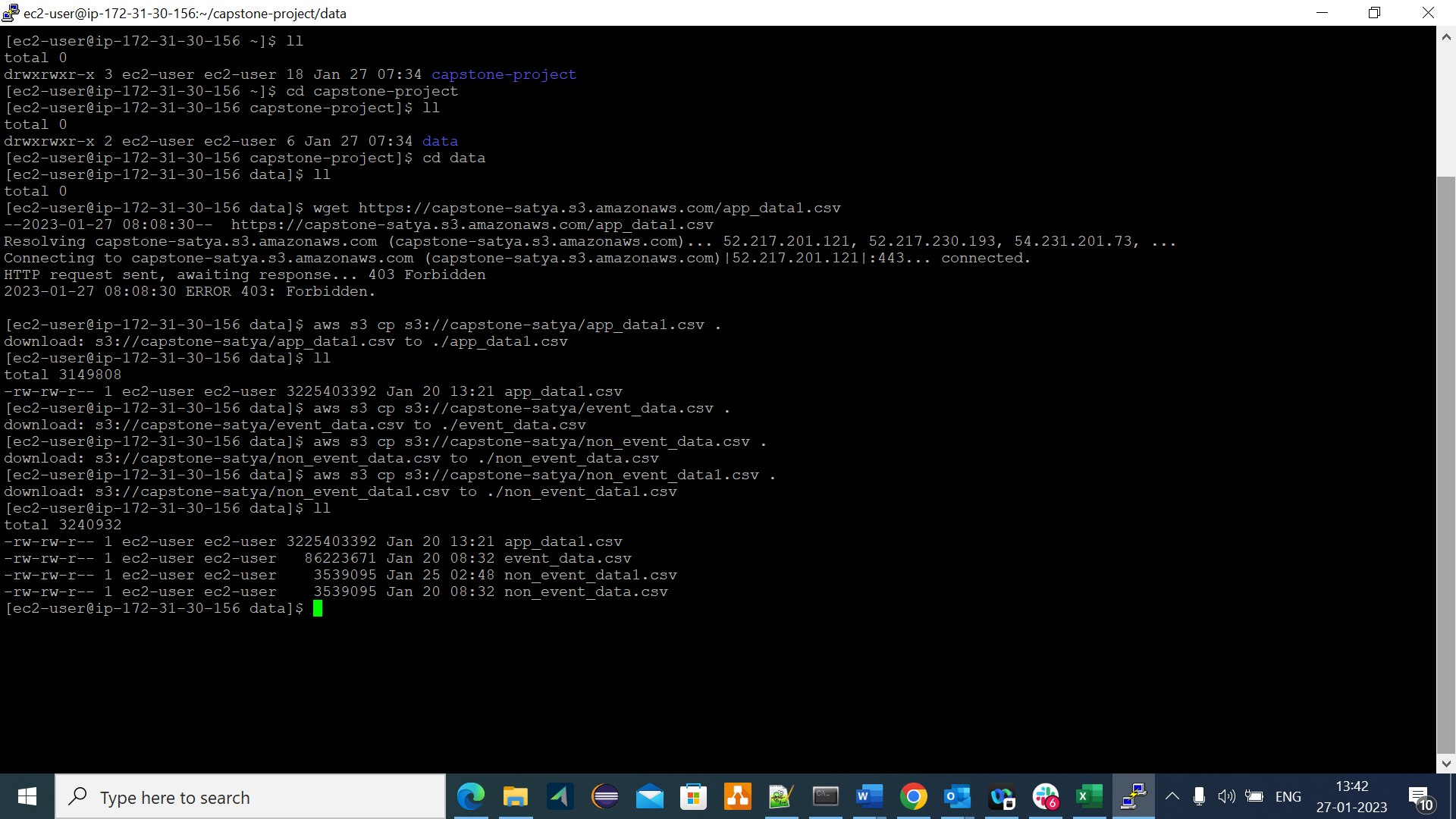
* Read the data from S3 in EC2

aws s3 cp s3://capstone-satya/app\_data1.csv .

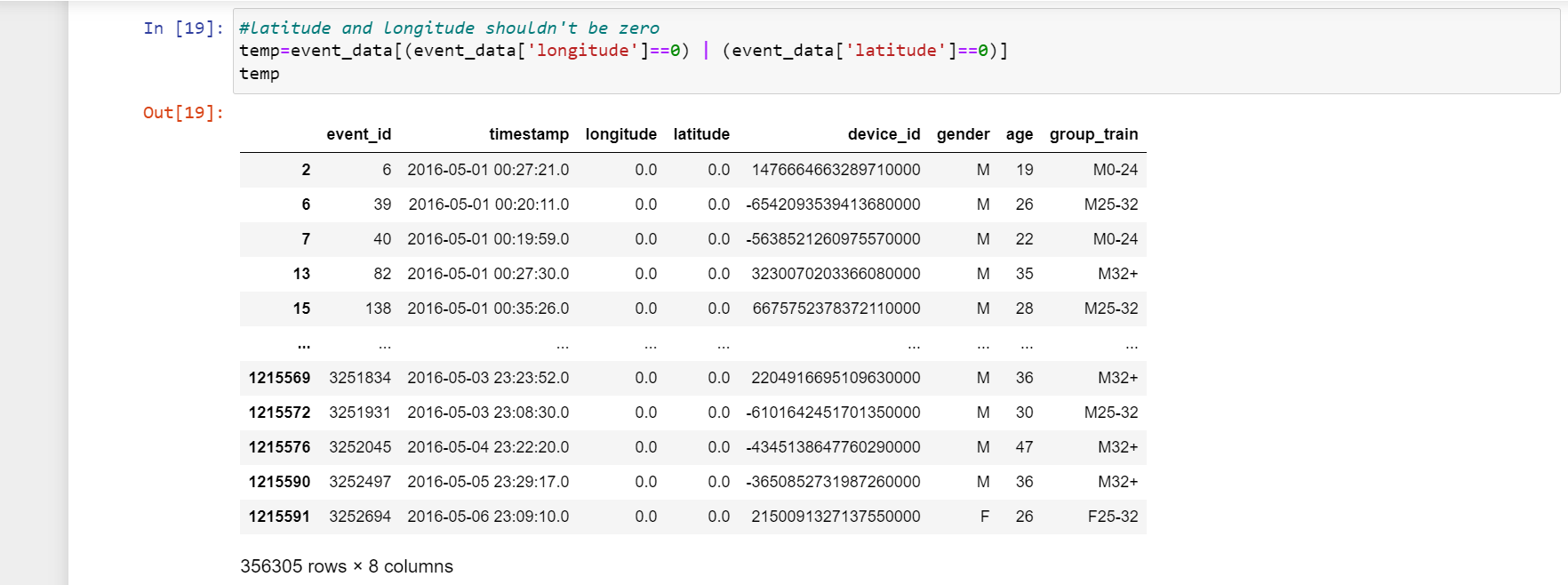
aws s3 cp s3://capstone-satya/event\_data.csv .

aws s3 cp s3://capstone-satya/non\_event\_data.csv .

aws s3 cp s3://capstone-satya/non\_event\_data1.csv .

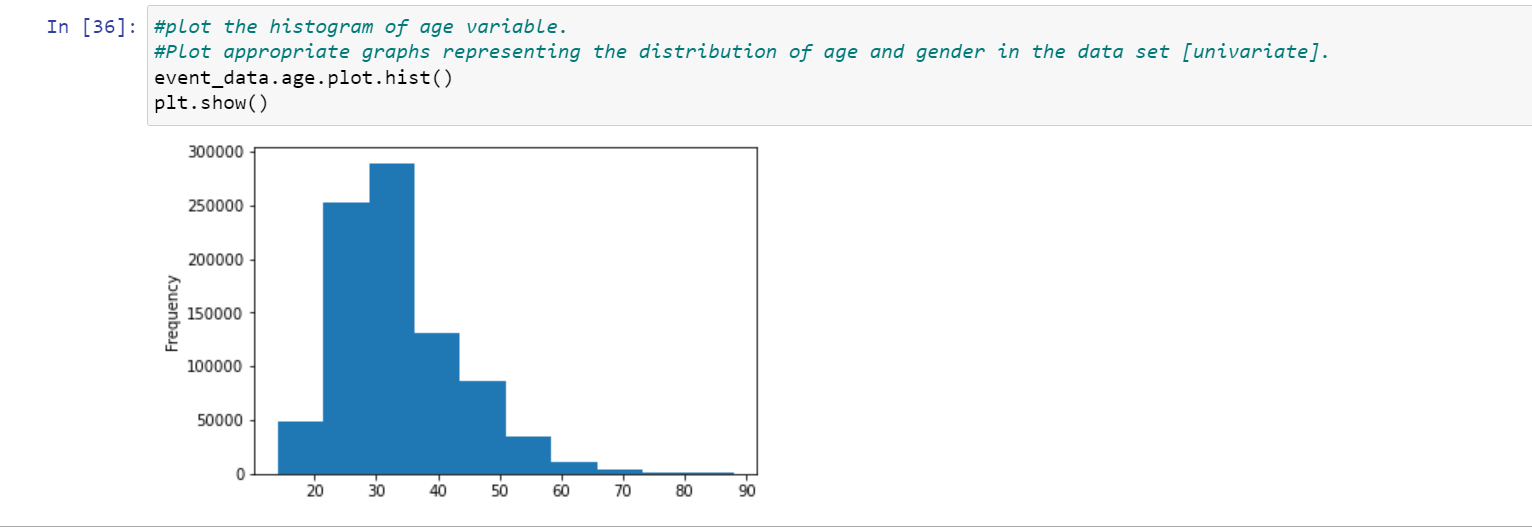


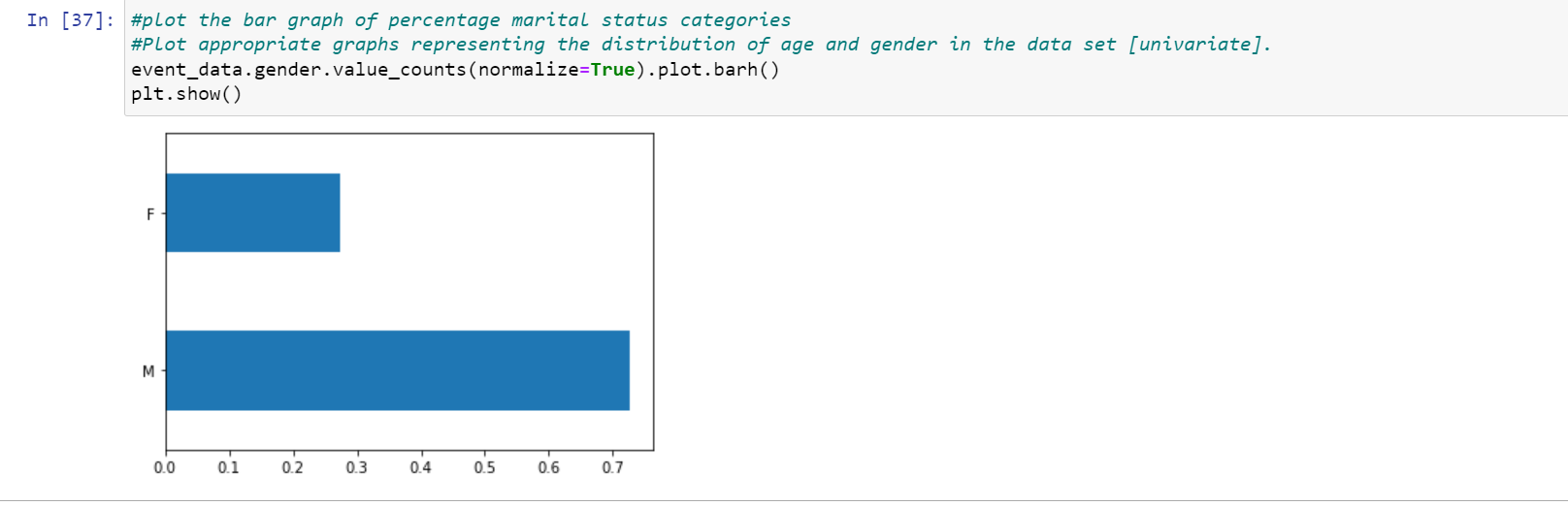
1. Cleaning data:



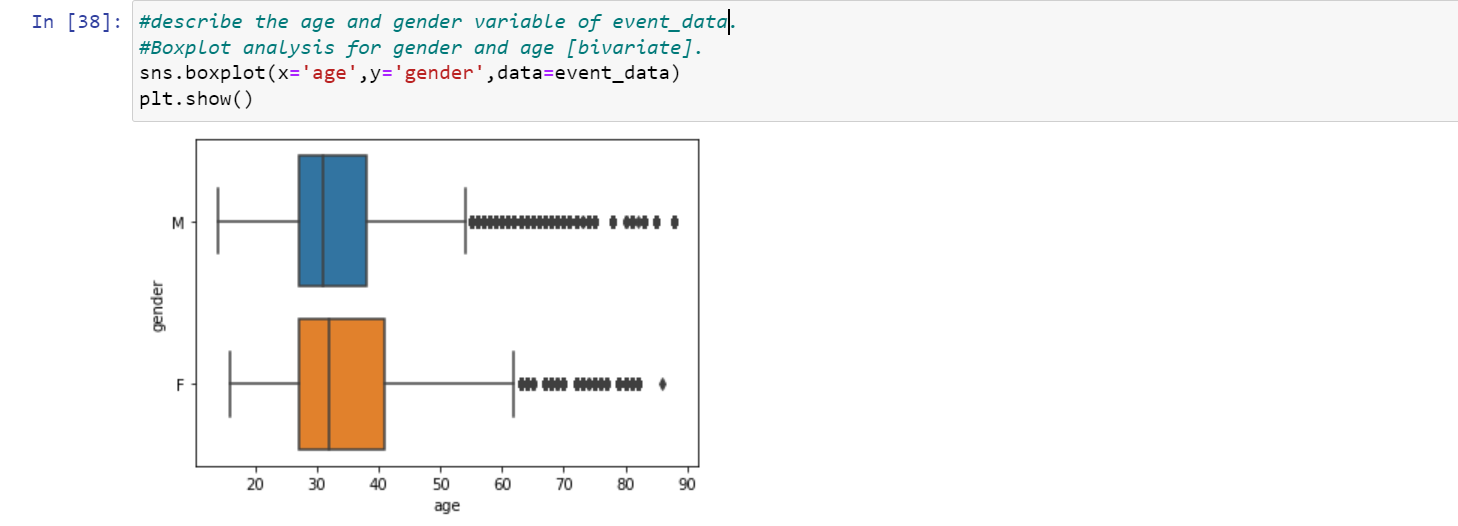
1. basic EDA and Visualisations

* Plot appropriate graphs representing the distribution of age and gender in the data set [univariate].

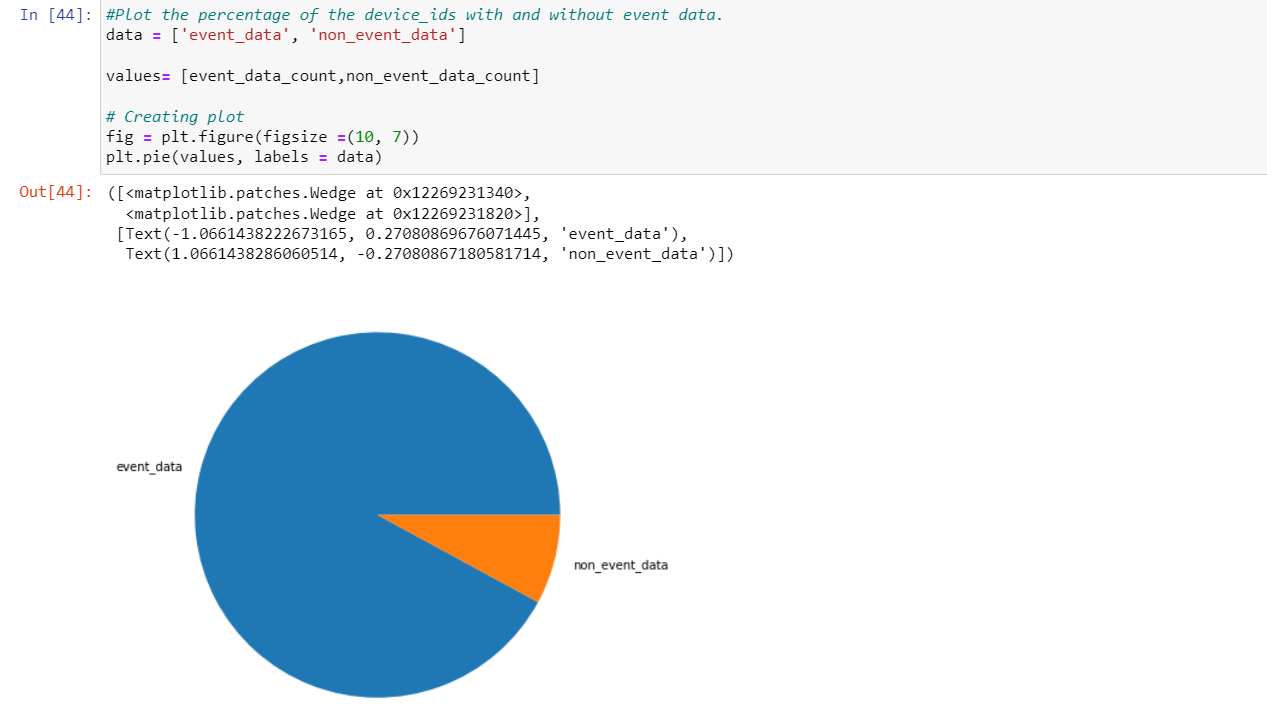




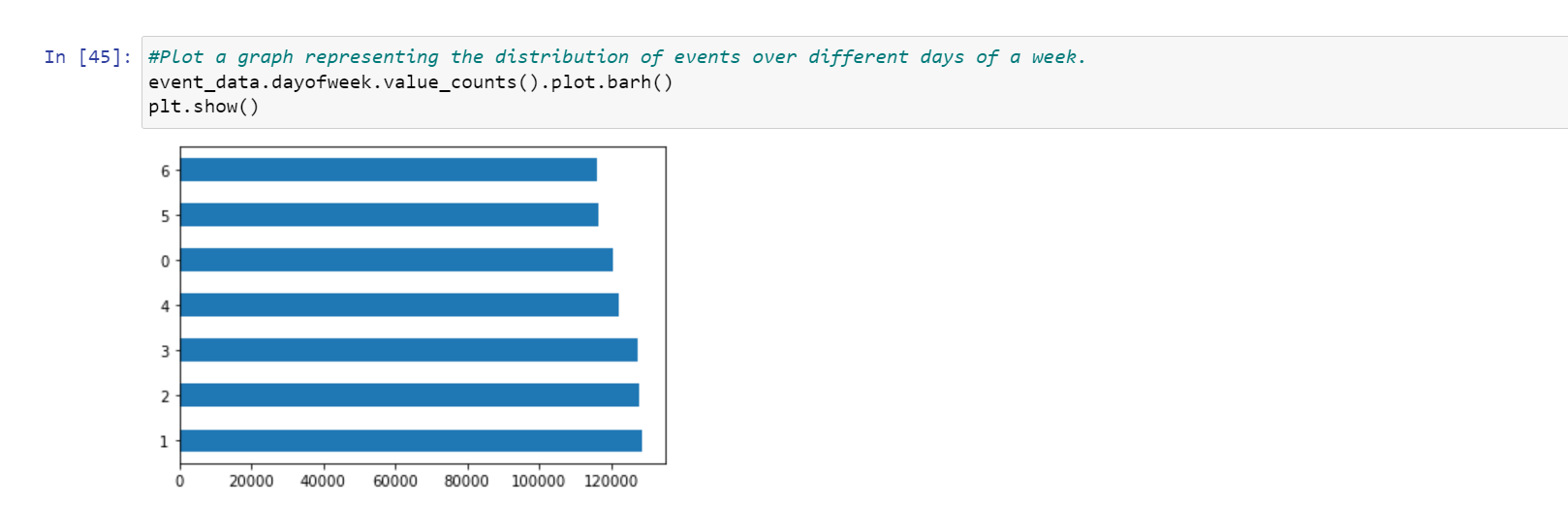
* Boxplot analysis for gender and age [bivariate].



* Plot the percentage of the device\_ids with and without event data.



* Plot a graph representing the distribution of events over different days of a week.



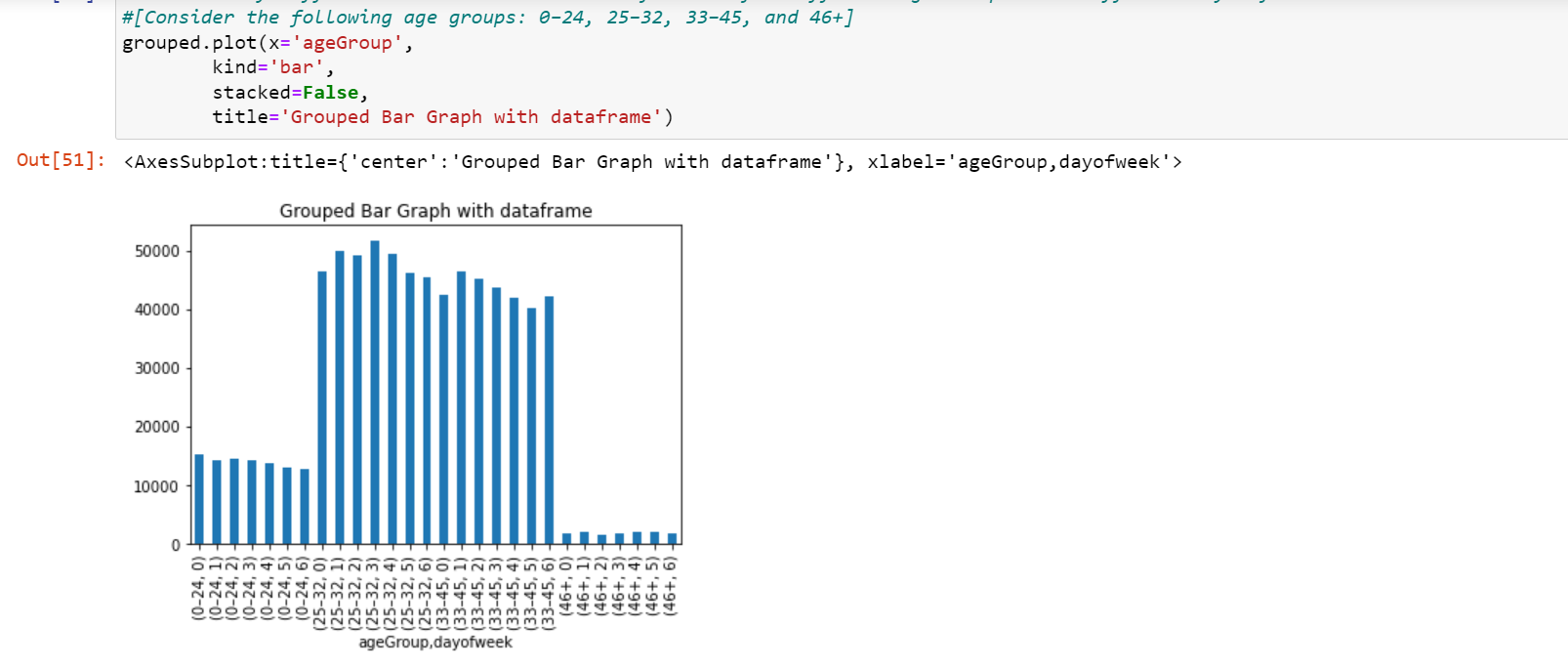
* Plot a graph representing the distribution of events per hour [for one-week data].

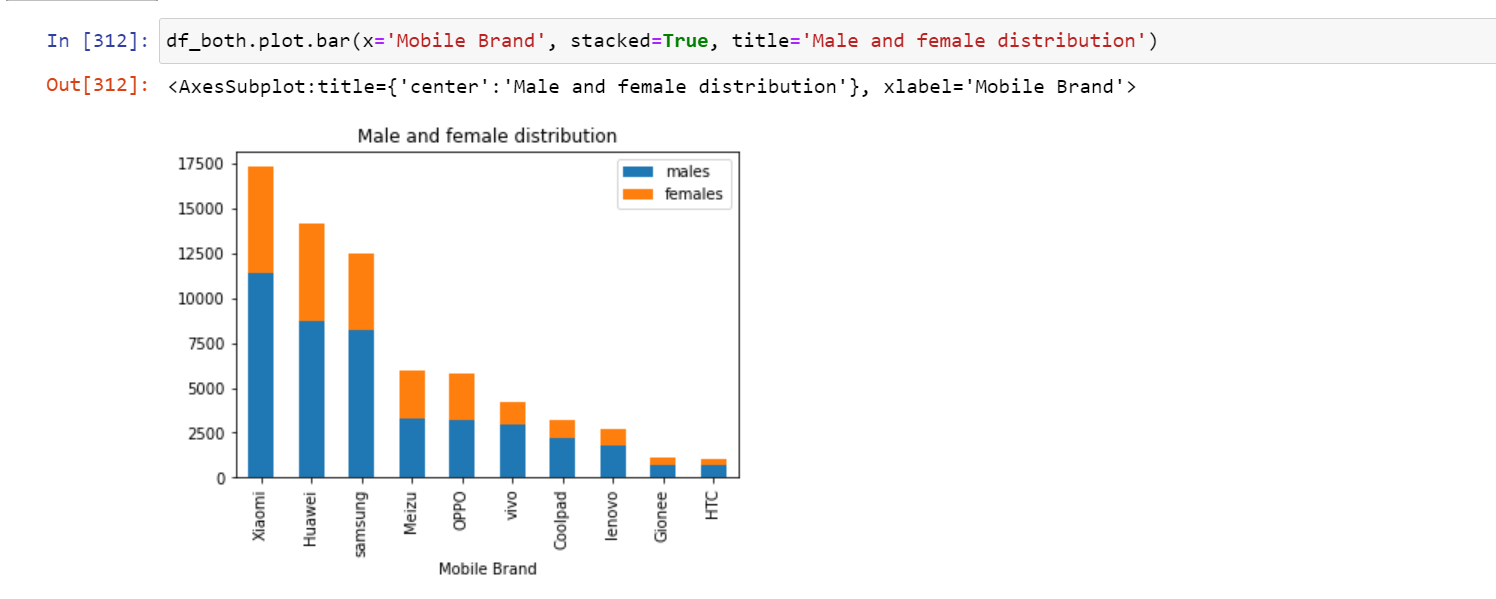


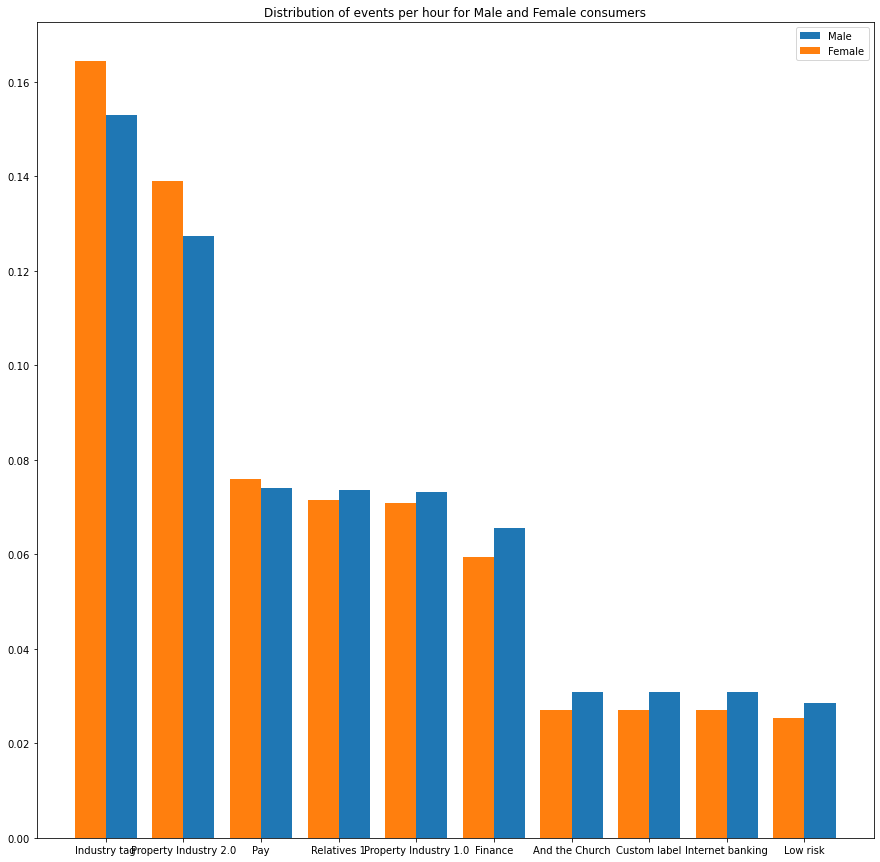
* The difference in the distribution of events per hour for Male and Female consumers. [Show the difference using an appropriate chart for one-week data.]



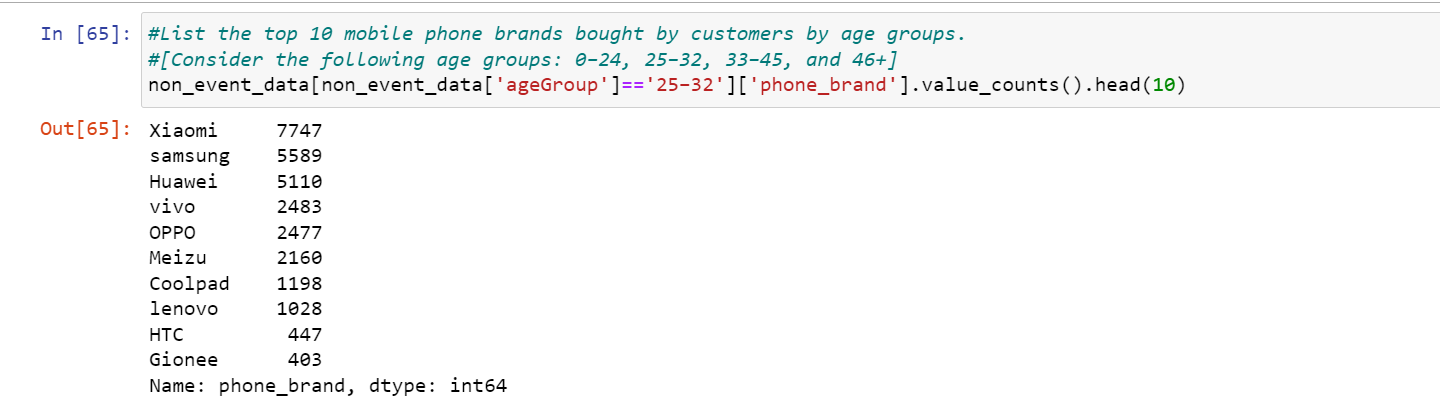
* Is there any difference in the distribution of Events for different Age Groups over different days of the week? [Consider the following age groups: 0–24, 25–32, 33–45, and 46+]



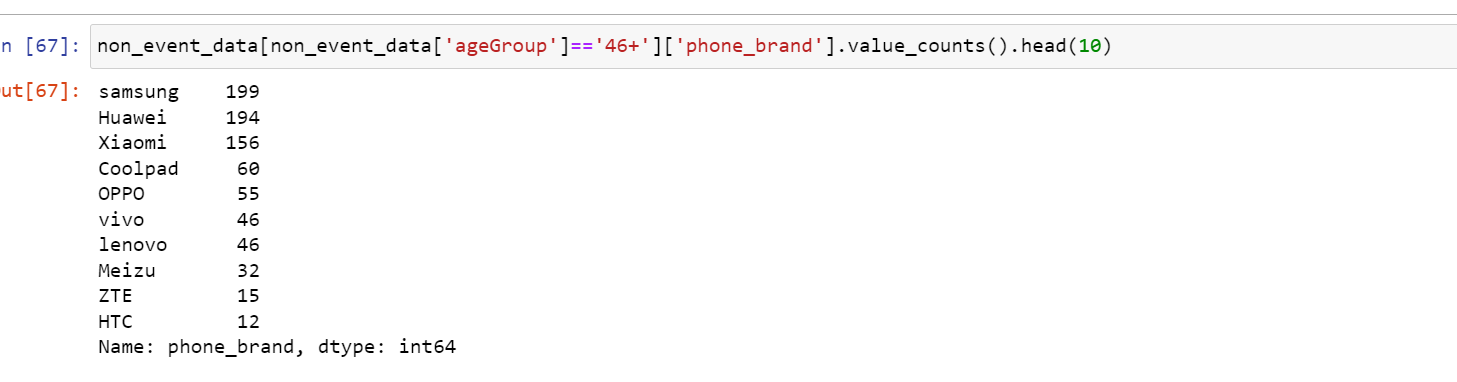
* Stacked bar chart for the top 10 mobile brands across male and female consumers.
* 
* Prepare a chart representing the ten frequently used applications and their respective male and female percentage.



* List the top 10 mobile phone brands bought by customers by age groups. [Consider the following age groups: 0–24, 25–32, 33–45, and 46+]



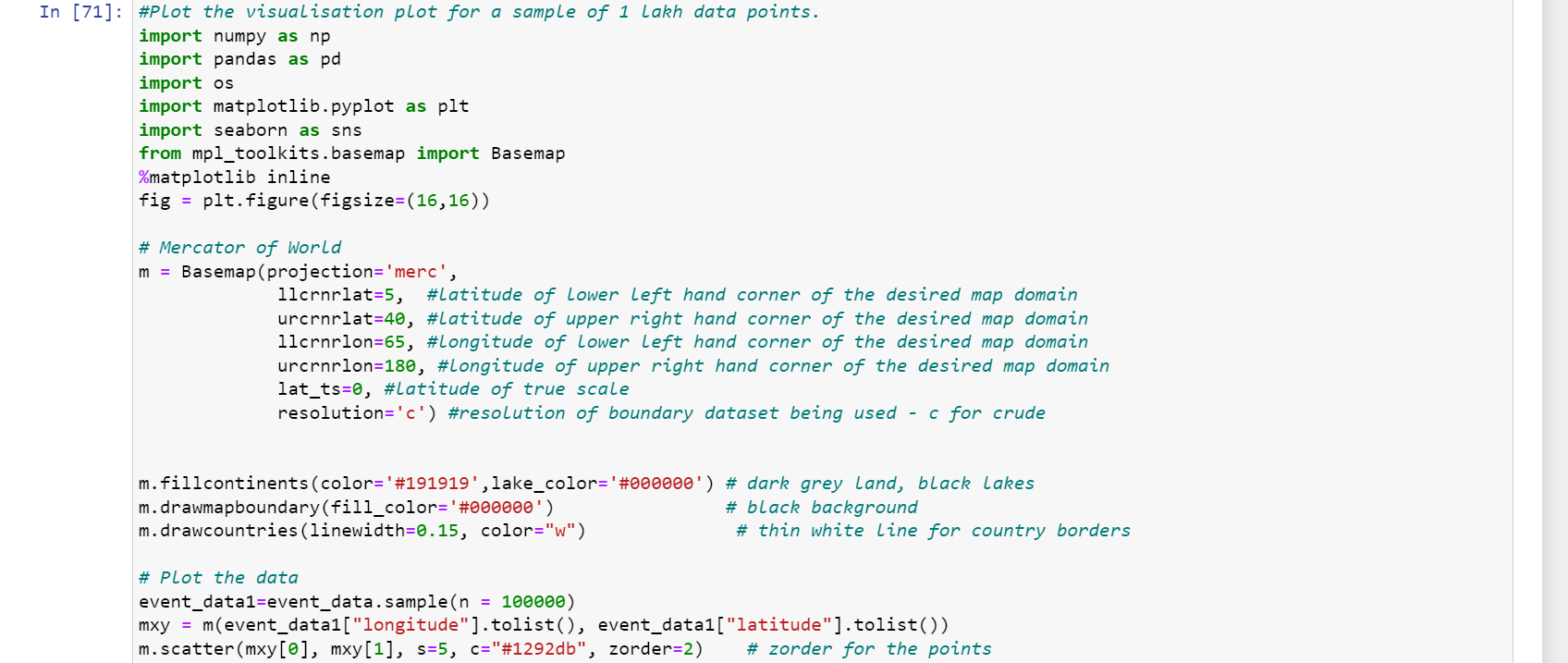


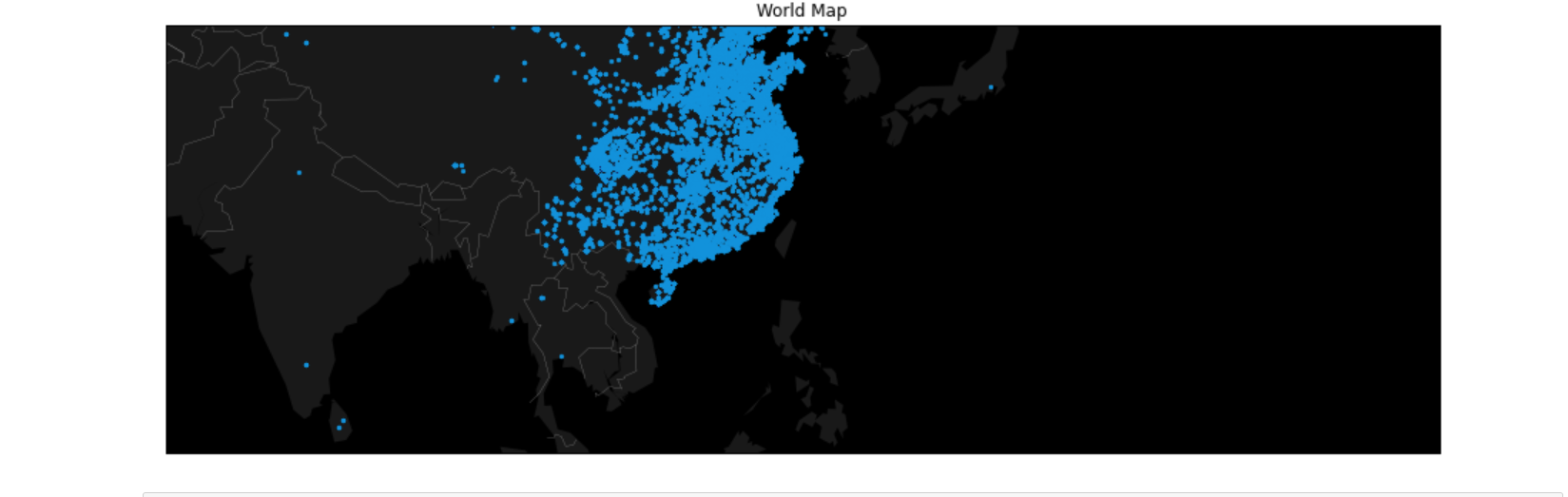




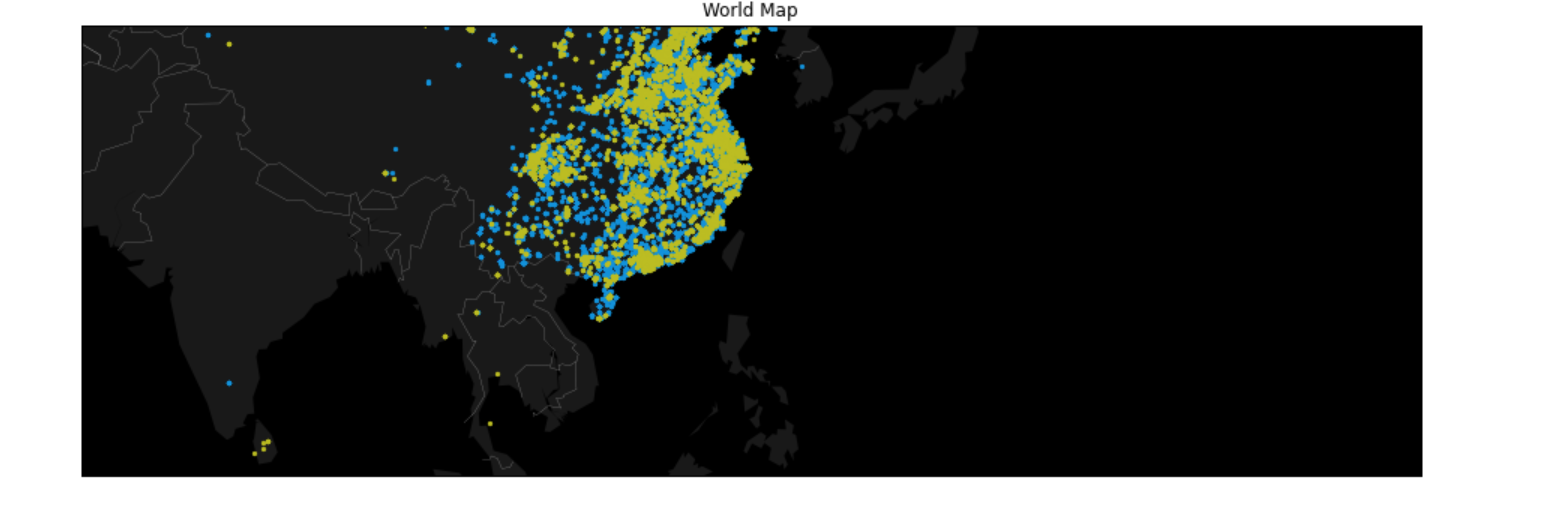
Advanced Visualization

* Plot the visualisation plot for a sample of 1 lakh data points.





* Compare the event visualisation plots based on the users' gender information. [This can be done on the sample of 1 lakh data points.]

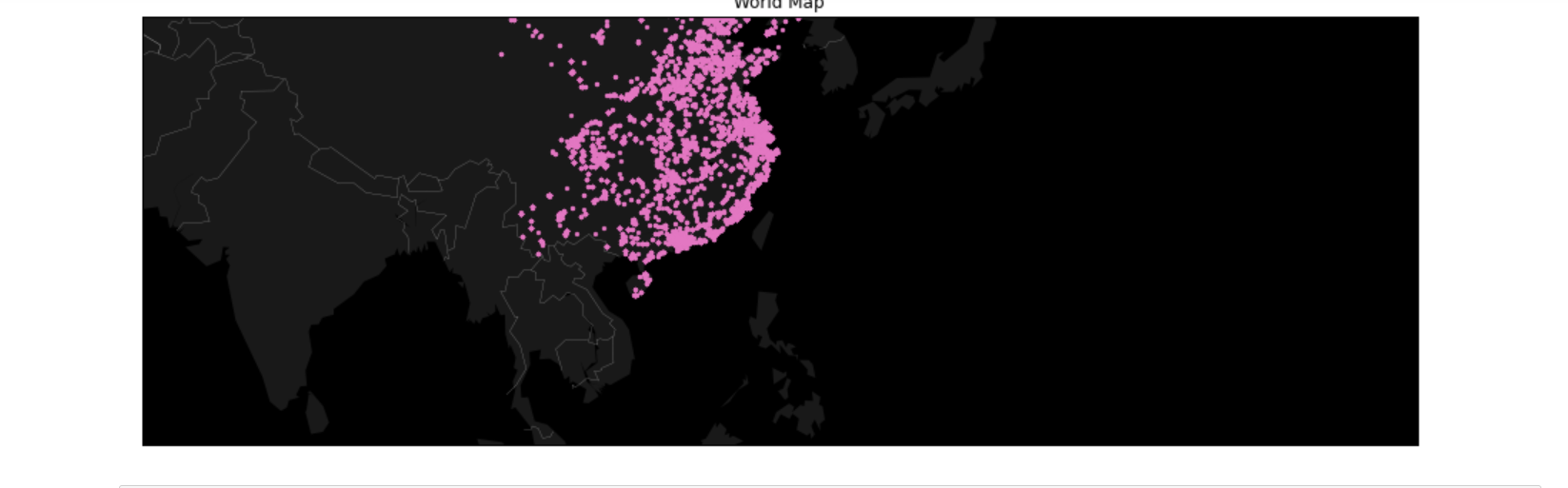


* Compare the event visualisation plots based on the following age groups:

0–24

25–32

32+



Clustering

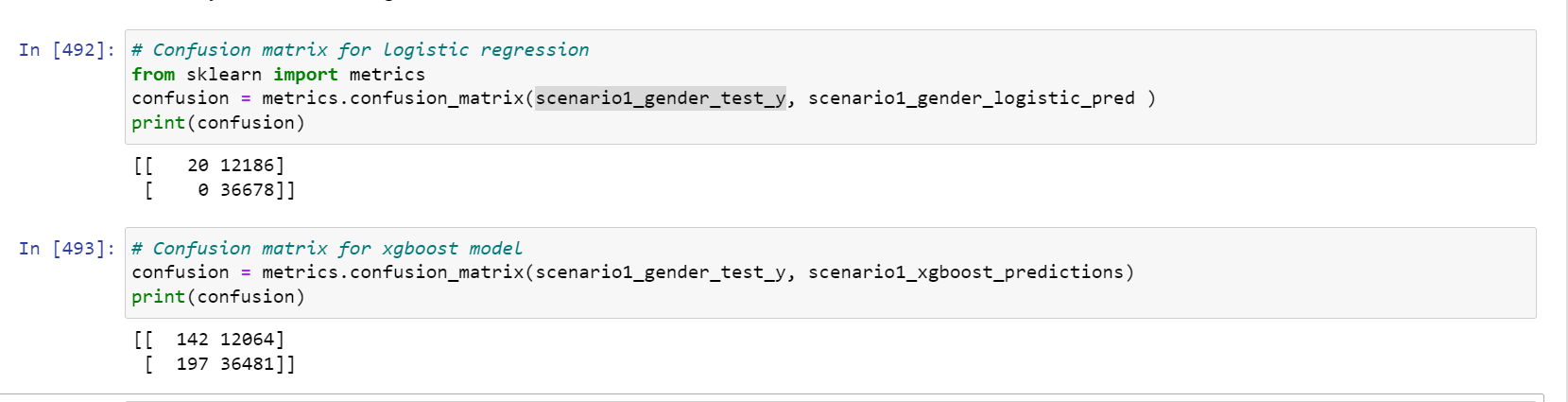


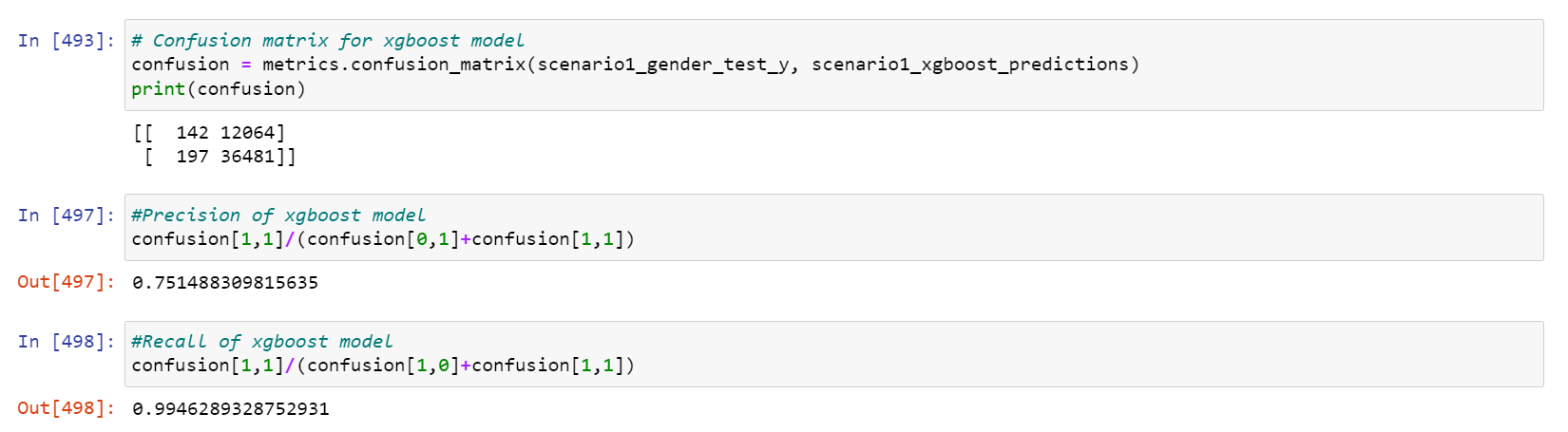
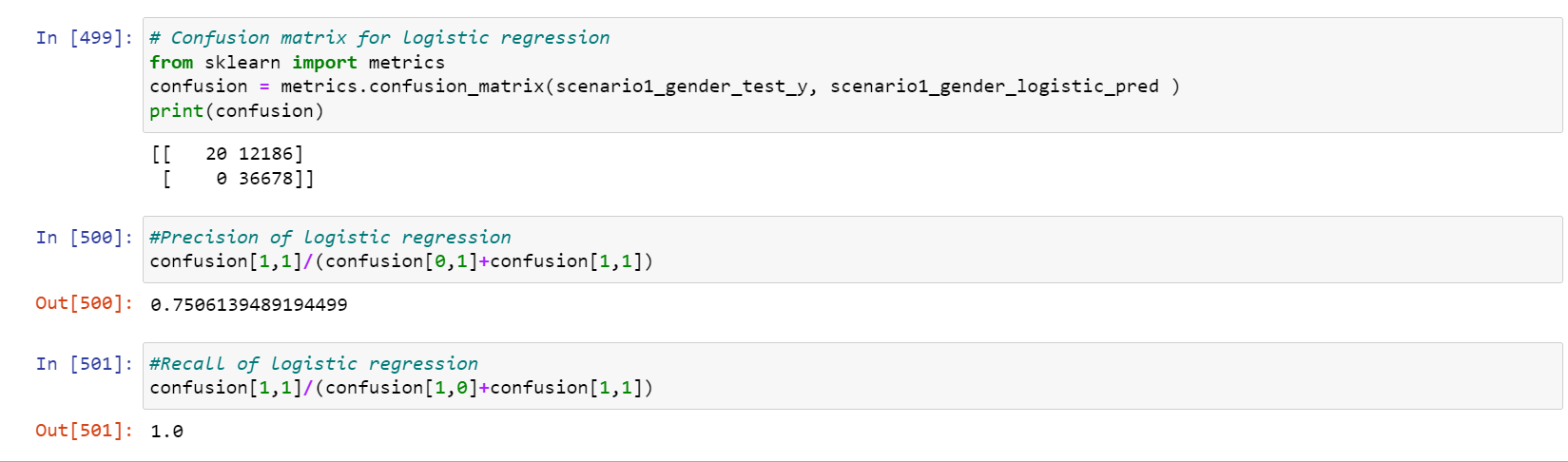


Accuracy



Confusion Matrix





ROC Curve



RMSE and R-square value

