TASK 3

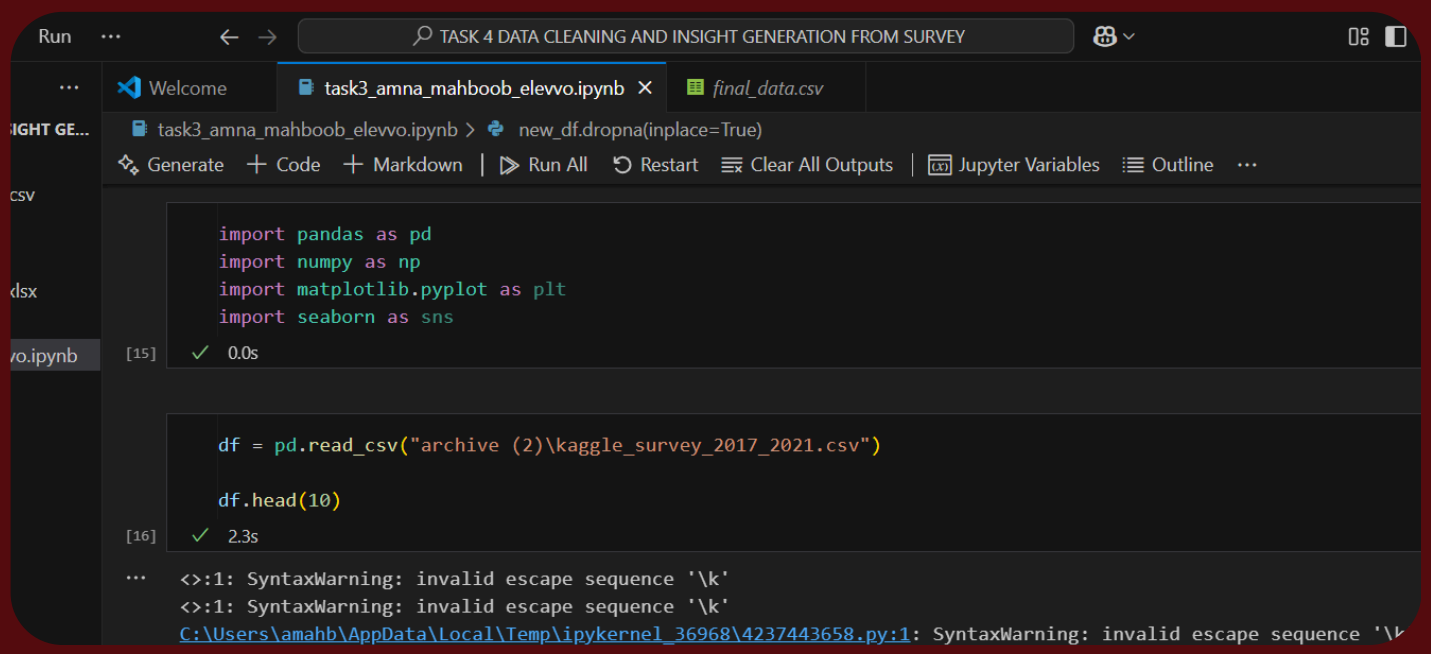
DATA CLEANING AND INSIGHT GENERATION FROM SURVEY DATA

AMNA MAHBOOB

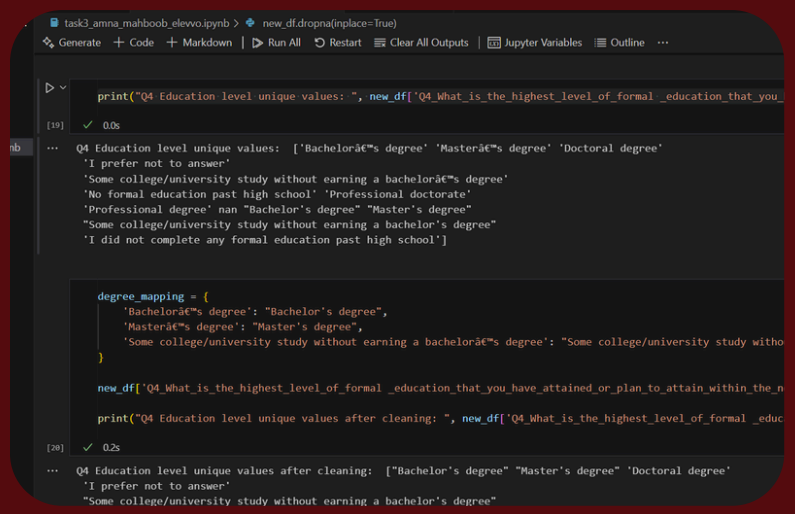
Elevvo
Data Analytics Intern

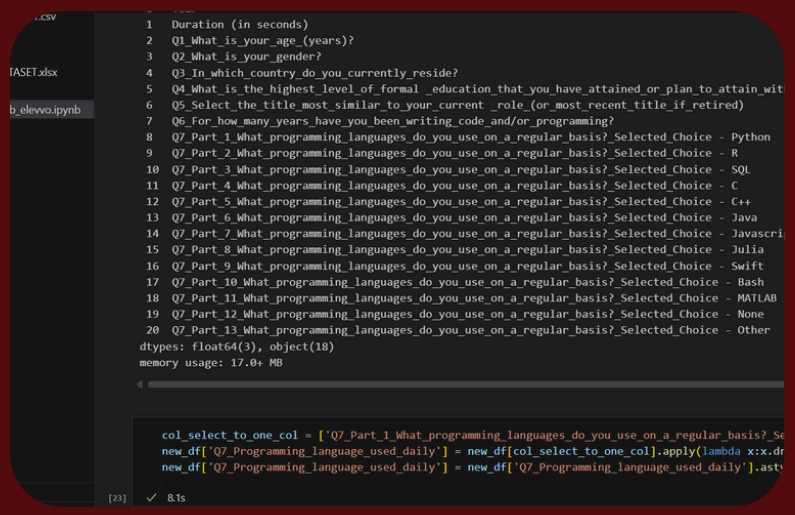
STEP 1: DEFINE PROJECT GOALS

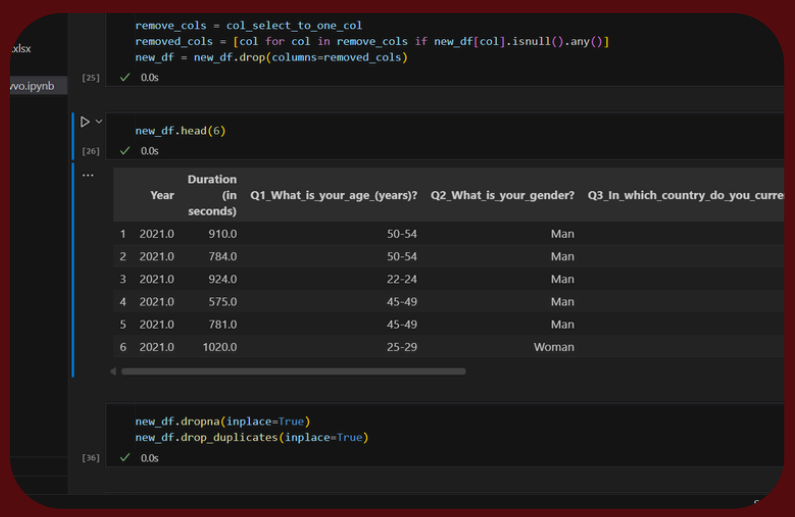
- Use the Kaggle Data Science Survey 2017-2021 to work with real-world survey data containing missing values, duplicates, and inconsistent formatting
- Clean the dataset and apply label encoding or mapping for categorical variables
- Extract meaningful insights about respondent (a person who answered the survey) behavior or preferences
- Create a summary dashboard or chart showing top 5 insights



STEP 2: CLEANING DATA

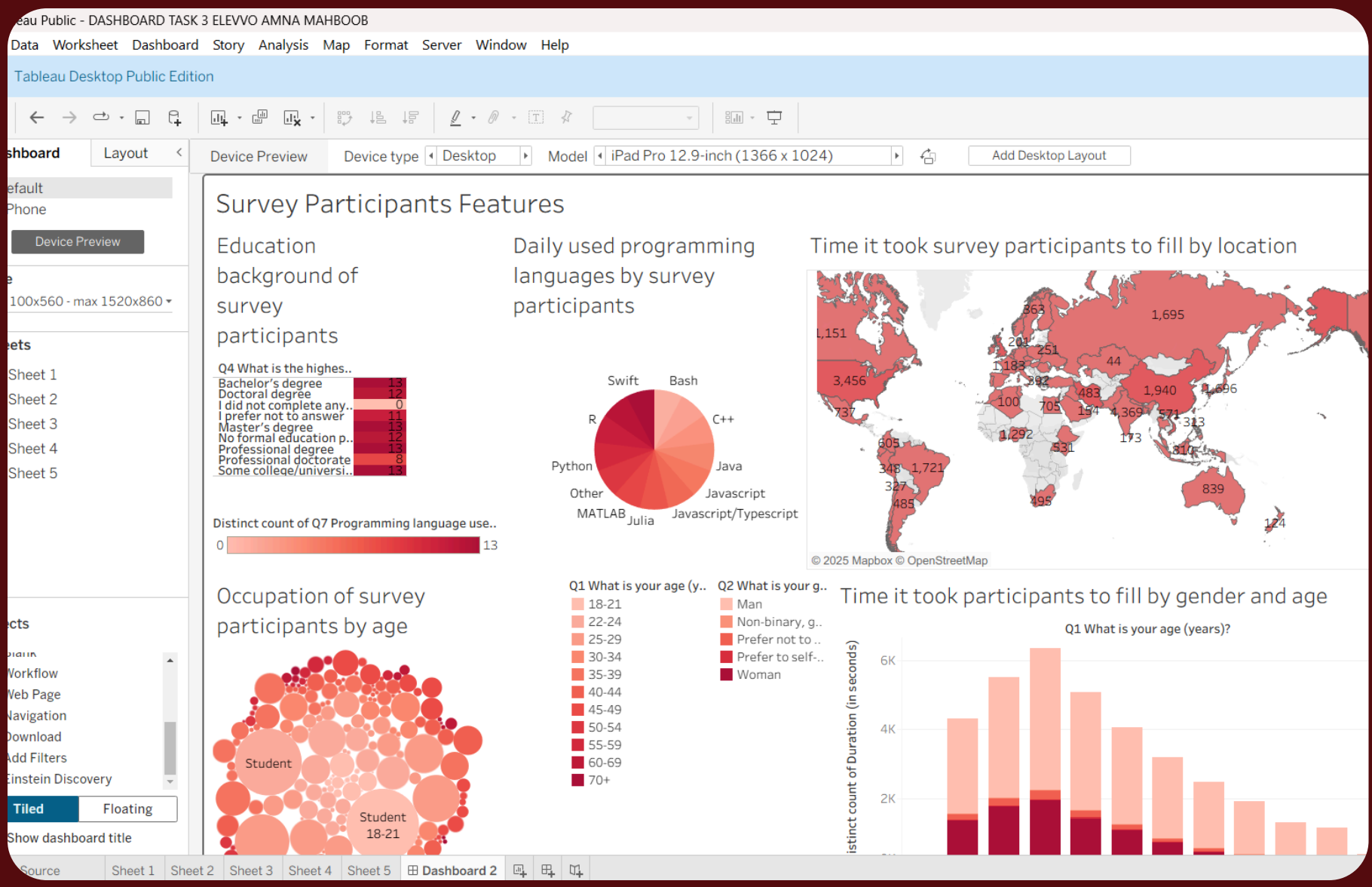






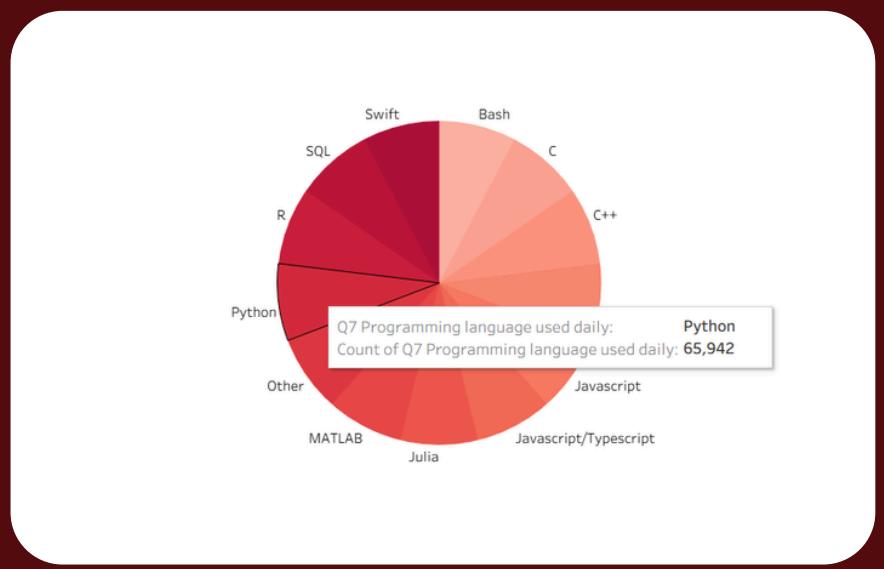
- Formatting of columns such as gender, education and so on by finding the unique values and removing those that are irrelevant or equating them to already existing ones
- Combining columns into one categorical column for the type of programming language most used
- Removing duplicates and null values after finishing the formatting and categorical handling

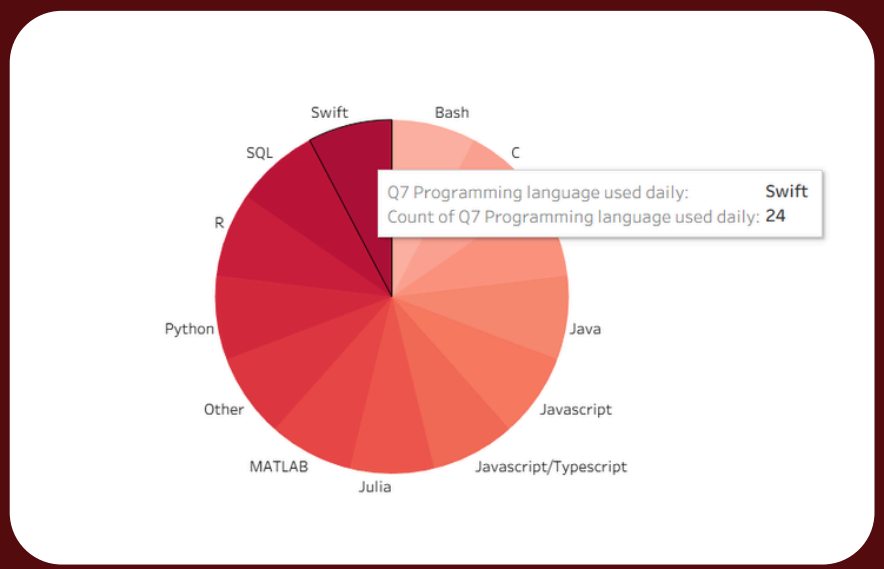
STEP 3: CREATING A DASHBOARD



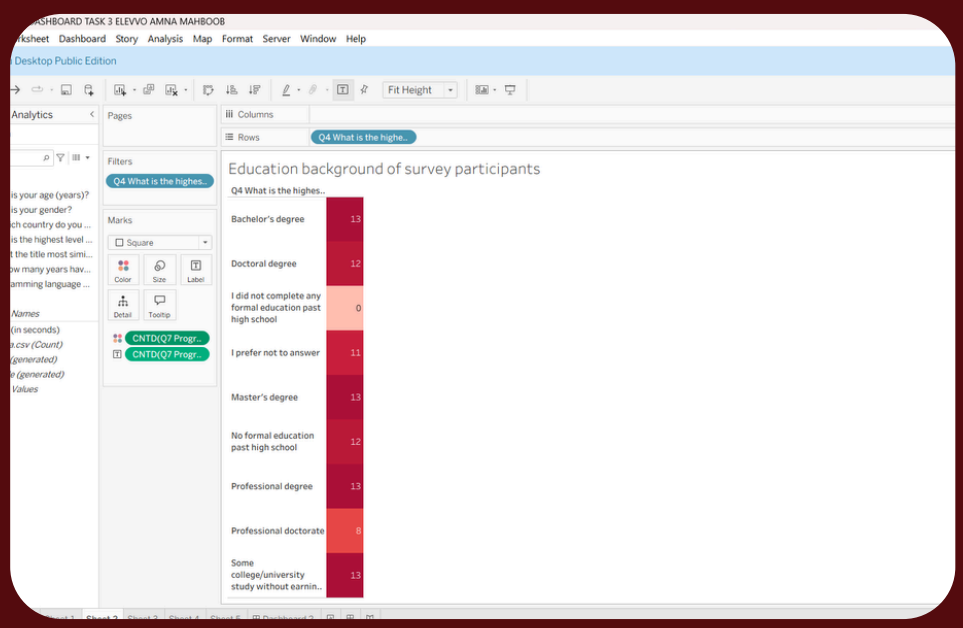
- Through Tableau we loaded the newly extracted dataset after cleaning
- Created separate infographics and merged it into 1 dashboard

STEP 3.1: VIEWING EACH INFOGRAPHIC



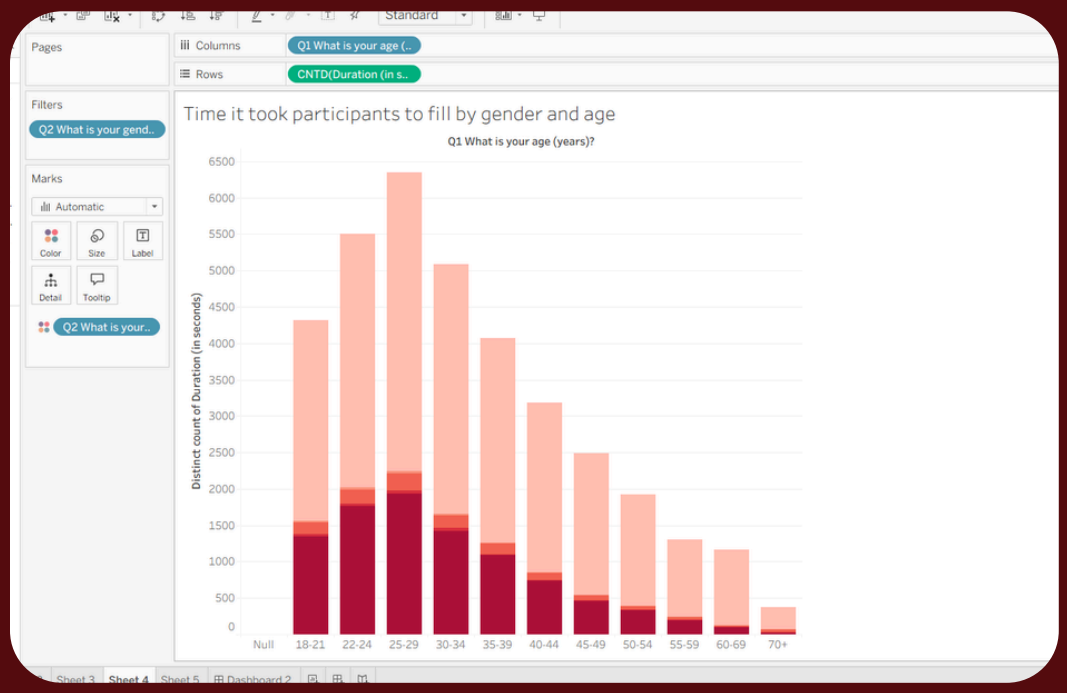


- Most used program = Python
- Least used program = Swift

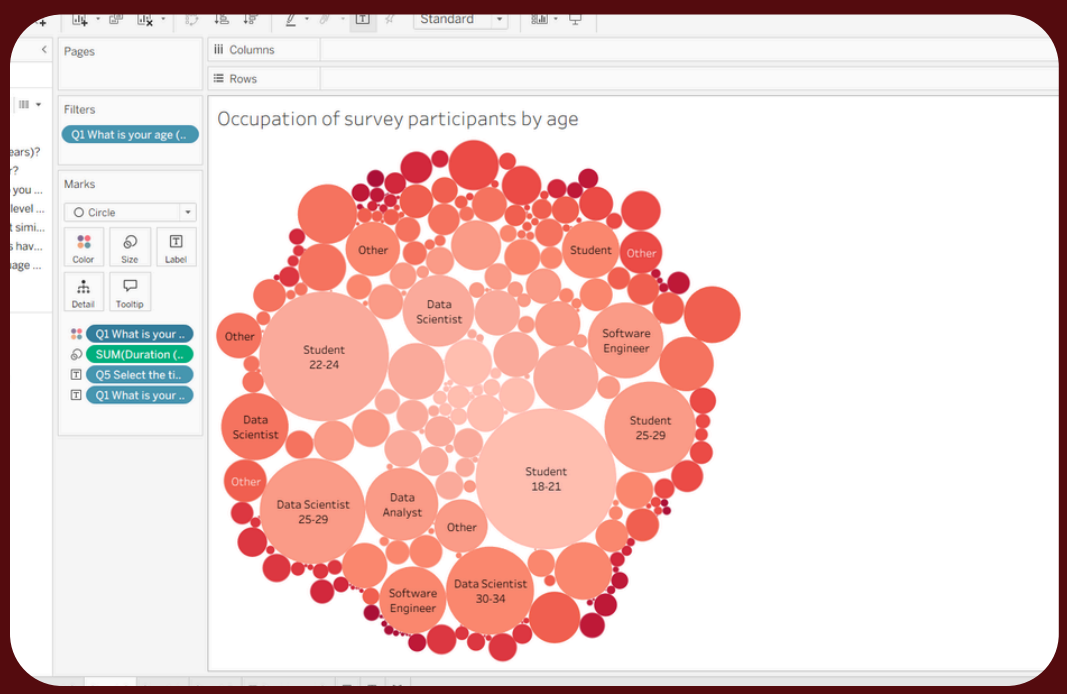


- Mainly master and bachelor students filled survey

STEP 3.2: VIEWING EACH INFOGRAPHIC

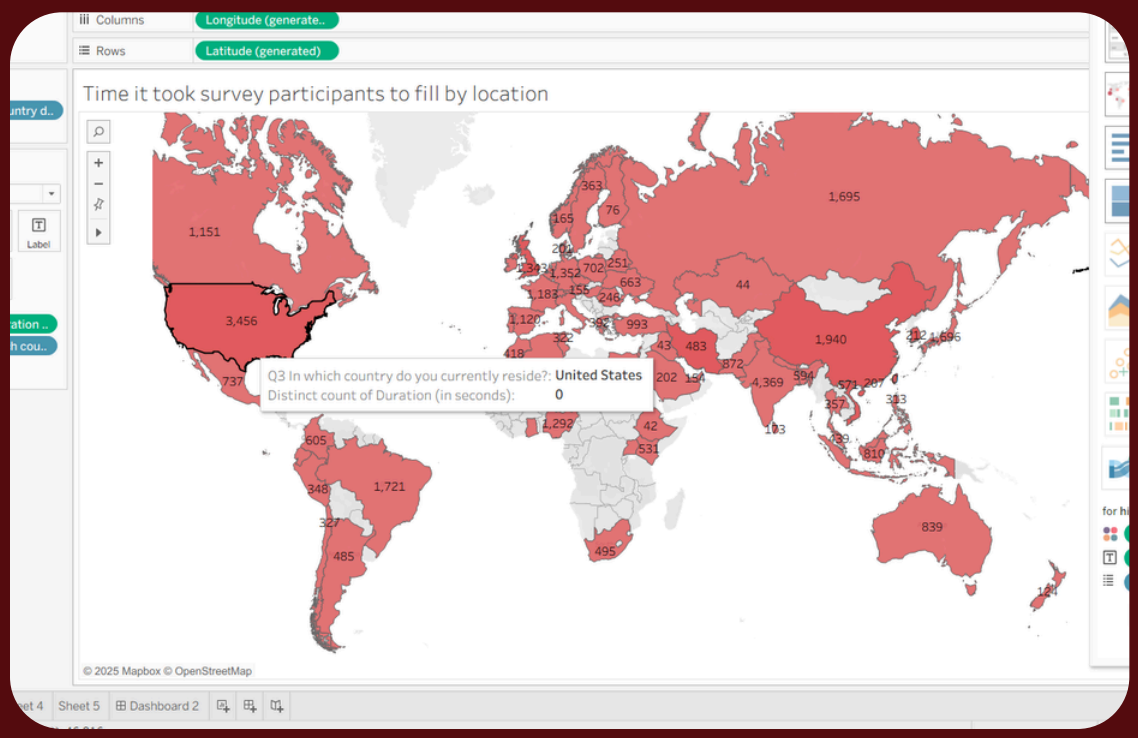


- The survey had mainly men fill it followed by women
- People from 25 to 29 filled this survey
- People above 70 had the least participation

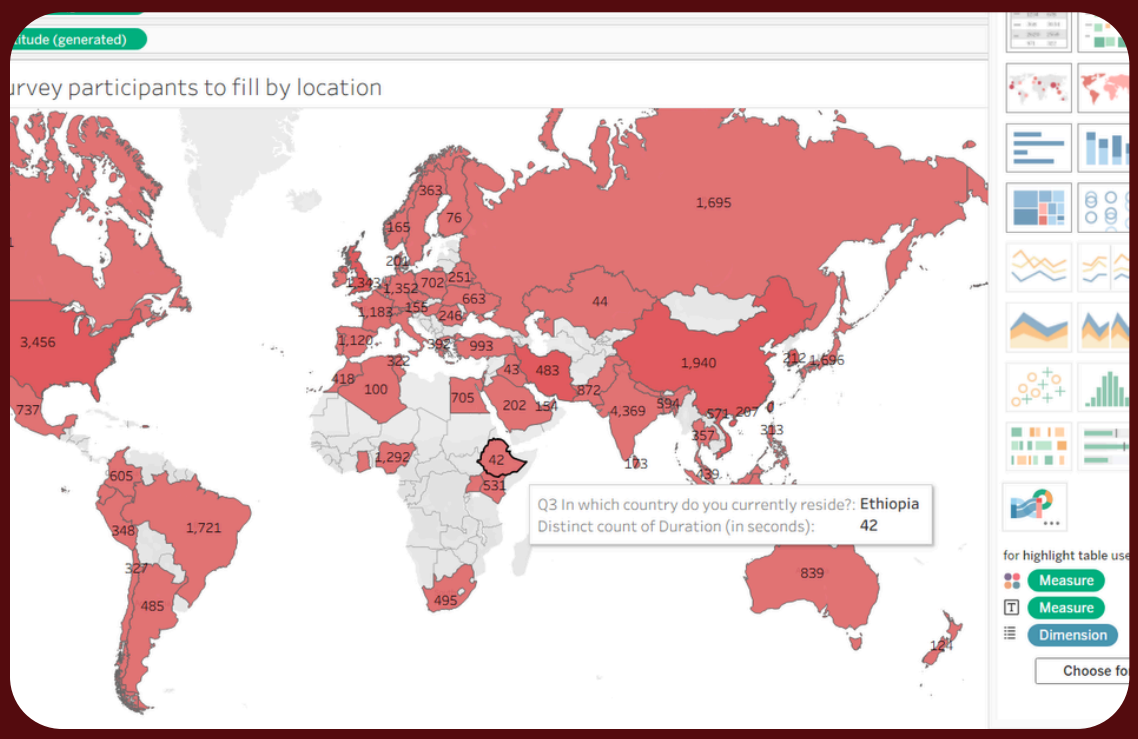


- Most users were students at the age of 18 to 21 or 22 to 24
- The 2nd most frequent group was data scientists of 25 to 29 years
- A couple of them were sales people and developer relations

STEP 3.3: VIEWING EACH INFOGRAPHIC



- Country that took the most to fill the survey = USA



- Country that took the least to fill the survey = Ethopia