

TASK 3

DATA CLEANING

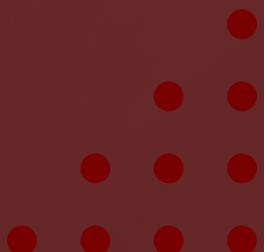
AND INSIGHT

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

The screenshot shows a Jupyter Notebook interface with the following details:

- Title Bar:** TASK 4 DATA CLEANING AND INSIGHT GENERATION FROM SURVEY
- File List:** Welcome, task3_amna_mahboob_eleovo.ipynb, final_data.csv
- Toolbar:** Generate, Code, Markdown, Run All, Restart, Clear All Outputs, Jupyter Variables, Outline, ...
- Code Cell 1:** [15]

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```
- Code Cell 2:** [16]

```
df = pd.read_csv("archive (2)\kaggle_survey_2017_2021.csv")
df.head(10)
```
- Output:** [16]

```
2.3s
```
- Warnings:** <>:1: SyntaxWarning: invalid escape sequence '\k'
<>:1: SyntaxWarning: invalid escape sequence '\k'
- File List (Left):** CSV, dsx, ipynb

STEP 2: CLEANING DATA

```
task3_anna.mahboob_eleovo.ipynb > new_df.dropna(inplace=True)
Generate + Code + Markdown | Run All | Restart | Clear All Outputs | Jupyter Variables | Outline ...
[19] ✓ 0s
nb ... Q4 Education level unique values: ['Bachelor's degree' 'Master's degree' 'Doctoral degree'
'I prefer not to answer'
'Some college/university study without earning a bachelor's degree'
'No formal education past high school' 'Professional doctorate'
'Professional degree' nan 'Bachelor's degree' "Master's degree"
"Some college/university study without earning a bachelor's degree"
'I did not complete any formal education past high school']

degree_mapping = [
    'Bachelor\'s degree': "Bachelor's degree",
    'Master\'s degree': "Master's degree",
    'Some college/university study without earning a bachelor\'s degree': "Some college/university study without earning a bachelor's degree"
]

new_df['Q4.What_is_the_highest_level_of_formal_education_that_you_have_attained_or_plan_to_attain_within_the_next_5_years'] = new_df['Q4.What_is_the_highest_level_of_formal_education_that_you_have_attained_or_plan_to_attain_within_the_next_5_years'].map(degree_mapping)

print("Q4 Education level unique values after cleaning: ", new_df['Q4.What_is_the_highest_level_of_formal_education_that_you_have_attained_or_plan_to_attain_within_the_next_5_years'])

[20] ✓ 0s
... Q4 Education level unique values after cleaning: ["Bachelor's degree" "Master's degree" 'Doctoral degree'
'I prefer not to answer'
"Some college/university study without earning a bachelor's degree"]
```

- 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

```
..CSV  
TASET.xlsx  
b_elevo.ipynb  
  
1 Duration (in seconds)  
2 Q1_What_is_your_age_(years)?  
3 Q2_What_is_your_gender?  
4 Q3_In_which_country_do_you_currently_reside?  
5 Q4_What_is_the_highest_level_of_formal_education_that_you_have_attained_or_plan_to_attain_wit  
6 Q5 Select the title most similar to your current role (or most recent title if retired)  
7 Q6 For how many years have you been writing code and/or programming?  
8 Q7_Part_1_What_programming_languages_do_you_use_on_a_regular_basis? Selected_Choice - Python  
9 Q7_Part_2_What_programming_languages_do_you_use_on_a_regular_basis? Selected_Choice - R  
10 Q7_Part_3_What_programming_languages_do_you_use_on_a_regular_basis? Selected_Choice - SQL  
11 Q7_Part_4_What_programming_languages_do_you_use_on_a_regular_basis? Selected_Choice - C  
12 Q7_Part_5_What_programming_languages_do_you_use_on_a_regular_basis? Selected_Choice - C++  
13 Q7_Part_6_What_programming_languages_do_you_use_on_a_regular_basis? Selected_Choice - Java  
14 Q7_Part_7_What_programming_languages_do_you_use_on_a_regular_basis? Selected_Choice - Javascript  
15 Q7_Part_8_What_programming_languages_do_you_use_on_a_regular_basis? Selected_Choice - Julia  
16 Q7_Part_9_What_programming_languages_do_you_use_on_a_regular_basis? Selected_Choice - Swift  
17 Q7_Part_10_What_programming_languages_do_you_use_on_a_regular_basis? Selected_Choice - Bash  
18 Q7_Part_11_What_programming_languages_do_you_use_on_a_regular_basis? Selected_Choice - MATLAB  
19 Q7_Part_12_What_programming_languages_do_you_use_on_a_regular_basis? Selected_Choice - None  
20 Q7_Part_13_What_programming_languages_do_you_use_on_a_regular_basis? Selected_Choice - Other  
dtypes: float64(3), object(18)  
memory usage: 17.0+ MB  
  
col_select_to_one_col = ['Q7_Part_1_What_programming_languages_do_you_use_on_a_regular_basis?',  
new_df['Q7_Programming_language_used_daily'] = new_df[col_select_to_one_col].apply(lambda x: x[0])  
new_df['Q7_Programming_language_used_daily'] = new_df['Q7_Programming_language_used_daily'].astype('category')  
  
[23] ✓ 8.1s
```

- Combining columns into one categorical column for the type of programming language most used

xlsx

[25] ✓ 0.0s

```
remove_cols = col_select_to_one_col
removed_cols = [col for col in remove_cols if new_df[col].isnull().any()]
new_df = new_df.drop(columns=removed_cols)
```

[26] ✓ 0.0s

...
Duration
Year (in seconds)
Q1_What_is_your_age_(years)? Q2_What_is_your_gender? Q3_In_which_country_do_you_curre

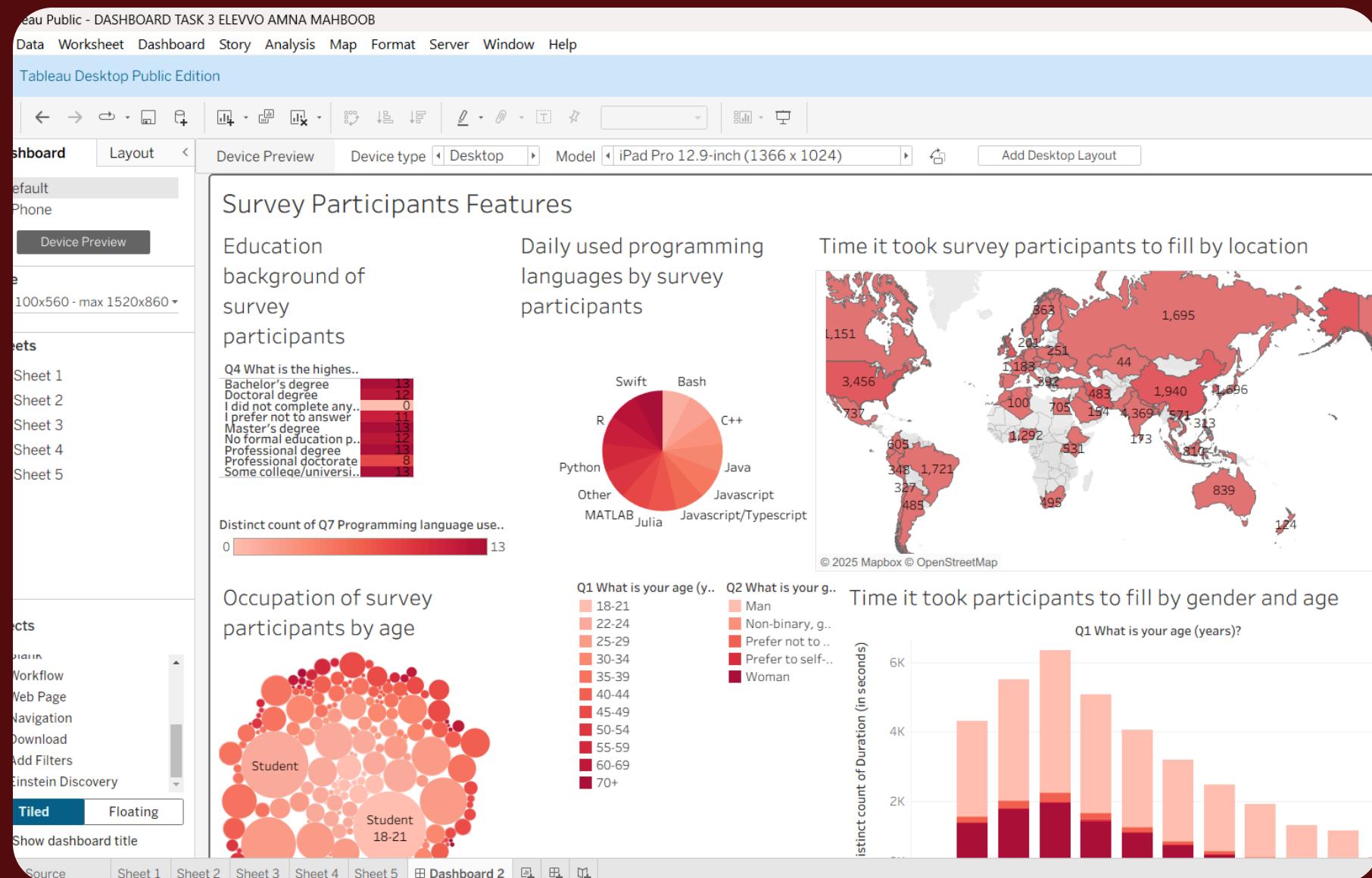
	Year	Duration (in seconds)	Q1_What_is_your_age_(years)?	Q2_What_is_your_gender?	Q3_In_which_country_do_you_curre
1	2021.0	910.0	50-54	Man	
2	2021.0	784.0	50-54	Man	
3	2021.0	924.0	22-24	Man	
4	2021.0	575.0	45-49	Man	
5	2021.0	781.0	45-49	Man	
6	2021.0	1020.0	25-29	Woman	

[36] ✓ 0.0s

```
new_df.dropna(inplace=True)
new_df.drop_duplicates(inplace=True)
```

- Removing duplicates and null values after finishing the formatting and categorical handling

STEP 3: CREATING A DASHBOARD

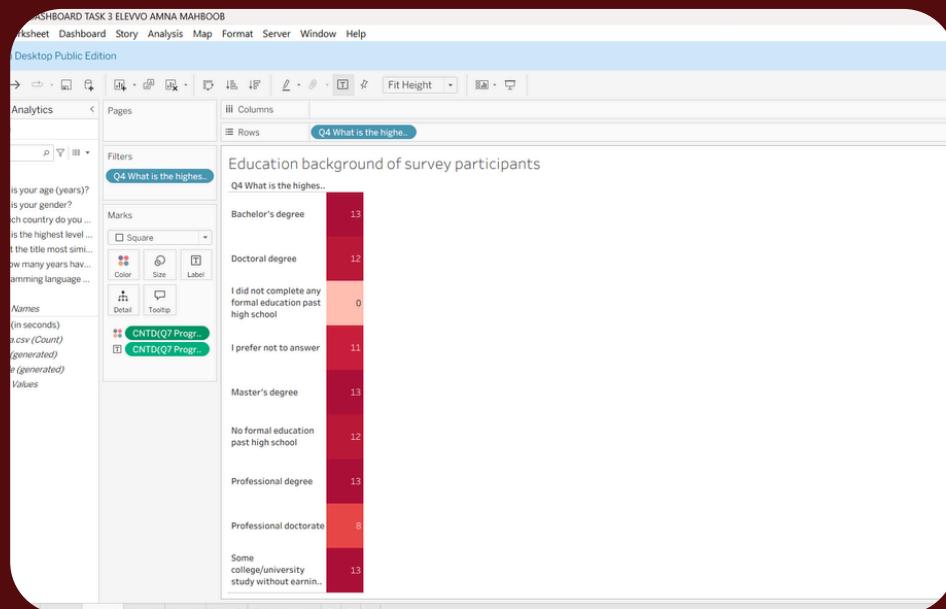


- Through Tableau we loaded the newly extracted dataset after cleaning
- Created separate infographics and merged it into 1 dashboard

STEP 3.1: VIEWING EACH INFOGRAPHIC

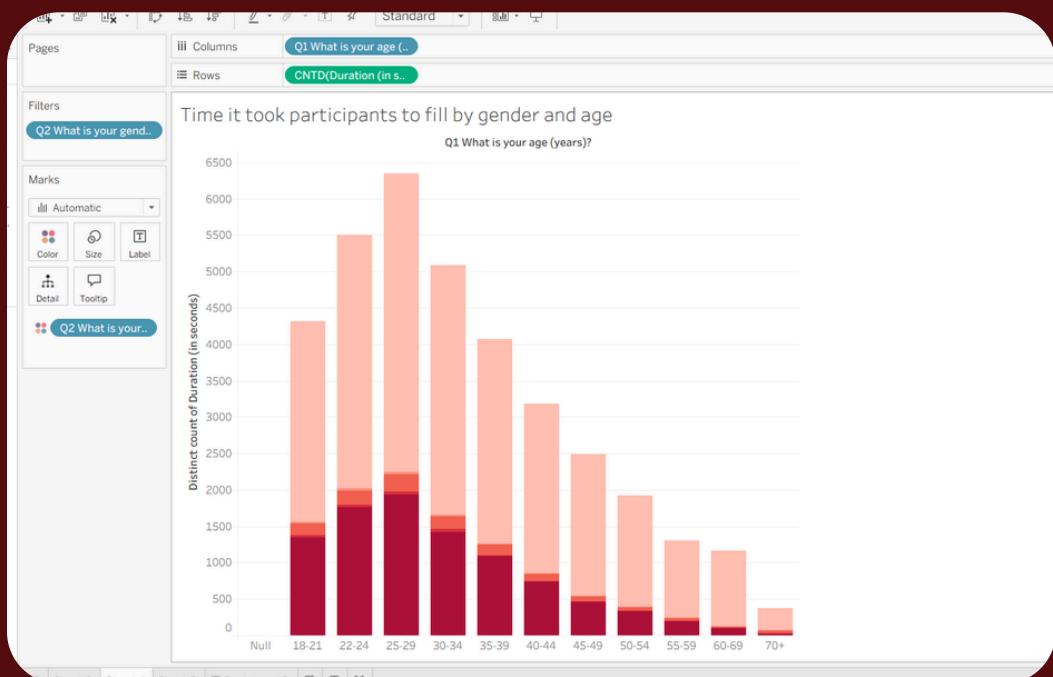


- Most used program = Python
- Least used program = Swift

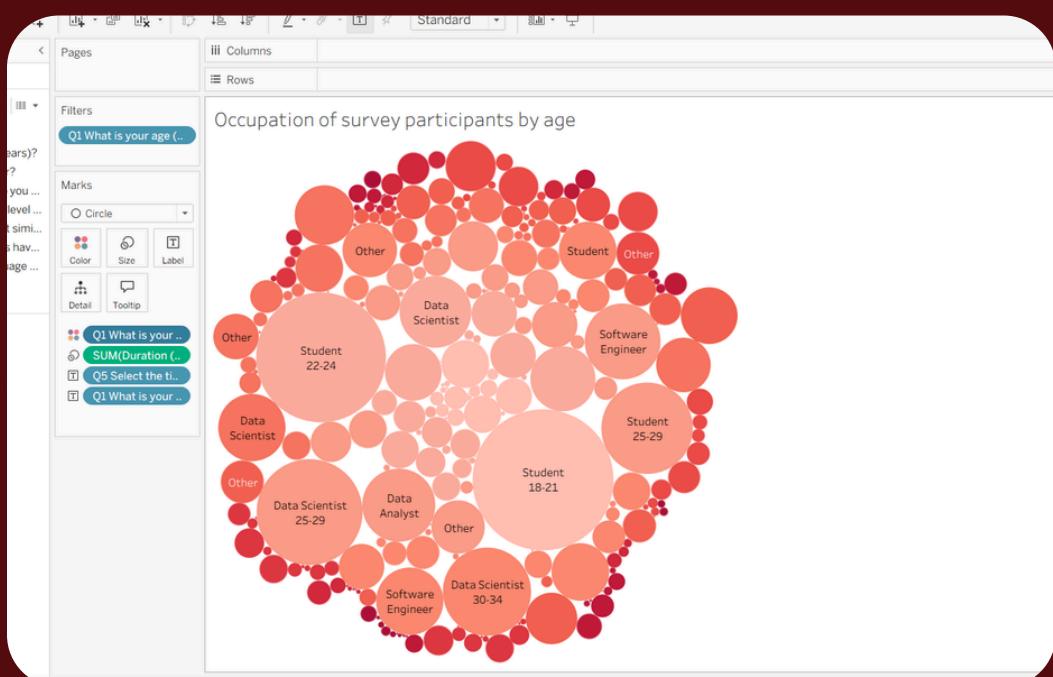


- Mainly master and bachelor students filled survey

STEP 3.2: VIEWING EACH INFOGRAPHIC

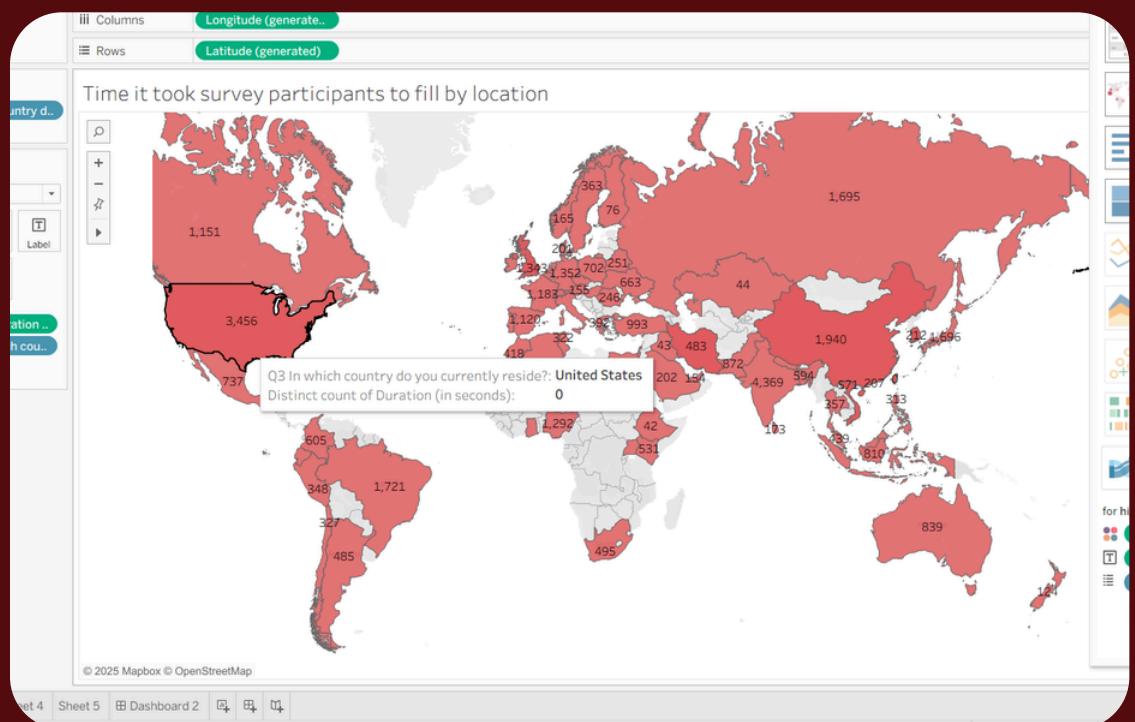


- The survey had mainly men fill it followed by women
- People from 25 to 29 filled this survey
- People above 70 had the least participation

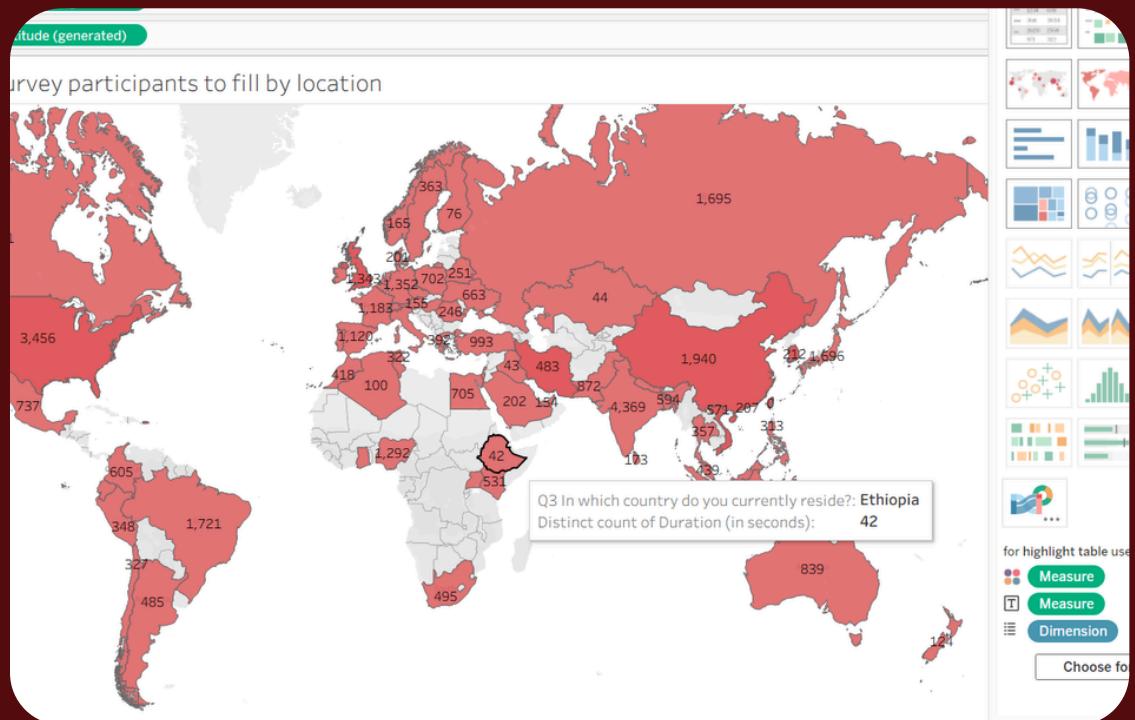


- Most users were students at the age of 18 to 21 or 22 to 24
- The 2nd most frequent group was data scientists of 25 to 29 years
- A couple of them were sales people and developer relations

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 = Ethiopia