DATA ENGINEERING

PROJECT REPORT

15th November 2024

Saumitra Agrawal (B22AI054) | Aatif Ahmad (B22AI002)

Project - FONT PREFERENCE ANALYSIS

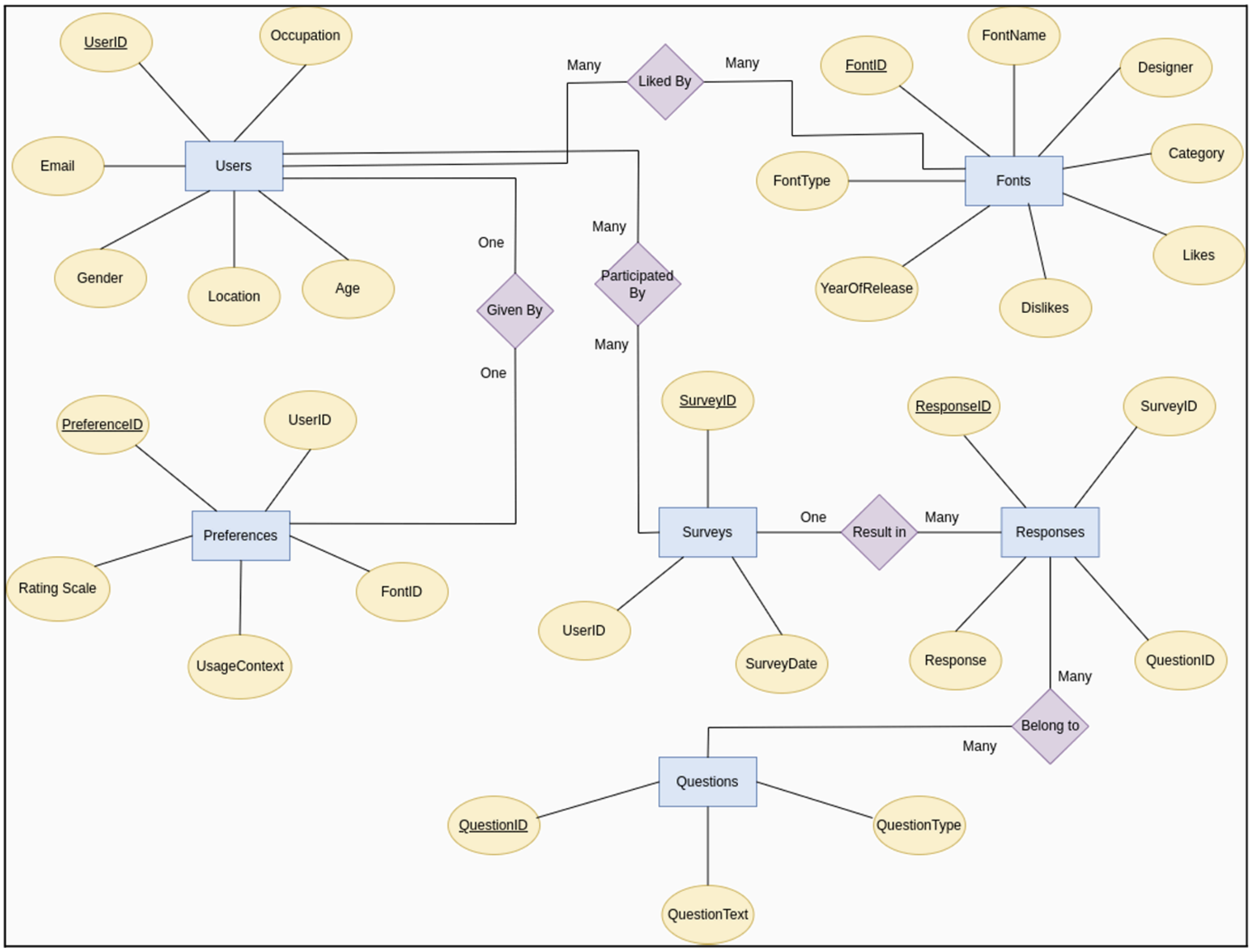
A. SOME IMPORTANT LINKS

- Google form for questions-Form
- Responses Sheet-Responses
- Github Repository- Github
- Colab Notebook Link-Colab

B. DATA SOURCES

- Conducting surveys (among peers)
- Datasets available online- Adobe Visual Font Recognition
- Google fonts analytics- Google Fonts Analytics
- Web Almanac by HTTP Archive- Web Almanac
- Synthetic Dataset(100k rows dataset with consistent features.

C. ER DIAGRAM OF THE DATABASE



- Users(UserID,Name,Email,Age,Gender,Location,Occupation)
- Fonts(FontID,FontName,FontType,Designer,YearOfRelease,Likes, Dislikes)
- Preferences(PreferenceID,UserID,FontID,UsageContext,Rating Scale)
- Surveys(SurveyID,UserID,SurveyDate)
- Responses(ResponseID,SurveyID,QuestionID,Response)
- Questions(QuestionID,QuestionText,QuestionType)

D. SURVEY QUESTIONS

- a. What is your age?
- b. What is your gender?
- c. What is your location?
- d. What is your occupation?

- e. How often do you use different fonts in your work? *
- f. In what context do you use mostly fonts? *
- g. What is your preferred font size? *
- h. How important is the choice of font important for your design/professional/educational work? *
- i. What factors influence your font preferences? *
- j. Which fonts do you like using in your work and which fonts do you dislike or avoid using in your work? (Give a list along with a rating scale. If you prefer using a font, give the usage context too.) *
  <List of fonts with the following rating scale>
  - Hate
  - Dislike
  - Neutral [Usage Context ?]
  - Like [Usage Context ?]
  - Love [Usage Context ?]
- k. Do you have any specific reason behind liking or preferring these fonts over others?
- l. Do you have any specific reason behind disliking or avoiding one or more of these fonts?

E. PIPELINE DESIGN

ETL (Extract-Transform-Load) pipeline

EXTRACTION USING GOOGLE FORMS --> TRANSFORMATIONS BY DATA WRANGLING --> LOADING INTO MYSQL

The pipeline gave us following advantages:
- Easy to interpret the table
- Suitable for structured data

But this pipeline gave the following drawbacks:

- Manual effort in transforming and loading data due to randomness
- Doesn’t support real-time processing.

F. DATA STORING TOOLS

The data storage for the project has been conducted in three ways:

a. Data collected through peers and online sources (Google Fonts Analytics)

The data collected through peers by floating a Google form is transformed and then processed through an ETL pipeline.

We ran an **SQL server** in a **docker container** and then connected it to **Azure Data Studio** where SQL tables are created for analysis.

This is how one of the tables appear:

	UserID	age	gender	location	occupation	email
1	1	19	Male	Jodhpur, Rajasthan	Student	b22ai002@iitj.ac.in
2	2	20	Male	Jodhpur, Rajasthan	Student	jagdishsuthar4581@gmail.com
3	3	21	Male	Jdohpur, Rajasthan	Student	b22ee004@iitj.ac.in
4	4	20	Male	Gorakhpur, Uttar Pradesh	Student	b22ai055@iitj.ac.in
5	5	20	Male	Jodhpur, Rajasthan	Student	b22ai061@iitj.ac.in
6	6	20	Male	Jodhpur, Rajasthan	Student	premkumarvks7@gmail.com
7	7	21	Male	Hyderabad, Telangana	Student	b22ai012@iitj.ac.in
8	8	19	Male	Nellore, Andhra Pradesh	Student	b22ai049@iitj.ac.in
9	9	19	Male	Jodhpur, Rajasthan	Student	b22ai052@iitj.ac.in
10	10	20	Male	Karimnagar, Telangana	Student	b22ai023@iitj.ac.in

The data from **Google Fonts Analytics** (specifically font_designers data) is stored in an index on an **elasticsearch** engine running locally.

The following is the result of a query in elasticsearch engine for getting all documents in the "**font_designers**" index:


```
{
  "took" : 202,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 25,
      "relation" : "eq"
    },
    "max_score" : 1.0,
    "hits" : [
      {
        "_index" : "font_designers",
        "_id" : "1",
        "_score" : 1.0,
        "_source" : {
          "FontID" : 1,
          "fontName" : "Roboto Mono",
          "designer" : "Google"
        }
      },
```


b. Dataset for Machine Learning tasks

Since the data collected through peers was inadequate for machine learning purposes, we created a mock dataset consisting of eight commonly used fonts Fabricate. This dataset comprises 100,000 rows.

c. Handling large dataset using spark

We used spark to find out the distribution of different fonts in our synthesized dataset.

The results are as follows:

```
+---------------+-----+
|font_name      |count|
+---------------+-----+
|Calibri        |12603|
|Helvetica      |12570|
|Arial          |12569|
|Roboto         |12565|
|Open Sans      |12474|
|Times New Roman|12471|
|Arima          |12437|
|Verdana        |12311|
+---------------+-----+

Execution Time: 16.771356183 seconds
```

G. DATA ANALYSIS AND USE OF ML MODELS

The synthetic generated data and the data received from peers was collected and collaborated to generate a dataset.
The chosen features for the dataset are {id,age, gender,usage_frequency,usage_context ,preferred_font_size}
The Target Column has been set to **Font_Name.**

id	age	gender	usage_frequency	usage_context	preferred_font_size	font_name
1	22	Male	Always	Personal	17	Times New Roman
2	48	Female	Rarely	Professional	13	Helvetica
3	35	Female	Rarely	Educational	20	Arima
4	51	Female	Often	Educational	16	Helvetica
5	44	Female	Rarely	Professional	8	Times New Roman
6	51	Male	Never	Personal	9	Arima
7	47	Female	Sometimes	Personal	32	Helvetica
8	42	Female	Sometimes	Educational	23	Times New Roman
9	31	Female	Never	Educational	15	Verdana
10	52	Female	Never	Professional	23	Calibri

Attached Snapshot of the Table

PREPROCESSING-

- The values of the column of usage frequency has been replaced by mappings
  'Never': 0,
  'Rarely': 1,
  'Sometimes': 2,
  'Often': 3,
  'Always': 4

- The Entries in the column of age group and Font size group have been grouped in the batches of 10 (1-10, 10-20,....) till the range of 100.

- Dropped rows with NaN values in `usage_frequency` or target column.

- Performed One Hot encoding of the categorical features like Gender and Usage Context.

TRAINING-

- Train Test Split Ratio = 80:20.
- Models Classified on are Decision Tree and Logistic Regression.
- Parameters for Decision Tree Classifier
  1. Maximum Height= 100.
  2. Prun after information gain = 0.7+.
- Parameters for Logistic Regression
  1. Max Iteration = 1000
  2. Learning Rate = 1e-3.
- Further Ensemble Learning has been applied to test the training model.

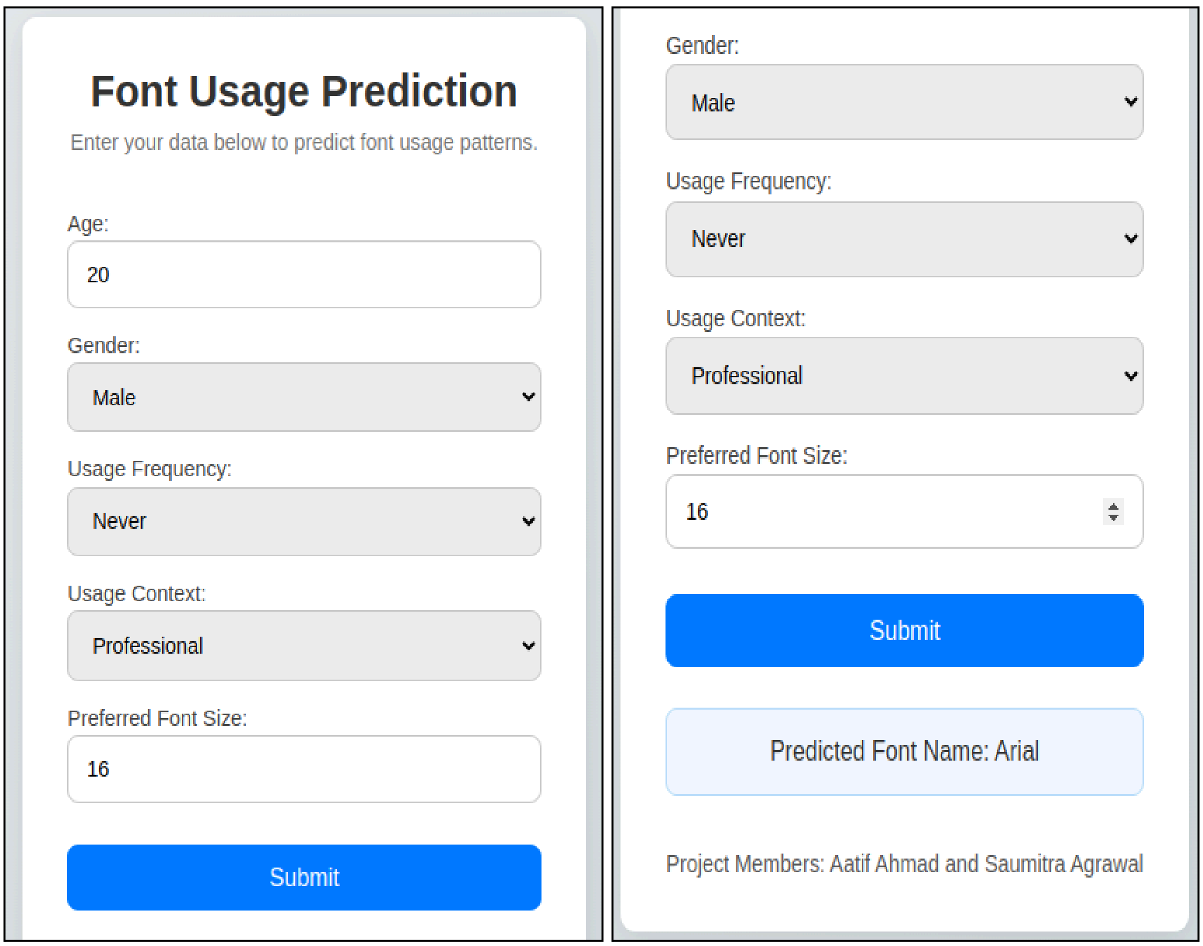
Accuracy Score=0.72
Precision=0.7
Recall=0.75
F1 score=0.7

H. APPLICATION FOR TESTING

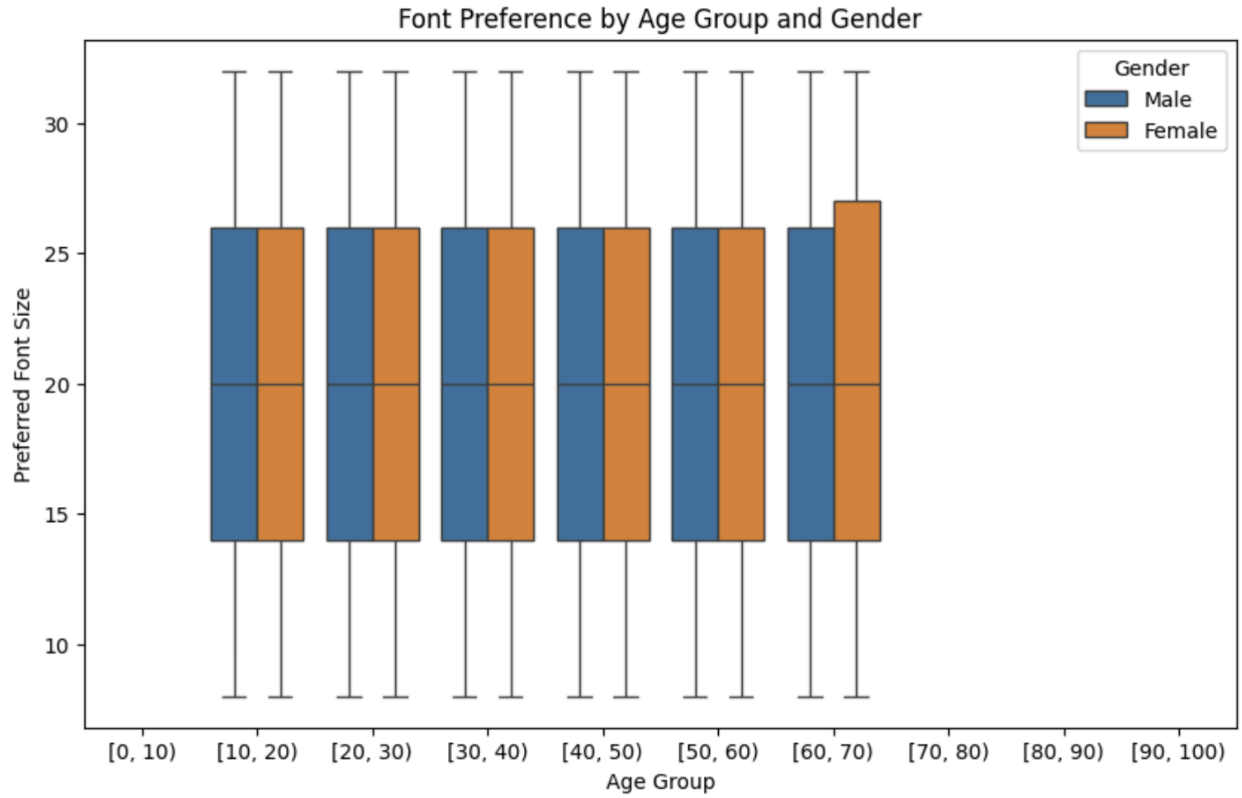
We created an application to test the machine learning model using the following tech stack:

a. **Frontend:** React
b. **Backend:** Flask

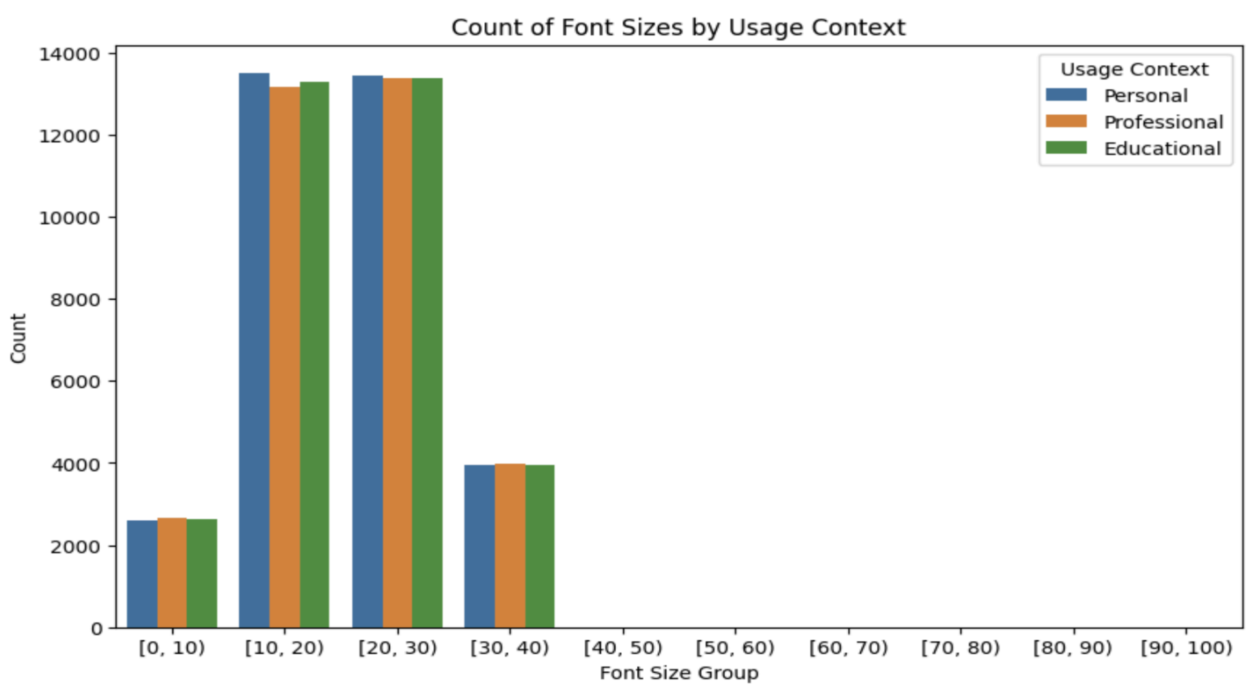
The application looks as follows:

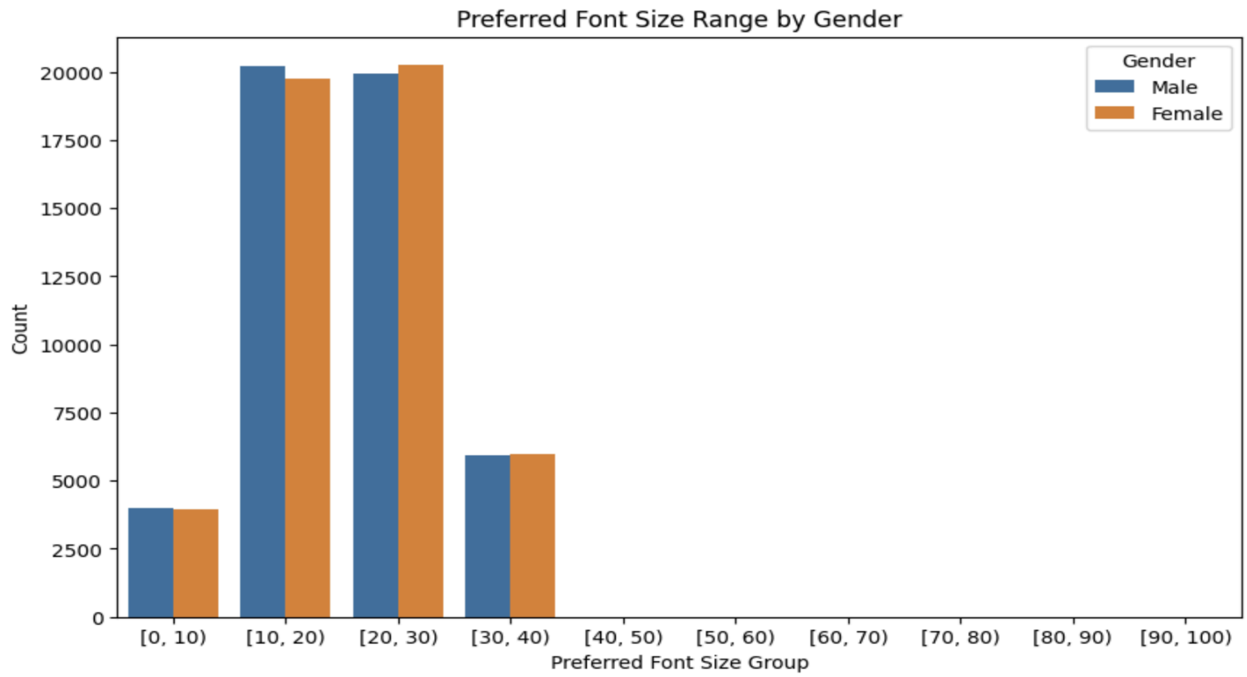


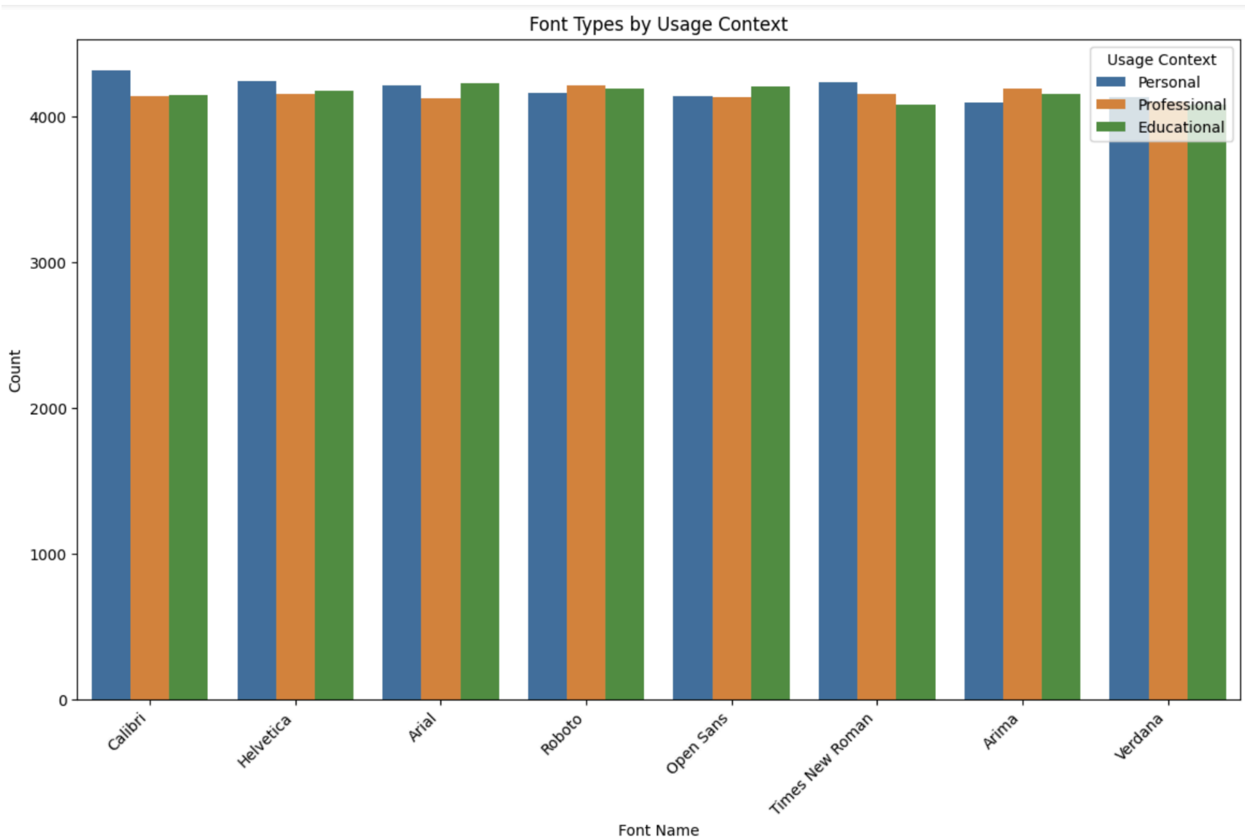
I. PLOTS FOR VISUALISING DATA

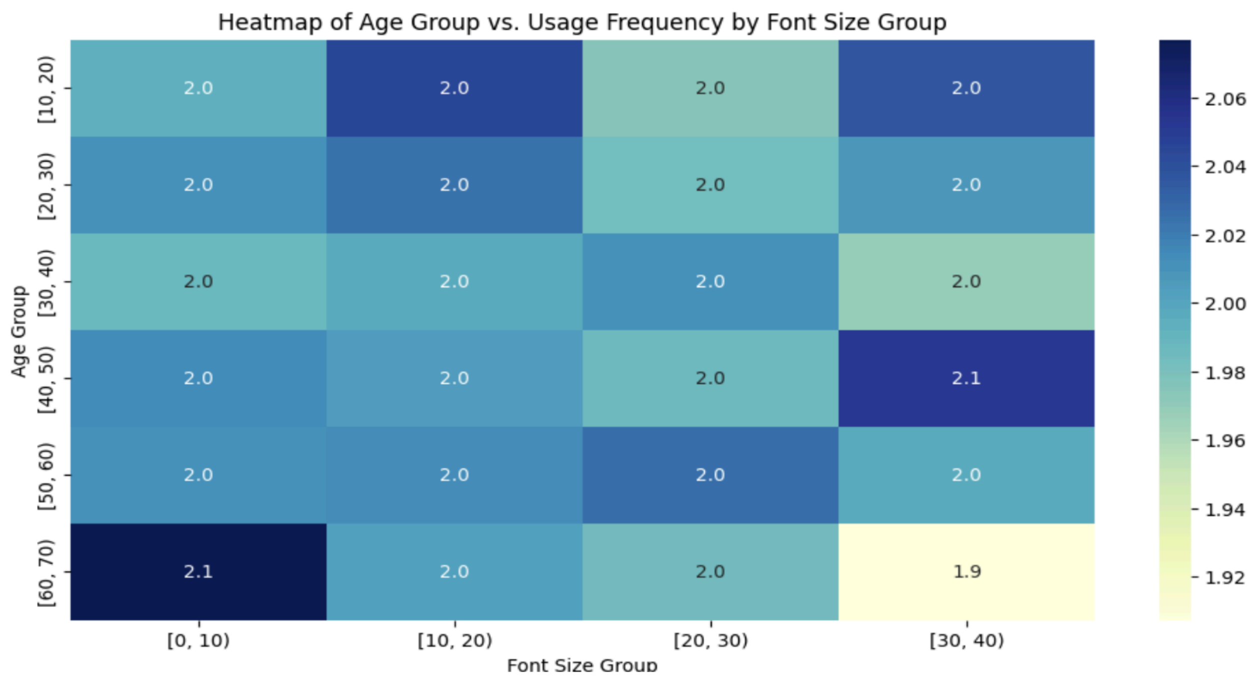


THIS DATA IS MORE OR LESS CONSISTENTLY DISTRIBUTED AMONG GENDERS




Preferred Font Size Range by Gender
Gender
Male
Female
Count
20000
17500
15000
12500
10000
7500
5000
2500
0
[0, 10)
[10, 20)
[20, 30)
[30, 40)
[40, 50)
[50, 60)
[60, 70)
[70, 80)
[80, 90)
[90, 100)
Preferred Font Size Group


Font Types by Usage Context
Usage Context
Personal
Professional
Educational
Count
4000
3000
2000
1000
0
Calibri
Helvetica
Arial
Roboto
Open Sans
Times New Roman
Arima
Verdana
Font Name



J. SOME TYPICAL PREDICTIONS

- Male Liked Fonts- Arial , Arima and Times New Roman.
- Female Liked Fonts- Open Sans and Helvetica.
- Most Professional Font- Roboto Mono
- Most Personal Font - Calibri