SDAIA Academy T5C04 Bootcamps: Data Science

Linear Regression and Web Scraping Module Project Presentation


# Predict the Number of Likes of a Linkedin Post

**Presented by: Nada Rambu**

01

# Problem and Solution

**Project background and objectives**

# Objective

## Predict the number of likes of a Linkedin post

02

# Data Acquisition

**Bringing the data from the its source to be processed**

# Features

- Likes - the number of "likes" the post received.  This represents the target variable.
- Followers - the number of followers in the network of the user who made the post.
- Age - the age of the user who made the post
- Gender - the gender of the user who made the post
- PostType - the type of post (Photo, Video, etc.)
- PostFrequency - the average number of posts made in a single day for the person making the post.
- Hashtags – the hashtags found with the post
- Education - the highest educational level of the individual making the post
- Date - the date of the post
- Time - the time of the post

# Features

- Likes - the number of "likes" the post received. This represents the target variable.
- Followers - the number of followers in the network of the user who made the post.
- ~~Age - the age of the user who made the post~~
- ~~Gender - the gender of the user who made the post~~
- PostType - the type of post (Photo, Video, etc.)
- PostFrequency - the average number of posts made in a single day for the person making the post.
- ~~Hashtags – the hashtags found with the post~~
- ~~Education – the highest educational level of the individual making the post~~
- Date - the date of the post
- Time - the time of the post.

# Features

## Total followers

users who has subscribed to the author account

## Posted since

days passed since posted

## Total posts

total number of posts for the author

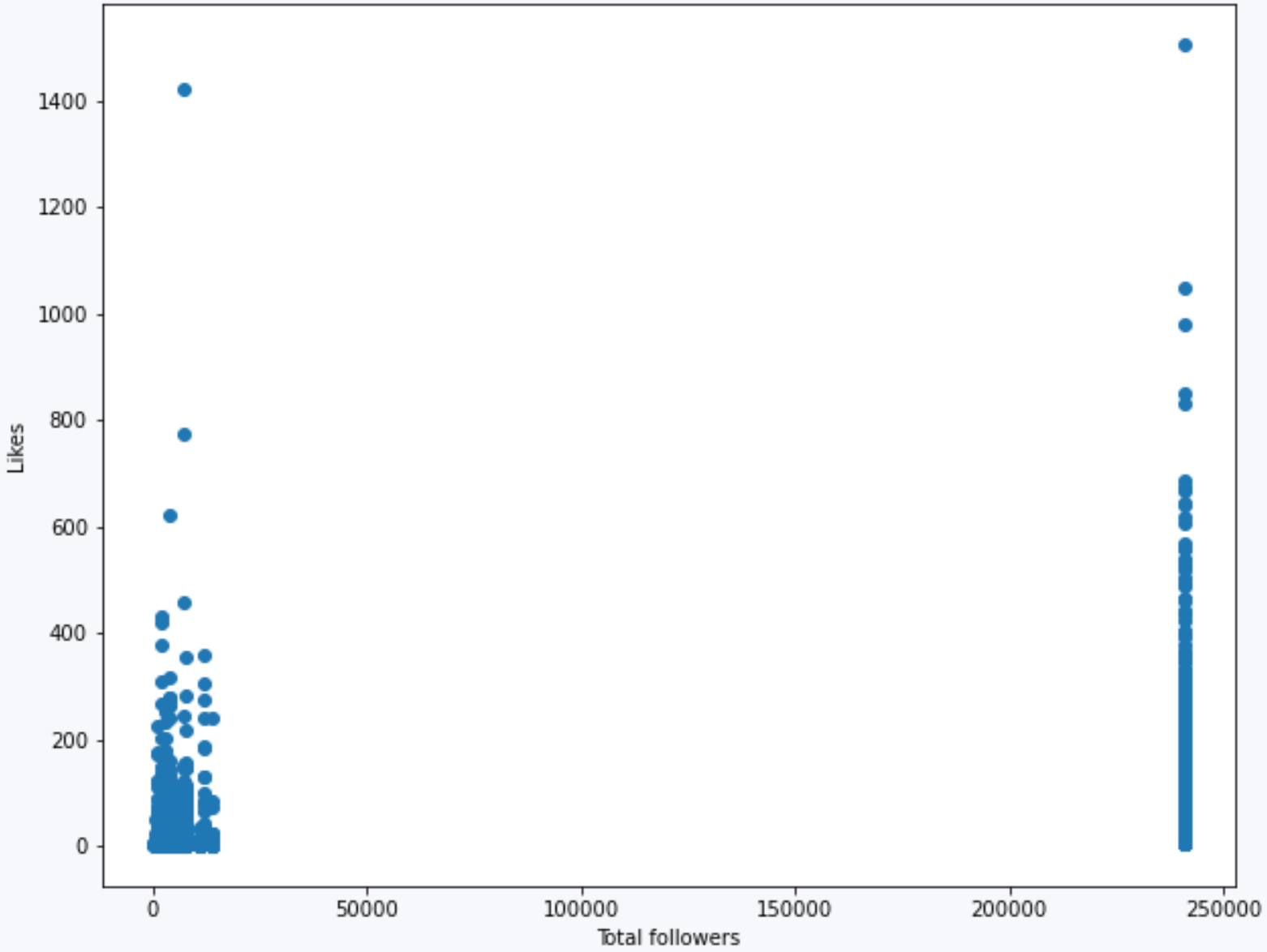
## Likes

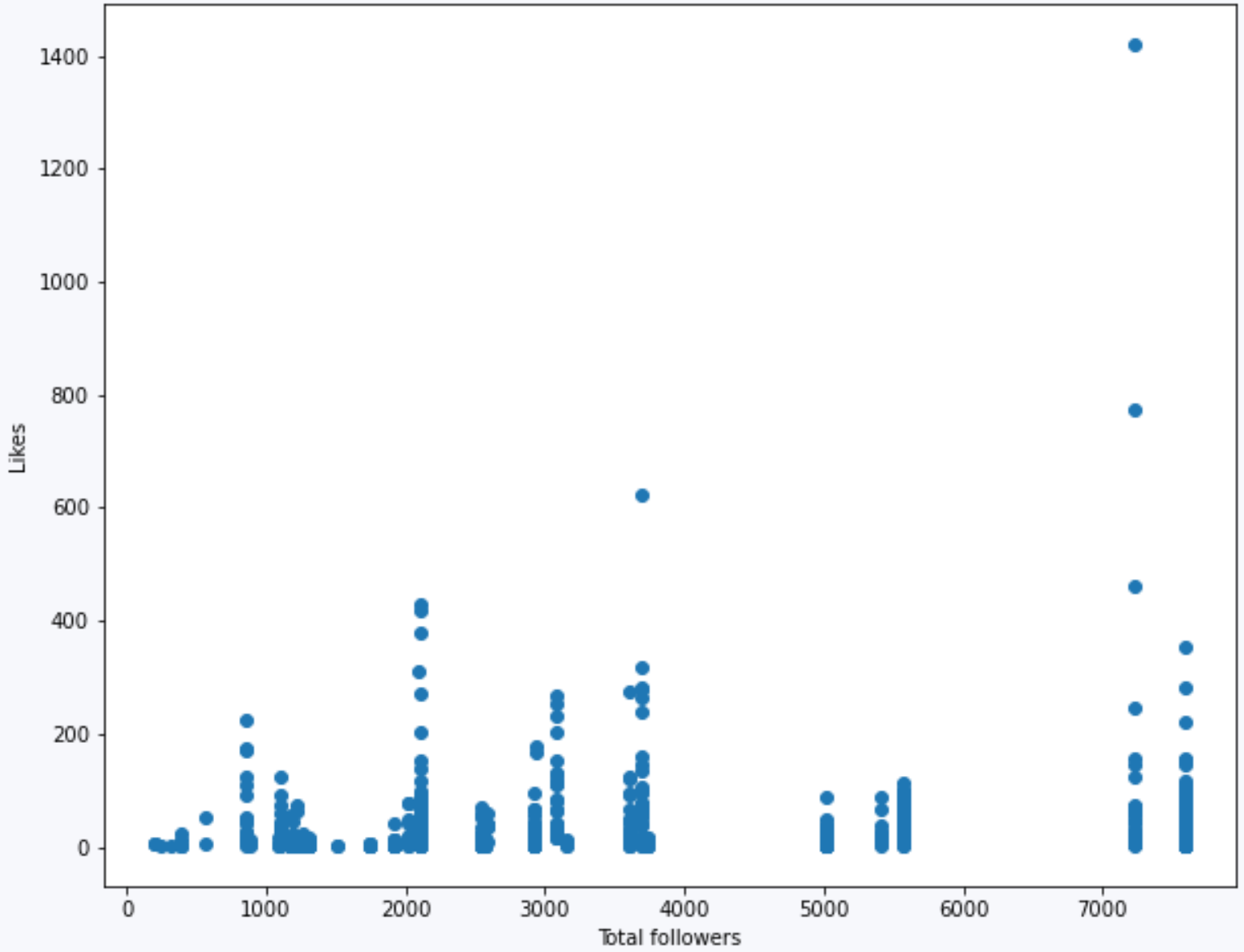
total likes received per post

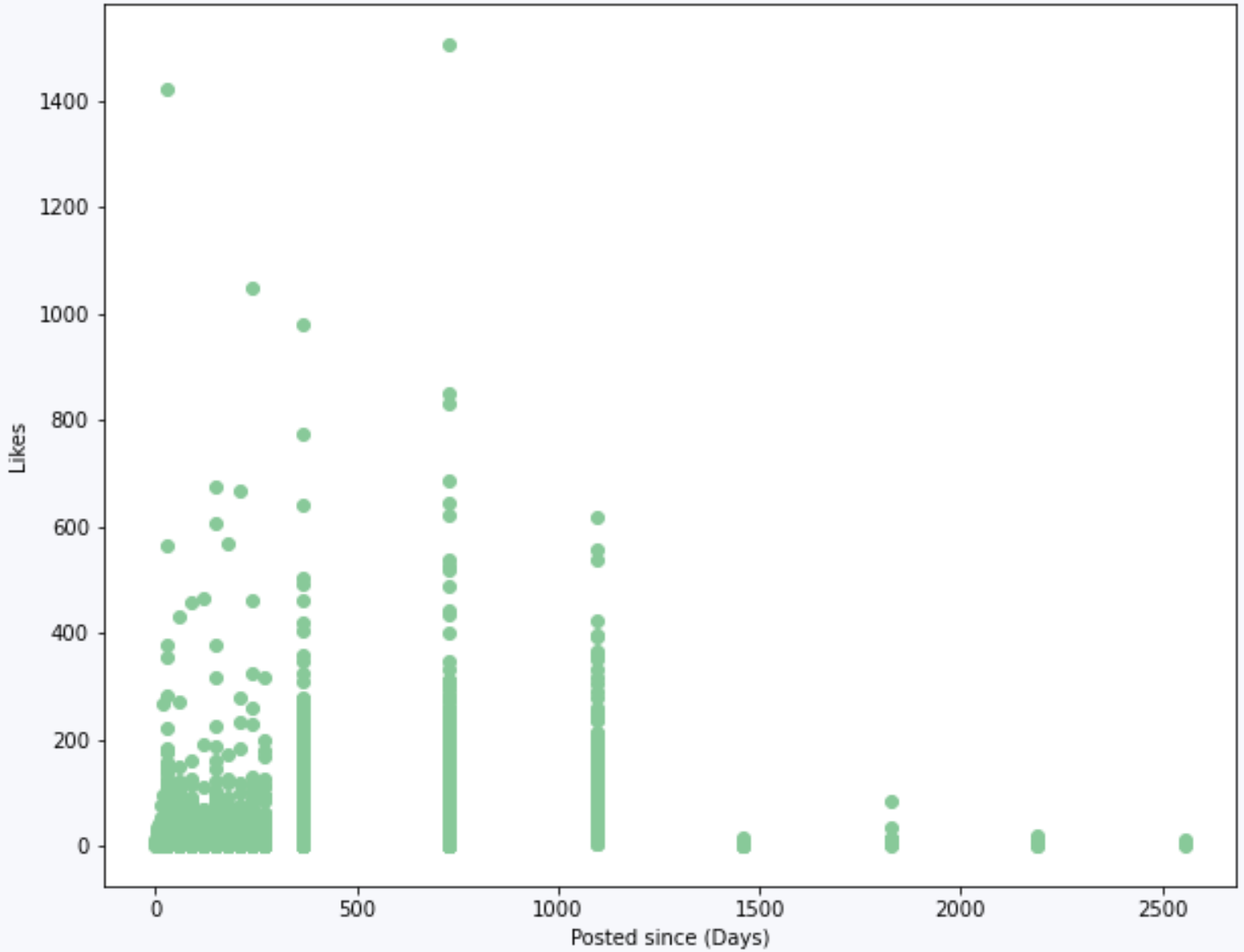
03

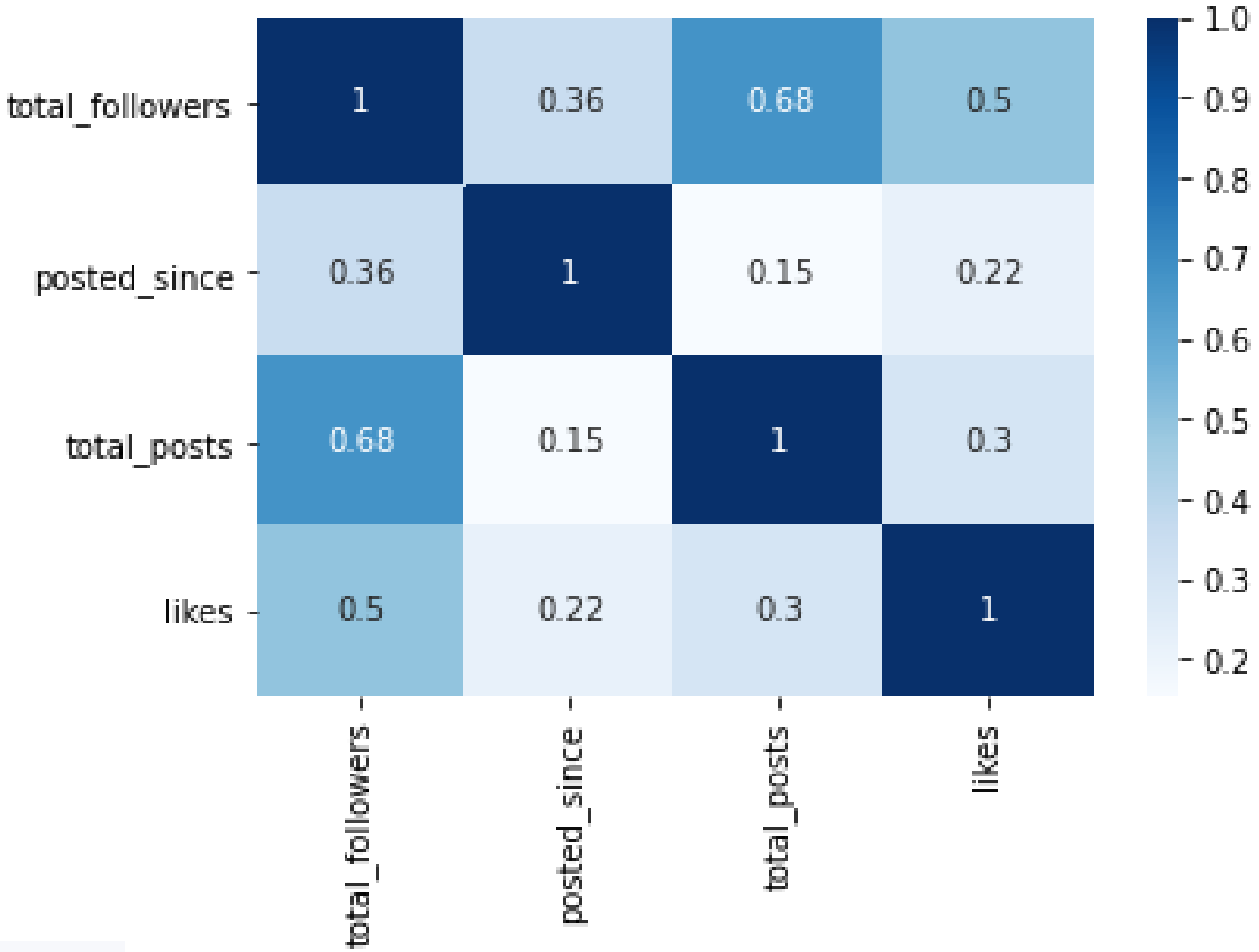
# Data Visualization

Gaining insights from the data


Likes
1400
1200
1000
800
600
400
200
0
0
50000
100000
150000
200000
250000
Total followers


1400
1200
1000
800
600
400
200
0
Likes
0
1000
2000
3000
4000
5000
6000
7000
Total followers


Likes
1400
1200
1000
800
600
400
200
0
0
500
1000
1500
2000
2500
Posted since (Days)


total_followers
posted_since
total_posts
likes
1
0.36
0.68
0.5
0.36
1
0.15
0.22
0.68
0.15
1
0.3
0.5
0.22
0.3
1
total_followers
posted_since
total_posts
likes
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

04

# Model Development

**Train and test the model to estimate results**

# Experiments

## Linear Regression

Accuracy **0.20**

MAE **50.8**

# Experiments

## Linear Regression

(with 3rd degree polynomial features)

Accuracy 0.22

MAE 49.4

# Experiments

Lasso
Ridge
ElasticNet

Accuracy 0.20

MAE 50.8

05

# Challenges

**Which made me cry…A LOT**

# Linkedin is

# INCONSISTENT

## AS HECK!!!!!!!

**University of Science and Technology, Yemen**

Bachelor's degree, Information Technology

2010 – 2014

**University of California, Berkeley**

B.A., Economics & Political Science

1999 – 2003

**Harvard Business School Executive Education**

2016 – 2016

**The University of British Columbia**

Bachelor of Education (B.Ed.)

1997 – 1998

**Linkedin why?**

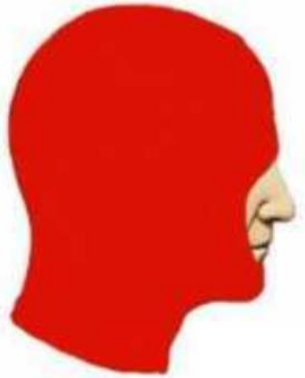
# Types of Headache

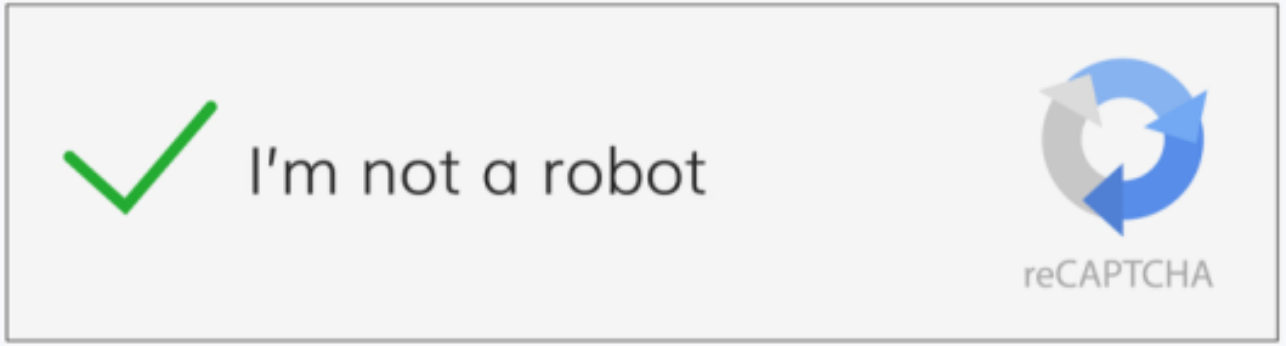
## Migraine

## Hypertension

## Stress

## Scraping LinkedIn




I'm not a robot
reCAPTCHA

# Thank You!