# CRISP DM METHODOLOGY

# REPORT ON PATA MCHUMBA DATING APP RECOMMENDER SYSTEM.

# BUSINESS UNDERSTANDING

With the current generation embracing technology and its applications, many people have become accustomed to the idea of using dating apps. Therefore, Pata Mchumba, a dating company, has approached us to create a recommendation system for their users to increase the effectiveness of matches based on their preferences. Moreover our recommender will focus mainly on emotional connection rather than physical appearance.

# BUSINESS OBJECTIVE

### Build a recommender system to successfully maximize user matches.

# ACCESSING THE CURRENT SITUATION

* Inventory of resources
  + Resources available for the project include:
    - Datasets:
      * OKCupid dataset. [[Link](https://www.tandfonline.com/doi/abs/10.1080/10691898.2015.11889737)]
  + Software:
    - An online IDE is known as google colab notebook.
    - Google drive - store the dataset.
    - JIRA - project management tool.
    - Canva - create presentation slides.
    - Google docs - prepare the data report.

# BUSINESS SUCCESS CRITERIA

* To get and outline major factors that make a great match.
* Answer the following questions:
  + Which has more importance? Physical attraction or Personality.
  + Which are the most common traits among users?
* To create a simple user interface for demonstration.

# PROCEDURE PLAN

| **Plan** | **Duration** |
| --- | --- |
| Research & Choosing a Project | 1 day |
| Choosing a dataset and perform EDA | 1 day |
| More EDA & Data Cleaning |  |
|  |  |
| Modeling & Evaluation |  |
| Conclusion |  |
|  |  |

# DATA UNDERSTANDING

## Describing the data

Data was acquired from the Okcupid website. This dataset is presented and made publicly available for use by dating sites. (Kirkegaard & Bjerrekær, 2016). The Dataset has 59946 rows and 31 features.

## Data Previewing

Previewing the first 5 rows in the dataset:

## Data exploration

First exploration of the general dataset was conducted. The table below displays a few things explored on the general dataset:

| Feature | No of unique values | No of null values | Null values  Percentage | Data type |
| --- | --- | --- | --- | --- |
| Age | 54 | 0 | 0% | int64 |
| Status | 5 | 0 | 0% | object |
| Sex | 2 | 0 | 0% | object |
| Orientation | 3 | 0 | 0% | object |
| Body\_type | 13 | 5296 | 8.8% | object |
| Diet | 19 | 24395 | 40.7% | object |
| Drinks | 7 | 14080\* | 4.98% | object |
| Drugs | 4 | 2985 | 23.487% | object |
| Education | 33 | 6628 | 11.06% | object |
| Ethnicity | 218 | 5680 | 9.48% | object |
| Height | 61 | 3 | 0.01% | float64 |
| Income | 13 | 0 | 0% | int64 |
| Job | 22 | 8198 | 13.68% | object |
| last\_online | 30123 | 0 | 0 | object |
| Location | 199 | 0 | 0 | object |
| Offspring | 16 | 35561 | 59.32% | object |
| Pets | 16 | 19921 | 33.23% | object |
| Religion | 46 | 20226 | 33.74% | object |
| Sign | 49 | 11056 | 18.44% | object |
| Smokes | 6 | 5512 | 9.19% | object |
| Speaks | 7648 | 50 | 0.083% | object |
| Essay0 | 54348 | 5488 | 9.15% | object |
| Essay1 | 51517 | 7572 | 12.63% | object |
| Essay2 | 48626 | 9638 | 16.08% | object |
| Essay3 | 43521 | 11476 | 19.14% | object |
| Essay4 | 49258 | 10537 | 17.58% | object |
| Essay5 | 48962 | 10850 | 18.10% | object |
| Essay6 | 43584 | 13771 | 22.97% | object |
| Essay7 | 45549 | 12451 | 20.77% | object |
| Essay8 | 39324 | 19225 | 32.07% | object |
| Essay9 | 45441 | 12603 | 21.02% | object |

## Verifying Data Integrity

1. The original dataset:
   1. The data is not complete. Contains some missing values
   2. The assumption that the data is correct and up to date

Once the integrity of the dataset was determined, cleaning of the columns followed and more EDA that involved making some visualizations too.

## Data Cleaning & EDA of Individual Columns

### 1.1 The Body Type Column

For this column, the missing values were filled with ‘unspecified’ as a placeholder. The user might have decided not to fill this column that has information on their body type.

### 1.2 The Diet Column

For this column we noticed that some of the options are similar. Therefore we will replace similar classes with the same class.

For example: dating\_df['diet'] = dating\_df['diet'].replace(['mostly anything','strictly anything'],'anything'). Here, we combine different forms of anything into one.

### 1.3 The Drinks Column

The missing values in the drinks column was replaced with ‘unspecified’ since some users may choose not to disclose that information.

### 1.4 The Drugs Column

As we did to the drinks column, null values were replaced with ‘unspecified’ as some users chose not to have that information public.

### 1.5 The Education Column

Firstly, we replaced the null values with ‘unspecified” since some users prefer not to tell their level of education.

### 1.6 The Ethnicity Column

Firstly, we replaced the null values with ‘unspecified” since some users prefer not to tell their race Ethnicity is also a physical feature therefore we might consider dropping..

### 1.7 The Height Column

Dropped the records with null values since they make 0.005% of the data.

### 1.8 The Job Column

Some users prefer not to disclose the type of work they do for a living therefore we filled the null values with unspecified.

### 1.9 The Offspring Column

The missing values are filled with unspecified since it is the user’s choice to disclose this information

### 2.0 The Religion Column

Firstly, we replaced the null values with ‘unspecified’. Then classified the religions into those who are serious and those who aren’t serious with religion.

The serious ones were under class 1. Those not serious were under class 0 and the unspecified represented by class 2.

Lastly, different columns were made to separate the religion and the class.

### 2.1 The Zodiac Column

Here the missing values were replaced with ‘uninterested’. It was figured this is not a popular topic to many thus the choice of the placeholder name.

To deal with the missing values in this section, we will fill the missing values with uninterested as zodiac signs do not matter to some people. We will further clean the column and only pick out the zodiac sign and remove the rest of the text in each row.

### 2.2 The Smokes Column

Some users prefer not to mention if they smoke or not therefore we fill the null values with ‘unspecified’

### 2.3 The Speaks Column

The missing values are filled with unspecified since some users prefer not to mention all languages they speak.

**2.4 Dropping columns**

Since our match making process will depend on social attraction(personality, shared interest and beliefs), we will drop columns that will affect our study in order to make it more superficial. The columns dropped are:Ethnicity, job, income, offspring, speaks, last online, height.

Speaks column was dropped since the data is from the United States, most people speak English.

**Filtering data**

The location column consisted of Cities and States as one string which we filtered to only specific states in the US and dropped the original column.

**Dealing with Outliers**

The age column was discovered to have outliers. Some ages went as high as 101 which was far from most of the values. We grouped the data into bins of 10e.g (20-30) and andy user above age 52 was categorized above 50.

**EXPLORATORY DATA ANALYSIS**

Deeper analysis was done on Gender, Age, Orientation, Drugs, Smokes, Alcohol, diet preferences and pet preferences. From the analysis we found:

* Most users were male averaging 60.1% and females 39.9%.
* The highest number of users are at the age of 26 and 27.
* On orientation, most users were straight where 52.2% of male were straight and 34.4%. of females were straight.About 3% of female were bisexual while 1.2% of male were bisexaul.
* 63.3% of the users have never used drugs before.
* About 73.2% of users do not smoke.
* On Alcohol consumption, 70.0% of users only drink at social events.
* Above 2500 users prefer to eat anything rather than be on a diet.
* Above 2000 users like both cats and dogs.

**TEXT PREPROCESSING IN NATURAL LANGUAGE PROCESSING(NLP)**

In NLP, text preprocessing is the first step in the process of building a model.

The various text preprocessing steps are:

* Tokenization
* Lower casing
* Stop words removal
* Stemming
* Lemmatization

These operations were done on the essay columns. This exercise was performed twice as for the first time it involved lemmatizing strings instead of individual words.

Moreover, apart from the above steps the following operations were performed:

* Dealing with null values and duplicates in the essay columns. Used the fillna(method).
* Removal of links and hyperlinks in the essays.
* EDA on the preprocessed essays data. Made word clouds out of popular words in each essay column.

**MODELING & EVALUATION**

One of the objectives of the project is to come up with a hybrid model which was done in this chapter. We created functions that enabled us to match people on our dataframe. The following was done towards the end of the project:

* Created a deep learning model to predict words based on essays.
* Tidying the dataset by renaming the columns to make it easier to read and analyze.
* Created a Word2Vec model.

Later on we created the functions to match users:

* Match through orientation.
* Match through gender.
* Match through zodiac signs.
* Match by age gap.

In the final model, we combined all the functions created above into one to come up with the final model.

# REFERENCES

Kirkegaard, E. O. W., & Bjerrekær, J. D. (2016, November 3). The OKCupid dataset: A very large public dataset of dating site users. *Open Differential Psychology*. https://doi.org/10.26775/odp.2016.11.03