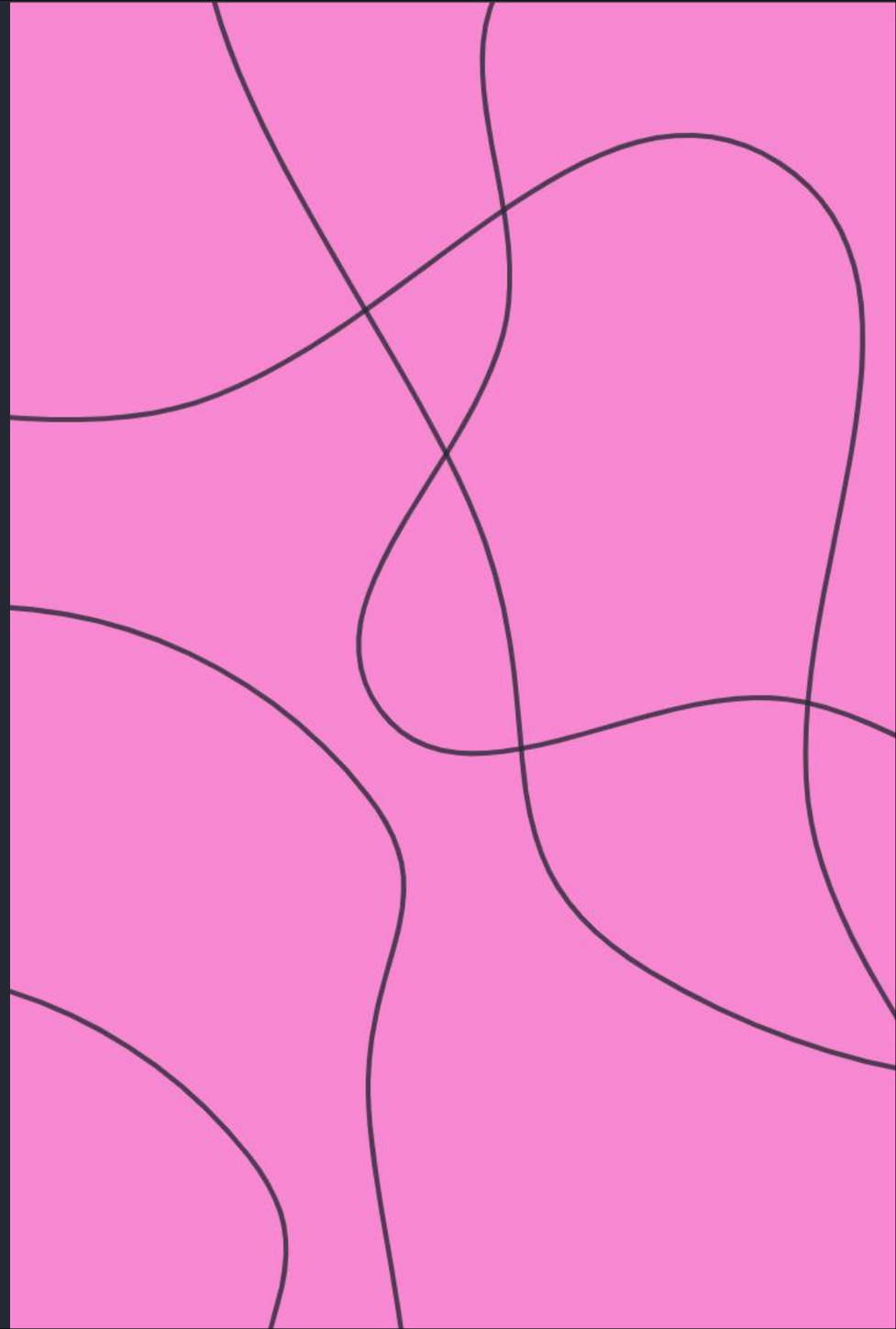
Data Analyst: Data Collection Pipeline

Welcome to the presentation on the Data Collection Pipeline, a key aspect of a Data Analyst's role. This pipeline is responsible for acquiring data and transforming it into meaningful insights. Let's explore the journey from data acquisition to storytelling!

by Debika Mukherjee

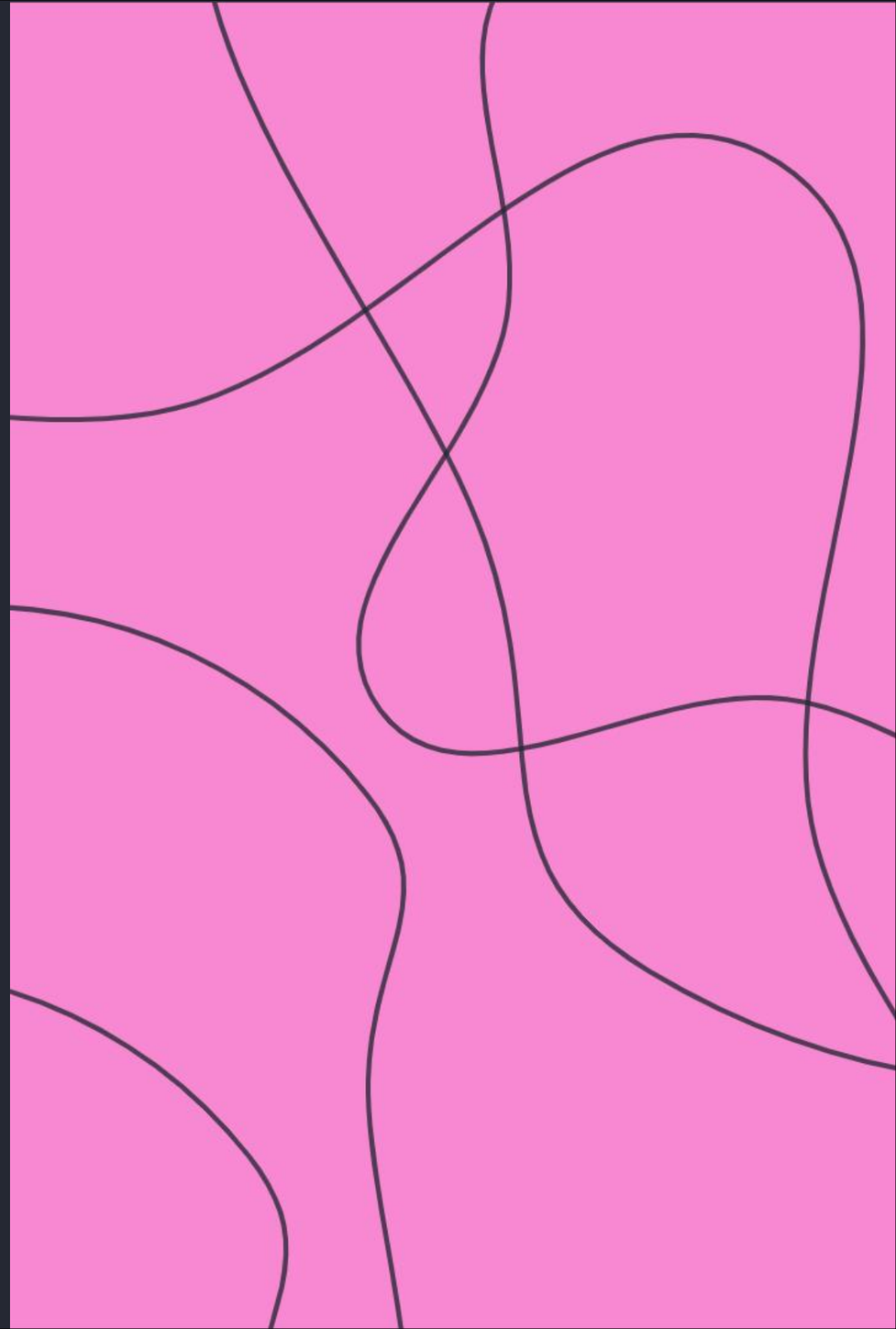
Debika Mukherjee

I am pleased to present my solo project on the Data Collection Pipeline. My name is Debika Mukherjee, and I am a Data Analyst based out of the United Kingdom. I hold a specialization in Data Analysis from Cardiff Metropolitan University.



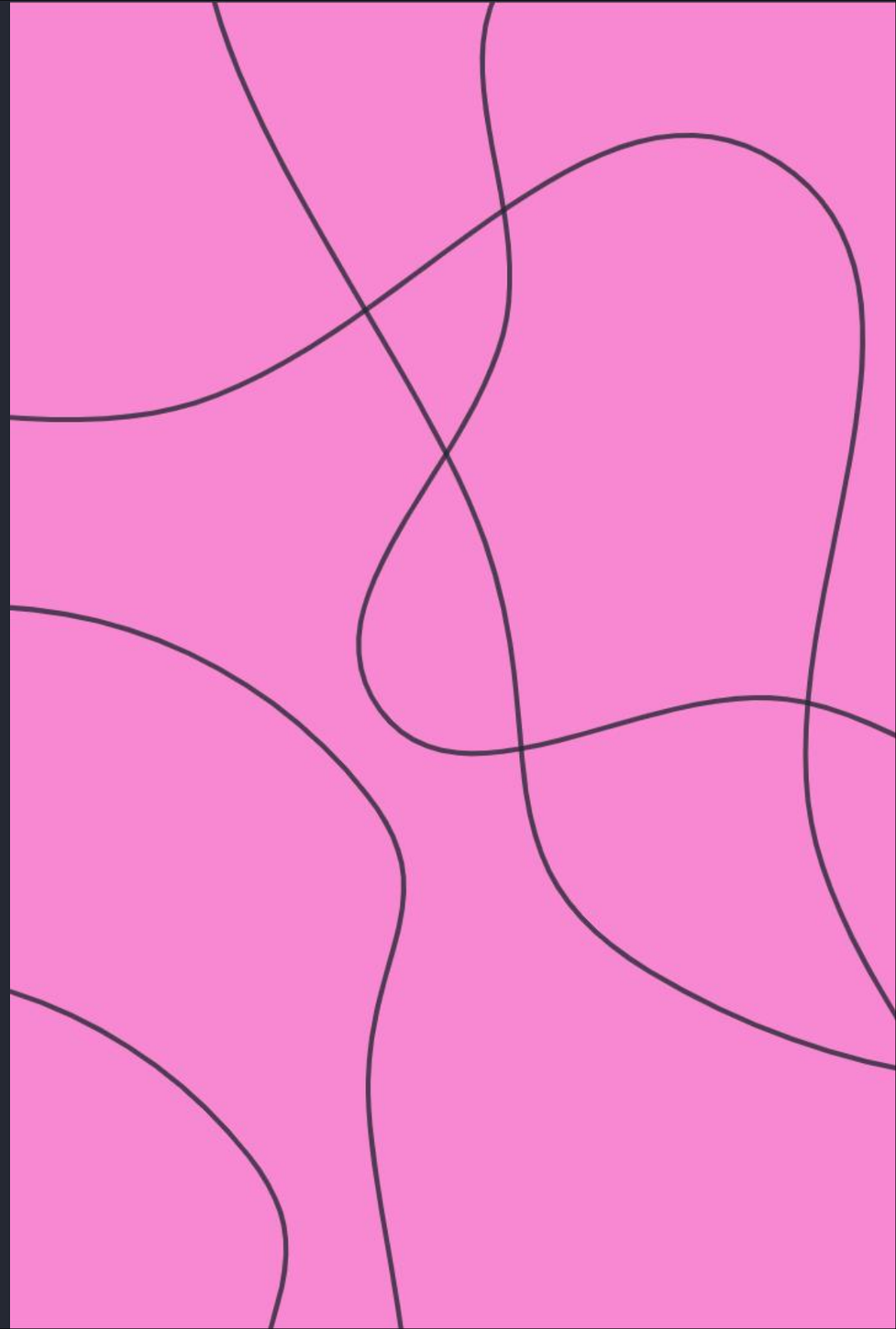
Problem Description: Data Collection Pipeline

Imagine XYZ company, which relies on collecting customer data through various means like Google Forms and Survey Monkey. Our task is to create a robust pipeline that collects and visualizes this data, ensuring its cleanliness and accuracy. We will address issues such as duplicate or irrelevant data and perform deduplication checks based on customer email IDs.



Github Repository

If you are interested in exploring the code and details of the Data Collection Pipeline, you can find it on my GitHub repository: debikaisms/Final-project: Data Collection Pipeline (Data Acquistion to Storytelling) Data Collection
Pipeline (Data Acquisition to Story telling) (github.com)



Data Overview & Key Variables

Let's take a closer look at the dataset and its key variables.

Number of Records: The dataset consists of 10 unique customer records.

Customer ID: An identifier for each customer, represented numerically.

Email Address: The email address of the customer.

Name: The name of the customer.

Age: The age of the customer.

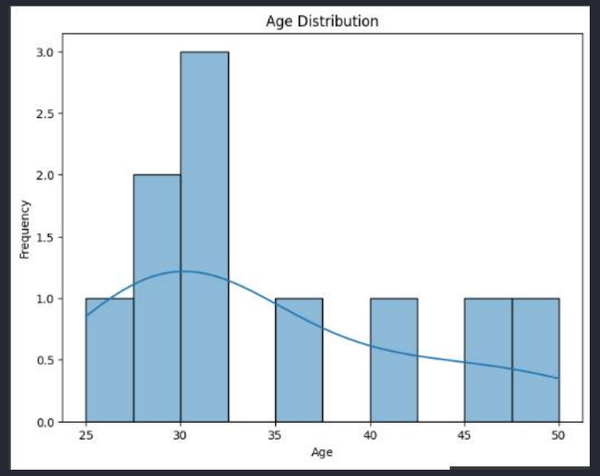
Gender: The gender of the customer, either Male or Female.

Satisfaction Rating (1-5): A numerical rating indicating the level of customer satisfaction, ranging from 1 to 5.

Product Feedback: Textual feedback provided by the customer.

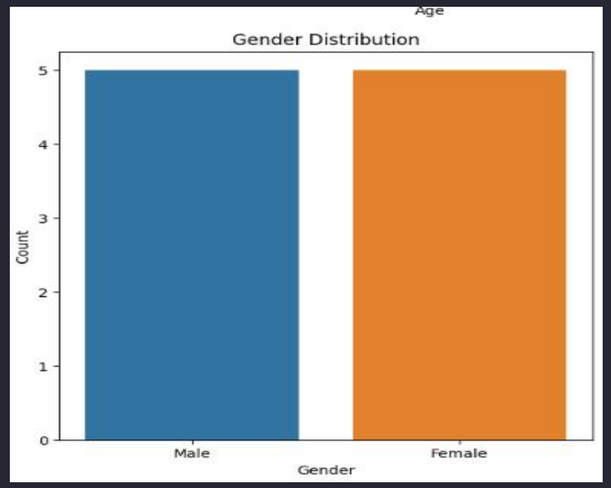
Duplicate Check: A categorical column indicating whether a duplicate customer record is present, with possible values of "No" or "Yes (Duplicate)".

Demographics



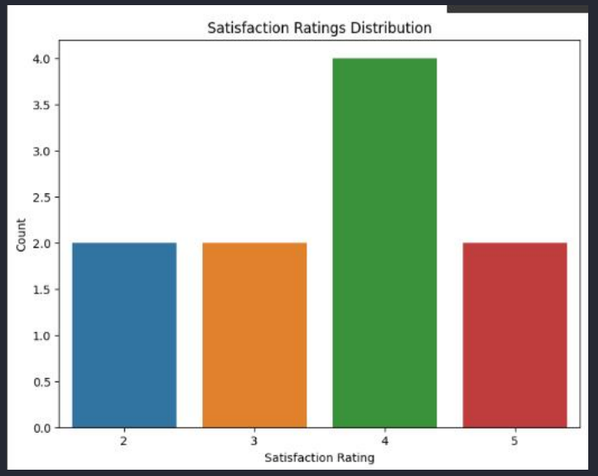
Age Distribution

Based on the data, the majority of XYZ Company's customers fall within the age range of 25 to 50.



Gender Distribution

The gender distribution shows a fairly even split between male and female customers.



Customer satisfaction

The majority of customers seem to have a positive experience, but there is room for improvement.

Recommendations for Technical Users

1) Regression Analysis
2) Classification model
3) Clustering
4) Natural Language Processing
5) Word cloud Analysis
6) Hyperparameter tuning