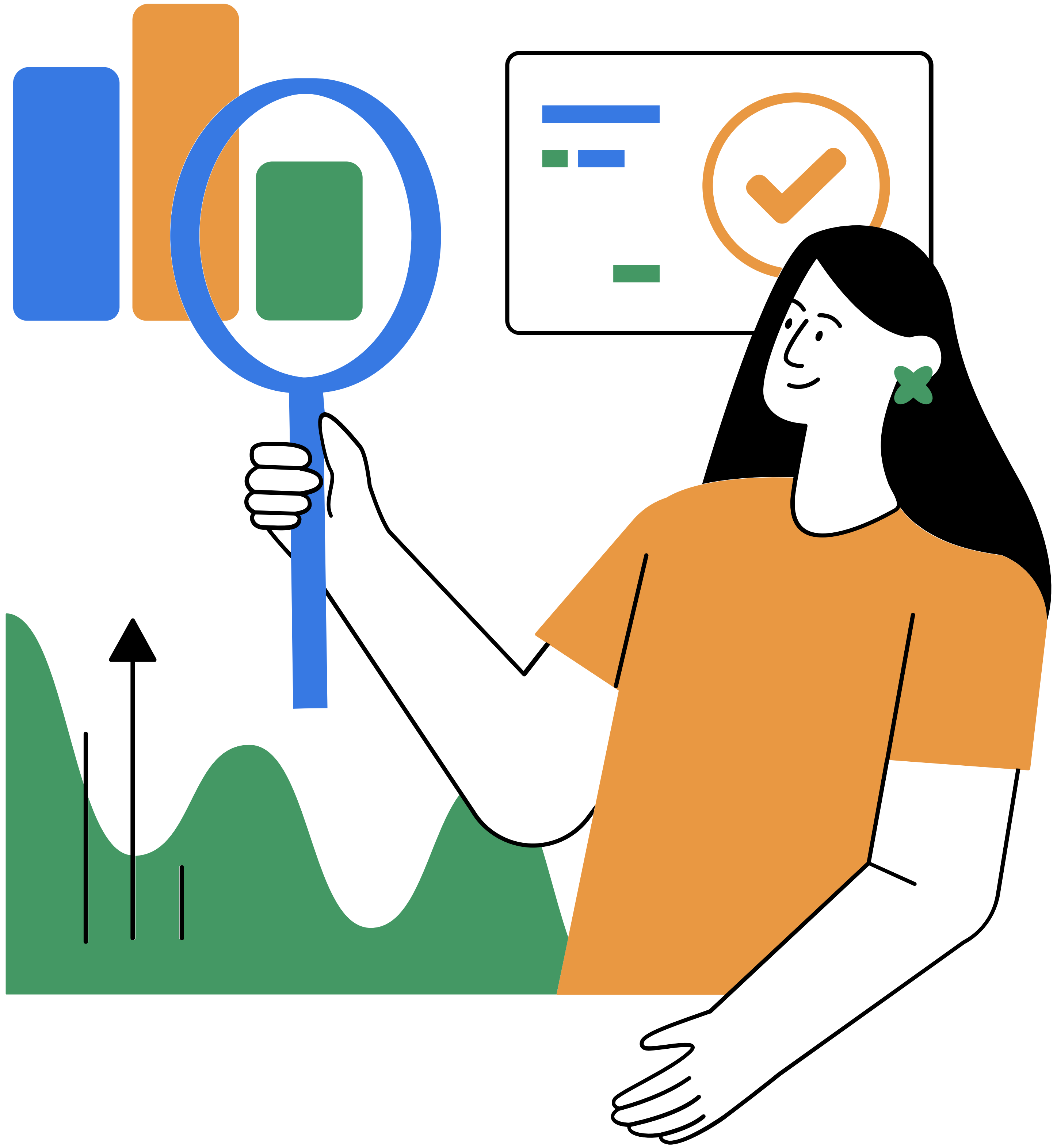
Data Analytics Pipeline

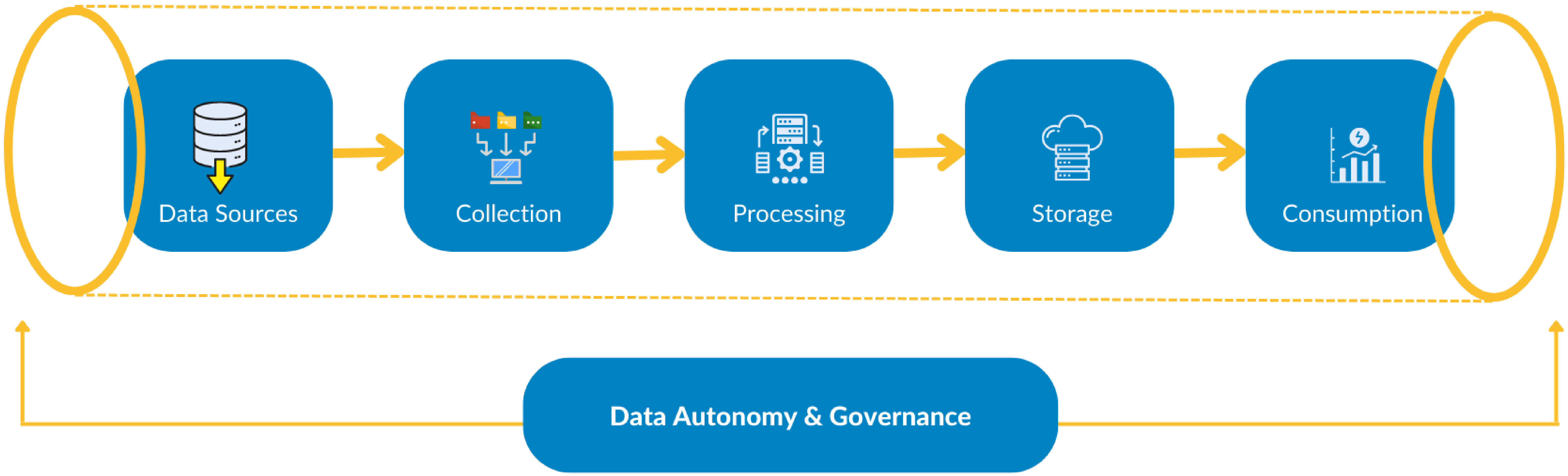


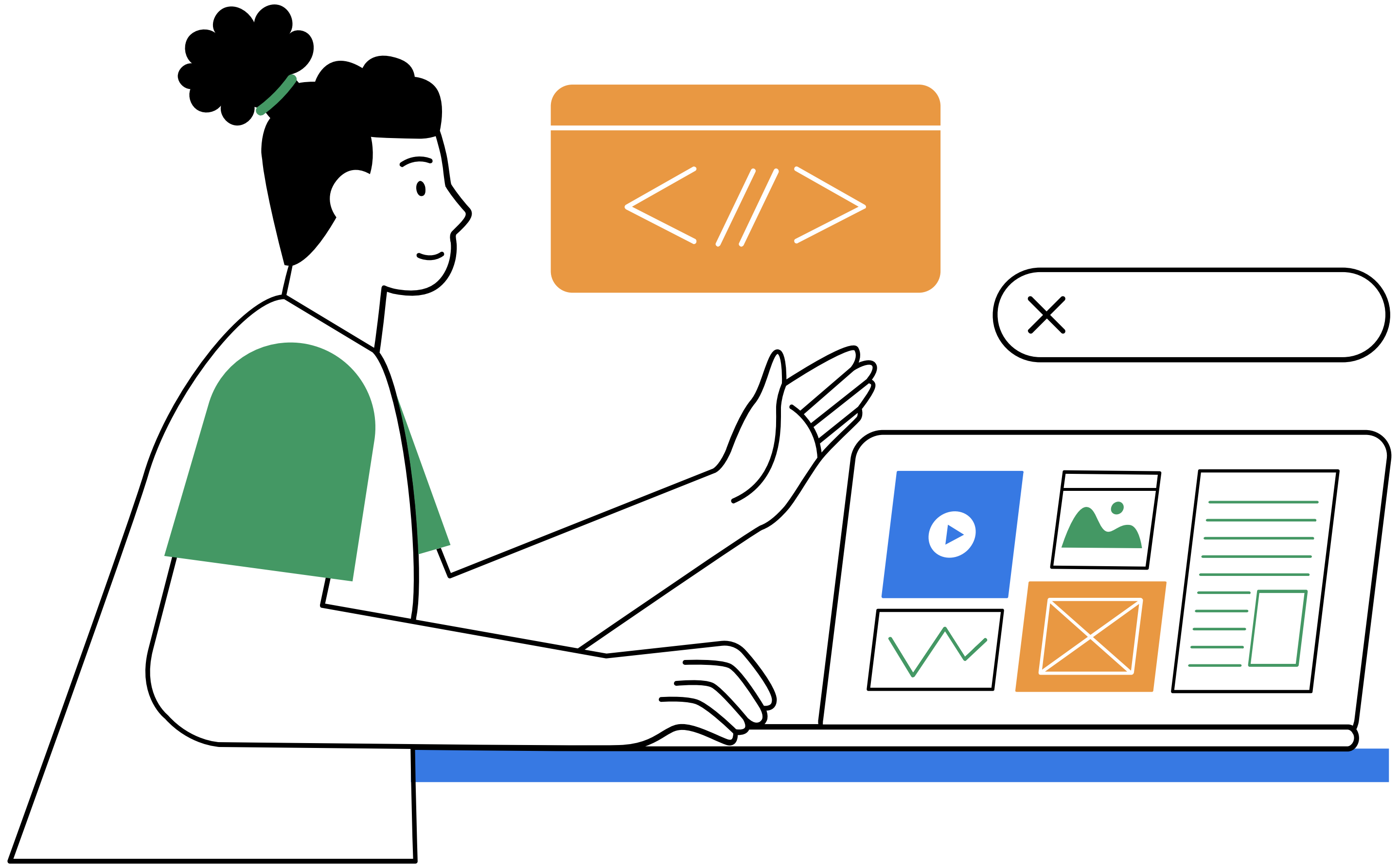
By- Khwnasat Giri Narzary

What is a Data Analytics Pipeline?

A data analytics pipeline is a structured workflow that processes raw data into meaningful insights. It automates **data collection**, **transformation**, **analysis**, and **visualization** to ensure accuracy and efficiency.

WHAT IS DATA PIPELINE?





Tools & Technologies

Data Collection(Extract)

- APIs & Connectors – Apache NiFi, Airbyte, Fivetran
- Streaming Tools – Apache Kafka, AWS Kinesis
- Database Connectors – MySQL, PostgreSQL, MongoDB

Data Cleaning(Transform)

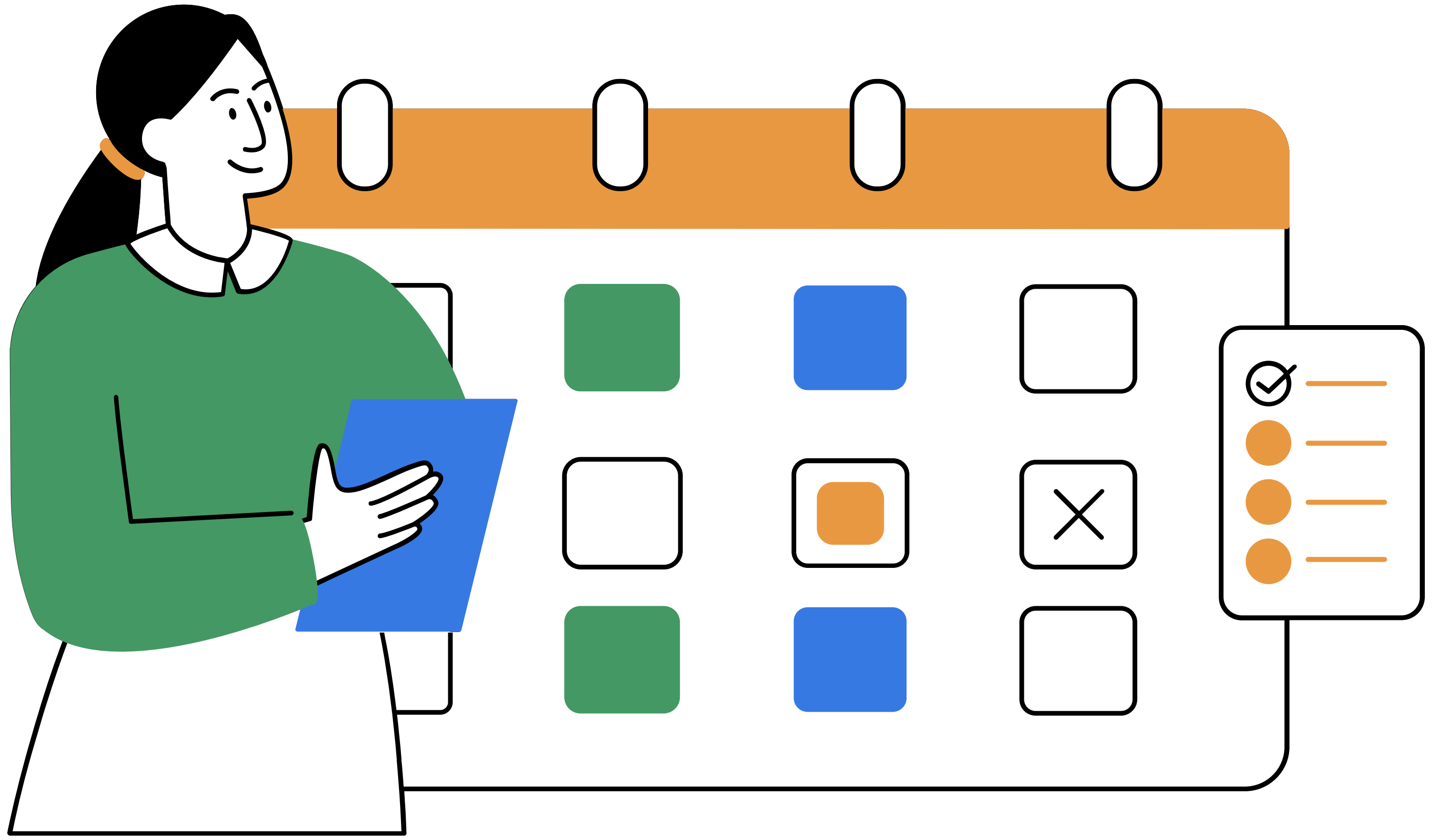
- Programming Languages – Python (Pandas), R, SQL
- ETL Tools – Apache Spark, dbt, Talend, Alteryx
- Cloud Services – AWS Glue, Google Dataflow, Azure Data Factory

Storage & Processing(Load)

- Data Warehouses – Snowflake, Amazon Redshift, Google BigQuery
- Data Lakes – Apache Hadoop, Amazon S3, Azure Data Lake
- Databases – PostgreSQL, MongoDB, Cassandra

A Practical Example of Data Analytics Pipeline Using Job Placement Dataset

Data Collection and Ingestion



Step 1

Collected raw Dataset from Kaggle in Csv format

→

Step 2

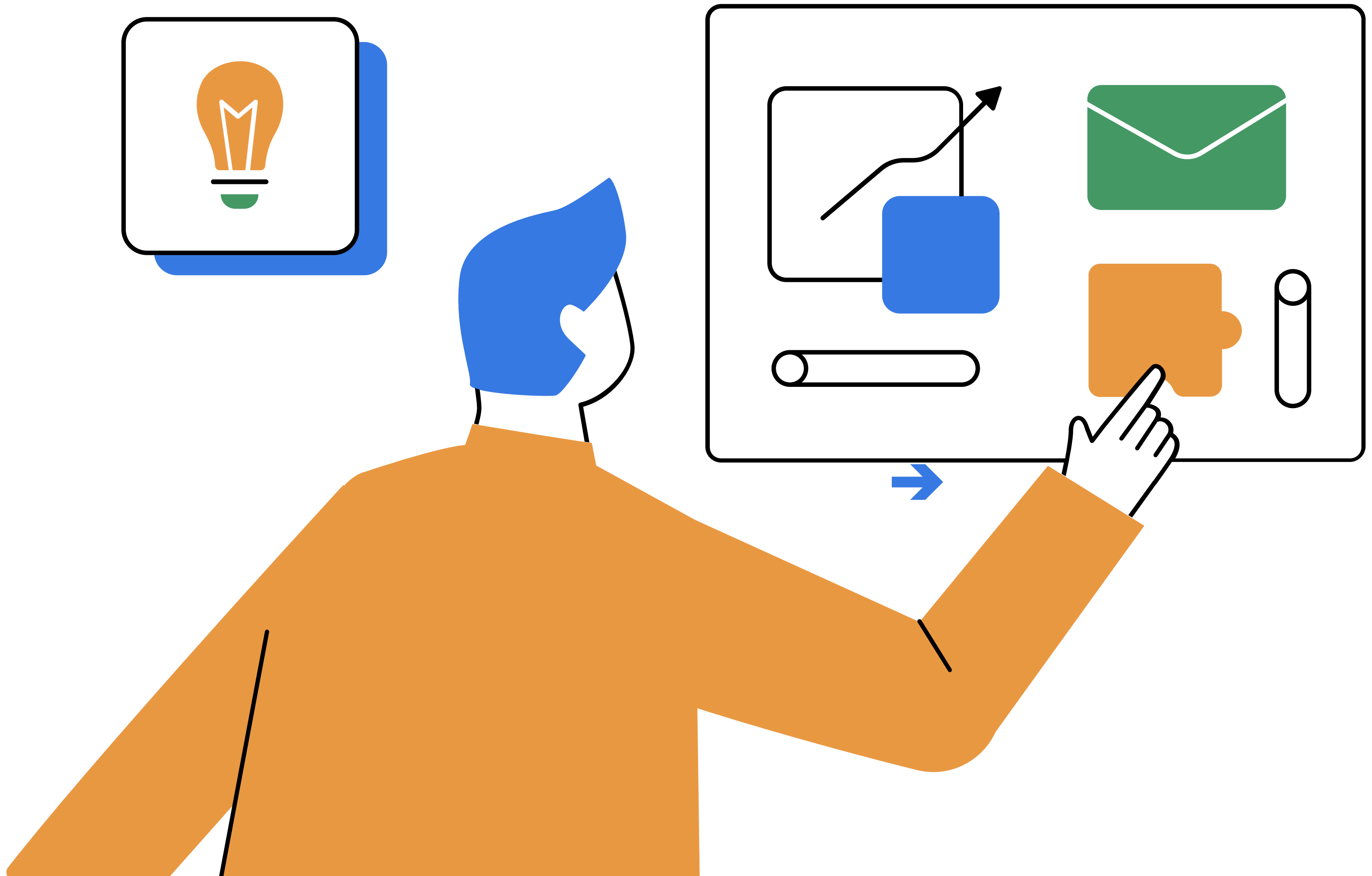
Loaded Dataset into Dataframe using Pandas

→

Step 3

Performed an intial Data Check

Data Cleaning & Transformation



1. Renamed Columns

- Renamed Long column names
  - Eg: years_of_experience to experience

2. Handled Duplicates

- Removed Duplicates

3. Standardizing Datatypes

- Converted objects to string and int wherever necessary

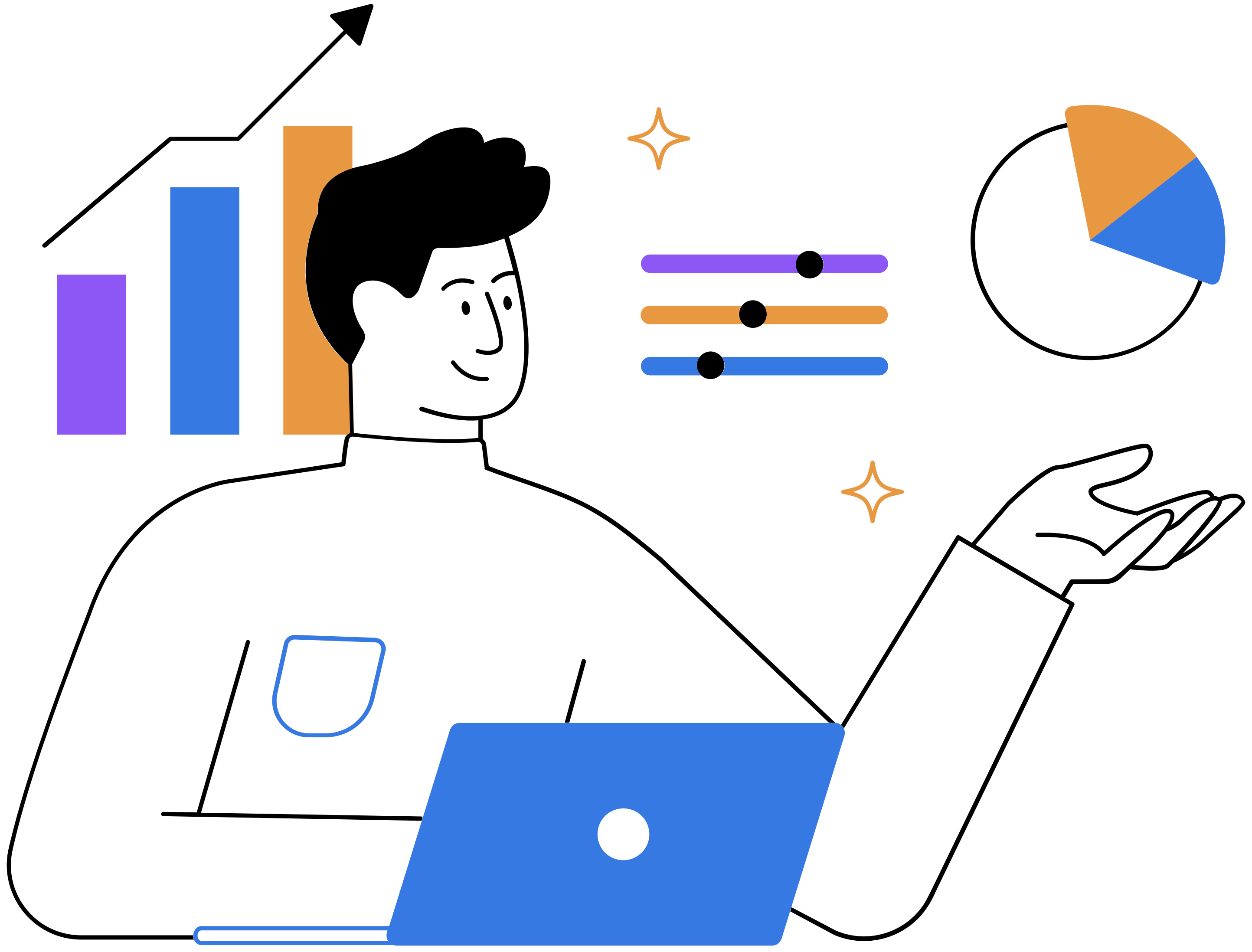
4. Missing Values

- Dropped row with empty value (single value)
- Filled Median values based on experience column (10%)
- Mode and Mean is also applicable in different scenarios

5. Handling Outliers

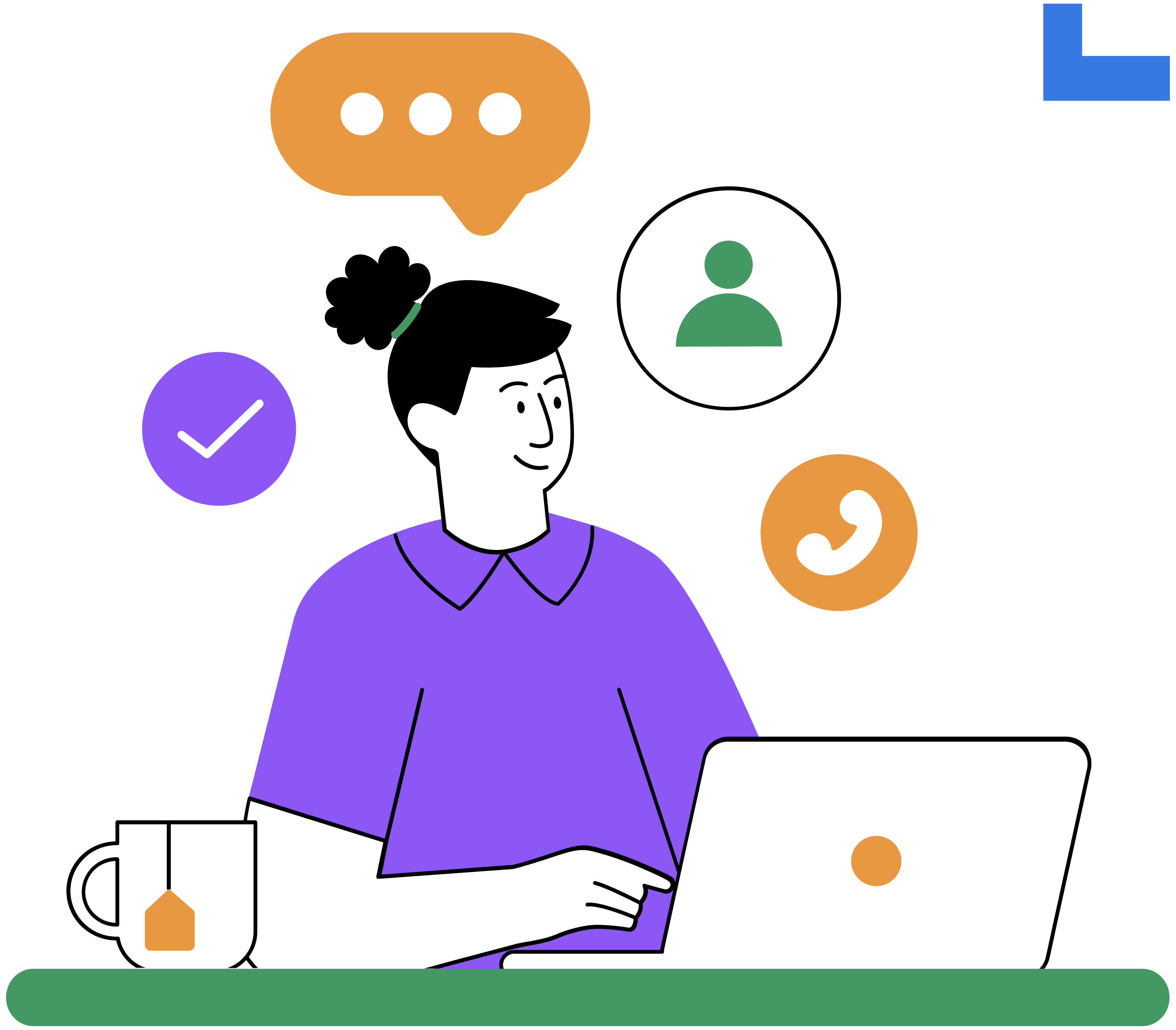
- Used Interquartile Range (IQR)

Handling Missing Values



6. Multivalued Columns

- Splitting multivalued columns to individual columns such as “skills”



Data Encoding & Feature Engineering

7. Gender

- Male = 1, Female = 0

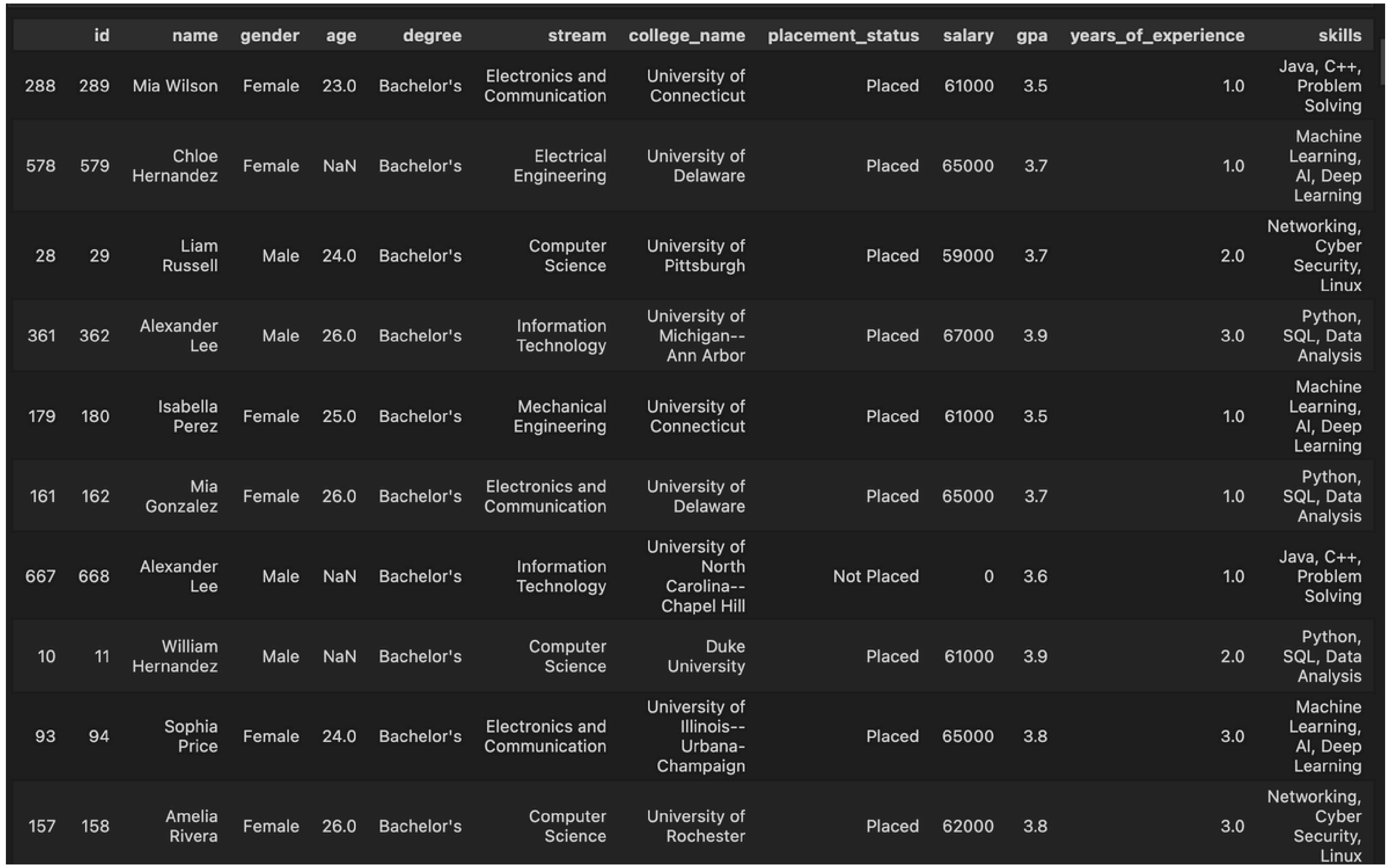
8. Stream

- Computer Science to CS

9. Dropped Columns

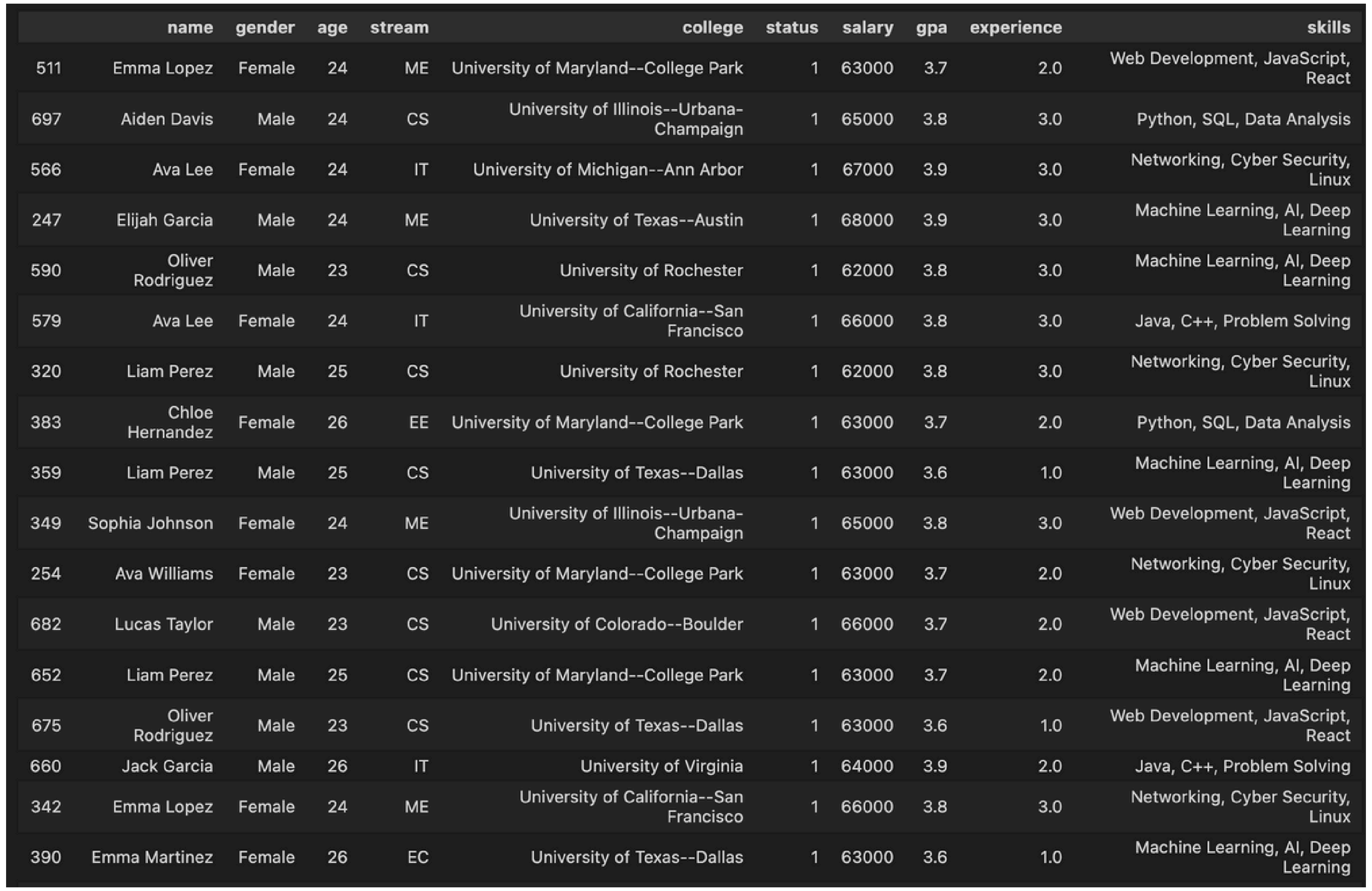
- Removed (id=completely unique, degree=same value)

Original Data

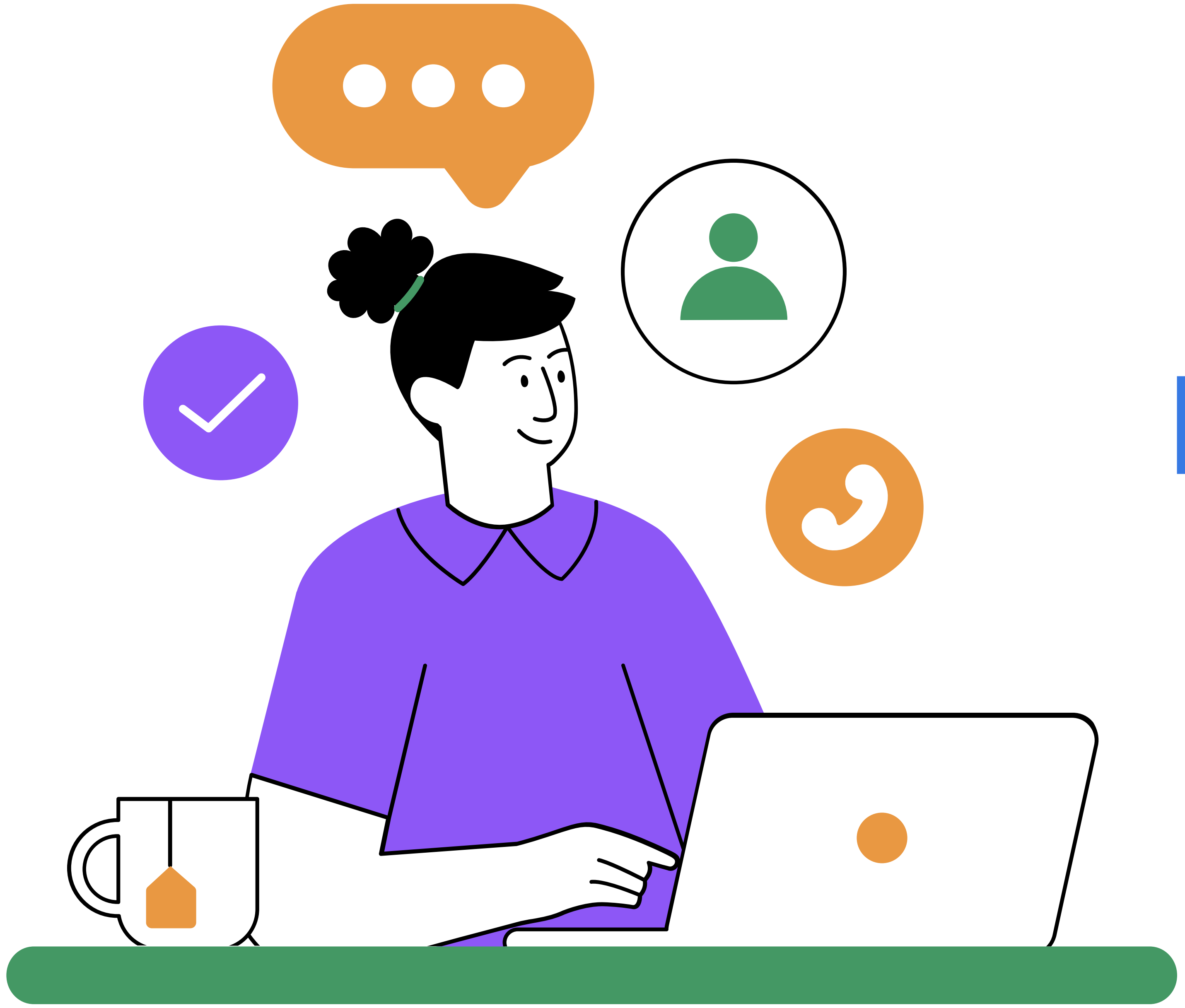


	id	name	gender	age	degree	stream	college_name	placement_status	salary	gpa	years_of_experience	skills
288	289	Mia Wilson	Female	23.0	Bachelor's	Electronics and Communication	University of Connecticut	Placed	61000	3.5	1.0	Java, C++, Problem Solving
578	579	Chloe Hernandez	Female	NaN	Bachelor's	Electrical Engineering	University of Delaware	Placed	65000	3.7	1.0	Machine Learning, AI, Deep Learning
28	29	Liam Russell	Male	24.0	Bachelor's	Computer Science	University of Pittsburgh	Placed	59000	3.7	2.0	Networking, Cyber Security, Linux
361	362	Alexander Lee	Male	26.0	Bachelor's	Information Technology	University of Michigan--Ann Arbor	Placed	67000	3.9	3.0	Python, SQL, Data Analysis
179	180	Isabella Perez	Female	25.0	Bachelor's	Mechanical Engineering	University of Connecticut	Placed	61000	3.5	1.0	Machine Learning, AI, Deep Learning
161	162	Mia Gonzalez	Female	26.0	Bachelor's	Electronics and Communication	University of Delaware	Placed	65000	3.7	1.0	Python, SQL, Data Analysis
667	668	Alexander Lee	Male	NaN	Bachelor's	Information Technology	University of North Carolina--Chapel Hill	Not Placed	0	3.6	1.0	Java, C++, Problem Solving
10	11	William Hernandez	Male	NaN	Bachelor's	Computer Science	Duke University	Placed	61000	3.9	2.0	Python, SQL, Data Analysis
93	94	Sophia Price	Female	24.0	Bachelor's	Electronics and Communication	University of Illinois--Urbana-Champaign	Placed	65000	3.8	3.0	Machine Learning, AI, Deep Learning
157	158	Amelia Rivera	Female	26.0	Bachelor's	Computer Science	University of Rochester	Placed	62000	3.8	3.0	Networking, Cyber Security, Linux

Processed Data

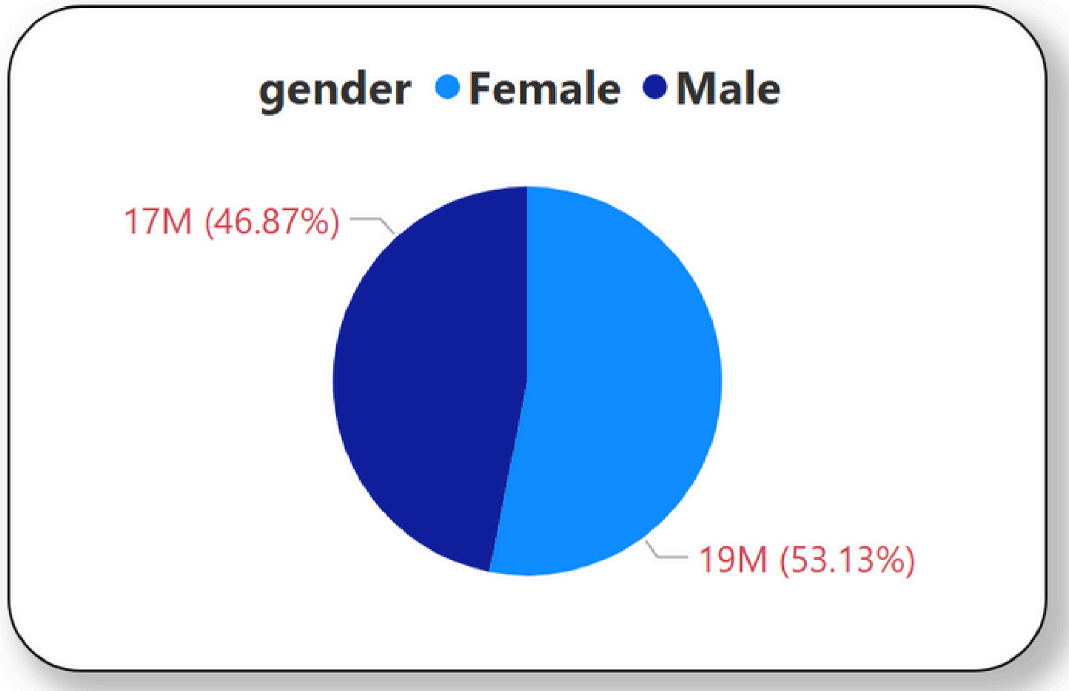


	name	gender	age	stream	college	status	salary	gpa	experience	skills
511	Emma Lopez	Female	24	ME	University of Maryland--College Park	1	63000	3.7	2.0	Web Development, JavaScript, React
697	Aiden Davis	Male	24	CS	University of Illinois--Urbana-Champaign	1	65000	3.8	3.0	Python, SQL, Data Analysis
566	Ava Lee	Female	24	IT	University of Michigan--Ann Arbor	1	67000	3.9	3.0	Networking, Cyber Security, Linux
247	Elijah Garcia	Male	24	ME	University of Texas--Austin	1	68000	3.9	3.0	Machine Learning, AI, Deep Learning
590	Oliver Rodriguez	Male	23	CS	University of Rochester	1	62000	3.8	3.0	Machine Learning, AI, Deep Learning
579	Ava Lee	Female	24	IT	University of California--San Francisco	1	66000	3.8	3.0	Java, C++, Problem Solving
320	Liam Perez	Male	25	CS	University of Rochester	1	62000	3.8	3.0	Networking, Cyber Security, Linux
383	Chloe Hernandez	Female	26	EE	University of Maryland--College Park	1	63000	3.7	2.0	Python, SQL, Data Analysis
359	Liam Perez	Male	25	CS	University of Texas--Dallas	1	63000	3.6	1.0	Machine Learning, AI, Deep Learning
349	Sophia Johnson	Female	24	ME	University of Illinois--Urbana-Champaign	1	65000	3.8	3.0	Web Development, JavaScript, React
254	Ava Williams	Female	23	CS	University of Maryland--College Park	1	63000	3.7	2.0	Networking, Cyber Security, Linux
682	Lucas Taylor	Male	23	CS	University of Colorado--Boulder	1	66000	3.7	2.0	Web Development, JavaScript, React
652	Liam Perez	Male	25	CS	University of Maryland--College Park	1	63000	3.7	2.0	Machine Learning, AI, Deep Learning
675	Oliver Rodriguez	Male	23	CS	University of Texas--Dallas	1	63000	3.6	1.0	Web Development, JavaScript, React
660	Jack Garcia	Male	26	IT	University of Virginia	1	64000	3.9	2.0	Java, C++, Problem Solving
342	Emma Lopez	Female	24	ME	University of California--San Francisco	1	66000	3.8	3.0	Networking, Cyber Security, Linux
390	Emma Martinez	Female	26	EC	University of Texas--Dallas	1	63000	3.6	1.0	Machine Learning, AI, Deep Learning



Loading Data to Power BI

JOB PLACEMENT EDA



Total Students Placed

569

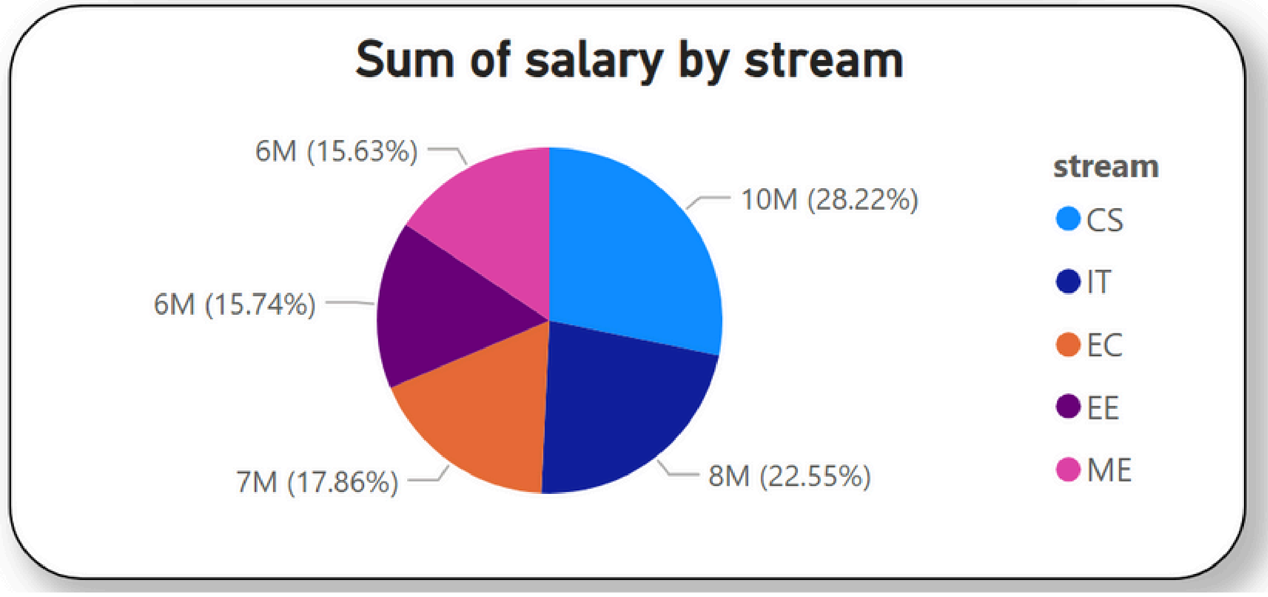
Male

267

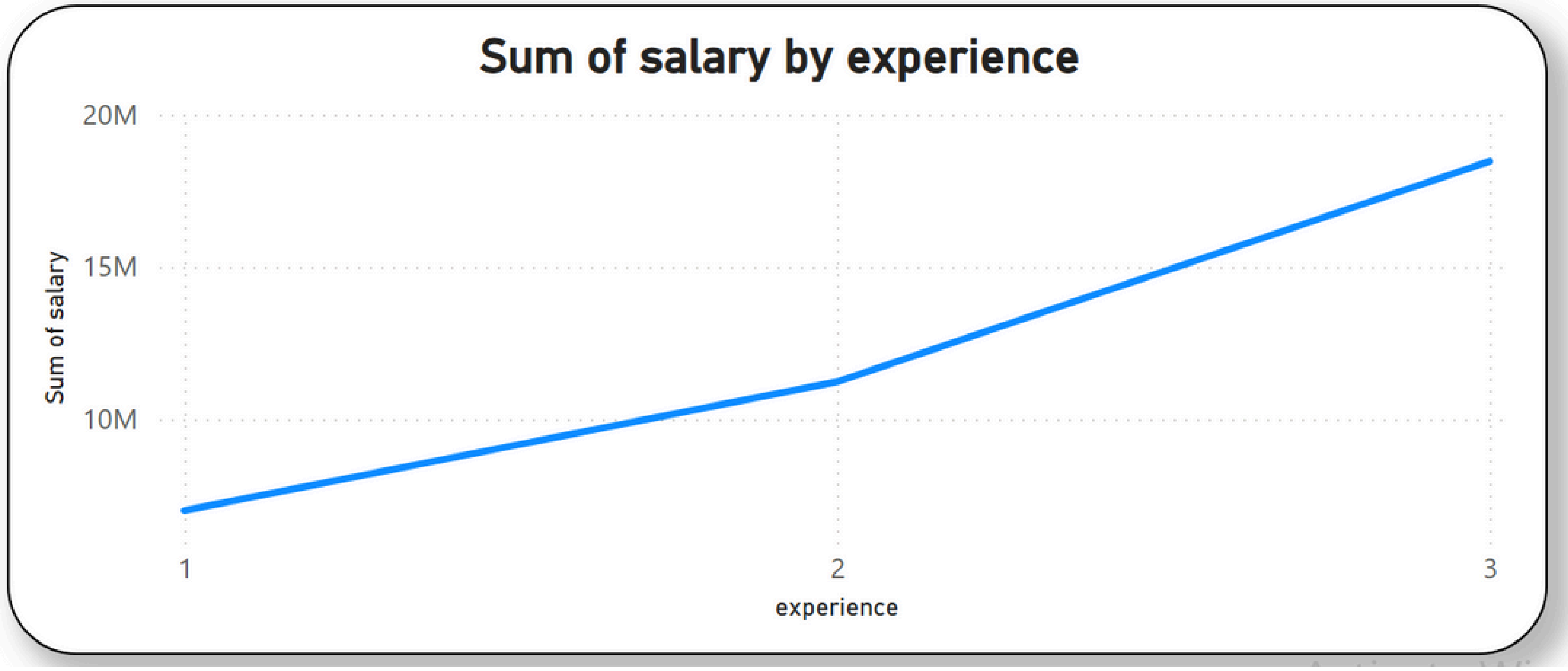
Female

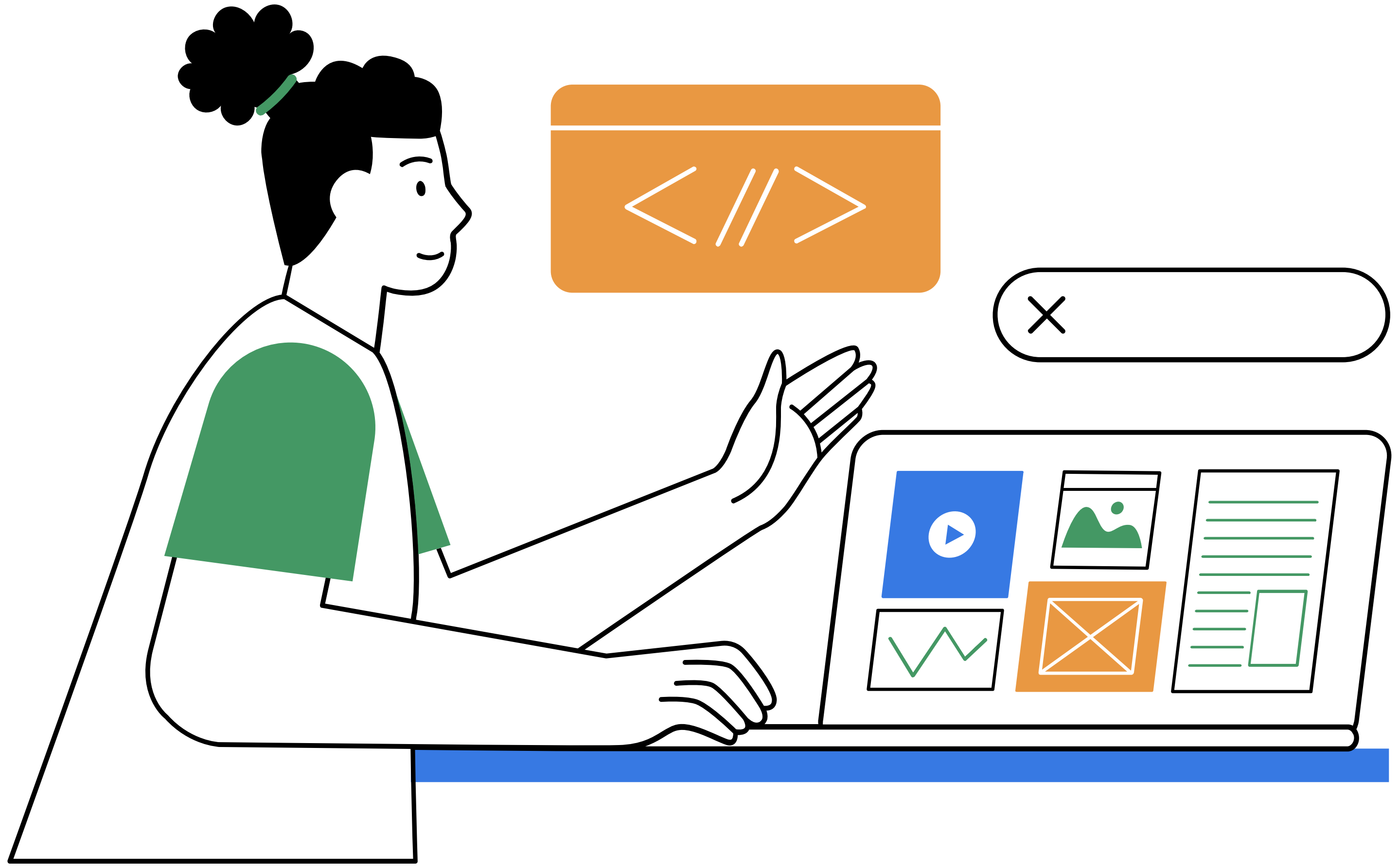
302

Sum of salary by stream



Sum of salary by experience





Key Takeaways and Future Scope

Placement Analysis

- Experience plays crucial role
- GPA & Experience lead to better salary package
- Females tend to have higher placement percentage by a slight margin

Next Steps in Analysis

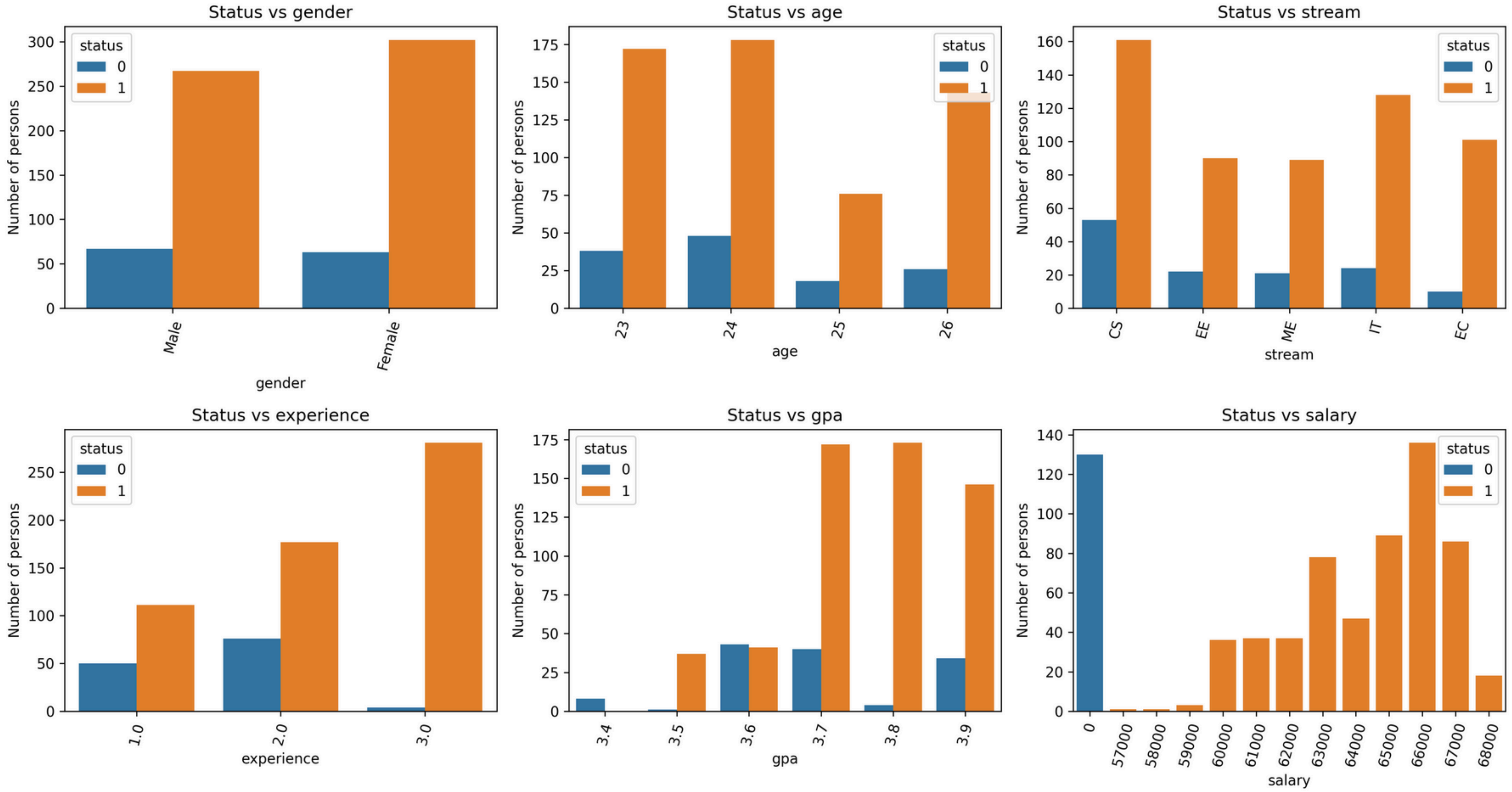
- Automate reporting and dashboard
- Conduct Predictive Modelling for placements
- Optimize data storage

Future Add ons

- Enhanced dataset with more features
- Career recommendation system based on profiles

Key Insights from Data Visualisation

Placement status (Placed = 1, Not Placed = 0)



Status vs gender

Status vs age

Status vs stream

Status vs experience

Status vs gpa

Status vs salary

Thank You