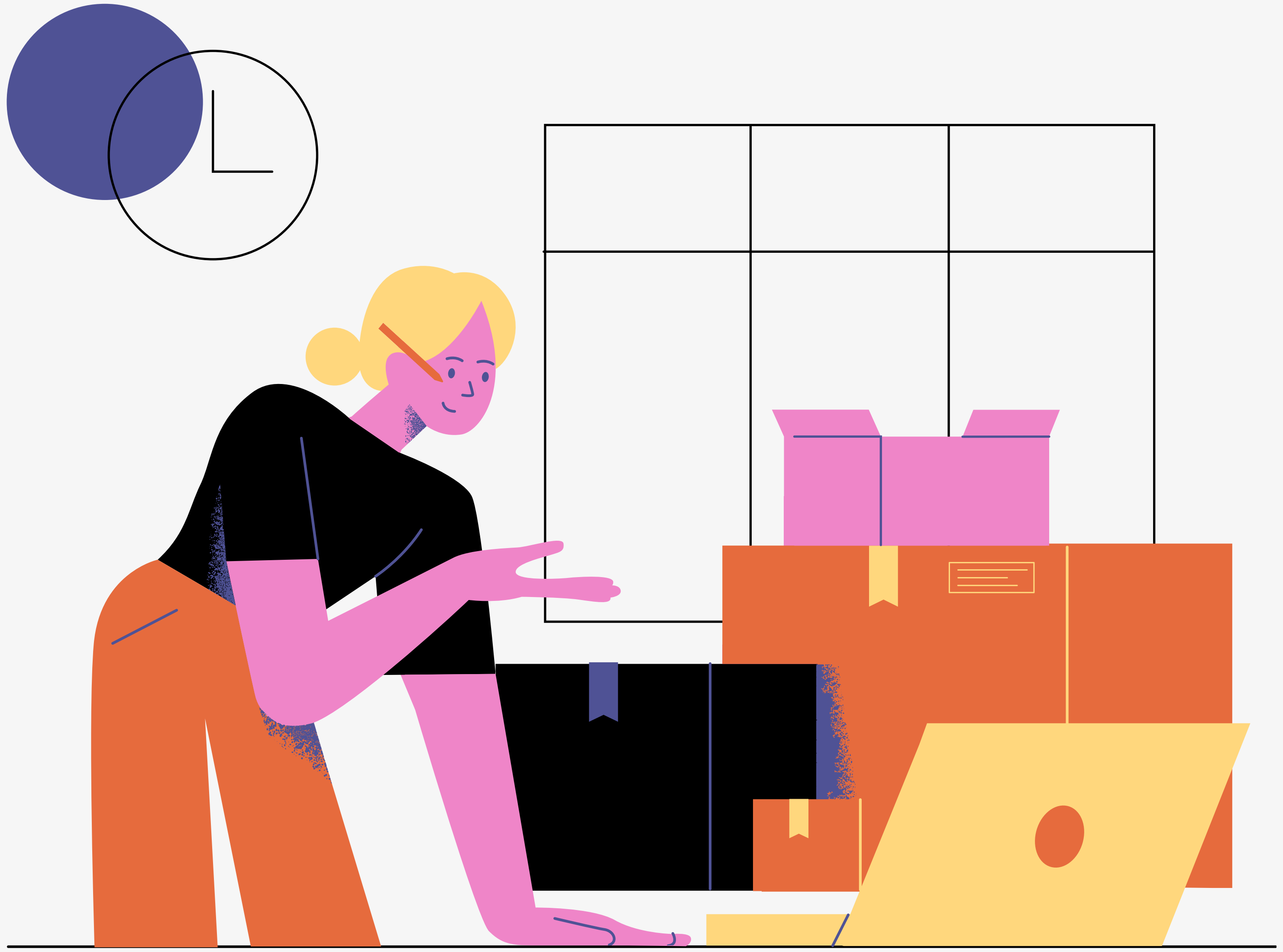
CUSTOMER ANALYTICS:

Data Preparation for Modeling

Overview



A frequent challenge in building models that extract business value from data is that datasets can be so extensive that it may take days for the model to produce predictions. It is essential to store the dataset as efficiently as possible to enable these models to operate on a more reasonable timeline without having to decrease the dataset's size.

A major online data science training provider does the project to optimizes one of the largest customer datasets. This dataset will eventually be utilized to predict whether the students are seeking new job opportunities, information they will then use to connect students with potential recruiters.

Problem Statement

The goal is to optimize the storage of a large customer dataset to improve model performance and reduce processing time. The dataset contains various information about students, including demographic details, educational background, work experience, and job-seeking intentions.

Goals

01 **Optimize Data Types**
Convert appropriate columns to more efficient data types (e.g., boolean, integer, float16, categorical).

02 **Filter Data**
Select a subset of the data based on specific criteria (e.g., experience, company size).

03 **Improve Data Storage**
Reduce memory usage by storing data in a more efficient format.

Dataset

The dataset is a CSV file named **customer_train.csv** containing information about online students. The source of this data is an online data science training provider.

The dataset includes the following columns:

- **student_id**: Unique identifier for each student.
- **city**: City code.
- **city_development_index**: Development index of the city.
- **gender**: Gender of the student.
- **relevant_experience**: Whether the student has relevant work experience.
- **enrolled_university**: Type of university course enrolled in.
- **education_level**: Highest education level.
- **major_discipline**: Major field of study.
- **experience**: Total work experience in years.
- **company_size**: Size of the current company.
- **company_type**: Type of company.
- **last_new_job**: Years since the last job change.
- **training_hours**: Hours of training completed.
- **job_change**: Whether the student is looking for a new job (1) or not (0).

Outputs

We create a DataFrame called **ds_jobs_transformed**, which optimizes the storage of data from **customer_train.csv** based on the following specifications:

01 Columns with only two categories should be stored as Booleans (**bool**).

02 Columns with integer values should be stored as 32-bit integers (**int32**).

03 Columns with floating-point values should be stored as 16-bit floats (**float16**).

04 Columns with nominal categorical data should be stored as the **category** data type.

05 Columns with ordinal categorical data should be stored as ordered categories, maintaining the natural order without converting them to numerical values.

06 The DataFrame should only include students with 10 or more years of experience, employed at companies with at least 1000 employees, as the recruiter base is focused on experienced professionals at large enterprises.

Data **Findings**

01 Data Type Optimization

The code converts columns with two categories to boolean, integer columns to int32, float columns to float16, and categorical columns to appropriate categorical data types.

02 Data Filtering

The code filters the dataset to include only students with 10 or more years of experience working at companies with 1000 or more employees.

03 Data Insights

The code provides insights into the distribution of categorical variables, such as gender, education level, and company size.

#01

```
city_103    4355
city_21     2702
city_16     1533
city_114    1336
city_160     845
            ...
city_129       3
city_111       3
city_121       3
city_140       1
city_171       1
Name: city, Length: 123, dtype: int64
```

This output shows the distribution of students across different cities. Each line represents a city and the corresponding number indicates the count of students from that city.

Here are some key observations:

- **City Diversity:** The dataset includes students from 123 different cities.
- **Uneven Distribution:** The distribution of students across cities is highly uneven. A few cities have a significantly larger number of students compared to others.
- **Dominant Cities:** Cities like **city_103** and **city_21** have a much higher number of students compared to the other cities. This could indicate that these cities are major hubs for online education or have a higher population of potential students.
- **Smaller Cities:** A large number of cities have only a few students, suggesting that the platform has a wider reach, even in smaller cities.

#02

```
Male       13221
Female      1238
Other        191
Name: gender, dtype: int64
```

This output shows the distribution of students based on their gender. It provides a count of students for each gender category:

- **Male:** 13,221 students
- **Female:** 12,38 students
- **Other:** 191 students

#03

```
Has relevant experience    13792
No relevant experience      5366
Name: relevant_experience, dtype: int64
```

This output shows the distribution of students based on their relevant work experience. It provides a count of students for each category:

- **Has relevant experience:** 13,792 students
- **No relevant experience:** 5,366 students

#04

```
no_enrollment       13817
Full time course     3757
Part time course     1198
Name: enrolled_university, dtype: int64
```

This output shows the distribution of students based on their enrollment status in a university. It provides a count of students for each category:

- **No enrollment:** 13,817 students
- **Full-time course:** 3,757 students
- **Part-time course:** 1,198 students

#05

```
Graduate          11598
Masters            4361
High School        2017
Phd                 414
Primary School      308
Name: education_level, dtype: int64
```

This output shows the distribution of students based on their highest level of education. It provides a count of students for each educational category:

- **Graduate:** 11,598 students
- **Masters:** 4,361 students
- **High School:** 2,017 students
- **PhD:** 414 students
- **Primary School:** 308 students

#06

```
STEM               14492
Humanities           669
Other                381
Business Degree      327
Arts                 253
No Major             223
Name: major_discipline, dtype: int64
```

This output shows the distribution of students based on their major discipline. It provides a count of students for each major discipline:

- **STEM:** 14,492 students
- **Humanities:** 669 students
- **Other:** 381 students
- **Business Degree:** 327 students
- **Arts:** 253 students
- **No Major:** 223 students

#07

```
>20    3286
5      1430
4      1403
3      1354
6      1216
2      1127
7      1028
10      985
9       980
8       802
15      686
11      664
14      586
1       549
<1      522
16      508
12      494
13      399
17      342
19      304
18      280
20      148
Name: experience, dtype: int64
```

This output shows the distribution of students based on their total work experience. Each line represents a specific experience level, and the number next to it indicates the count of students with that level of experience.

Here are some key observations:

- **Experience Range:** The data includes students with a wide range of experience, from less than one year to over 20 years.

- **Most Common Experience Levels:** The most common experience levels are 5, 4, and 3 years, with over 1,300 students in each category.

- **Decreasing Frequency:** As the experience level increases, the number of students generally decreases. This is likely due to the fact that fewer people have extensive work experience.

#08

```
50-99         3083
100-499       2571
10000+        2019
10-49         1471
1000-4999     1328
<10           1308
500-999        877
5000-9999      563
Name: company_size, dtype: int64
```

This output shows the distribution of students based on the size of their current company. Each line represents a company size range, and the number next to it indicates the count of students working in companies of that size.

Here are some key observations:

- **Most Common Company Sizes:** The most common company sizes are 50-99 employees and 100-499 employees, with over 2,500 students in each category.
- **Large Companies:** A significant number of students work in large companies with 10,000 or more employees.
- **Smaller Companies:** A considerable number of students also work in smaller companies with fewer than 10 employees.

#09

```
Pvt Ltd                9817
Funded Startup         1001
Public Sector           955
Early Stage Startup     603
NGO                     521
Other                   121
Name: company_type, dtype: int64
```

This output shows the distribution of students based on the type of company they are currently working for. Each line represents a company type, and the number next to it indicates the count of students working in that type of company.

Here are some key observations:

- **Pvt Ltd:** This is the most common company type, with 9817 students working in private limited companies.
- **Funded Startup:** 1001 students work in funded startups.
- **Public Sector:** 955 students work in the public sector.
- **Early Stage Startup:** 603 students work in early-stage startups.
- **NGO:** 521 students work in non-governmental organizations.
- **Other:** 121 students work in other types of companies.

#10

```
1        8040
>4       3290
2        2900
never    2452
4        1029
3        1024
Name: last_new_job, dtype: int64
```

This output shows the distribution of students based on the number of years since their last job change. Each line represents a time period, and the number next to it indicates the count of students with that specific time since their last job change.

Here are some key observations:

- **Recent Job Changes:** The majority of students had their last job change within the past year (8040 students).
- **Longer Gaps:** A significant number of students (3290) have not changed jobs in more than 4 years.
- **Intermediate Gaps:** There are also substantial numbers of students who changed jobs 2, 3, or 4 years ago.

Insights

By optimizing data types and filtering the dataset, the code effectively addresses the problem of storing large datasets efficiently. This optimized dataset can be used to train machine learning models more quickly and efficiently, leading to improved model performance and reduced computational costs.