A common problem when creating models to generate business value from data is that the datasets can be so large that it can take days for the model to generate predictions. Ensuring that your dataset is stored as efficiently as possible is crucial for allowing these models to run on a more reasonable timescale without having to reduce the size of the dataset.

You've been hired by a major online data science training provider called *Training Data Ltd.* to clean up one of their largest customer datasets. This dataset will eventually be used to predict whether their students are looking for a new job or not, information that they will then use to direct them to prospective recruiters.

You've been given access to `customer_train.csv`, which is a subset of their entire customer dataset, so you can create a proof-of-concept of a much more efficient storage solution. The dataset contains anonymized student information, and whether they were looking for a new job or not during training:

Column	Description
`student_id`	A unique ID for each student.
`city`	A code for the city the student lives in.
`city_development_index`	A scaled development index for the city.
`gender`	The student's gender.
`relevant_experience`	An indicator of the student's work relevant experience.
`enrolled_university`	The type of university course enrolled in (if any).
`education_level`	The student's education level.
`major_discipline`	The educational discipline of the student.
`experience`	The student's total work experience (in years).
`company_size`	The number of employees at the student's current employer.
`last_new_job`	The number of years between the student's current and previous jobs.
`training_hours`	The number of hours of training completed.
`job_change`	An indicator of whether the student is looking for a new job (`1`) or not (`0`).

In [1]:
```
# Start your code here!
import pandas as pd
#Load ds_jobs dataframe
ds_jobs = pd.read_csv("customer_train.csv")
```

In [2]:
```
print(ds_jobs.head())
print(ds_jobs.info())
```

```
   student_id      city  city_development_index gender  \
0        8949  city_103                   0.920   Male
1       29725   city_40                   0.776   Male
2       11561   city_21                   0.624    NaN
3       33241  city_115                   0.789    NaN
4         666  city_162                   0.767   Male

       relevant_experience enrolled_university education_level  \
0  Has relevant experience       no_enrollment        Graduate
1   No relevant experience       no_enrollment        Graduate
2   No relevant experience    Full time course        Graduate
3   No relevant experience                NaN        Graduate
4  Has relevant experience       no_enrollment         Masters

  major_discipline experience company_size    company_type last_new_job  \
0             STEM        >20          NaN             NaN            1
1             STEM         15        50-99         Pvt Ltd           >4
2             STEM          5          NaN             NaN        never
3  Business Degree         <1          NaN         Pvt Ltd        never
4             STEM        >20        50-99  Funded Startup            4

   training_hours  job_change
0              36           1
1              47           0
2              83           0
3              52           1
4               8           0
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19158 entries, 0 to 19157
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   student_id              19158 non-null  int64
 1   city                    19158 non-null  object
 2   city_development_index  19158 non-null  float64
 3   gender                  14650 non-null  object
 4   relevant_experience     19158 non-null  object
 5   enrolled_university     18772 non-null  object
 6   education_level         18698 non-null  object
 7   major_discipline        16345 non-null  object
 8   experience              19093 non-null  object
 9   company_size            13220 non-null  object
 10  company_type            13018 non-null  object
 11  last_new_job            18735 non-null  object
 12  training_hours          19158 non-null  int64
 13  job_change              19158 non-null  int64
dtypes: float64(1), int64(3), object(10)
memory usage: 2.0+ MB
None
```

In [3]:
```
#student_id should be integer, which is already the case
ds_jobs_clean= pd.DataFrame()
ds_jobs_clean['student_id'] = ds_jobs['student_id'].astype('int32')
```

In [4]:
```
#City should be a nomial categorical data valiables
ds_jobs_clean['city'] = ds_jobs['city'].astype('category')
```

In [5]:
```
#City_development_index should be float, which is the case already
ds_jobs_clean['city_development_index'] = ds_jobs.city_development_index.astype("float16")
```

In [6]:
```
#Gender could be a nomial categorical variable
ds_jobs_clean['gender'] = ds_jobs['gender'].astype('category')
```

In [7]:
```
#inspect relevant_experience
print(ds_jobs.relevant_experience.unique())
my_series0 = pd.Categorical(ds_jobs.relevant_experience, categories=["No relevant experience", "Has relevant experience"], ordered = True)
```

```
['Has relevant experience' 'No relevant experience']
```

In [8]:
```
#revelant_expereince should be a categorical variable
ds_jobs_clean['relevant_experience'] = my_series0
```

In [9]:
```
#inspect enrolled_university
ds_jobs.enrolled_university.unique()
my_series1 = pd.Categorical(ds_jobs.enrolled_university, categories = ["no_enrollment", "Part time course", "Full time course"], ordered= True)
```

In [10]:
```
#Enrolled_in_University should be nominal category
ds_jobs_clean['enrolled_university'] = my_series1
```

In [11]:
```
#inspect education_level
ds_jobs.education_level.unique()
my_series2 = pd.Categorical(ds_jobs.education_level,categories= ["Primary School", "High School", "Graduate", "Masters", "Phd"], ordered = True)
```

In [12]:
```
#Education_level should be nomial categorical variable
ds_jobs_clean['education_level'] = my_series2
```

In [13]:
```
#Major discpline should be nomial categorical variable
ds_jobs_clean['major_discipline'] = ds_jobs.major_discipline.astype('category')
```

In [14]:
```
#Experience should be ordinal categorical data
my_series = pd.Categorical(ds_jobs.experience, categories=['0','<1', '1','2','3','4','5','6','7','8','9','10',\
                                                           '11','12','13','14','15','16','17','18','19','20','>20'], ordered = True)

ds_jobs_clean['experience'] = my_series
```

In [15]:
```
#Company_size should be ordinal categorical data
ds_jobs.company_size.unique()
my_series1 = pd.Categorical(ds_jobs.company_size, categories=['<10', '10-49', '50-99','100-499', '500-999', '1000-4999', '5000-9999', '10000+'], ordered = True)
ds_jobs_clean['company_size'] = my_series1
```

In [16]:
```
#Compnay type should be nominal categorical data
ds_jobs_clean['company_type'] = ds_jobs.company_type.astype('category')
```

In [17]:
```
#last_new_job should be ordinal categorical data
ds_jobs.last_new_job.unique()
my_series2 = pd.Categorical(ds_jobs.last_new_job, categories=['never', '1', '2', '3', '4', '>4'], ordered = True)
ds_jobs_clean['last_new_job'] = my_series2
```

In [18]:
```
#Training hours should be int, which is already the case
ds_jobs_clean['training_hours'] = ds_jobs.training_hours.astype('int32')
```

In [19]:
```
#job_change should be nominal categorical data
ds_jobs_clean['job_change'] = ds_jobs.job_change.astype('int32')
```

In [20]:
```
ds_jobs_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19158 entries, 0 to 19157
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   student_id              19158 non-null  int32
 1   city                    19158 non-null  category
 2   city_development_index  19158 non-null  float16
 3   gender                  14650 non-null  category
 4   relevant_experience     19158 non-null  category
 5   enrolled_university     18772 non-null  category
 6   education_level         18698 non-null  category
 7   major_discipline        16345 non-null  category
 8   experience              19093 non-null  category
 9   company_size            13220 non-null  category
 10  company_type            13018 non-null  category
 11  last_new_job            18735 non-null  category
 12  training_hours          19158 non-null  int32
 13  job_change              19158 non-null  int32
dtypes: category(10), float16(1), int32(3)
memory usage: 456.5 KB
```

In [21]:
```
ds_jobs_clean = ds_jobs_clean[(ds_jobs_clean['experience'] >= '10') & (ds_jobs_clean['company_size'] >= '1000-4999') ]
ds_jobs_clean
```

Out[21]:

	student_id	city	city_development_index	gender	relevant_experience	enrolled_university	education_level	major_discipline	experience	company_size	company_type	last_new_job	training_hours	job_change
9	699	city_103	0.919922	NaN	Has relevant experience	no_enrollment	Graduate	STEM	17	10000+	Pvt Ltd	>4	123	0
12	25619	city_61	0.913086	Male	Has relevant experience	no_enrollment	Graduate	STEM	>20	1000-4999	Pvt Ltd	3	23	0
31	22293	city_103	0.919922	Male	Has relevant experience	Part time course	Graduate	STEM	19	5000-9999	Pvt Ltd	>4	141	0
34	26494	city_16	0.910156	Male	Has relevant experience	no_enrollment	Graduate	Business Degree	12	5000-9999	Pvt Ltd	3	145	0
40	2547	city_114	0.925781	Female	Has relevant experience	Full time course	Masters	STEM	16	1000-4999	Public Sector	2	14	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
19097	25447	city_103	0.919922	Male	Has relevant experience	no_enrollment	Graduate	STEM	>20	1000-4999	Pvt Ltd	>4	57	0
19101	6803	city_16	0.910156	Male	Has relevant experience	no_enrollment	High School	NaN	10	10000+	Pvt Ltd	1	89	0
19103	32932	city_10	0.895020	Male	Has relevant experience	Part time course	Masters	Other	>20	1000-4999	Pvt Ltd	>4	18	0
19128	3365	city_16	0.910156	NaN	Has relevant experience	no_enrollment	Graduate	Humanities	>20	1000-4999	Pvt Ltd	>4	23	0
19143	33047	city_103	0.919922	Male	Has relevant experience	no_enrollment	Graduate	STEM	>20	10000+	Pvt Ltd	>4	18	0

2201 rows × 14 columns