PROJECT: CUSTOMER ANALYTICS: PREPARING DATA FOR MODELING

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.
`company_type`	The type of company employing the student.
`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`).

```
# Import necessary libraries
import pandas as pd

# Load the dataset
ds_jobs = pd.read_csv("customer_train.csv")

# View the dataset
ds_jobs.head()
```

	s...		city_development_ind...		relevant_experience	enrolled_universi...	education...
0	8949	city_103	0.92	Male	Has relevant experience	no_enrollment	Graduate
1	29725	city_40	0.776	Male	No relevant experience	no_enrollment	Graduate
2	11561	city_21	0.624	null	No relevant experience	Full time course	Graduate
3	33241	city_115	0.789	null	No relevant experience	null	Graduate
4	666	city_162	0.767	Male	Has relevant experience	no_enrollment	Masters

Rows: 5 Expand

```
# Create a copy of ds_jobs for transforming
ds_jobs_transformed = ds_jobs.copy()
```

Store the data much more efficiently

You have to store the data in ds_jobs_transformed, following the requirements

1. Columns containing categories with two factors must be stores as booleans
2. Columns containing integers only must be store as int32
3. Columns containing floats must be stored as float16
4. Columns containing nominal categorical data must be stored as category dtype
5. Columns with ordinal data must be stored as an ordered category
6. Df should be filtered to only students with 10 or more years of experience at companies with at least 1000 employees.

```
ds_jobs_transformed.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  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  float64
dtypes: float64(2), int64(2), object(10)
memory usage: 2.0+ MB
```

Columns containing categories with two factors must be stores as booleans

The columns with just two categories are:

- job_change
- relevant_experience

```
ds_jobs_transformed['job_change'].value_counts(dropna = False)

# Converting the column to boolean

ds_jobs_transformed.loc[ds_jobs_transformed['job_change'] == 1, 'job_change'] = True
ds_jobs_transformed.loc[ds_jobs_transformed['job_change'] == 0, 'job_change'] = False

ds_jobs_transformed.loc[ds_jobs_transformed['relevant_experience'] == 'Has relevant experience', 'relevant_experience'] = True
ds_jobs_transformed.loc[ds_jobs_transformed['relevant_experience'] == 'No relevant experience', 'relevant_experience'] = False

ds_jobs_transformed['job_change'] = ds_jobs_transformed['job_change'].astype('bool')
ds_jobs_transformed['relevant_experience'] = ds_jobs_transformed['relevant_experience'].astype('bool')
```

```
ds_jobs_transformed['job_change'].value_counts(dropna = False)
```

	j…
False	14381
True	4777

Rows: 2 Expand

```
ds_jobs_transformed['relevant_experience'].value_counts(dropna = False)
```

	relevant_experie…
True	13792
False	5366

Rows: 2 Expand

Columns containing integers only must be store as int32

The columns with integers are:

- student_id
- training_hours

```
# Converting the columns to dtype int32

ds_jobs_transformed['student_id'] = ds_jobs_transformed['student_id'].astype('int32')
ds_jobs_transformed['training_hours'] = ds_jobs_transformed['training_hours'].astype('int32')

ds_jobs_transformed.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  object
 2   city_development_index  19158 non-null  float64
 3   gender                  14650 non-null  object
 4   relevant_experience     19158 non-null  bool
 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  int32
 13  job_change              19158 non-null  bool
dtypes: bool(2), float64(1), int32(2), object(9)
memory usage: 1.6+ MB
```

Columns containing floats must be stored as float16

The columns with floats are:

- city_development_index

```
ds_jobs_transformed['city_development_index'] = ds_jobs_transformed['city_development_index'].astype('float16')

ds_jobs_transformed.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  object
 2   city_development_index  19158 non-null  float16
 3   gender                  14650 non-null  object
 4   relevant_experience     19158 non-null  bool
 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  int32
 13  job_change              19158 non-null  bool
dtypes: bool(2), float16(1), int32(2), object(9)
memory usage: 1.5+ MB
```

Columns containing nominal categorical data must be stored as category dtype

The columns with nominal categorical data are:

- city
- gender
- major_discipline

```
ds_jobs_transformed[['city', 'gender', 'major_discipline', 'company_type']] = ds_jobs_transformed[['city', 'gender',
'major_discipline', 'company_type']].astype('category')

ds_jobs_transformed.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  bool
 5   enrolled_university     18772 non-null  object
 6   education_level         18698 non-null  object
 7   major_discipline        16345 non-null  category
 8   experience              19093 non-null  object
 9   company_size            13220 non-null  object
 10  company_type            13018 non-null  category
 11  last_new_job            18735 non-null  object
 12  training_hours          19158 non-null  int32
 13  job_change              19158 non-null  bool
dtypes: bool(2), category(4), float16(1), int32(2), object(5)
memory usage: 1.0+ MB
```

Columns with ordinal data must be stored as an ordered category

The columns with ordinal categorical data are:

- last_new_job
- experience
- education_level
- enrolled_university
- company_size

```
ds_jobs_transformed['enrolled_university'].unique()
```

```
array(['no_enrollment', 'Full time course', nan, 'Part time course'],
      dtype=object)
```

```
ds_jobs_transformed[['last_new_job', 'experience', 'education_level', 'enrolled_university', 'company_size']] =
ds_jobs_transformed[['last_new_job', 'experience', 'education_level', 'enrolled_university', 'company_size']].astype('category')
```

```
# last new job
ds_jobs_transformed['last_new_job'] = ds_jobs_transformed['last_new_job'].cat.set_categories(
                                                                    new_categories = ['never','1', '2', '3', '4',
'>4'],
                                                                    ordered = True) #Nan

# experience
ds_jobs_transformed['experience'] = ds_jobs_transformed['experience'].cat.set_categories(
                                                                new_categories = ['<1', '1', '2', '3', '4', '5', '6',
'7',
                                                                                  '8', '9', '10', '11', '12','13', '14',
'15',
                                                                                  '16', '17', '18', '19', '20', '>20'],
                                                                ordered = True) #Nan

# education_level
ds_jobs_transformed['education_level'] = ds_jobs_transformed['education_level'].cat.set_categories(
                                                                    new_categories = ['Primary School','High School',
                                                                                      'Graduate', 'Masters', 'Phd'],
                                                                    ordered = True) #Nan

# enrolled_university
ds_jobs_transformed['company_size'] = ds_jobs_transformed['company_size'].cat.set_categories(
                                                                    new_categories = ['<10','10-49',
                                                                                      '50-99', '100-499', '1000-4999',
                                                                                     '5000-9999', '10000+'],
                                                                    ordered = True) #Nan

# company_size
ds_jobs_transformed['enrolled_university'] = ds_jobs_transformed['enrolled_university'].cat.set_categories(
                                                                    new_categories = ['no_enrollment','Part time
course',
                                                                                      'Full time course'],
                                                                    ordered = True) #Nan

ds_jobs_transformed.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  bool
 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            12343 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  bool
dtypes: bool(2), category(9), float16(1), int32(2)
memory usage: 400.2 KB
```

Df should be filtered to only students with 10 or more years of experience at companies with at least 1000 employees.

```
# filtering the dataset
filter_exp = ['10', '11', '12','13', '14', '15','16', '17', '18', '19', '20', '>20']
filter_comp = ['1000-4999','5000-9999', '10000+']

ds_jobs_transformed = ds_jobs_transformed[(ds_jobs_transformed['experience'].isin(filter_exp)) &
(ds_jobs_transformed['company_size'].isin(filter_comp))]
```

```
ds_jobs_transformed
```

	s...		city_development_ind...		relevant_experie...	enrolled_universi...	education...
9	699	city_103	0.919921875		True	no_enrollment	Graduate
12	25619	city_61	0.9130859375	Male	True	no_enrollment	Graduate
31	22293	city_103	0.919921875	Male	True	Part time course	Graduate
34	26494	city_16	0.91015625	Male	True	no_enrollment	Graduate
40	2547	city_114	0.92578125	Female	True	Full time course	Masters
47	25987	city_103	0.919921875	Other	True	no_enrollment	Graduate
104	1180	city_16	0.91015625	Male	True	no_enrollment	Graduate
108	25349	city_16	0.91015625	Male	True	no_enrollment	Graduate
115	20576	city_97	0.9248046875	Male	True	no_enrollment	Graduate
130	3921	city_36	0.8930664062		False	no_enrollment	Phd
144	24796	city_103	0.919921875	Male	True	no_enrollment	Graduate
146	22718	city_157	0.7690429688		True	no_enrollment	Graduate
154	12154	city_16	0.91015625		True	no_enrollment	Masters
157	28817	city_89	0.9248046875	Male	True	Full time course	Graduate
159	7280	city_103	0.919921875	Male	True	no_enrollment	Graduate
160	16903	city_103	0.919921875	Male	True	no_enrollment	Graduate

Rows: 2,201 Expand

```
ds_jobs_transformed.shape
```

```
(2201, 14)
```

```
# Final information of the data
ds_jobs_transformed.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2201 entries, 9 to 19143
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   student_id              2201 non-null   int32
 1   city                    2201 non-null   category
 2   city_development_index  2201 non-null   float16
 3   gender                  1821 non-null   category
 4   relevant_experience     2201 non-null   bool
 5   enrolled_university     2185 non-null   category
 6   education_level         2184 non-null   category
 7   major_discipline        2097 non-null   category
 8   experience              2201 non-null   category
 9   company_size            2201 non-null   category
 10  company_type            2144 non-null   category
 11  last_new_job            2184 non-null   category
 12  training_hours          2201 non-null   int32
 13  job_change              2201 non-null   bool
dtypes: bool(2), category(9), float16(1), int32(2)
memory usage: 134.1 KB
```