In [212…]
```
# Import necessary libraries
import pandas as pd
import numpy as np


ds_jobs = pd.read_csv("customer_train.csv")

ds_jobs.head()
```

Out[212…]

	student_id	city	city_development_index	gender	relevant_experience	enrolled_univ
0	8949	city_103	0.920	Male	Has relevant experience	no_enrol
1	29725	city_40	0.776	Male	No relevant experience	no_enrol
2	11561	city_21	0.624	NaN	No relevant experience	Full time c
3	33241	city_115	0.789	NaN	No relevant experience	
4	666	city_162	0.767	Male	Has relevant experience	no_enrol

In [213…]
```
# Create a copy of ds_jobs for transforming
ds_jobs_transformed = ds_jobs.copy()
```

In [214…]
```
display(ds_jobs_transformed.dtypes)
```

```
student_id                int64
city                     object
city_development_index    float64
gender                   object
relevant_experience      object
enrolled_university      object
education_level          object
major_discipline         object
experience               object
company_size             object
company_type             object
last_new_job             object
training_hours            int64
job_change              float64
dtype: object
```

In [215…]
```
#Transform columns containing categories with only two factors to Booleans
```

In [216…]
```
columns_with_2_factors = [column for column in ds_jobs_transformed.columns if len(d
```

In [217…]
```
display(columns_with_2_factors)
```

```
['relevant_experience', 'job_change']
```

In [219…
```
ds_jobs_transformed["relevant_experience"] = np.where(ds_jobs_transformed["relevant
```

In [220…
```
ds_jobs_transformed["relevant_experience"] = ds_jobs_transformed["relevant_experien
```

In [221…
```
display(ds_jobs_transformed["job_change"].value_counts())
```

```
0.0    14381
1.0     4777
Name: job_change, dtype: int64
```

In [222…
```
ds_jobs_transformed["job_change"] = ds_jobs_transformed["job_change"].astype("bool"
```

In [223…
```
#Columns containing integers only must be stored as 32-bit integers
```

In [224…
```
ds_jobs_transformed["student_id"] = ds_jobs_transformed["student_id"].astype("int32
```

In [225…
```
ds_jobs_transformed["training_hours"] = ds_jobs_transformed["training_hours"].astyp
```

In [226…
```
#Columns containing floats must be stored as 16-bit floats
```

In [227…
```
ds_jobs_transformed["city_development_index"] = ds_jobs_transformed["city_developme
```

In [228…
```
#Columns containing nominal categorical data must be stored as the category data ty
```

In [229…
```
nominals = ["city","gender","major_discipline", "company_type"]
for cat in nominals:
    ds_jobs_transformed[cat] = ds_jobs_transformed[cat].astype("category")
```

In [230…
```
#Columns containing ordinal categorical data must be stored as ordered categories,
```

In [231…
```
ordinals = ["education_level", "experience", "company_size", "last_new_job","enroll
```

In [233…
```
experience = ["<1"]
for x in range(1,21):
    experience.append(str(x))
experience.append(">20")
```

In [234…
```
last_new_job = ["never"]
for x in range(1,5):
    last_new_job.append(str(x))
last_new_job.append(">4")
```

In [235…
```
categories_by_columns = {
    "education_level":["Primary School","High School","Graduate","Masters","Phd"],
    "experience":experience,
    "company_size":["<10","10-49","50-99","100-499","500-999","1000-4999","5000-999
    "last_new_job":last_new_job,
    "enrolled_university":["no_enrollment","Part time course","Full time course"]
}
```

In [236…
```
for column,cats in categories_by_columns.items():
    print(f"changing {column}")
    ds_jobs_transformed[column] = ds_jobs_transformed[column].astype("category")
    ds_jobs_transformed[column].cat.reorder_categories(
        new_categories = cats,
        ordered = True,
        inplace = True
    )
```

```
changing education_level
changing experience
changing company_size
changing last_new_job
changing enrolled_university
```

In [238…
```
ds_jobs_transformed = ds_jobs_transformed[(ds_jobs_transformed["experience"].isin(c
```

In [239…
```
ds_jobs.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
```

In [240…
```
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: 69.5 KB
```