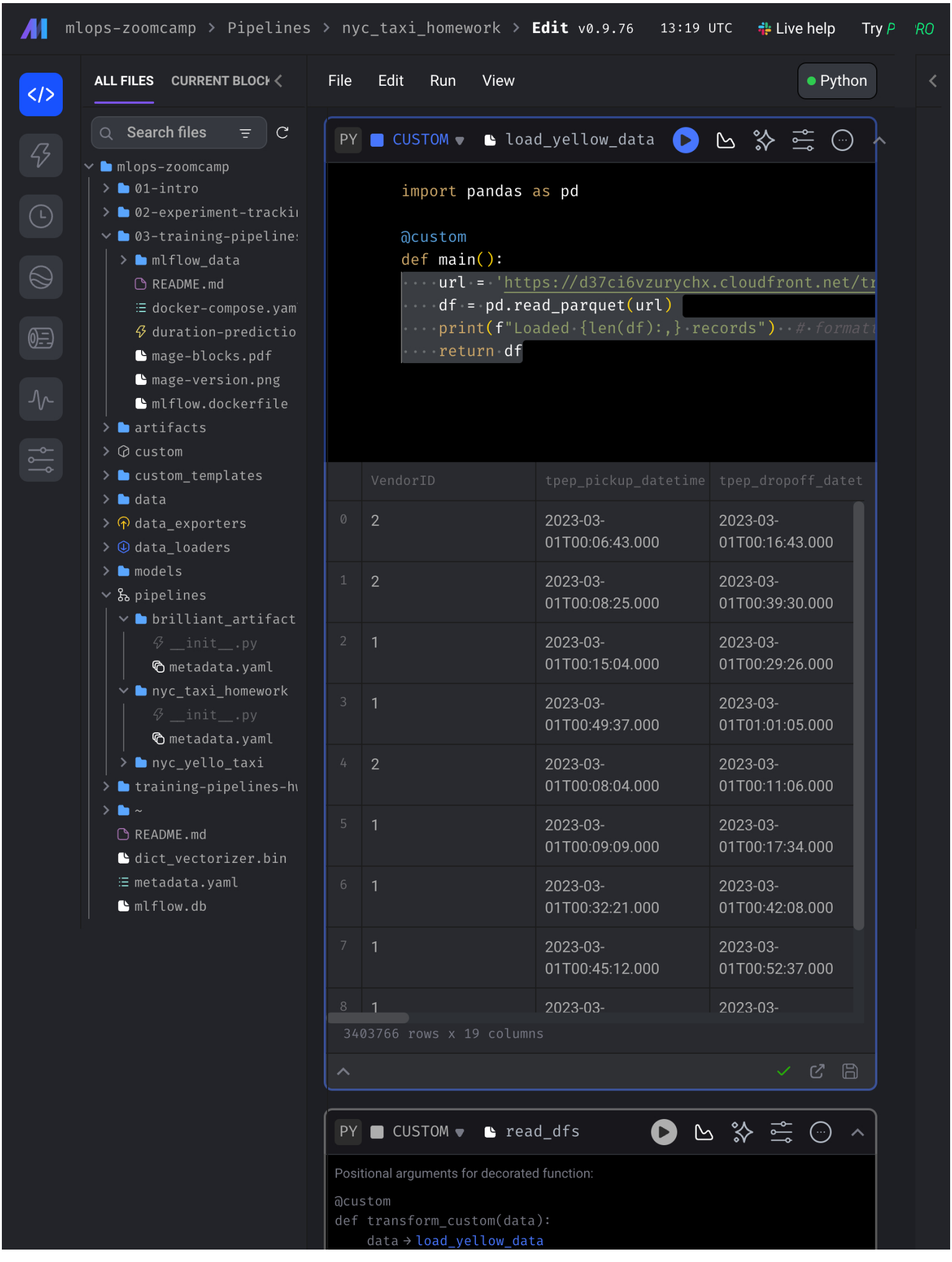

mlops-zoomcamp > Pipelines > nyc_taxi_homework > Edit v0.9.76 13:19 UTC Live help
ALL FILES CURRENT BLOCK
File Edit Run View
Python
Search files
mlops-zoomcamp
01-intro
02-experiment-tracki
03-training-pipeline
mlflow_data
README.md
docker-compose.yam
duration-predictio
mage-blocks.pdf
mage-version.png
mlflow.dockerfile
artifacts
custom
custom_templates
data
data_exporters
data_loaders
models
pipelines
brilliant_artifact
__init__.py
metadata.yaml
nyc_taxi_homework
__init__.py
metadata.yaml
nyc_yello_taxi
training-pipelines-h
~
README.md
dict_vectorizer.bin
metadata.yaml
mlflow.db
PY CUSTOM load_yellow_data
import pandas as pd
@custom
def main():
url = 'https://d37ci6vzurychx.cloudfront.net/tr
df = pd.read_parquet(url)
print(f"Loaded {len(df):,} records")
return df
VendorID
tpep_pickup_datetime
tpep_dropoff_datet
0 2 2023-03-01T00:06:43.000 2023-03-01T00:16:43.000
1 2 2023-03-01T00:08:25.000 2023-03-01T00:39:30.000
2 1 2023-03-01T00:15:04.000 2023-03-01T00:29:26.000
3 1 2023-03-01T00:49:37.000 2023-03-01T01:01:05.000
4 2 2023-03-01T00:08:04.000 2023-03-01T00:11:06.000
5 1 2023-03-01T00:09:09.000 2023-03-01T00:17:34.000
6 1 2023-03-01T00:32:21.000 2023-03-01T00:42:08.000
7 1 2023-03-01T00:45:12.000 2023-03-01T00:52:37.000
8 1 2023-03-
3403766 rows x 19 columns
PY CUSTOM read_dfs
Positional arguments for decorated function:
@custom
def transform_custom(data):
data → load_yellow_data

```python
import pandas as pd
import pickle

@custom
def main(df):
    df.tpep_dropoff_datetime = pd.to_datetime(df.tp
    df.tpep_pickup_datetime = pd.to_datetime(df.tpe

    df['duration'] = df.tpep_dropoff_datetime - df.
    df['duration'] = df.duration.dt.total_seconds()

    df = df[(df.duration >= 1) & (df.duration <= 60

    categorical = ['PULocationID', 'DOLocationID']
    df[categorical] = df[categorical].astype(str)

    print(f"Filtered dataset shape: {df.shape}")
    return df
```

	VendorID	tpep_pickup_datetime	tpep_dropoff_datet
0	2	2023-03-01T00:06:43.000	2023-03-01T00:16:43.000
1	2	2023-03-01T00:08:25.000	2023-03-01T00:39:30.000
2	1	2023-03-01T00:15:04.000	2023-03-01T00:29:26.000
3	1	2023-03-01T00:49:37.000	2023-03-01T01:01:05.000
4	2	2023-03-01T00:08:04.000	2023-03-01T00:11:06.000
5	1	2023-03-01T00:09:09.000	2023-03-01T00:17:34.000
6	1	2023-03-01T00:32:21.000	2023-03-01T00:42:08.000
7	1	2023-03-01T00:45:12.000	2023-03-01T00:52:37.000
8	1	2023-03-	2023-03-

3316216 rows x 20 columns

PY CUSTOM train_linear_model

Positional arguments for decorated function:

```
@custom
def transform_custom(data):
    data → read_dfs
```

```
from sklearn.feature_extraction import DictVectoriz
from sklearn.linear_model import LinearRegression

@custom
def train_linear_model(df):
    df['PULocationID'] = df['PULocationID'].astype(
    df['DOLocationID'] = df['DOLocationID'].astype(

    # Prepare input features
    categorical = ['PULocationID', 'DOLocationID']
    dicts = df[categorical].to_dict(orient='records

    # Fit DictVectorizer
    dv = DictVectorizer()
    X = dv.fit_transform(dicts)

    # Target variable
    y = df['duration'].values

    # Train Linear Regression
    model = LinearRegression()
    model.fit(X, y)

    print(f"Intercept: {model.intercept_:.2f}")
    print(f"Vectorizer: {dv}")
    print(f"Model: {model}")

    return dv, model
```

OUTPUT 0 OUTPUT 1

▾ DictVectorizer

```
DictVectorizer()
```

PY DATA EXPORTER save

Positional arguments for decorated function:

```
@data_exporter
def export_data(data):
    data → train_linear_model
```

```
from pathlib import Path
```

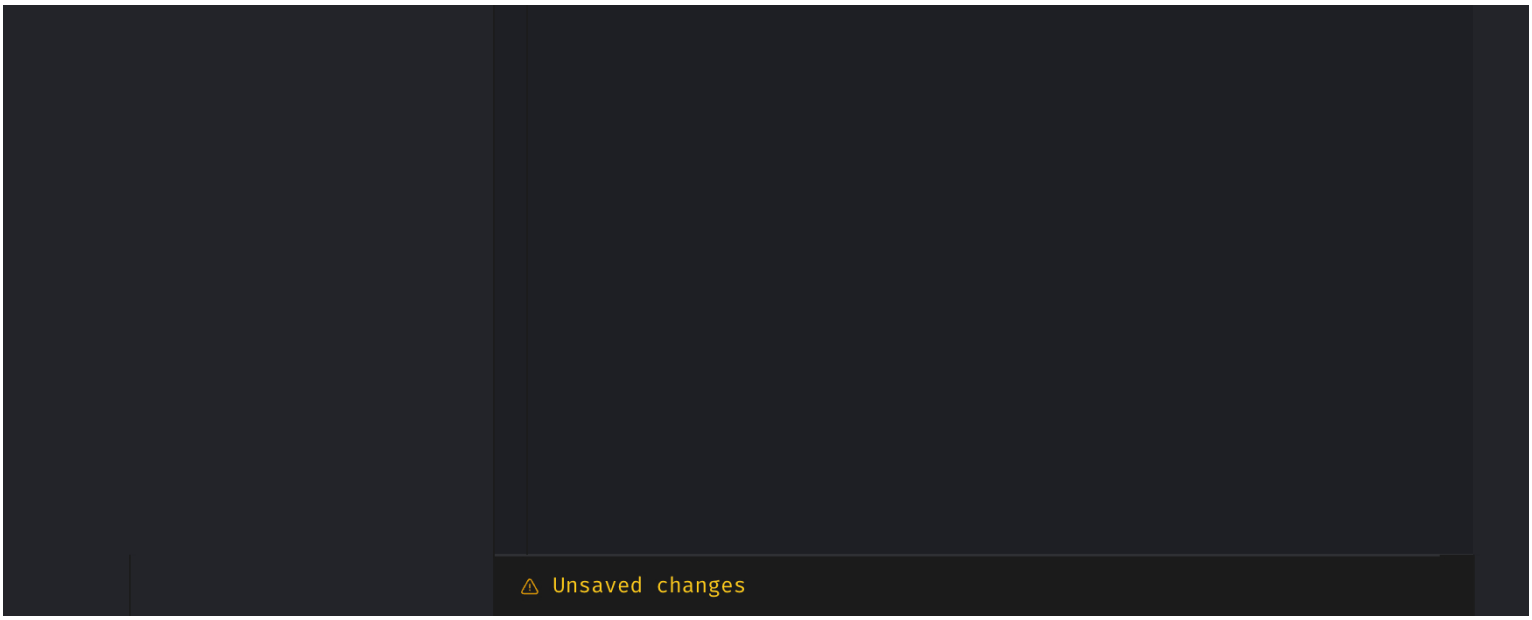
```
import pickle
import mlflow


mlflow.set_tracking_uri("http://localhost:5054")
mlflow.set_experiment("nyc-taxi-experiment")

models_folder = Path('models')
models_folder.mkdir(exist_ok=True)

if 'data_exporter' not in globals():
    from mage_ai.data_preparation.decorators import


@data_exporter
def export_data(data, *args, **kwargs):
    """
    Exports data to some source.

    Args:
        data: The output from the upstream parent b
        args: The output from any additional upstre

    Output (optional):
        Optionally return any object and it'll be l
        displayed when inspecting the block run.
    """
    dv, lr = data
    with mlflow.start_run():
        with open('dict_vectorizer.bin', 'wb') as f
            pickle.dump(dv, f_out)
        mlflow.log_artifact('dict_vectorizer.bin')

        mlflow.sklearn.log_model(lr, 'model')

    print('OK')
```

```
WARNING:urllib3.connectionpool:Retrying (Retry(total=
6, connect=6, read=7, redirect=7, status=7)) after con
nection broken by 'NewConnectionError('<urllib3.connec
tion.HTTPConnection object at 0×161ffa570>: Failed to
establish a new connection: [Errno 61] Connection refu
sed')': /api/2.0/mlflow/experiments/get-by-name?experi
ment_name=nyc-taxi-experiment
WARNING:urllib3.connectionpool:Retrying (Retry(total=
5, connect=5, read=7, redirect=7, status=7)) after con
nection broken by 'NewConnectionError('<urllib3.connec
tion.HTTPConnection object at 0×34a87bd10>: Failed to
```

```
establish a new connection: [Errno 61] Connection refu
sed')': /api/2.0/mlflow/experiments/get-by-name?experi
ment_name=nyc-taxi-experiment
WARNING:urllib3.connectionpool:Retrying (Retry(total=
4, connect=4, read=7, redirect=7, status=7)) after con
nection broken by 'NewConnectionError('<urllib3.connec
tion.HTTPConnection object at 0×34a87a090>: Failed to
establish a new connection: [Errno 61] Connection refu
sed')': /api/2.0/mlflow/experiments/get-by-name?experi
ment_name=nyc-taxi-experiment
WARNING:urllib3.connectionpool:Retrying (Retry(total=
3, connect=3, read=7, redirect=7, status=7)) after con
nection broken by 'NewConnectionError('<urllib3.connec
tion.HTTPConnection object at 0×34a754410>: Failed to
establish a new connection: [Errno 61] Connection refu
sed')': /api/2.0/mlflow/experiments/get-by-name?experi
ment_name=nyc-taxi-experiment
WARNING:urllib3.connectionpool:Retrying (Retry(total=
2, connect=2, read=7, redirect=7, status=7)) after con
nection broken by 'NewConnectionError('<urllib3.connec
tion.HTTPConnection object at 0×34b1a5ee0>: Failed to
establish a new connection: [Errno 61] Connection refu
sed')': /api/2.0/mlflow/experiments/get-by-name?experi
ment_name=nyc-taxi-experiment
2025/06/04 16:19:08 INFO mlflow.tracking.fluent: Exper
iment with name 'nyc-taxi-experiment' does not exist.
Creating a new experiment.
2025/06/04 16:19:21 WARNING mlflow.models.model: Model
logged without a signature and input example. Please s
et `input_example` parameter when logging the model to
auto infer the model signature.
🏃 View run shivering-mink-679 at: http://localhost:505
4/#/experiments/1/runs/177594a22c3e4438988084e42185a61
5
🧪 View experiment at: http://localhost:5054/#/experime
nts/1
OK
```

68.654s

All blocks Custom Markdown Search for ⌘ + /


⚠ Unsaved changes