SORU

https://github.com/erkansirin78/datasets/raw/master/dirty_store_transactions.csv veri setini kullanarak aşağıdaki işleri yapan bir python uygulamasını jenkins ve ci_cd kullanarak deploy ediniz.

1. veriyi temizlesin ve yapısal hale getirsin.
2. Temiz veriyi /tmp/clean_store_transactions.csv olarak yazsın.
3. Python uygulaması günlük olarak çalışmalıdır.

CEVAP

Veri setini indiriniz

```
wget -O /tmp/dirty_store_transactions.csv \
https://github.com/erkansirin78/datasets/raw/master/dirty_store_transactions.csv
```

Veri temizliği yapan python uygulaması

- /tmp/02_dirty_data_clean_answer.py

```
import pandas as pd
import re

df = pd.read_csv("/tmp/dirty_store_transactions.csv")

def clean_store_location(st_loc):
    return re.sub(r'[^\w\s]', '', st_loc).strip()

def clean_product_id(pd_id):
    matches = re.findall(r'\d+', pd_id)
    if matches:
        return matches[0]
    return pd_id

def remove_dollar(amount):
    return float(amount.replace('$', ''))

df['STORE_LOCATION'] = df['STORE_LOCATION'].map(lambda x: clean_store_location(x))
df['PRODUCT_ID'] = df['PRODUCT_ID'].map(lambda x: clean_product_id(x))

for to_clean in ['MRP', 'CP', 'DISCOUNT', 'SP']:
    df[to_clean] = df[to_clean].map(lambda x: remove_dollar(x))

df.to_csv('/tmp/clean_store_transactions.csv', index=False)
```

Jenkins projesi oluşturunuz

- New Item -> Freestyle project -> Proje ismi girin
- Build Triggers -> Build periodically -> H * * * *
- Build -> Execute shell -> Command

```
pip3 install pandas && python3 /tmp/02_dirty_data_clean_answer.py
```

Kontrol

```
(venvspark) [train@localhost spark-dirty-data]$ ll /tmp/ | grep jenkins
-rw-r--r--. 1 jenkins jenkins 2501063 Feb 19 23:59 clean_store_transactions.csv
```