JDS Final Task - Dilan's Travel Guide - PART 1: Data downloading and preparation

1. **BASH**

1.1 I download the dilans_data.csv dataset to my remote server, I check the number of lines and the look of the dataset then I put it into a new directory named 'dilan'.

```
wget ........../dilans_data.csv
ls -v
cat dilans_data.csv|head
cat dilans_data.csv|tail
cat dilans_data.csv|wc -l (number of lines: 597 902)
mkdir dilan
cp dilans_data.csv dilan
cd dilan
cd ..
rm dilans_data.csv
```

1.2 I create three smaller datasets from the original dilans_data.csv, according to the event type: read, subscribe and buy. I check that I have all the lines.

```
grep "read" dilans_data.csv > dilans_read
grep "subscribe" dilans_data.csv > dilans_subs
grep "buy" dilans_data.csv > dilans_buy
ls -v
cd dilans_read
cat dilans_read|head
cat dilans_read|wc -l #(number of lines: 581 877)
cat dilans_subs|head
cat dilans_subs|wc -l #(number of lines: 7 618)
cat dilans_buy|head
cat dilans_buy|wc -l #(number of lines: 8 407)
```

1.3 I divide dilans_read dataset into first readers (first_read) and returned readers (ret_read). Again I check the number of lines in all files.

```
grep "AdWords" dilans_read > dilans_adwords
grep "Reddit" dilans_read > dilans_reddit
grep "SEO" dilans_read > dilans_seo
ls -v
cat dilans_adword| head
cat dilans_adwords|wc -l (number of lines: 63 065)
cat dilans_reddit| head
cat dilans_reddit|wc -l (number of lines: 105 216)
cat dilans_seo| head
cat dilans_seo|wc -l (number of lines: 41 742)
cat dilans_adwords dilans_seo dilans_reddit > dilans_first_read
```

```
ls -v
cat dilans_first_read|head
cat dilans_first_read|tail
cat dilans_first_read|wc -l (number of lines: 210 023)
grep -v "AdWords\|Reddit\|SEO" dilans_read > dilans_ret_read
ls -v
cat dilans_ret_read|head
cat dilans_ret_read|wc -l (number of lines: 371 854)
```

2. **SQL**

2.1 I give myself useruser privileges to create new tables.

```
sudo -u postgres -i
psql
ALTER USER dorisimon WITH SUPERUSER;
\q
exit
```

2.2 I create the new tables and I fill them with data. Then I check that I still have all the lines in the new tables.

```
psql -U dorisimon -d postgres

CREATE TABLE first_read (
my_time TIMESTAMP,
event_type TEXT,
country TEXT,
user_id TEXT,
source TEXT,
topic TEXT);

CREATE TABLE ret_read (
my_time TIMESTAMP,
event_type TEXT,
country TEXT,
user_id TEXT,
topic TEXT);

CREATE TABLE adwords (
my_time TIMESTAMP,
event_type TEXT,
country TEXT,
user_id TEXT,
source TEXT,
topic TEXT);

CREATE TABLE reddit (
my_time TIMESTAMP,
```

```
event_type TEXT,
country TEXT,
user_id TEXT,
source TEXT,
topic TEXT);

CREATE TABLE seo (
my_time TIMESTAMP,
event_type TEXT,
country TEXT,
user_id TEXT,
source TEXT,
topic TEXT);

CREATE TABLE subs (
my_time TIMESTAMP,
event_type TEXT,
user_id TEXT);

CREATE TABLE buy (
my_time TIMESTAMP,
event_type TEXT,
user_id TEXT,
price INTEGER);

COPY first_read FROM '/home/dorisimon/dilan/dilans_first_read' DELIMITER ';';
COPY ret_read FROM '/home/dorisimon/dilan/dilans_ret_read' DELIMITER ';';
COPY adwords FROM '/home/dorisimon/dilan/dilans_adwords' DELIMITER ';';
COPY reddit FROM '/home/dorisimon/dilan/dilans_reddit' DELIMITER ';';
COPY seo FROM '/home/dorisimon/dilan/dilans_seo' DELIMITER ';';
COPY subs FROM '/home/dorisimon/dilan/dilans_subs' DELIMITER ';';
COPY buy FROM '/home/dorisimon/dilan/dilans_buy' DELIMITER ';';

cat dilans_adwords|wc -l
SELECT COUNT(*) FROM adwords;
cat dilans_buy|wc -l
SELECT COUNT(*) FROM buy;
cat dilans_first_read|wc -l
SELECT COUNT(*) FROM firts_read;
cat dilans_reddit|wc -l
SELECT COUNT(*) FROM reddit;
cat dilans_ret_read|wc -l
SELECT COUNT(*) FROM ret_read;
cat dilans_seo|wc -l
SELECT COUNT(*) FROM seo;
cat dilans_subs|wc -l
SELECT COUNT(*) FROM subs;
```