business_up_top("etl_project")

Group: business_on_top()
Proposal: ETL on Store Data and Dow Data
Transformation Type: (see below)
DataBase Type: Relational

Extract

Data Source: Both files came from separate sources on Kaggle
Data Format: .csv

Transform

1. Convert dates in excel to a consistent format
   - Converted the order dates for the store data in the format of MM/DD/YYYY. Some of the entries had been erroneously entered as DD/MM/YYYY. Used Excel functions to clean up the incorrect dates.
2. Read in both CSV files into JN separately
3. Dow Jones Data Clean Up:
   - Use str.lower() to rename all column names to lowercase to match the column names in PGAdmin
   - Rename column name "vol." to 'vol(m)" and "change" to "change_percent" (in the next step we remove "M" and the "%" sign so we put those details in the column names)
   - Use .replace() to remove commas, the letter M from volume column, and the percent sign. These were removed convert all numbers to float
   - Convert all numbers to float (except for date)
4. Store Data Clean Up:
   - Use str.lower() to rename all column names to lowercase to match the column names in PGAdmin and use .replace() to replace "-" and " " with "_" to match the column names in PGAdmin
   - Select only the columns needed for analysis ['order_id', 'order_date', 'city', 'state', 'postal_code', 'region', 'product_id', 'category', 'sub_category', 'product_name', 'sales']
5. Convert all date(s) in both files to Pandas DateTime for consistency
   - Used the to_datetime to convert from the existing date format to a consistent format for the two different files. The store data had the format of %m/%d/%Y and the DJIA data had the format of %b %d,%Y
6. Select only the columns needed
   - Stripped only the columns needed for the store data into a new Dataframe
7. Set up schema in pgAdmin
   - Created the two tables in a etl_db with SQL create functions.
8. Connect PGAdmin to Jupyter notebook
   - Dependencies to connect: **from** sqlalchemy **import** create_engine
   - Connect JN and PGAdmin with password, postgres, and localhost
   - Print out the table names to test connection

- Use python to load CSV data from dataframes into the table
- Use python to read tables and test load

9. Join tables in PGAdmin
    - on store.order_date and dow.date. Created queries to look at store.sales versus dow.vol or dow.change.
    - Round sum to two decimal places using ROUND(,2) in SQL

Load

Final DB

etl_db

DB Tables

```
CREATE TABLE dow (
        id serial PRIMARY KEY,
        date date,
        price decimal,
        open decimal,
        high decimal,
        low decimal,
        vol_m decimal,
        change_percent decimal
);
```

```
CREATE TABLE store (
        id serial PRIMARY KEY,
        order_id varchar(250),
        order_date date,
        ship_date date,
        city varchar (250),
        state varchar(250),
        postal_code   varchar(250),
        region varchar(250),
        product_id varchar (250),
        category varchar(250),
        sub_category varchar(250),
        product_name varchar(250),
        sales decimal
);
```

Why

The purpose of this DB is compare daily sales orders against daily DJIA prices to measure the correlation between the 2 entities