Homework 4 - Launch Plan for Retail Startup

Lily Annen

Chapman University

Argyros School of Business and Economics

MGSC 410-01, Applied Business Analytics

Analysis Plan

Goal: create an analytical plan for the launch of targeted retail locations to determine the best 10 markets to compete effectively against an established competitor.

Action List of Steps

1. **Open the datasets**: from google drive folder provided, first use excel to open the data and determine data quality and structure.
2. **Data exploration and set-up**: look at all datasets, explain the variables, and account for any data quality/concerns.
3. **Clean the data**: rename file names and columns for consistency, count missing/null values, drop columns, decide to drop rows if needed.
4. **Merge datasets**: merge features, sales, stores_DMA based on similar columns (Store(#), Date, IsHoliday).
5. **Break dataset down by DMA**: once features, sales, and stores_DMA are merged, break store information down further by DMA location, or city.
6. **Export to Drive**: export each location dataframe as a csv file, download, and open in tableau for modeling and visualizations.
7. **Generate list of questions**: what do I want to show through my analysis? what ideas for models and which variables to use to tell a story?
8. **Create visualizations**: how do I best use modeling and visualizations to answer questions about the data for a launch plan? Plan out what graphs to use and how to encorporate visualizations on easy to interpret dashboards.
9. **Explain and interpret findings**: make sure visualizations are clear, organized, consistent, double check for data quality errors, and put final analysis into words relating to the graphs and charts.
10. **Seek out other sources**: take the time to identify other sources relevant to my recommendation plan. Dive into DMA_dashboard dataset and region datasets.
11. **Overall analysis and explanation**: ~~what do I want to show through my analysis? what ideas for models and which variables to use to tell a story?~~
12. **Retail Launch Recommendation**: export visualizations from dashboards on tableau, write a detailed retail launch recommendation for the startup and include models and graphs.

Data Exploration

Goal: conduct analysis on provided datasets, seek out and analyze additional relevant datasets that can help determine launch strategy.

DMA Information

A demographic dataset at the DMA (Designated Market Area aka major city or metropolitan area). You can use this data to analysis the population, household, income and ethnicity of the various DMA's as potential locations.

MSA Information

MSA (Metropolitan Statistical Areas) are similar in nature to DMA's (note; they are not 1 to 1 likeness, but the purposes of this analysis they should be assumed to be the same). These files provide input into consumer expenditure in major categories.

Retail Data Information

Historical sales data for 45 competitive stores located in different major cites, some major markets have more than one store. Each store contains a number of departments.

Stores w/ DMA Dataset

This file contains information about the 45 stores, indicating the type and size of store and location. Within this file are the following fields:

- ***store***: the store number
- ***type***: A, B, or C
- ***size***: size of the city?
- ***DMA***: Designated Market Area aka major city or metropolitan area

Sales Dataset

This is the historical training data, which covers to 2010-02-05 to 2012-11-01. Within this file there are the following fields:

- ***store***: the store number
- ***dept***: the department number
- ***date***: the week
- ***weekly_Sales***: sales for the given department in the given store
- ***isHoliday***: whether the week is a special holiday week

Features Dataset

Additional data related to the store, department, and regional activity for the given dates. It contains the following fields:

- ***store***: the store number
- ***date***: the week
- ***temperature***: average temperature in the region
- ***fuel_price***: cost of fuel in the region
- ***markdown1-5***: anonymized data related to promotional markdowns that the competitor is running (only available after Nov 2011, and is not available for all stores all the time)
- ***CPI***: the consumer price index
- ***unemployment***: the unemployment rate
- ***isHoliday***: whether the week is a special holiday week

Additional Information

Four holidays fall within the following weeks in the dataset (not all holidays are in the data):

- ***Super Bowl***: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
- ***Labor Day***: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13
- ***Thanksgiving***: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13
- ***Christmas*** 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

The competitor runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted ***five times higher*** in the evaluation than non-holiday weeks.

An optional part of the challenge presented by this analysis is modeling the ***effects of markdowns*** on these holiday weeks in the absence of complete/ideal historical data to ensure that the impact of these markdowns is accounted for in the recommendation.

Data Assessment

The google drive folder contains 9 datasets. These include: Stores with DMA Dataset (2), Sales Dataset (.xlsx & .csv), Features Dataset (.xlsx & .csv), PI-18803 DMAs, midwest, northeast, south, and west. There is a mix between Excel files (.xlsx) and CSV files (.csv). All of the datasets seem to be unique at first glace, but contain some of the same information and are somewhat unorganized and messy.

Stores with DMA

I started with the Stores with DMA dataset first, this was the smallest dataset with 45 rows and 9 columns. Originally, many columns did not have headers or descriptions. One column had only "?" characters, so I discarded it. I figured out that the unlabeled values were the count of the number of stores per city (Atlanta-Tampa), and the computed values were the overall store size divided by the number of stores. I added labels to these columns and restructured the data to make more sense visually. I added the 'outlier' columns to a new file, ***Stores with City Ratios***. The city with the most stores is Los Angeles (7 stores), followed by Atlanta, Cleveland, Dallas, and Houston (4 stores), Austin, Chicago, and Denver (3 stores), Charlotte, Kansas City, Philadelphia, Salt Lake City, San Diego, Tampa (2 stores), and Orlando (1 store). There were obvious spelling errors as well, so I fixed those. Additionally, many DMA city names were lengthy (for example 'Orlando-Daytona Brach-Melbourne FL') so I shortened those to single words ('Orlando') for simplicity. A second 'Stores with DMA' dataset was included, but contained repeat information about the store number, type, size, and DMA as the prior dataset. I discarded this data for my analysis. I also recorded the number of null values per variable in this dataset, which was 0.

Sales

Next, I moved on to the Sales dataset. This data looked more clean, and included 421,571 rows and 5 columns. I didn't spot any missing values, and all columns looked to be formatted well. I ran code again to calculate the number of null values per variable in this dataset, which was 0.

Features

The Features dataset included 8,191 rows and 12 columns. I noticed some overlap in variables compared to the Sales dataset, which were Store (number), Date (YYYY-MM-DD), and IsHoliday (T/F). The markdown 1-5 variable included a large amount of 'NA' values, so another data cleaning step could be to get rid of any rows that are 'NA', or discard the variables altogether.

DMA Dashboards

Formally named PI-18803 DMAs, I renamed this file to DMA Dashboards. This included three excel sheets: 18801-Dashboard, 18802-Dashboard, and DMAs. The 18801 sheet is comprised of a dashboard with a list of DMAs and four different figures. These figures tell information about vehicle ownership, average household expense overall, on mass transit, on taxi, and on public transportation, broken down by DMA. The 18802 sheet details the population composition by some numerical value (unknown) and race, again broken down by DMA. The last sheet included 212 rows and 26 columns, and looks to be all the data utilized by the dashboards. Of these features, the most notable for the analysis are population 18+, median household income, households with no vehicles, households with 1-2 vehicles, households with 2+ vehicles, and household breakdowns of race and expenses on public transportation.

Midwest, Northeast, South, West

These four datasets are comprised of 30+ rows and 5-10 columns (depending on the region). The data details the selected region's metropolitan statistical areas, average annual expenditures and characteristics from the Consumer Expenditure Survey (2016-2017). These files were much different in structure than the rest, and appear to have more of a table view or various population characteristics for each city.

Data Cleaning

Import Packages

```
# import necessary packages
import warnings
warnings.filterwarnings('ignore')

import pandas as pd
import numpy as np

from plotnine import *

%matplotlib inline

from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

Stores_DMA

```
# load stores with DMA file
stores_DMA = pd.read_excel('/content/drive/My Drive/Y4/Colab
Notebooks/aba/Stores w DMA.xlsx')
stores_DMA.head()
```

```
   Store Type    Size     DMA
0     24    A  203819  Atlanta
1     31    A  203750  Atlanta
2     44    C   39910  Atlanta
3     33    A   39690  Atlanta
4     35    B  103681   Austin
```

```
# check stores_DMA for missing values
i = 0
print("# of null values (stores_DMA),", "size:", stores_DMA.shape)

for column in stores_DMA:
  print(" ", column, " :", stores_DMA[column].isna().sum())
  i+=1

  if i == 4:
    break
```

```
# of null values (stores_DMA), size: (45, 4)
  Store  : 0
  Type  : 0
  Size  : 0
  DMA   : 0
```

```
# check column type
stores_DMA.dtypes
```

```
Store     int64
Type     object
Size      int64
DMA      object
dtype: object
```

Stores_City

```
# load stores with city ratio file
stores_city = pd.read_excel('/content/drive/My Drive/Y4/Colab
Notebooks/aba/Stores w city ratios.xlsx')
stores_city.head()
```

```
        City  sum of size  num of stores  size / num of stores
0    Atlanta       487169              4         121792.250000
1     Austin       178466              3          59488.666667
2  Charlotte       248271              2         124135.500000
3    Chicago       389985              3         129995.000000
4  Cleveland       557646              4         139411.500000
```

```
# check stores_city for missing values
i = 0
print("# of null values (stores_city),", "size:", stores_city.shape)

for column in stores_city:
  print(" ", column, " :", stores_city[column].isna().sum())
  i+=1

  if i == 4:
    break
```

```
# of null values (stores_city), size: (15, 4)
  City  : 0
  sum of size  : 0
  num of stores  : 0
  size / num of stores  : 0
```

```
# check column type
stores_city.dtypes
```

```
City                    object
sum of size              int64
num of stores            int64
```

```
size / num of stores    float64
dtype: object
```

Sales

```
# load sales excel file
sales = pd.read_excel('/content/drive/My Drive/Y4/Colab
Notebooks/aba/Sales.xlsx')
sales.head()
```

```
   Store  Dept       Date Weekly_Sales  IsHoliday
0    1.0   1.0 2010-02-05      24924.5      False
1    1.0   1.0 2010-02-12     46039.49       True
2    1.0   1.0 2010-02-19     41595.55      False
3    1.0   1.0 2010-02-26     19403.54      False
4    1.0   1.0 2010-03-05      21827.9      False
```

```
# check sales for missing values
i = 0
print("# of null values (sales),", "size:", sales.shape)

for column in sales:
  print(" ", column, " :", sales[column].isna().sum())
  i+=1

  if i == 5:
    break
```

```
# of null values (sales), size: (421570, 5)
  Store  : 0
  Dept  : 0
  Date  : 0
  Weekly_Sales  : 0
  IsHoliday  : 0
```

```
# check column types
sales.dtypes
```

```
Store                  float64
Dept                   float64
Date            datetime64[ns]
Weekly_Sales            object
IsHoliday                 bool
dtype: object
```

```
# change 'Weekly_Sales' to numeric
sales["Weekly_Sales"] = sales["Weekly_Sales"].apply(lambda x:
pd.to_numeric(x, errors='coerce')).dropna()

# check column types again
sales.dtypes
```

```
Store                  float64
Dept                   float64
Date            datetime64[ns]
Weekly_Sales           float64
IsHoliday                 bool
dtype: object
```

```
# change 'Store' and 'Dept' to int
sales = sales.astype({"Store": int})
sales = sales.astype({"Dept": int})
```

Features

```
# load features excel file
features = pd.read_excel('/content/drive/My Drive/Y4/Colab
Notebooks/aba/Features.xlsx')
features.head()
```

```
   Store       Date Temperature Fuel_Price MarkDown1 MarkDown2
MarkDown3  \
0      1 2010-02-05       42.31       2572       NaN       NaN
NaN
1      1 2010-02-12       38.51       2548       NaN       NaN
NaN
2      1 2010-02-19       39.93       2514       NaN       NaN
NaN
3      1 2010-02-26       46.63       2561       NaN       NaN
NaN
4      1 2010-03-05        46.5       2625       NaN       NaN
NaN

  MarkDown4 MarkDown5           CPI Unemployment  IsHoliday
0       NaN       NaN  2.110964e+09         8106      False
1       NaN       NaN  2.112422e+09         8106       True
2       NaN       NaN  2.112891e+09         8106      False
3       NaN       NaN  2.113196e+09         8106      False
4       NaN       NaN  2.113501e+09         8106      False
```

```
# check features for missing values
i = 0
print("# of null values (features),", "size:", features.shape)

for column in features:
  print(" ", column, " :", features[column].isna().sum())
  i+=1

  if i == 12:
    break
```

```
# of null values (features), size: (8190, 12)
  Store  : 0
```

```
  Date  : 0
  Temperature  : 0
  Fuel_Price  : 0
  MarkDown1  : 4158
  MarkDown2  : 5269
  MarkDown3  : 4577
  MarkDown4  : 4726
  MarkDown5  : 4140
  CPI  : 585
  Unemployment  : 585
  IsHoliday  : 0
```

```
# check column types
features.dtypes
```

```
Store                    int64
Date            datetime64[ns]
Temperature             object
Fuel_Price              object
MarkDown1               object
MarkDown2               object
MarkDown3               object
MarkDown4               object
MarkDown5               object
CPI                    float64
Unemployment            object
IsHoliday                 bool
dtype: object
```

```
# change 'MarkDown1,2,3,4,5', 'temperature' to numeric
features["MarkDown1"] = features["MarkDown1"].apply(lambda x:
pd.to_numeric(x, errors='coerce')).dropna()
features["MarkDown2"] = features["MarkDown2"].apply(lambda x:
pd.to_numeric(x, errors='coerce')).dropna()
features["MarkDown3"] = features["MarkDown3"].apply(lambda x:
pd.to_numeric(x, errors='coerce')).dropna()
features["MarkDown4"] = features["MarkDown4"].apply(lambda x:
pd.to_numeric(x, errors='coerce')).dropna()
features["MarkDown5"] = features["MarkDown5"].apply(lambda x:
pd.to_numeric(x, errors='coerce')).dropna()
features["Temperature"] = features["Temperature"].apply(lambda x:
pd.to_numeric(x, errors='coerce')).dropna()
```

```
# check column type again
features.dtypes
```

```
Store                    int64
Date            datetime64[ns]
Temperature            float64
Fuel_Price              object
MarkDown1              float64
```

```
MarkDown2              float64
MarkDown3              float64
MarkDown4              float64
MarkDown5              float64
CPI                    float64
Unemployment            object
IsHoliday                 bool
dtype: object
```

```
# drop missing values?
# features.dropna(inplace=True)
# features =
features.dropna(subset=['MarkDown1','MarkDown2','MarkDown3','MarkDown4
','MarkDown5','CPI','Unemployment'])

# check size
# features.shape  # now 2069 - TOO LOW?
```

Merge Datasets

```
# compare size of features, sales, and stores_DMA
print("SIZE")                                  # KEY COLUMNS
print("  features:", features.shape)           # Store, Date, IsHoliday
print("  sales:", sales.shape)                 # Store, Date, IsHoliday
print("  stores_DMA:", stores_DMA.shape)       # Store
```

```
SIZE
  features: (8190, 12)
  sales: (421570, 5)
  stores_DMA: (45, 4)
```

Outer Merge (sales & features)

```
# create outer merge
outer_merged = pd.merge(
    sales, features, how="outer", on=["Store", "Date", "IsHoliday"]
    )

# check size
outer_merged.shape
```

```
(423325, 14)
```

```
# merge features and sales
features_sales0 = pd.DataFrame(outer_merged)
features_sales0.head()
```

```
   Store  Dept       Date  Weekly_Sales  IsHoliday  Temperature
Fuel_Price  \
0      1   1.0 2010-02-05      24924.50      False        42.31
2572
1      1   2.0 2010-02-05      50605.27      False        42.31
```

```
2572
2      1   3.0 2010-02-05      13740.12      False        42.31
2572
3      1   4.0 2010-02-05      39954.04      False        42.31
2572
4      1   5.0 2010-02-05      32229.38      False        42.31
2572

   MarkDown1  MarkDown2  MarkDown3  MarkDown4  MarkDown5           CPI
\
0        NaN        NaN        NaN        NaN        NaN  2.110964e+09

1        NaN        NaN        NaN        NaN        NaN  2.110964e+09

2        NaN        NaN        NaN        NaN        NaN  2.110964e+09

3        NaN        NaN        NaN        NaN        NaN  2.110964e+09

4        NaN        NaN        NaN        NaN        NaN  2.110964e+09


  Unemployment
0         8106
1         8106
2         8106
3         8106
4         8106
```

Inner Merge (sales & features)

```
# create inner merge
inner_merged = pd.merge(
    sales, features, on=["Store", "Date", "IsHoliday"]
    )

# check size
inner_merged.shape
```

```
(421570, 14)
```

```
# merge features and sales
features_sales = pd.DataFrame(inner_merged)
features_sales.head()
```

```
   Store  Dept       Date  Weekly_Sales  IsHoliday  Temperature
Fuel_Price  \
0      1     1 2010-02-05      24924.50      False        42.31
2572
1      1     2 2010-02-05      50605.27      False        42.31
2572
```

```
2      1     3 2010-02-05      13740.12      False        42.31
2572
3      1     4 2010-02-05      39954.04      False        42.31
2572
4      1     5 2010-02-05      32229.38      False        42.31
2572

   MarkDown1  MarkDown2  MarkDown3  MarkDown4  MarkDown5           CPI
\
0        NaN        NaN        NaN        NaN        NaN  2.110964e+09

1        NaN        NaN        NaN        NaN        NaN  2.110964e+09

2        NaN        NaN        NaN        NaN        NaN  2.110964e+09

3        NaN        NaN        NaN        NaN        NaN  2.110964e+09

4        NaN        NaN        NaN        NaN        NaN  2.110964e+09


  Unemployment
0         8106
1         8106
2         8106
3         8106
4         8106
```

Inner Merge (sales, features & stores_DMA)

```
# create inner merge
inner_merged2 = pd.merge(
    features_sales, stores_DMA, on=["Store"]
    )

# check size
inner_merged2.shape
```

```
(421570, 17)
```

```
# merge features_sales and stores_DMA
features_sales_stores = pd.DataFrame(inner_merged2)
features_sales_stores.head()
```

```
   Store  Dept       Date  Weekly_Sales  IsHoliday  Temperature
Fuel_Price  \
0      1     1 2010-02-05      24924.50      False        42.31
2572
1      1     2 2010-02-05      50605.27      False        42.31
2572
2      1     3 2010-02-05      13740.12      False        42.31
```

```
2572
3      1     4 2010-02-05      39954.04      False        42.31
2572
4      1     5 2010-02-05      32229.38      False        42.31
2572

   MarkDown1  MarkDown2  MarkDown3  MarkDown4  MarkDown5           CPI
\
0        NaN        NaN        NaN        NaN        NaN  2.110964e+09

1        NaN        NaN        NaN        NaN        NaN  2.110964e+09

2        NaN        NaN        NaN        NaN        NaN  2.110964e+09

3        NaN        NaN        NaN        NaN        NaN  2.110964e+09

4        NaN        NaN        NaN        NaN        NaN  2.110964e+09


  Unemployment Type    Size     DMA
0         8106    A  151315  Houston
1         8106    A  151315  Houston
2         8106    A  151315  Houston
3         8106    A  151315  Houston
4         8106    A  151315  Houston
```

```
# check features_sales_stores for missing values
i = 0
print("# of null values (features_sales_stores)")

for column in features_sales_stores:
  print(" ", column, " :", features_sales_stores[column].isna().sum())
  i+=1

  if i == 17:
    break
```

```
# of null values (features_sales_stores)
  Store  : 0
  Dept  : 0
  Date  : 0
  Weekly_Sales  : 30117
  IsHoliday  : 0
  Temperature  : 7197
  Fuel_Price  : 0
  MarkDown1  : 289301
  MarkDown2  : 319874
  MarkDown3  : 304146
  MarkDown4  : 297528
  MarkDown5  : 289330
```

```
  CPI  : 0
  Unemployment  : 0
  Type  : 0
  Size  : 0
  DMA   : 0
```

```
# remove NA for weekly_sales
features_sales_stores.dropna(inplace=True)
features_sales_stores =
features_sales_stores.dropna(subset=["Weekly_Sales"])

# remove markdown columns?
features_sales_stores0 =
features_sales_stores.drop(features_sales_stores.columns[[7,8,9,10,11]
], axis=1)
features_sales_stores0.head()
```

```
   Store  Dept       Date  Weekly_Sales  IsHoliday  Temperature
Fuel_Price  \
0      1     1 2010-02-05      24924.50      False        42.31
2572
1      1     2 2010-02-05      50605.27      False        42.31
2572
2      1     3 2010-02-05      13740.12      False        42.31
2572
3      1     4 2010-02-05      39954.04      False        42.31
2572
4      1     5 2010-02-05      32229.38      False        42.31
2572

            CPI Unemployment Type    Size     DMA
0  2.110964e+09         8106    A  151315  Houston
1  2.110964e+09         8106    A  151315  Houston
2  2.110964e+09         8106    A  151315  Houston
3  2.110964e+09         8106    A  151315  Houston
4  2.110964e+09         8106    A  151315  Houston
```

```
# remove NA for weekly_sales
features_sales_stores0.dropna(inplace=True)
features_sales_stores0 =
features_sales_stores0.dropna(subset=["Weekly_Sales"])
```

Group by DMA

```
features_sales_stores.columns =
features_sales_stores.columns.str.strip()
features_sales_stores.head()
```

```
      Store  Dept       Date  Weekly_Sales  IsHoliday  Temperature
Fuel_Price  \
6587      1     1 2011-11-11      18689.54      False        59.11
3297
```

```
6588      1     2 2011-11-11      44936.47      False        59.11
3297
6589      1     3 2011-11-11       9959.64      False        59.11
3297
6590      1     4 2011-11-11      36826.52      False        59.11
3297
6591      1     5 2011-11-11      31002.65      False        59.11
3297

      MarkDown1  MarkDown2  MarkDown3  MarkDown4  MarkDown5
CPI  \
6587    10382.9    6115.67     215.07    2406.62    6551.42
2.179981e+09
6588    10382.9    6115.67     215.07    2406.62    6551.42
2.179981e+09
6589    10382.9    6115.67     215.07    2406.62    6551.42
2.179981e+09
6590    10382.9    6115.67     215.07    2406.62    6551.42
2.179981e+09
6591    10382.9    6115.67     215.07    2406.62    6551.42
2.179981e+09

     Unemployment Type    Size      DMA
6587         7866    A  151315  Houston
6588         7866    A  151315  Houston
6589         7866    A  151315  Houston
6590         7866    A  151315  Houston
6591         7866    A  151315  Houston
```

```
dma = features_sales_stores.groupby('DMA')
dma.first()
```

```
                Store  Dept       Date  Weekly_Sales  IsHoliday
Temperature  \
DMA

Atlanta            24     1 2011-11-11      13578.93      False
46.78
Austin              5     2 2011-11-18      11568.77      False
64.33
Charlotte          17     1 2011-11-11      17072.88      False
27.61
Chicago            12     1 2011-11-11      13386.97      False
48.76
Cleveland           2     1 2011-11-18      23923.11      False
62.01
Dallas              9     1 2011-12-23      26975.98      False
44.43
Denver              3     1 2011-11-11       6525.18      False
61.70
```

```
Houston             1     1 2011-11-11      18689.54      False
59.11
Kansas City        25     1 2011-11-11      16508.85      False
44.81
Los Angeles        11     1 2011-11-25      19008.41       True
70.03
Oralando            6     1 2011-11-11      17870.11      False
61.33
Philadelphia       19     1 2011-11-11      18790.30      False
48.22
Salt Lake City     21     1 2011-11-18      12800.42      False
61.90
San Diego           8     1 2011-11-18      11044.58      False
51.72
Tampa              20     1 2011-11-18      34233.84      False
49.30

               Fuel_Price  MarkDown1  MarkDown2  MarkDown3  MarkDown4
\
DMA

Atlanta              3719    9391.17   20595.49     686.11    5950.16

Austin               3308     952.21     116.11      95.01     295.13

Charlotte            3513    3935.00    4172.99     225.67    2515.61

Chicago              3824   18049.87    7939.68     234.64   10463.07

Cleveland            3308    6490.92    1217.76     152.12     873.82

Dallas               3112     719.14       0.24     318.75      17.72

Denver               3297    8787.59    2006.62     200.01     570.57

Houston              3297   10382.90    6115.67     215.07    2406.62

Kansas City          3.53    7642.65    8851.65     196.09    3286.86

Los Angeles          3236     531.09      74.35   71081.98       3.00

Oralando             3297   12590.41    2173.66     123.20    3108.55

Philadelphia         3719   27064.58   16590.75     502.96    8568.62

Salt Lake City       3308    8864.34     545.17      58.71     747.43

San Diego            3308    6839.45     114.32     166.32     868.79
```

```
Tampa                3.53    4817.96    1673.96     658.55    1043.37


                MarkDown5           CPI         Unemployment Type
Size
DMA

Atlanta           6034.65  1.364618e+09                 8454    A
203819
Austin            4368.45  2.187939e+08  2022-03-06 00:00:00    B
34875
Charlotte         3009.79  1.298167e+09                 6617    B
93188
Chicago           5588.33  1.298167e+09                12.89    B
112238
Cleveland         7656.42  2.178670e+09                 7441    A
202307
Dallas            2357.19  2.230661e+09                 6054    B
125833
Denver            1005.33  2.214118e+09                 7197    B
37392
Houston           6551.42  2.179981e+09                 7866    A
151315
Kansas City       3287.36  2.109810e+09                 7082    B
128107
Los Angeles       1676.85  2.219011e+09                 7197    A
207499
Oralando          5416.96  2.195631e+09                 6551    A
202505
Philadelphia      6349.89  1.364618e+09                 7866    A
203819
Salt Lake City    4961.44  2.178670e+09                 7441    B
140167
San Diego         4442.66  2.216912e+09                 6123    A
155078
Tampa             6545.16  2.111847e+09                 7082    A
203742
```

```
atlanta = dma.get_group('Atlanta')
atlanta.head()
```

```
        Store  Dept       Date  Weekly_Sales  IsHoliday
Temperature  \
233439     24     1 2011-11-11      13578.93      False        46.78

233440     24     2 2011-11-11      39623.52      False        46.78

233442     24     4 2011-11-11      29425.46      False        46.78

233443     24     5 2011-11-11      36768.72      False        46.78
```

```
233444     24     6 2011-11-11       3182.87      False        46.78


       Fuel_Price  MarkDown1  MarkDown2  MarkDown3  MarkDown4
MarkDown5  \
233439        3719    9391.17   20595.49     686.11    5950.16
6034.65
233440        3719    9391.17   20595.49     686.11    5950.16
6034.65
233442        3719    9391.17   20595.49     686.11    5950.16
6034.65
233443        3719    9391.17   20595.49     686.11    5950.16
6034.65
233444        3719    9391.17   20595.49     686.11    5950.16
6034.65

                CPI Unemployment Type    Size      DMA
233439  1.364618e+09         8454    A  203819  Atlanta
233440  1.364618e+09         8454    A  203819  Atlanta
233442  1.364618e+09         8454    A  203819  Atlanta
233443  1.364618e+09         8454    A  203819  Atlanta
233444  1.364618e+09         8454    A  203819  Atlanta
```

```
austin = dma.get_group('Austin')
austin.head()
```

```
       Store  Dept       Date  Weekly_Sales  IsHoliday  Temperature  \
45630      5     2 2011-11-18      11568.77      False        64.33
45631      5     3 2011-11-18       3582.18      False        64.33
45632      5     4 2011-11-18       9424.31      False        64.33
45633      5     5 2011-11-18       8633.96      False        64.33
45634      5     6 2011-11-18       -101.26      False        64.33

      Fuel_Price  MarkDown1  MarkDown2  MarkDown3  MarkDown4
MarkDown5  \
45630       3308     952.21     116.11      95.01     295.13
4368.45
45631       3308     952.21     116.11      95.01     295.13
4368.45
45632       3308     952.21     116.11      95.01     295.13
4368.45
45633       3308     952.21     116.11      95.01     295.13
4368.45
45634       3308     952.21     116.11      95.01     295.13
4368.45

               CPI         Unemployment Type   Size     DMA
45630  218793912.0  2022-03-06 00:00:00    B  34875  Austin
45631  218793912.0  2022-03-06 00:00:00    B  34875  Austin
```

```
45632  218793912.0  2022-03-06 00:00:00    B  34875  Austin
45633  218793912.0  2022-03-06 00:00:00    B  34875  Austin
45634  218793912.0  2022-03-06 00:00:00    B  34875  Austin
```

```
charlotte = dma.get_group('Charlotte')
charlotte.head()
```

```
        Store  Dept       Date  Weekly_Sales  IsHoliday
Temperature  \
163768     17     1 2011-11-11      17072.88      False        27.61

163769     17     2 2011-11-11      42967.10      False        27.61

163770     17     3 2011-11-11      12968.44      False        27.61

163771     17     4 2011-11-11      22926.47      False        27.61

163772     17     5 2011-11-11      41308.18      False        27.61


       Fuel_Price  MarkDown1  MarkDown2  MarkDown3  MarkDown4
MarkDown5  \
163768       3513     3935.0    4172.99     225.67    2515.61
3009.79
163769       3513     3935.0    4172.99     225.67    2515.61
3009.79
163770       3513     3935.0    4172.99     225.67    2515.61
3009.79
163771       3513     3935.0    4172.99     225.67    2515.61
3009.79
163772       3513     3935.0    4172.99     225.67    2515.61
3009.79

                CPI Unemployment Type   Size        DMA
163768  1.298167e+09         6617    B  93188  Charlotte
163769  1.298167e+09         6617    B  93188  Charlotte
163770  1.298167e+09         6617    B  93188  Charlotte
163771  1.298167e+09         6617    B  93188  Charlotte
163772  1.298167e+09         6617    B  93188  Charlotte
```

```
chicago = dma.get_group('Chicago')
chicago.head()
```

```
        Store  Dept       Date  Weekly_Sales  IsHoliday
Temperature  \
114112     12     1 2011-11-11      13386.97      False        48.76

114113     12     2 2011-11-11      76102.74      False        48.76

114114     12     3 2011-11-11      11553.95      False        48.76
```

```
114115     12     4 2011-11-11      26921.57      False        48.76

114116     12     5 2011-11-11      30585.90      False        48.76


       Fuel_Price  MarkDown1  MarkDown2  MarkDown3  MarkDown4
MarkDown5  \
114112       3824   18049.87    7939.68     234.64   10463.07
5588.33
114113       3824   18049.87    7939.68     234.64   10463.07
5588.33
114114       3824   18049.87    7939.68     234.64   10463.07
5588.33
114115       3824   18049.87    7939.68     234.64   10463.07
5588.33
114116       3824   18049.87    7939.68     234.64   10463.07
5588.33

                CPI Unemployment Type    Size      DMA
114112  1.298167e+09        12.89    B  112238  Chicago
114113  1.298167e+09        12.89    B  112238  Chicago
114114  1.298167e+09        12.89    B  112238  Chicago
114115  1.298167e+09        12.89    B  112238  Chicago
114116  1.298167e+09        12.89    B  112238  Chicago

cleveland = dma.get_group('Cleveland')
cleveland.head()

       Store  Dept       Date  Weekly_Sales  IsHoliday  Temperature  \
16891      2     1 2011-11-18      23923.11      False        62.01
16892      2     2 2011-11-18      61442.83      False        62.01
16893      2     3 2011-11-18      11235.57      False        62.01
16894      2     4 2011-11-18      47936.90      False        62.01
16895      2     5 2011-11-18      33040.97      False        62.01

      Fuel_Price  MarkDown1  MarkDown2  MarkDown3  MarkDown4
MarkDown5  \
16891       3308    6490.92    1217.76     152.12     873.82
7656.42
16892       3308    6490.92    1217.76     152.12     873.82
7656.42
16893       3308    6490.92    1217.76     152.12     873.82
7656.42
16894       3308    6490.92    1217.76     152.12     873.82
7656.42
16895       3308    6490.92    1217.76     152.12     873.82
7656.42

                CPI Unemployment Type    Size        DMA
```

```
16891  2.178670e+09         7441    A  202307  Cleveland
16892  2.178670e+09         7441    A  202307  Cleveland
16893  2.178670e+09         7441    A  202307  Cleveland
16894  2.178670e+09         7441    A  202307  Cleveland
16895  2.178670e+09         7441    A  202307  Cleveland
```

```
dallas = dma.get_group('Dallas')
dallas.head()
```

```
       Store  Dept       Date  Weekly_Sales  IsHoliday  Temperature  \
84702      9     1 2011-12-23      26975.98      False        44.43
84703      9     2 2011-12-23      34028.54      False        44.43
84704      9     3 2011-12-23       8351.74      False        44.43
84705      9     4 2011-12-23      20918.69      False        44.43
84706      9     5 2011-12-23      48433.90      False        44.43

      Fuel_Price  MarkDown1  MarkDown2  MarkDown3  MarkDown4
MarkDown5  \
84702       3112     719.14       0.24     318.75      17.72
2357.19
84703       3112     719.14       0.24     318.75      17.72
2357.19
84704       3112     719.14       0.24     318.75      17.72
2357.19
84705       3112     719.14       0.24     318.75      17.72
2357.19
84706       3112     719.14       0.24     318.75      17.72
2357.19

                CPI Unemployment Type    Size     DMA
84702  2.230661e+09         6054    B  125833  Dallas
84703  2.230661e+09         6054    B  125833  Dallas
84704  2.230661e+09         6054    B  125833  Dallas
84705  2.230661e+09         6054    B  125833  Dallas
84706  2.230661e+09         6054    B  125833  Dallas
```

```
denver = dma.get_group('Denver')
denver.head()
```

```
       Store  Dept       Date  Weekly_Sales  IsHoliday  Temperature  \
26273      3     1 2011-11-11       6525.18      False         61.7
26274      3     2 2011-11-11      16030.50      False         61.7
26275      3     3 2011-11-11       4611.85      False         61.7
26276      3     4 2011-11-11       8223.47      False         61.7
26277      3     5 2011-11-11      14194.63      False         61.7

      Fuel_Price  MarkDown1  MarkDown2  MarkDown3  MarkDown4
MarkDown5  \
26273       3297    8787.59    2006.62     200.01     570.57
1005.33
26274       3297    8787.59    2006.62     200.01     570.57
```

```
1005.33
26275       3297    8787.59    2006.62     200.01     570.57
1005.33
26276       3297    8787.59    2006.62     200.01     570.57
1005.33
26277       3297    8787.59    2006.62     200.01     570.57
1005.33

                CPI Unemployment Type   Size     DMA
26273  2.214118e+09         7197    B  37392  Denver
26274  2.214118e+09         7197    B  37392  Denver
26275  2.214118e+09         7197    B  37392  Denver
26276  2.214118e+09         7197    B  37392  Denver
26277  2.214118e+09         7197    B  37392  Denver
```

```
houston = dma.get_group('Houston')
houston.head()
```

```
      Store  Dept       Date  Weekly_Sales  IsHoliday  Temperature
Fuel_Price  \
6587      1     1 2011-11-11      18689.54      False        59.11
3297
6588      1     2 2011-11-11      44936.47      False        59.11
3297
6589      1     3 2011-11-11       9959.64      False        59.11
3297
6590      1     4 2011-11-11      36826.52      False        59.11
3297
6591      1     5 2011-11-11      31002.65      False        59.11
3297

      MarkDown1  MarkDown2  MarkDown3  MarkDown4  MarkDown5
CPI  \
6587    10382.9    6115.67     215.07    2406.62    6551.42
2.179981e+09
6588    10382.9    6115.67     215.07    2406.62    6551.42
2.179981e+09
6589    10382.9    6115.67     215.07    2406.62    6551.42
2.179981e+09
6590    10382.9    6115.67     215.07    2406.62    6551.42
2.179981e+09
6591    10382.9    6115.67     215.07    2406.62    6551.42
2.179981e+09

     Unemployment Type    Size      DMA
6587         7866    A  151315  Houston
6588         7866    A  151315  Houston
6589         7866    A  151315  Houston
6590         7866    A  151315  Houston
6591         7866    A  151315  Houston
```

```
kansasCity = dma.get_group('Kansas City')
kansasCity.head()
```

```
        Store  Dept       Date  Weekly_Sales  IsHoliday
Temperature  \
243393     25     1 2011-11-11      16508.85      False        44.81

243394     25     2 2011-11-11      34976.37      False        44.81

243395     25     3 2011-11-11       9391.93      False        44.81

243396     25     4 2011-11-11      21866.25      False        44.81

243397     25     5 2011-11-11      17941.96      False        44.81


       Fuel_Price  MarkDown1  MarkDown2  MarkDown3  MarkDown4
MarkDown5  \
243393       3.53    7642.65    8851.65     196.09    3286.86
3287.36
243394       3.53    7642.65    8851.65     196.09    3286.86
3287.36
243395       3.53    7642.65    8851.65     196.09    3286.86
3287.36
243396       3.53    7642.65    8851.65     196.09    3286.86
3287.36
243397       3.53    7642.65    8851.65     196.09    3286.86
3287.36

                 CPI Unemployment Type    Size          DMA
243393  2.109810e+09         7082    B  128107  Kansas City
243394  2.109810e+09         7082    B  128107  Kansas City
243395  2.109810e+09         7082    B  128107  Kansas City
243396  2.109810e+09         7082    B  128107  Kansas City
243397  2.109810e+09         7082    B  128107  Kansas City
```

```
losAngeles = dma.get_group('Los Angeles')
losAngeles.head()
```

```
        Store  Dept       Date  Weekly_Sales  IsHoliday
Temperature  \
104418     11     1 2011-11-25      19008.41       True        70.03

104419     11     2 2011-11-25      54977.82       True        70.03

104420     11     3 2011-11-25      10322.33       True        70.03

104421     11     4 2011-11-25      35035.34       True        70.03

104422     11     5 2011-11-25     113522.16       True        70.03
```

```
       Fuel_Price  MarkDown1  MarkDown2  MarkDown3  MarkDown4
MarkDown5  \
104418       3236     531.09      74.35   71081.98        3.0
1676.85
104419       3236     531.09      74.35   71081.98        3.0
1676.85
104420       3236     531.09      74.35   71081.98        3.0
1676.85
104421       3236     531.09      74.35   71081.98        3.0
1676.85
104422       3236     531.09      74.35   71081.98        3.0
1676.85

                CPI Unemployment Type    Size          DMA
104418  2.219011e+09         7197    A  207499  Los Angeles
104419  2.219011e+09         7197    A  207499  Los Angeles
104420  2.219011e+09         7197    A  207499  Los Angeles
104421  2.219011e+09         7197    A  207499  Los Angeles
104422  2.219011e+09         7197    A  207499  Los Angeles
```

```
# it's spelled wrong I know, too late now to go back and fix  :(
orlando = dma.get_group('Oralando')
orlando.head()
```

```
       Store  Dept       Date  Weekly_Sales  IsHoliday  Temperature  \
55348      6     1 2011-11-11      17870.11      False        61.33
55349      6     2 2011-11-11      48783.32      False        61.33
55350      6     3 2011-11-11      11049.41      False        61.33
55351      6     4 2011-11-11      33771.93      False        61.33
55352      6     5 2011-11-11      39401.36      False        61.33

      Fuel_Price  MarkDown1  MarkDown2  MarkDown3  MarkDown4
MarkDown5  \
55348       3297   12590.41    2173.66      123.2    3108.55
5416.96
55349       3297   12590.41    2173.66      123.2    3108.55
5416.96
55350       3297   12590.41    2173.66      123.2    3108.55
5416.96
55351       3297   12590.41    2173.66      123.2    3108.55
5416.96
55352       3297   12590.41    2173.66      123.2    3108.55
5416.96

                CPI Unemployment Type    Size       DMA
55348  2.195631e+09         6551    A  202505  Oralando
55349  2.195631e+09         6551    A  202505  Oralando
55350  2.195631e+09         6551    A  202505  Oralando
```

```
55351  2.195631e+09         6551    A  202505  Oralando
55352  2.195631e+09         6551    A  202505  Oralando
```

```
philadelphia = dma.get_group('Philadelphia')
philadelphia.head()
```

```
        Store  Dept       Date  Weekly_Sales  IsHoliday
Temperature  \
183713     19     1 2011-11-11      18790.30      False        48.22

183714     19     2 2011-11-11      50399.44      False        48.22

183715     19     3 2011-11-11      12251.42      False        48.22

183716     19     4 2011-11-11      32289.01      False        48.22

183717     19     5 2011-11-11      35444.35      False        48.22


       Fuel_Price  MarkDown1  MarkDown2  MarkDown3  MarkDown4
MarkDown5  \
183713       3719   27064.58   16590.75     502.96    8568.62
6349.89
183714       3719   27064.58   16590.75     502.96    8568.62
6349.89
183715       3719   27064.58   16590.75     502.96    8568.62
6349.89
183716       3719   27064.58   16590.75     502.96    8568.62
6349.89
183717       3719   27064.58   16590.75     502.96    8568.62
6349.89

                CPI Unemployment Type    Size           DMA
183713  1.364618e+09         7866    A  203819  Philadelphia
183714  1.364618e+09         7866    A  203819  Philadelphia
183715  1.364618e+09         7866    A  203819  Philadelphia
183716  1.364618e+09         7866    A  203819  Philadelphia
183717  1.364618e+09         7866    A  203819  Philadelphia
```

```
saltLakeCity = dma.get_group('Salt Lake City')
saltLakeCity.head()
```

```
        Store  Dept       Date  Weekly_Sales  IsHoliday
Temperature  \
203749     21     1 2011-11-18      12800.42      False         61.9

203750     21     2 2011-11-18      50625.93      False         61.9

203751     21     3 2011-11-18       9852.59      False         61.9
```

```
203752     21     4 2011-11-18      19597.43      False         61.9

203753     21     5 2011-11-18      14902.38      False         61.9


       Fuel_Price  MarkDown1  MarkDown2  MarkDown3  MarkDown4
MarkDown5  \
203749       3308    8864.34     545.17      58.71     747.43
4961.44
203750       3308    8864.34     545.17      58.71     747.43
4961.44
203751       3308    8864.34     545.17      58.71     747.43
4961.44
203752       3308    8864.34     545.17      58.71     747.43
4961.44
203753       3308    8864.34     545.17      58.71     747.43
4961.44

                CPI Unemployment Type    Size             DMA
203749  2.178670e+09         7441    B  140167  Salt Lake City
203750  2.178670e+09         7441    B  140167  Salt Lake City
203751  2.178670e+09         7441    B  140167  Salt Lake City
203752  2.178670e+09         7441    B  140167  Salt Lake City
203753  2.178670e+09         7441    B  140167  Salt Lake City
```

```
sanDiego = dma.get_group('San Diego')
sanDiego.head()
```

```
       Store  Dept       Date  Weekly_Sales  IsHoliday  Temperature  \
75192      8     1 2011-11-18      11044.58      False        51.72
75193      8     2 2011-11-18      35177.25      False        51.72
75195      8     4 2011-11-18      22130.51      False        51.72
75196      8     5 2011-11-18      22551.77      False        51.72
75197      8     6 2011-11-18       1734.44      False        51.72

      Fuel_Price  MarkDown1  MarkDown2  MarkDown3  MarkDown4
MarkDown5  \
75192       3308    6839.45     114.32     166.32     868.79
4442.66
75193       3308    6839.45     114.32     166.32     868.79
4442.66
75195       3308    6839.45     114.32     166.32     868.79
4442.66
75196       3308    6839.45     114.32     166.32     868.79
4442.66
75197       3308    6839.45     114.32     166.32     868.79
4442.66

                CPI Unemployment Type    Size        DMA
75192  2.216912e+09         6123    A  155078  San Diego
```

```
75193  2.216912e+09         6123    A  155078  San Diego
75195  2.216912e+09         6123    A  155078  San Diego
75196  2.216912e+09         6123    A  155078  San Diego
75197  2.216912e+09         6123    A  155078  San Diego
```

```
tampa = dma.get_group('Tampa')
tampa.head()
```

```
        Store  Dept       Date  Weekly_Sales  IsHoliday
Temperature  \
193968     20     1 2011-11-18      34233.84      False         49.3

193969     20     2 2011-11-18      73054.31      False         49.3

193970     20     3 2011-11-18      12186.03      False         49.3

193971     20     4 2011-11-18      51656.07      False         49.3

193972     20     5 2011-11-18      45800.43      False         49.3


       Fuel_Price  MarkDown1  MarkDown2  MarkDown3  MarkDown4
MarkDown5  \
193968       3.53    4817.96    1673.96     658.55    1043.37
6545.16
193969       3.53    4817.96    1673.96     658.55    1043.37
6545.16
193970       3.53    4817.96    1673.96     658.55    1043.37
6545.16
193971       3.53    4817.96    1673.96     658.55    1043.37
6545.16
193972       3.53    4817.96    1673.96     658.55    1043.37
6545.16

                CPI Unemployment Type    Size    DMA
193968  2.111847e+09         7082    A  203742  Tampa
193969  2.111847e+09         7082    A  203742  Tampa
193970  2.111847e+09         7082    A  203742  Tampa
193971  2.111847e+09         7082    A  203742  Tampa
193972  2.111847e+09         7082    A  203742  Tampa
```

Export Data for Visualizations

```
atlanta.to_csv('atlanta.csv')
!cp atlanta.csv "drive/My Drive/Y4/Colab Notebooks/aba"

austin.to_csv('austin.csv')
!cp austin.csv "drive/My Drive/Y4/Colab Notebooks/aba"

charlotte.to_csv('charlotte.csv')
```

```
!cp charlotte.csv "drive/My Drive/Y4/Colab Notebooks/aba"

chicago.to_csv('chicago.csv')
!cp chicago.csv "drive/My Drive/Y4/Colab Notebooks/aba"

cleveland.to_csv('cleveland.csv')
!cp cleveland.csv "drive/My Drive/Y4/Colab Notebooks/aba"

dallas.to_csv('dallas.csv')
!cp dallas.csv "drive/My Drive/Y4/Colab Notebooks/aba"

denver.to_csv('denver.csv')
!cp denver.csv "drive/My Drive/Y4/Colab Notebooks/aba"

houston.to_csv('houston.csv')
!cp houston.csv "drive/My Drive/Y4/Colab Notebooks/aba"

kansasCity.to_csv('kansasCity.csv')
!cp kansasCity.csv "drive/My Drive/Y4/Colab Notebooks/aba"

losAngeles.to_csv('losAngeles.csv')
!cp losAngeles.csv "drive/My Drive/Y4/Colab Notebooks/aba"

orlando.to_csv('orlando.csv')
!cp orlando.csv "drive/My Drive/Y4/Colab Notebooks/aba"

philadelphia.to_csv('philadelphia.csv')
!cp philadelphia.csv "drive/My Drive/Y4/Colab Notebooks/aba"

saltLakeCity.to_csv('saltLakeCity.csv')
!cp saltLakeCity.csv "drive/My Drive/Y4/Colab Notebooks/aba"

sanDiego.to_csv('sanDiego.csv')
!cp sanDiego.csv "drive/My Drive/Y4/Colab Notebooks/aba"

tampa.to_csv('tampa.csv')
!cp tampa.csv "drive/My Drive/Y4/Colab Notebooks/aba"
```

Overall Data Findings

Merging the datasets features, sales, and stores_DMA was successful. When I broke the data down by DMA (city) location, I noticed there are many unusual values in some variables that I did not catch before. This went unseen due to the mass size of the original dataset. ~~Ideally with more computing power, knowledge and time these values could be filtered out.~~

Analyze Data/Insights

Questions for Analysis

- What is the ranking of cities most desirable to launch in?
- What city/region has the top weekly sales?
- What region is the highest performing?
- How do holidays impact sales and store performance?
- What impact do temperature, fuel price, CPI, and unemployment have on store performance?
- What does the data tell us about Walmart/Target's retail strategy?
- How do income levels affect consumer trends?
- Which categories (and consequently) which spend levels are the most attractive?
- What market assumptions are made in the data?

Launch Recommendation for Retail Startup

Goal: produce an analysis with a clear recommendation on which markets to choose for launch of new store.

Launch Plan

Also included in submission as a PDF.

Link: docs.google.com/launch-plan

References/Links

Original Kaggle Dataset: kaggle.com/walmart-recruiting-store-sales-forcasting-data

Homework 4 Data: drive.google.com/hw4-data

Stack Overflow: stackoverflow.com/load-xlsx-file, stackoverflow.com/groupby-keyerror,

stackoverflow.com/remove-rows

Real Python: realpython.com/pandas-merge