Week 1 Application Assignment 2

August 21, 2020

0.1 Week 1 Application Assignment 2: Data Visualization

0.2 Note to peer reviewer: I am not expecting you understand Python code. You only need to read the descriptions and results presented inside the report.

0.2.1 Data Description

The dataset in the file contains the customer reward program information for the top 100 US retailers based on annual sales. Several aspects of the customer reward programs are shown in this dataset, including customer reward program adoption, the size of customer reward, and the expiration term of customer rewards. The dataset contains the following columns:

- Retailer: name of the retailer.
- Salerank: sales rank of the retailer based on 2013 US sales.
- X2013USSales: total US sales amount in the year 2013
- X2013WorldSales: total worldwide sales amount in the year 2013
- NumStores: number of stores
- Industry: the industry of the retailer.
- ProfitMargin: profit margin of the retailer.
- Reward: whether the retailer has a customer reward program, Yes=1, and No=0.
- ProgramName: name of the customer reward program, if any.
- RewardType: type of the reward.
- RewardStructure: a description of the structure of the reward.
- RewardSize: reward as a percentage of purchase amount.
- ExpirationMonth: number of months before earned rewards expire. 999 indicates no expiration.
- IndustryType: a column created based on Industry.

Part 1. Create a plot to show at least one interesting aspect of the data. You can use Excel or another visualization software (such as Tableau). Make sure your file is less than 10MB to ensure your file gets uploaded correctly.

Part 2. Briefly explain why the aspect(s) of the data that you chose to show in the plot is interesting.

[1]:

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import datetime
```

```
%matplotlib inline
sns.set_style('dark')
sns.set(font_scale=1.2)

import warnings
warnings.filterwarnings('ignore')

pd.set_option('display.max_columns',None)
#pd.set_option('display.max_rows',None)
```

[2]:
```
df = pd.read_csv("crp_cleandata.csv")
```

[3]:
```
df
```

[3]:
```
                               Retailer  Salerank  X2013USSales  \
0                                   A&P        74         5.831
1                            Albertsons        21        19.452
2                                  Aldi        38        10.898
3   Alimentation Couche Tard (Circle K)        82         4.755
4                          Apple Stores        15        26.648
..                                  ...       ...           ...
95                          Toys "R" Us        57         7.525
96                       Tractor Supply        77         5.165
97                           True Value        27        16.330
98                             Walgreen         6        68.068
99                       WilliamsSonoma        89         4.163

    X2013WorldSales  ProfitMargin  NumStores                  Industry  \
0             5.831         48.85      0.277  Discount, Variety Stores
1            19.452         69.02      1.024            Grocery Stores
2            10.650         69.41      1.328            Grocery Stores
3             8.551         68.03      3.826            Grocery Stores
4            30.736         11.07      0.254       ElectronicEquipment
..              ...           ...        ...                       ...
95           13.307         79.33      0.868   Specialty Retail, Other
96            5.165          2.65      1.276   Specialty Retail, Other
97           16.330         38.69      4.494   Home Improvement Stores
98           70.096         23.47      7.998               Drug Stores
99            4.388         30.09      0.553    Home Furnishing Stores

    Reward                           ProgramName    RewardType  \
0        0                    No rewards program           NaN
1        0                    No rewards program           NaN
2        0                    No rewards program           NaN
3        0                    No rewards program           NaN
4        0                    No rewards program           NaN
..     ...                                   ...           ...
```

```
95       1                         Rewards"R" Us  store credit
96       1                       Neighbor's Club  store credit
97       1                     TrueValue Rewards  store credit
98       1                       Balance Rewards  store credit
99       1  WilliamsSonoma Visa??Signature??Card  store credit

                                      RewardStructure  RewardSize  \
0                                                 NaN         NaN
1                                                 NaN         NaN
2                                                 NaN         NaN
3                                                 NaN         NaN
4                                                 NaN         NaN
..                                                …         …
95  Spend $125 earn $5 Reward Dollars for future p…        4.00
96  Spend $150 X3 times earn seasonal (quarterly) …        1.11
97  Spend $250 earn 2,500 points = $5 reward certi…        2.00
98  500 points on every prescription filled in the…       10.00
99      Spend $3,333 earn 10,000 pts = $100 gift card        3.00

    ExpirationMonth IndustryType
0               NaN     Discount
1               NaN      Grocery
2               NaN      Grocery
3               NaN      Grocery
4               NaN    Specialty
..              …          …
95             12.0    Specialty
96              3.0    Specialty
97             24.0    Specialty
98              6.0    Specialty
99            999.0    Specialty

[100 rows x 14 columns]
```

0.2.2 Exploratory Data Analysis

```
[4]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Retailer         100 non-null    object
 1   Salerank         100 non-null    int64
 2   X2013USSales     100 non-null    float64
 3   X2013WorldSales  100 non-null    float64
 4   ProfitMargin     100 non-null    float64
 5   NumStores        100 non-null    float64
 6   Industry         100 non-null    object
 7   Reward           100 non-null    int64
 8   ProgramName      100 non-null    object
 9   RewardType       59 non-null     object
 10  RewardStructure  59 non-null     object
 11  RewardSize       56 non-null     float64
 12  ExpirationMonth  56 non-null     float64
 13  IndustryType     100 non-null    object
dtypes: float64(6), int64(2), object(6)
memory usage: 11.1+ KB
```

[5]:

```
df.describe(include='all')
```

[5]:

```
                      Retailer    Salerank  X2013USSales  X2013WorldSales  \
count                      100  100.000000    100.000000       100.000000
unique                     100         NaN           NaN              NaN
top     Good Neighbor Pharmacy         NaN           NaN              NaN
freq                         1         NaN           NaN              NaN
mean                       NaN   50.500000     18.373500        24.131540
std                        NaN   29.011492     36.476003        50.845864
min                        NaN    1.000000      3.600000         3.600000
25%                        NaN   25.750000      5.206750         6.108250
50%                        NaN   50.500000      8.348500         9.629000
75%                        NaN   75.250000     16.841000        22.131500
max                        NaN  100.000000    334.302000       473.979000

        ProfitMargin   NumStores                  Industry  Reward  \
count      100.00000  100.000000                       100  100.00
unique           NaN         NaN                        17     NaN
top              NaN         NaN  Discount, Variety Stores     NaN
freq             NaN         NaN                        18     NaN
mean        45.27300    2.698760                       NaN    0.55
std         29.23139    3.997641                       NaN    0.50
min          1.02000    0.000000                       NaN    0.00
25%         19.44500    0.338500                       NaN    0.00
50%         42.02000    1.331500                       NaN    1.00
75%         69.57250    3.519750                       NaN    1.00
max         99.20000   26.644000                       NaN    1.00

               ProgramName    RewardType  \
count                  100            59
unique                  58            10
top     No rewards program  store credit
freq                    40            37
mean                   NaN           NaN
```

```
std                     NaN           NaN
min                     NaN           NaN
25%                     NaN           NaN
50%                     NaN           NaN
75%                     NaN           NaN
max                     NaN           NaN

                                          RewardStructure  RewardSize  \
count                                                  59   56.000000
unique                                                 56         NaN
top     Get 10% discount in rewards for shopping paper…         NaN
freq                                                    2         NaN
mean                                                  NaN    6.582857
std                                                   NaN    6.929193
min                                                   NaN    1.000000
25%                                                   NaN    2.000000
50%                                                   NaN    3.415000
75%                                                   NaN   10.000000
max                                                   NaN   33.330000

        ExpirationMonth IndustryType
count         56.000000          100
unique              NaN            5
top                 NaN    Specialty
freq                NaN           43
mean         166.607143          NaN
std          367.614592          NaN
min            1.000000          NaN
25%            2.000000          NaN
50%            6.000000          NaN
75%           12.000000          NaN
max          999.000000          NaN
```

[6]:

```
df.shape
```

[6]:

```
(100, 14)
```

[ ]:

```
```

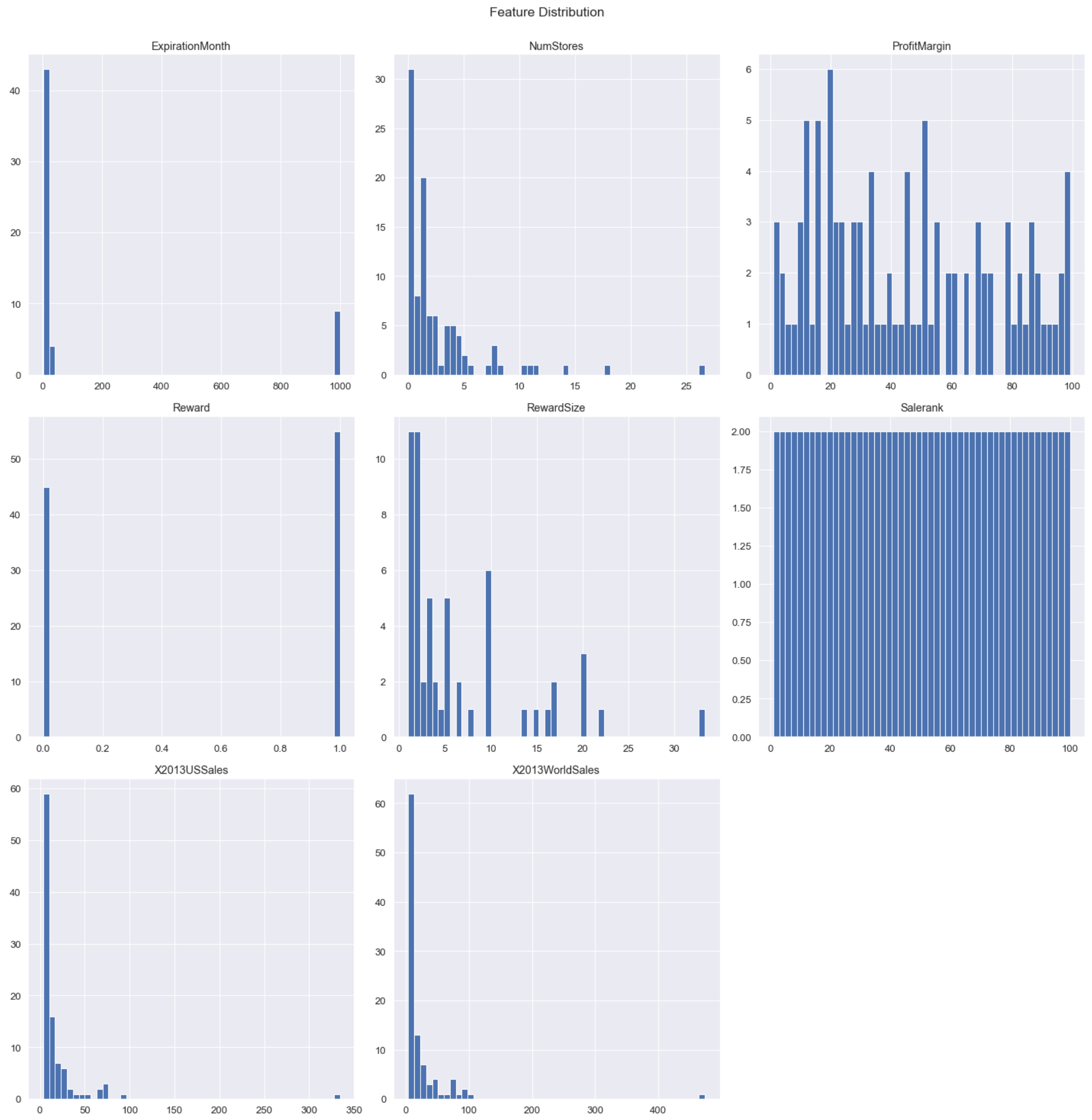
0.2.3 Data Visualization

0.2.4 Univariate Data Exploration

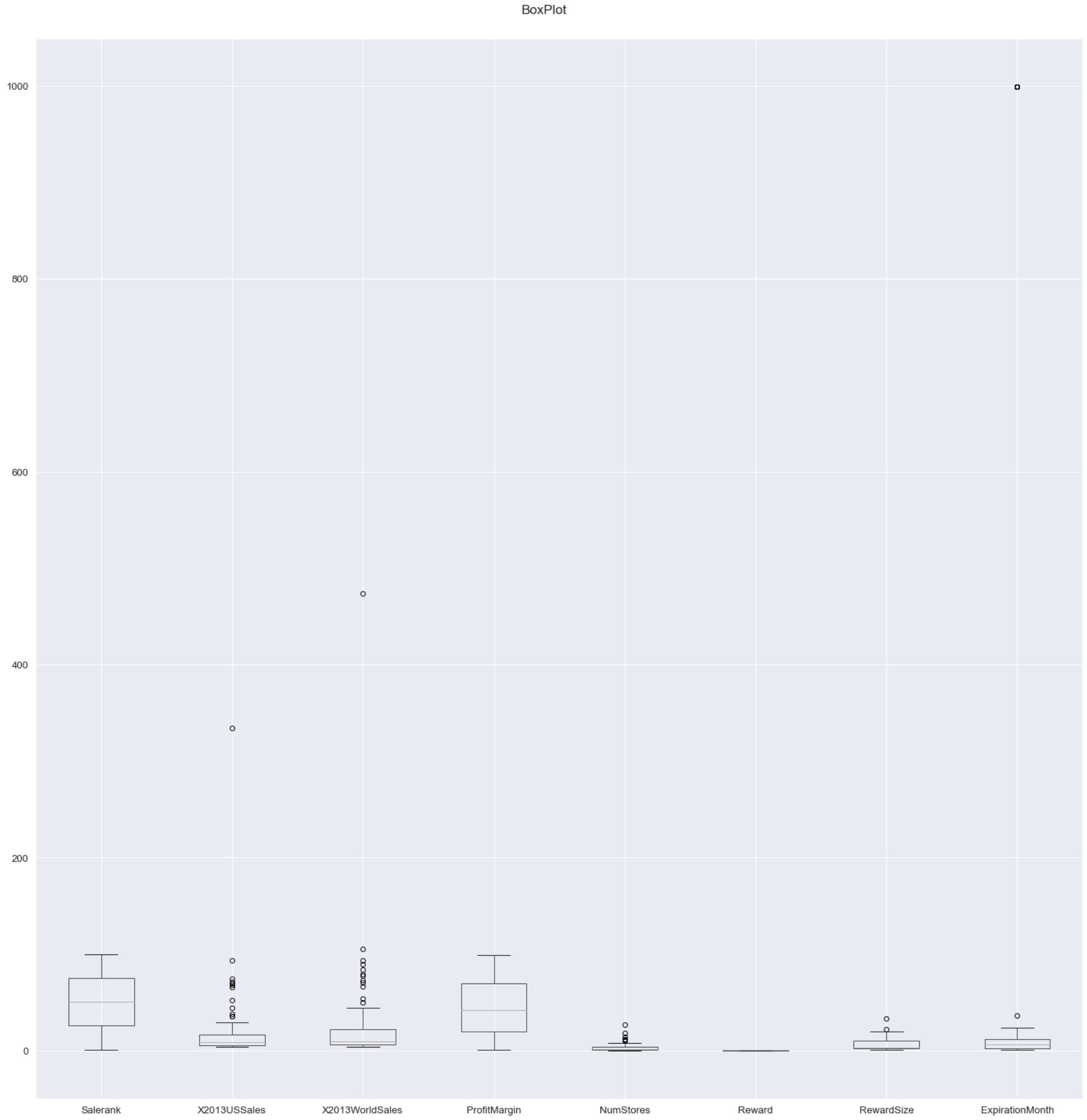
[7]:

```
df.hist(bins=50, figsize=(20,20))

plt.suptitle('Feature Distribution', x=0.5, y=1.02, ha='center',␣
 ↪fontsize='large')
```

```
plt.tight_layout()

plt.show()
```

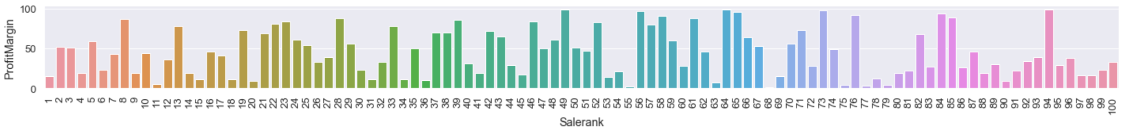


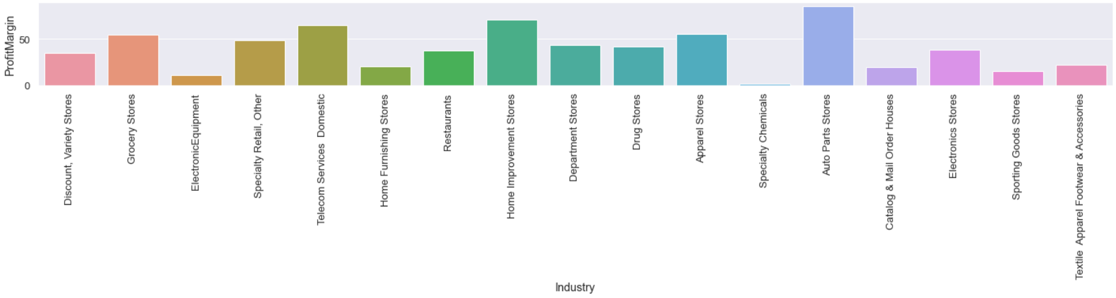
[8]:
```
df.boxplot(figsize=(20,20))

plt.suptitle('BoxPlot', x=0.5, y=1.02, ha='center', fontsize='large')

plt.tight_layout()

plt.show()
```

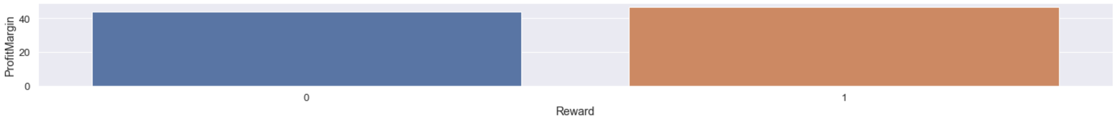


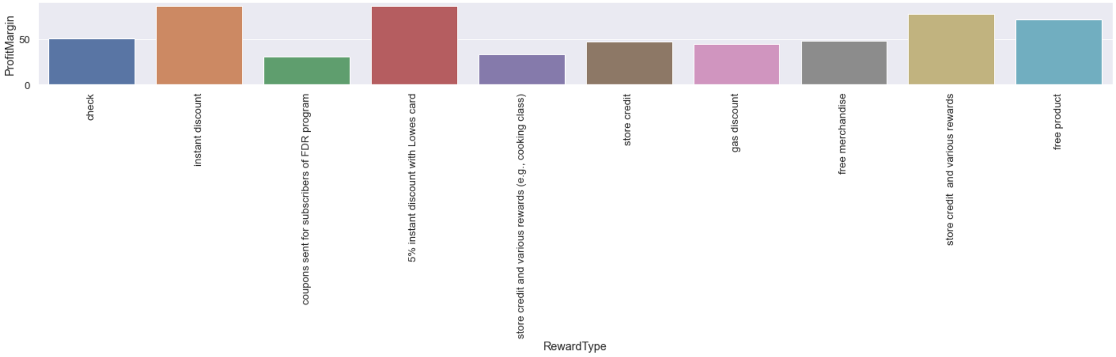
0.2.5 Observation 1: Below are Profit Margins based om SaleRank, Industry, Rewards, RewardType and IndustryType. The results are quite varied.

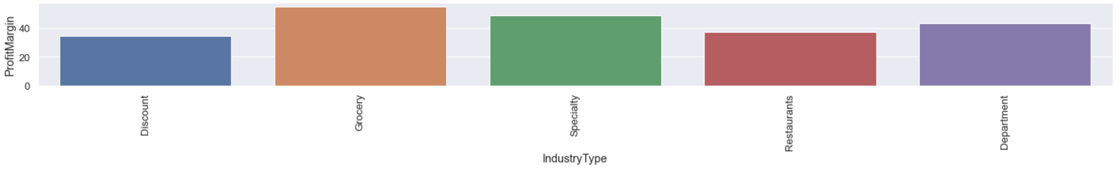
```
[9]: fig = plt.figure(figsize=(20,40))

plt.subplot(7,1,1)
plt.xticks(rotation=90)
plt.title("")
sns.barplot(x='Salerank',y='ProfitMargin',data=df,ci=None)

plt.subplot(7,1,2)
plt.title("")
plt.xticks(rotation=90)
sns.barplot(x='Industry',y='ProfitMargin',data=df,ci=None)

plt.subplot(7,1,3)
plt.title("")
sns.barplot(x='Reward',y='ProfitMargin',data=df,ci=None)

plt.subplot(7,1,4)
plt.title("")
plt.xticks(rotation=90)
sns.barplot(x='RewardType',y='ProfitMargin',data=df,ci=None)

plt.subplot(7,1,5)
plt.title("")
plt.xticks(rotation=90)
sns.barplot(x='IndustryType',y='ProfitMargin',data=df,ci=None)


plt.tight_layout()
plt.show()
```


ProfitMargin
100
50
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
Salerank


ProfitMargin
50
0
Discount, Variety Stores
Grocery Stores
ElectronicEquipment
Specialty Retail, Other
Telecom Services Domestic
Home Furnishing Stores
Restaurants
Home Improvement Stores
Department Stores
Drug Stores
Apparel Stores
Specialty Chemicals
Auto Parts Stores
Catalog & Mail Order Houses
Electronics Stores
Sporting Goods Stores
Textile Apparel Footwear & Accessories
Industry

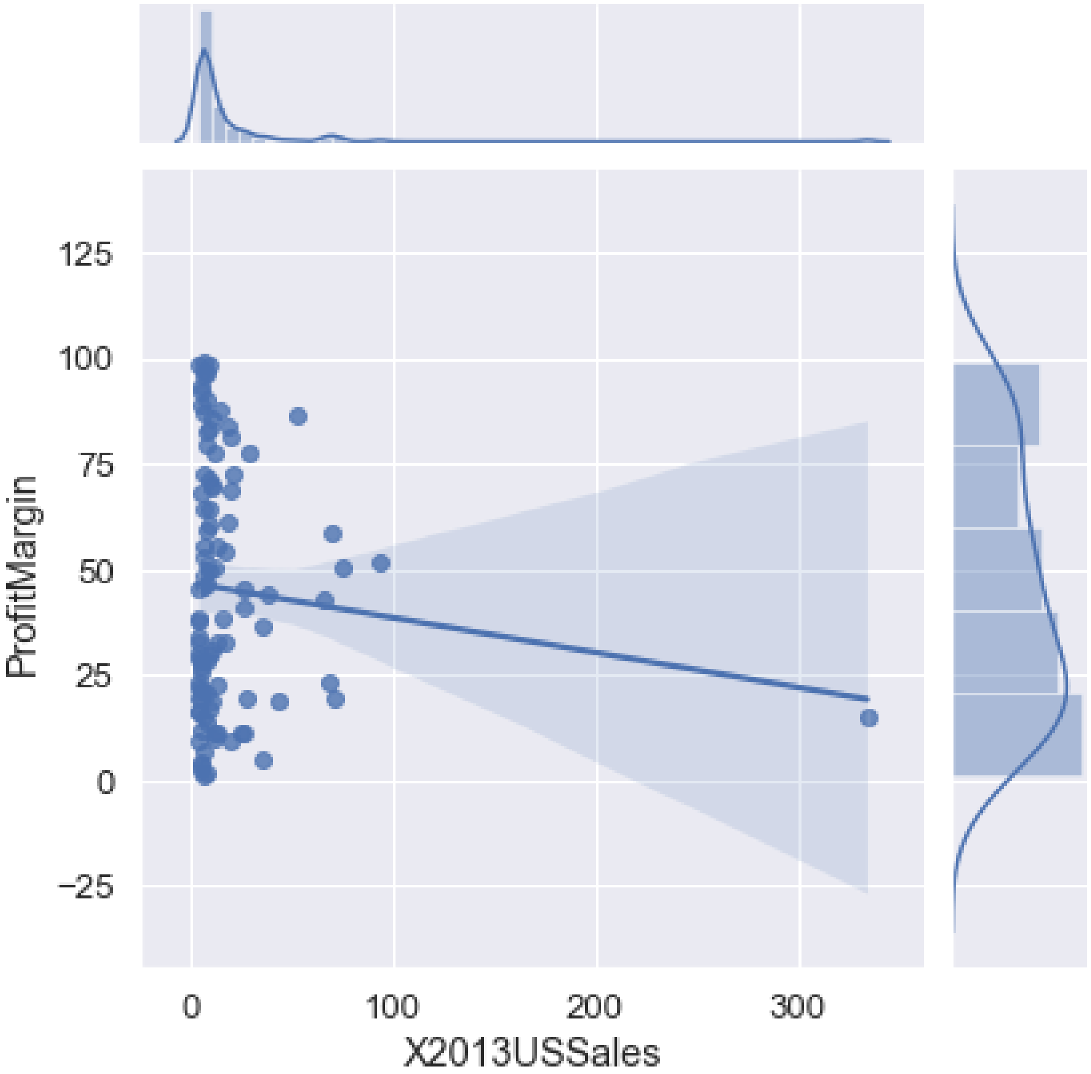

ProfitMargin
40
20
0
0
1
Reward

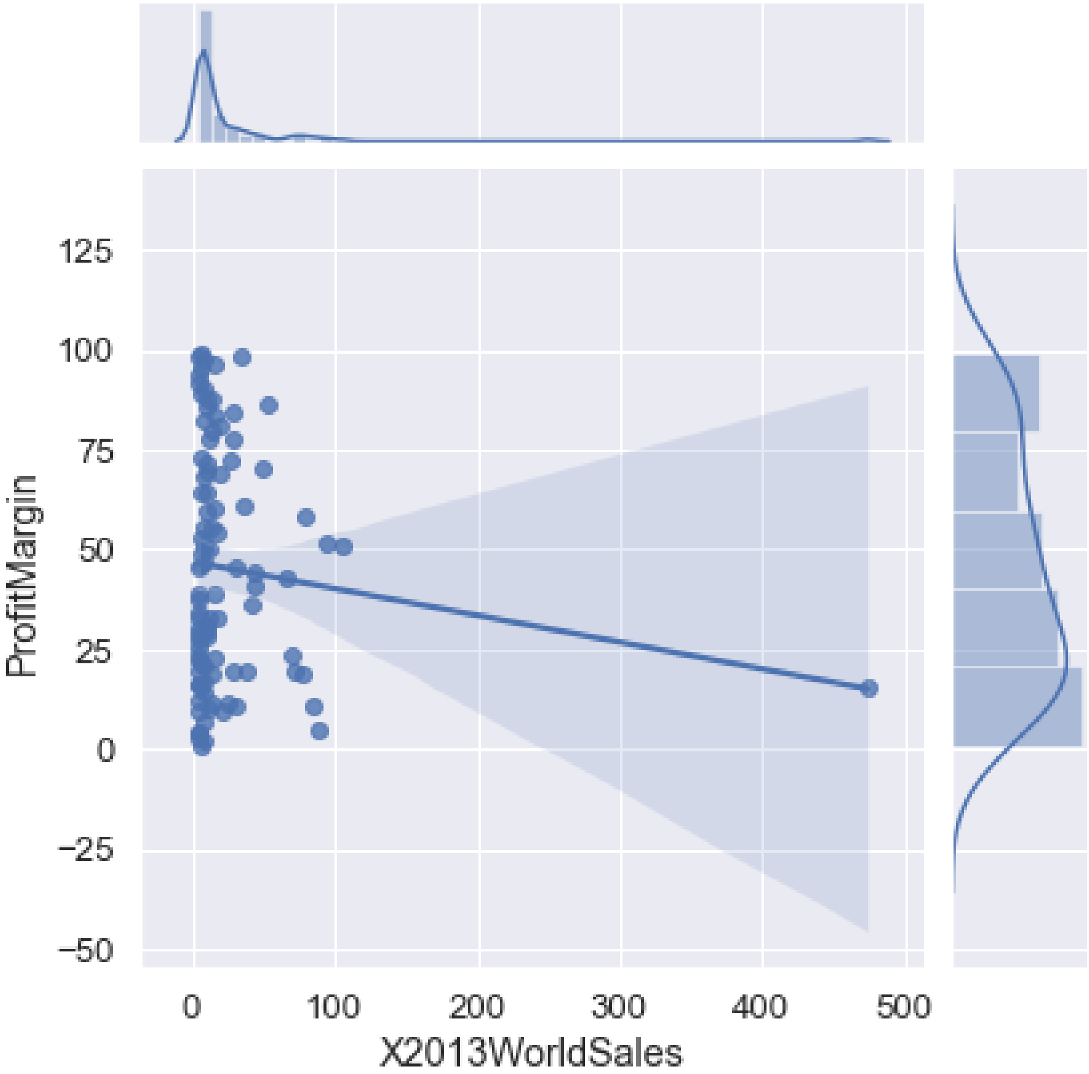

ProfitMargin
50
0
check
instant discount
coupons sent for subscribers of FDR program
5% instant discount with Lowes card
store credit and various rewards (e.g., cooking class)
store credit
gas discount
free merchandise
store credit and various rewards
free product
RewardType

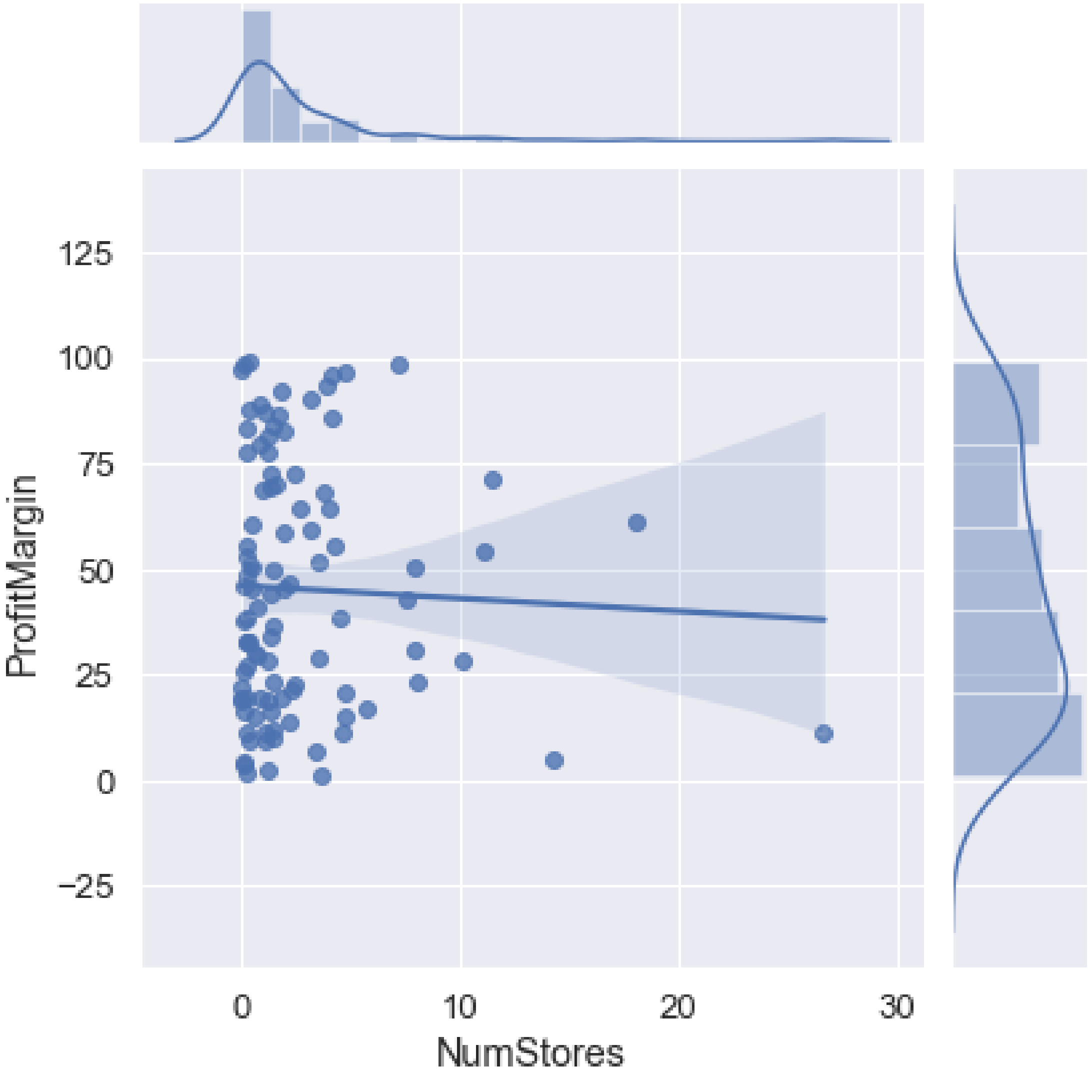

ProfitMargin
40
20
0
Discount
Grocery
Specialty
Restaurants
Department
IndustryType

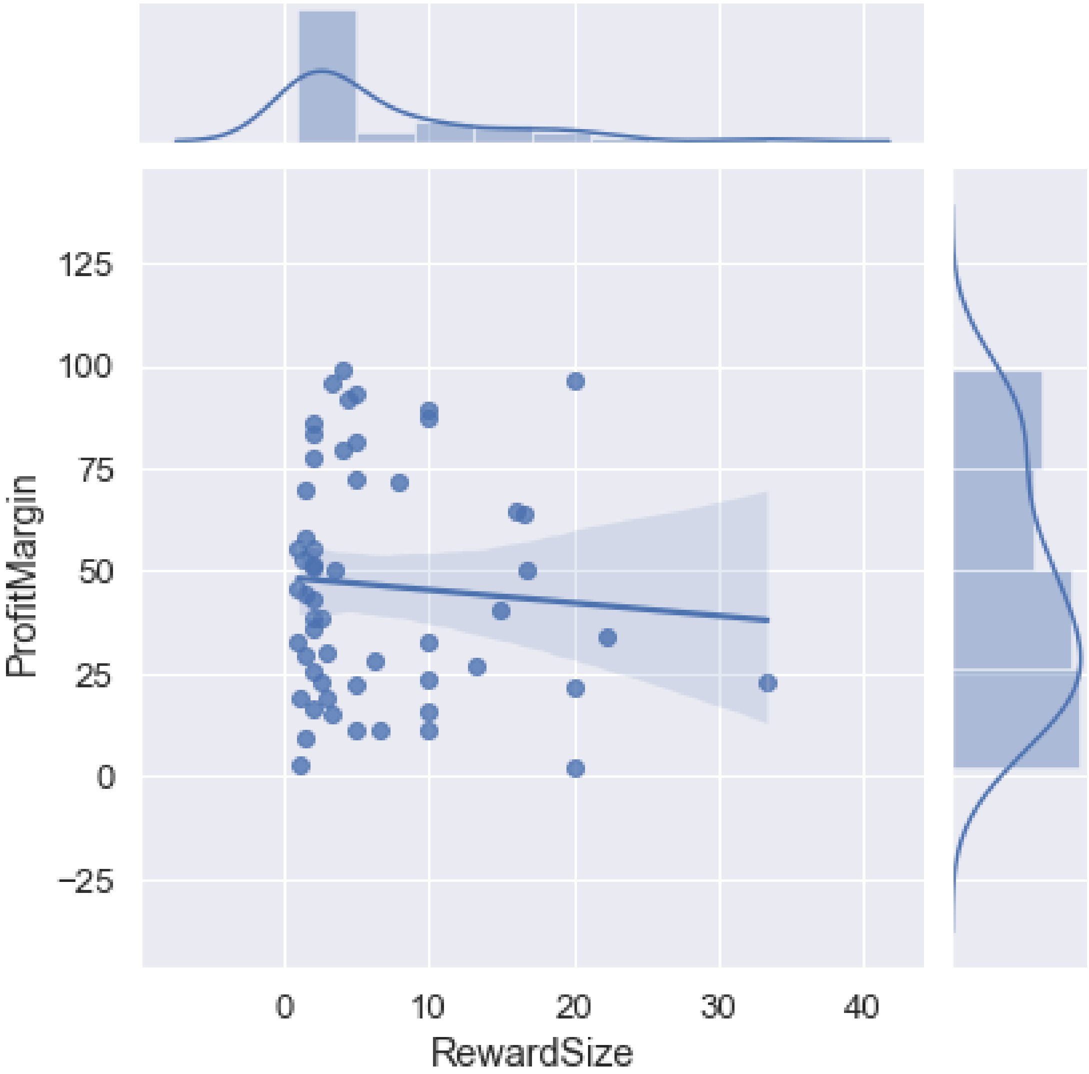
0.2.6 Bivariate Data Exploration

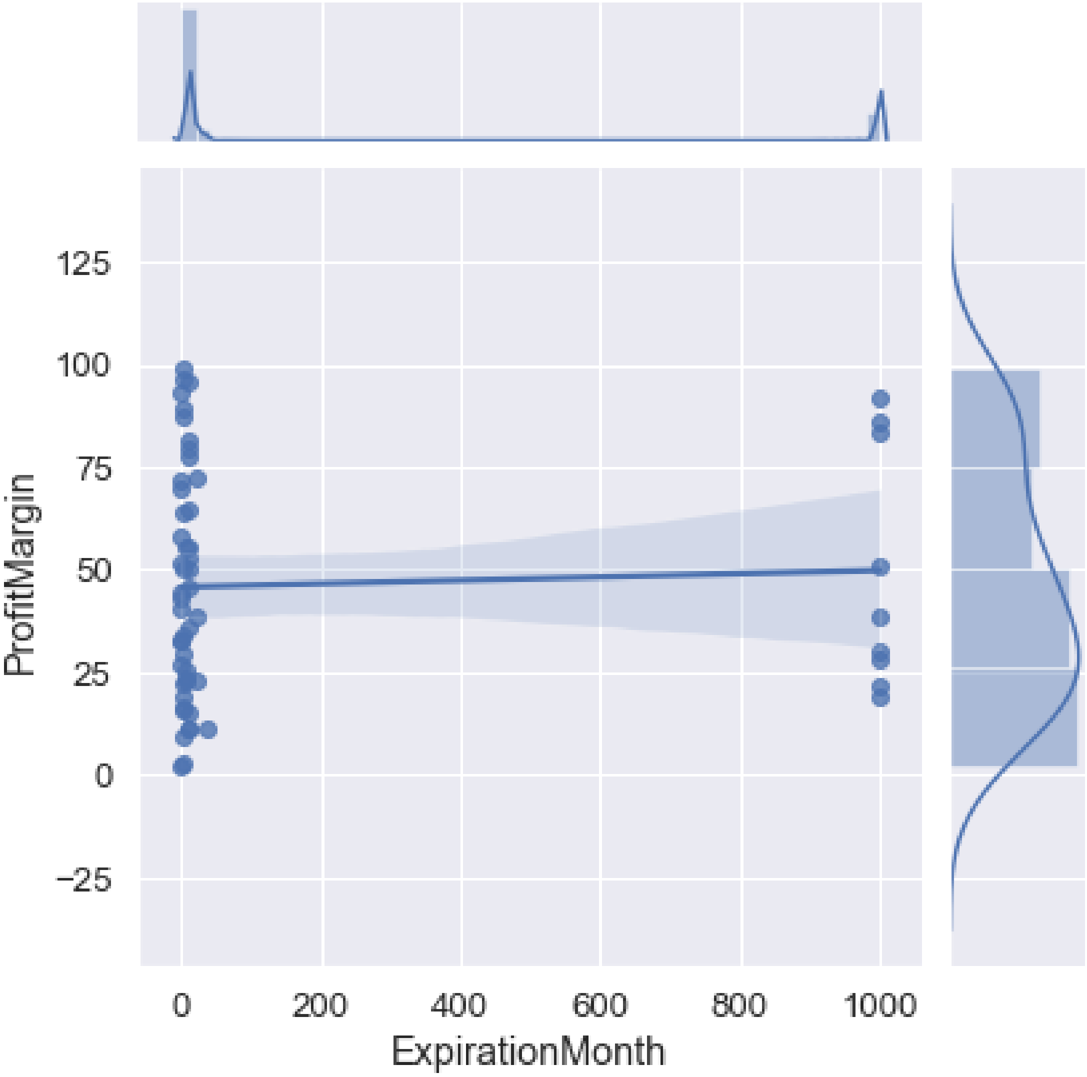
[10]:
```
sns.jointplot(x='X2013USSales', y='ProfitMargin',data=df, kind='reg')

sns.jointplot(x='X2013WorldSales', y='ProfitMargin',data=df, kind='reg')

sns.jointplot(x='NumStores', y='ProfitMargin',data=df, kind='reg')

sns.jointplot(x='RewardSize', y='ProfitMargin',data=df, kind='reg')

sns.jointplot(x='ExpirationMonth', y='ProfitMargin',data=df, kind='reg')

plt.show()
```




ProfitMargin
125
100
75
50
25
0
−25
−50
0
100
200
300
400
500
X2013WorldSales


125
100
75
50
25
0
−25
ProfitMargin
0
10
20
30
NumStores


125
100
75
50
25
0
−25
ProfitMargin
0
10
20
30
40
RewardSize



0.2.7 Observation 2: There is a slight decrease of profit margin with more sales or stores or reward sizes

```
[ ]:
```