UDACITY NANODEGREE Predictive Analytics for Business
Predictive Analytics for Business

Daiane Ucceli Kreitlow

August 29, 2023.

Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions need to be made?

We need to build predictive sales profit based on some data and then decide if we should send a catalog to new customers.

The minimum value to profit is about $10,000 for 250 new customers.

If we meet the conditions, so we will send the catalogs to these new customers.

2. What data is needed to inform those decisions?

We have data about customers - 1 year.

Using this data we will predict:

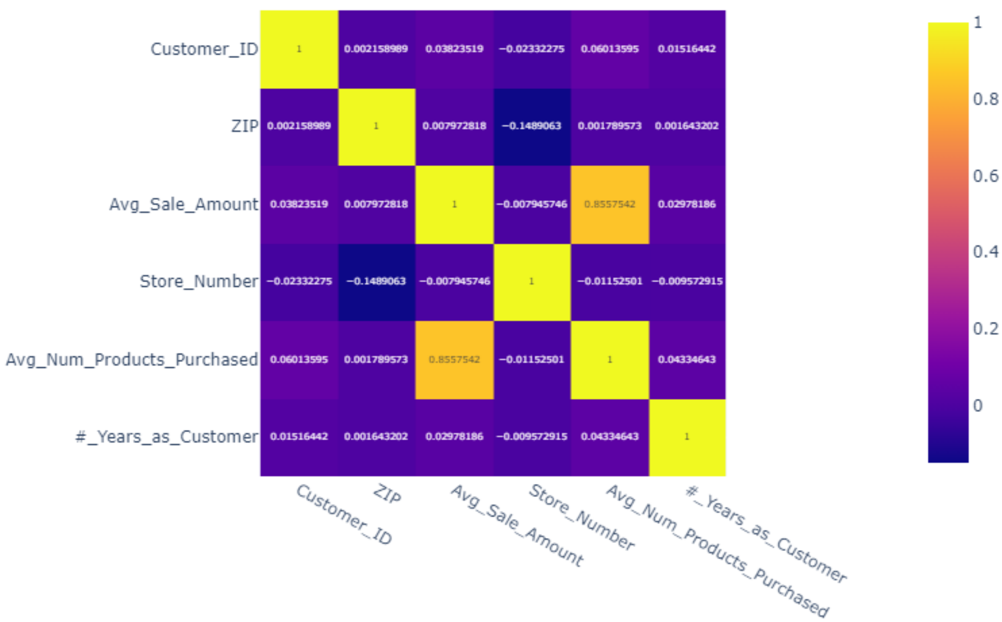
Expected profit that must be higher than $10,000;
Consider gross margin 50%;
Consider printing costs of $6.5 per catalog.

Step 2: Analysis, Modeling, and Validation

1. How and why did you select the predictor variables in your model? You must explain how the continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

I used a correlation matrix to decide which variables to grab to build a linear prediction.

Below we can see the plot and some considerations:

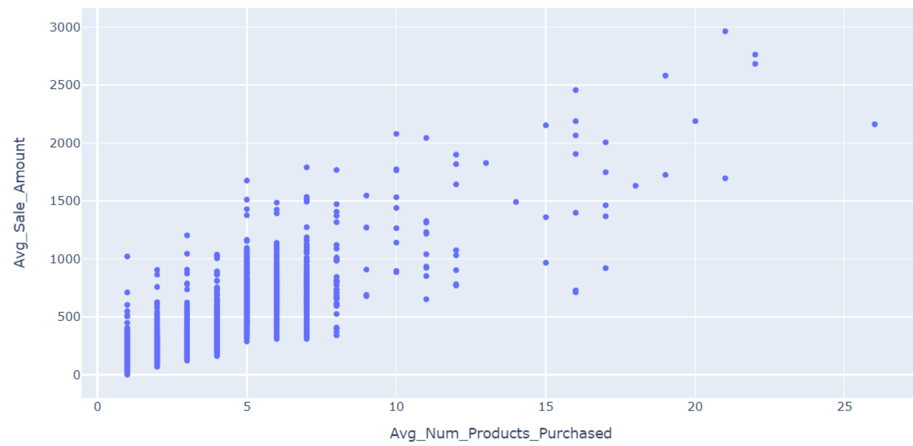


As we can see, there is a strong relationship between **Avg_Sale_Amount** and **Avg_Num_Products_Purchased**. So I decided to plot and show the linear relation between them.

In [13]:
```python
# using plotly

import plotly.express as px

fig = px.scatter(df, x="Avg_Num_Products_Purchased", y="Avg_Sale_Amount")
fig.show()
```



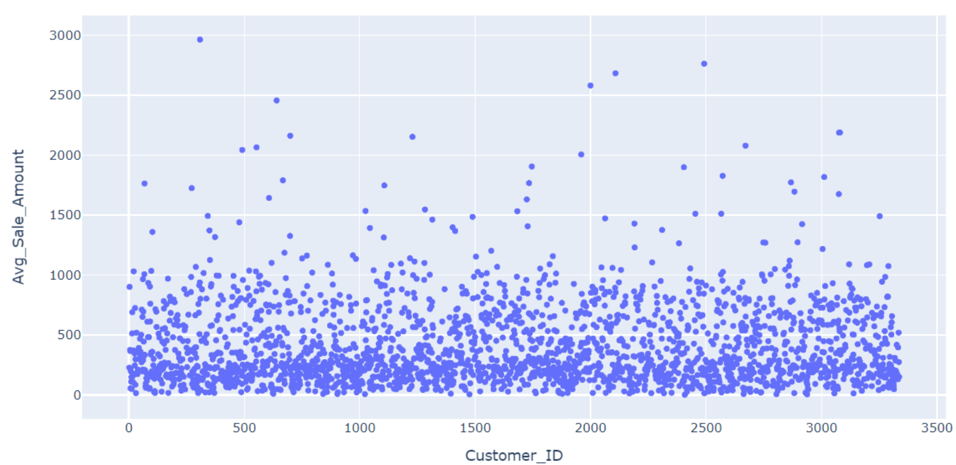
The scatter plot above shows this relationship and the importance of this feature to the model we want to build.

The other numerical variables have no important correlation with our target variable (**Avg_Sale_Amount**). So we must not use them.

I verified the relationship between them by scatterplot too:

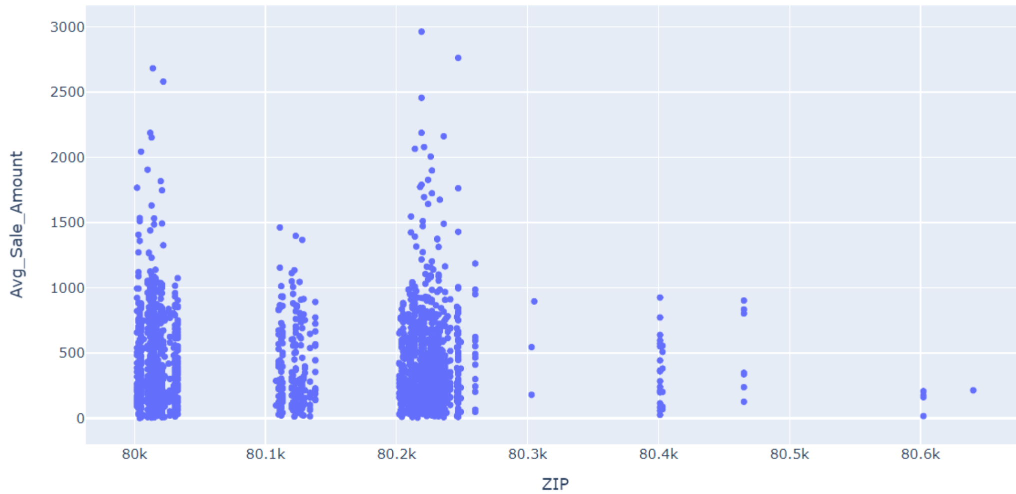
- **Customer_ID:**

In [14]:
```python
fig = px.scatter(df, x="Customer_ID", y="Avg_Sale_Amount")
fig.show()
```



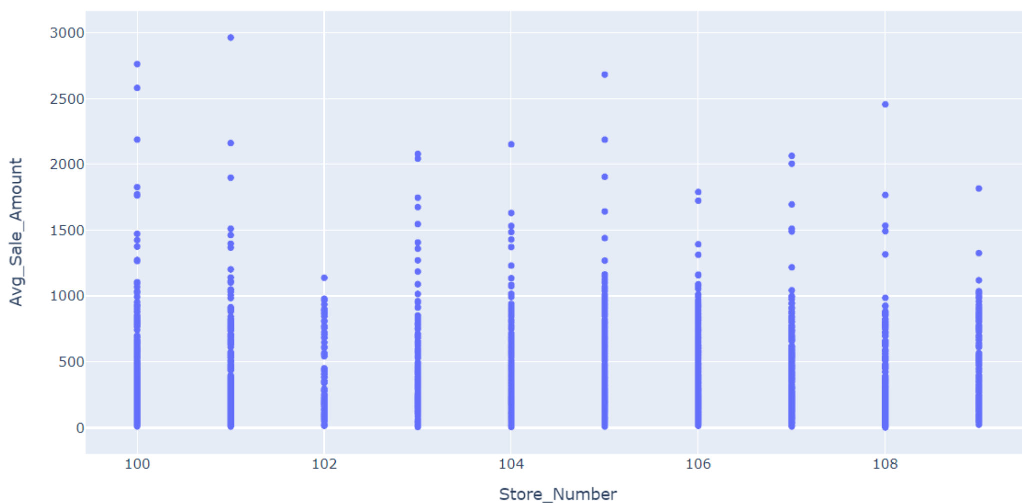
- ## ZIP:

In [15]:
```
fig = px.scatter(df, x="ZIP", y="Avg_Sale_Amount")
fig.show()
```



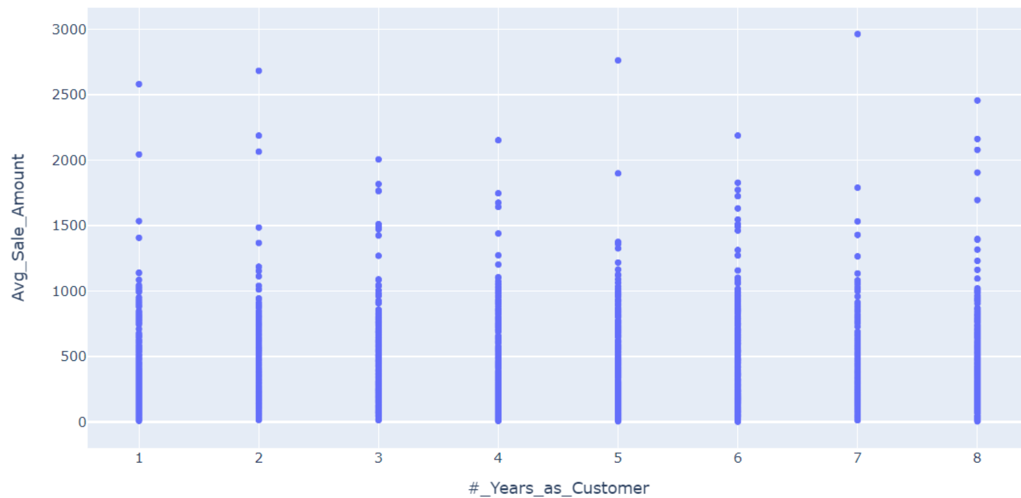
- ## Store_Number:

In [16]:
```
fig = px.scatter(df, x="Store_Number", y="Avg_Sale_Amount")
fig.show()
```



- **#_Years_as_Customer:**

```
In [53]: fig = px.scatter(df, x="#_Years_as_Customer", y="Avg_Sale_Amount")
         fig.show()
```



The small correlation is evidenced.

There is another variable **Customer_Segment,** that is a categorical variable, and I decided to evaluate de correlation between it and our target variable, but for that, I needed to dive into a dummiezation process:

Our data has some categorical variables and we must avaible they.

```
In [9]: df['Customer_Segment'].value_counts()

Out[9]: Store Mailing List              1108
        Loyalty Club Only                579
        Credit Card Only                 494
        Loyalty Club and Credit Card     194
        Name: Customer_Segment, dtype: int64
```

```
In [10]: list_to_convert = ['Customer_Segment']

         df_with_dummies = pd.get_dummies(df, columns = list_to_convert)

         df_with_dummies.head()

Out[10]:
```

vg_Num_Products_Purchased	#_Years_as_Customer	Customer_Segment_Credit Card Only	Customer_Segment_Loyalty Club Only	Customer_Segment_Loyalty Club and Credit Card	Customer_Segment_Store Mailing List
1	6	0	0	0	1
1	6	0	0	0	1
1	3	0	0	0	1
1	6	0	0	0	1
1	2	0	0	0	1

After including dummies I dropped from the data frame:

'Name'
'Customer_ID'
'Address'
'City'
'State'
'ZIP'
'Store_Number'
'#_Years_as_Customer'
'Responded_to_Last_Catalog'

Finally we obtain our data frame:

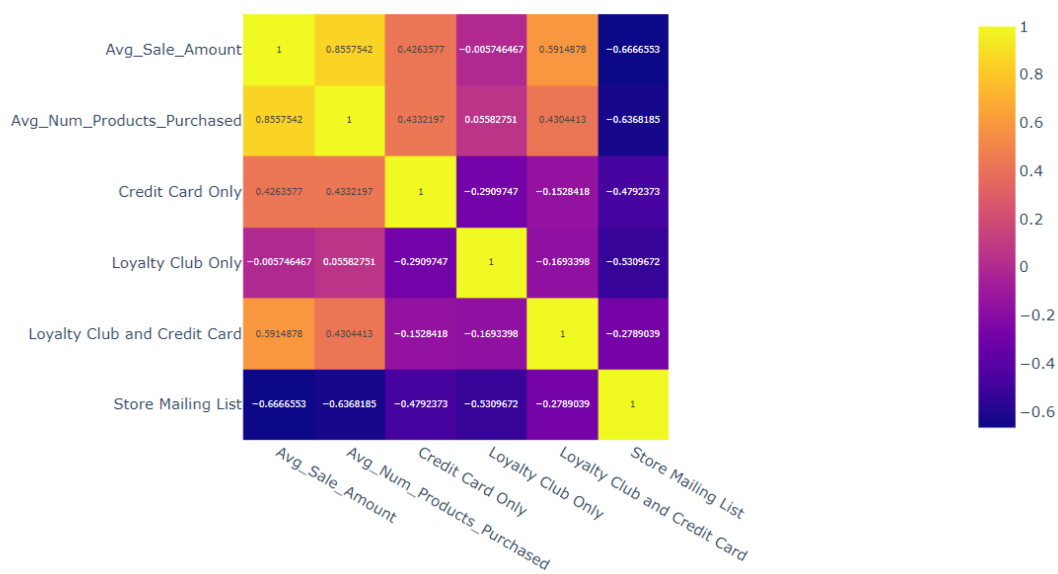
In [17]:
```
#drop the columns that we don't need in linear regression model

customers_after = df_with_dummies.drop(['Name','Customer_ID','Address','City','State','ZIP','Store_Number','#_Years_as_Customer',
customers_after.head()
```

Out[17]:

	Avg_Sale_Amount	Avg_Num_Products_Purchased	Credit Card Only	Loyalty Club Only	Loyalty Club and Credit Card	Store Mailing List
0	227.90	1	0	0	0	1
1	55.00	1	0	0	0	1
2	212.57	1	0	0	0	1
3	195.31	1	0	0	0	1
4	110.55	1	0	0	0	1

And then the final correlation matrix to the model:



2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.

I didn't use Alterix for these project, so I used ***sklearn*** to build a model.

I will show the process:

- **split input and output variables as X and Y:**

Building the model:

In [22]:
```python
#split input and output variables as X and Y

X = customers_after.drop('Avg_Sale_Amount',axis = 1)
Y = customers_after[['Avg_Sale_Amount']]
```

In [23]:
```python
X.head()
```

Out[23]:

	Avg_Num_Products_Purchased	Credit Card Only	Loyalty Club Only	Loyalty Club and Credit Card	Store Mailing List
0	1	0	0	0	1
1	1	0	0	0	1
2	1	0	0	0	1
3	1	0	0	0	1
4	1	0	0	0	1

In [24]:
```python
Y.head()
```

Out[24]:

	Avg_Sale_Amount
0	227.90
1	55.00
2	212.57
3	195.31
4	110.55

- **split train and test data frames, applying the model and getting the intercept and coefficients:**

In [25]:
```python
from sklearn.model_selection import train_test_split

X_train, X_test, Y_train, Y_test = train_test_split(X,Y,test_size=0.2,random_state=1)
regression_model = LinearRegression()
regression_model.fit(X_train, Y_train)
```

Out[25]:
```
LinearRegression()
```

In [26]:
```python
#grab the intercept and the coef
intercept = regression_model.intercept_[0]
intercept
```

Out[26]:
```
289.8075462158939
```

In [27]:
```python
print('The intercept: {:.4}'.format(intercept))
```

```
The intercept: 289.8
```

In [28]:
```python
#loop through the dictionary and print the data
for coef in zip(X.columns, regression_model.coef_[0]):
    print('Coefficient for {} is {:.4}'.format(coef[0],coef[1]))
```

```
Coefficient for Avg_Num_Products_Purchased is 62.74
Coefficient for Credit Card Only is 36.28
Coefficient for Loyalty Club Only is -120.6
Coefficient for Loyalty Club and Credit Card is 308.2
Coefficient for Store Mailing List is -223.8
```

So, the formula is

Avg_Sale_Amount = 289.8 + (62.74 * Avg_Num_Products_Purchased) + (36.28 * If Credit Card Only) + (-120.6 * If Loyalty Club Only) + (308.2 * If Loyalty Club and Credit Card) + (-223.8 * If Store Mailing List)

Implemented predict:

```
In [29]: Y_pred = regression_model.predict(X_test)
```

```
In [30]: Y_pred
```

```
Out[30]: array([[ 420.1433135 ],
                [ 128.74416024],
                [ 974.41858518],
                [ 231.91628333],
                [ 128.74416024],
                [ 545.62800029],
                [ 577.05927681],
                [ 482.8856569 ],
                [1099.90327196],
                [ 974.41858518],
                [ 482.8856569 ],
                [ 254.22884703],
                [ 191.48650363],
                [ 294.65862672],
                [ 128.74416024],
                [ 545.62800029],
                [ 974.41858518],
                [ 231.91628333],
                [ 231.91628333],
```

Evaluation:

```
In [34]: from sklearn.metrics import r2_score

         r2 = r2_score(Y_test, Y_pred)

         print("R-squared score:", r2)

R-squared score: 0.8464770964994399
```

R-squared scored near to 1.

```
In [33]: # Get p-values for the coefficients
p_values = model.pvalues

print(p_values)
```

```
const                             5.381494e-227
Avg_Num_Products_Purchased        7.989877e-312
Credit Card Only                   2.981342e-38
Loyalty Club Only                  7.002317e-36
Loyalty Club and Credit Card      1.606332e-237
Store Mailing List                1.100780e-173
dtype: float64
```

P-values are much smaller than 0.05.

3. What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal (ex: 1.28)

Avg_Sale_Amount = 289.8 + (62.74 * Avg_Num_Products_Purchased) + (36.28 * If Credit Card Only) + (-120.6 * If Loyalty Club Only) + (308.2 * If Loyalty Club and Credit Card) + (-223.8 * If Store Mailing List)

Step 3: Presentation/Visualization

1. What is your recommendation? Should the company send the catalog to these 250 customers?

    Applying the model and according to the outputs from the model, the answer is Yes, We should send the catalogs to the new clients, ‘cause there is a profit higher than $10,000.

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

    I used Python to evaluate the situation, First choose the target variable Avg_Sales_Amount, because we must predict the profit to send the catalogs.

Then the analyses with scatterplots and linear regression confirm that predictor variables have low P-values. Next, with predicted values, multiply Score_Yes by Sale_Amount and then Gross margin percentage.

3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?

```
In [47]: df_drop['New_Predicted_Sales'] = df_drop['Predicted_Sales']*df_drop['Score_Yes']
         df_drop.head()
Out[47]:
```

	Avg_Num_Products_Purchased	Score_No	Score_Yes	Credit Card Only	Loyalty Club Only	Loyalty Club and Credit Card	Store Mailing List	Predicted_Sales	New_Predicted_Sales
0	3	0.694964	0.305036	0	1	0	0	357.42	109.025898
1	6	0.527275	0.472725	0	0	1	0	974.44	460.641698
2	7	0.421118	0.578882	0	1	0	0	608.38	352.180140
3	2	0.694862	0.305138	0	1	0	0	294.68	89.918010
4	4	0.612294	0.387706	0	1	0	0	420.16	162.898492

```
In [48]: total_sales = df_drop['New_Predicted_Sales'].sum()
         total_sales
Out[48]: 46952.4566095738

In [49]: total_cost = 6.50 * 250
         total_cost
Out[49]: 1625.0

In [50]: gross_margin = 0.5 * total_sales
         gross_margin
Out[50]: 23476.2283047869

In [51]: profit = gross_margin - total_cost
         profit
Out[51]: 21851.2283047869
```

The final profit found is the double amount expected as a minimum condition to send catalogs to the new customers: $21,851.228.