Analyzing Amazon Sales Data Using Linear Regression Techniques to predict Revenue

Jidapa Pooljan

Digital Business Transformation, College of Innovation,

Thammasat University, Thailand

August 2024

Abstract:

This study presents a multiple linear regression model developed to predict total revenue from Amazon sales data, focusing on unit price and unit cost as primary features. The dataset underwent thorough preprocessing, including feature selection through correlation analysis, to ensure the inclusion of significant predictors. A pipeline was created, incorporating scaling and regression, and the model was trained and tested using a 90/10 train-test split. The model's performance was evaluated with a model evaluation R-squared value of 0.7451, indicating a strong relationship between the selected features and total revenue. The results demonstrate the model's potential in forecasting revenue, offering actionable insights for optimizing pricing and cost strategies in e-commerce. Future work is recommended to enhance predictive accuracy through additional feature engineering and more advanced modeling techniques.

Keywords: Amazon Sales Dataset, Linear Regression, Feature Selection, Revenue Prediction, E-commerce

……

Introduction

Problem Statement:

During and after the pandemic, the growth of e-commerce has been rapid and substantial. To continue serving customers effectively and sustain revenue growth while maintaining a competitive edge, e-commerce companies need to predict revenue and forecast sales accurately. This capability will enable support for business growth especially online store (Kumar, 2023). Effective of revenue forecasting can significantly enhance business operations and profitability

Solution:

Using historical sales data and linear regression model, can develop pricing strategies and adjust cost rates (Pan, 2024). Multiple linear regression can be used to predict company revenue and determine the impact of related variables (Pahmi, 2018). Also, in this project use Multiple linear regression to predict total revenue with impact unit cost and unit price data.

Benefit: Accurate revenue predictions to improve better inventory management, pricing strategies, and overall business planning, leading to sustained revenue growth and increased efficiency and profitability.

Objective:

- To select related features or variables to use in Linear Regression model
- Apply linear regression model to predict total revenue from unit price and unit cost.

Literature review

Feature selection is a critical preprocessing step in building effective linear regression models. By carefully selecting relevant features, practitioners can enhance model performance, interpretability, and computational efficiency. Including irrelevant or redundant features can lead to overfitting, where a model becomes overly complex and performs poorly on unseen data.

To mitigate these challenges, various feature selection techniques have been developed. Statistical methods such as correlation analysis and stepwise selection can identify features that are strongly associated with the target variable (James et al., 2013). Additionally, filter methods, which assess feature importance independently of the model, can be employed to rank features based on their statistical properties (Chandrashekar & Sahin, 2014).

Visual tools like correlation matrices can provide valuable insights into feature relationships, aiding in the selection process (Wagavkar, 2023). By understanding these interconnections, practitioners can identify potential multicollinearity issues and make informed decisions about feature inclusion.

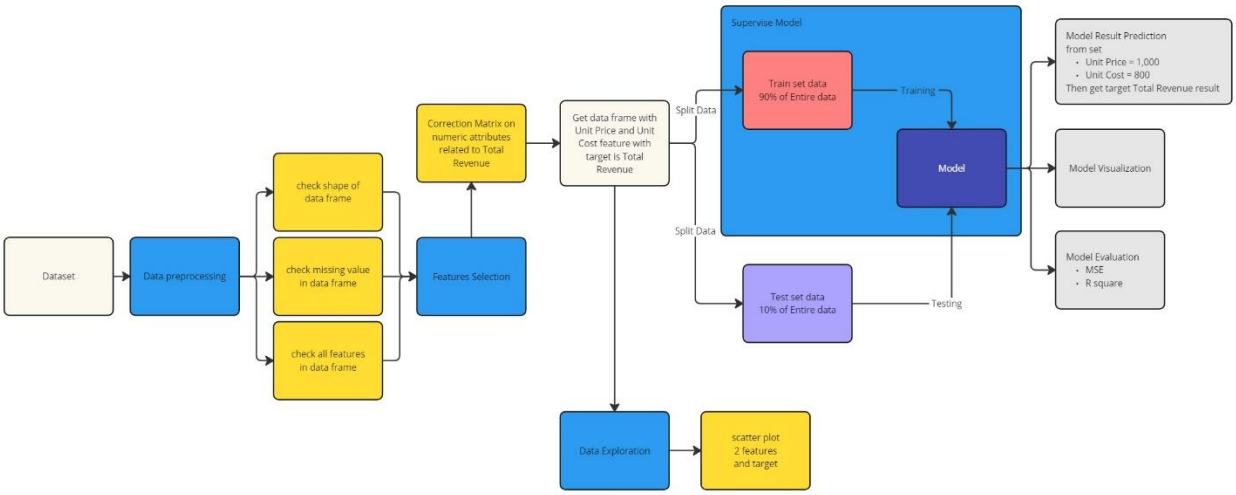
Ultimately, the choice of feature selection method depends on the specific dataset, modeling goals, and computational resources available. A combination of techniques often yields the best results.

To apply linear regression model in this study. Linear regression is a statistical method used to estimate the relationship between a dependent variable and one or more independent variables (IBM, n.d.). In the context of business, it can be applied to predict total revenue based on factors such as unit price (Cooper & Schindler, 2014). By analyzing historical data, businesses can identify patterns between unit price and total revenue, allowing for the creation of a predictive model (Chatterjee & Hadi, 2015). This model can then be used to estimate potential revenue outcomes under different pricing scenarios.

It's important to note that linear regression assumes a linear relationship between variables, which may not always hold true in real-world situations (Hair, Black,

Babin, & Anderson, 2014). However, it remains a valuable tool for initial analysis and can provide valuable insights into revenue trends.

Conceptual Model



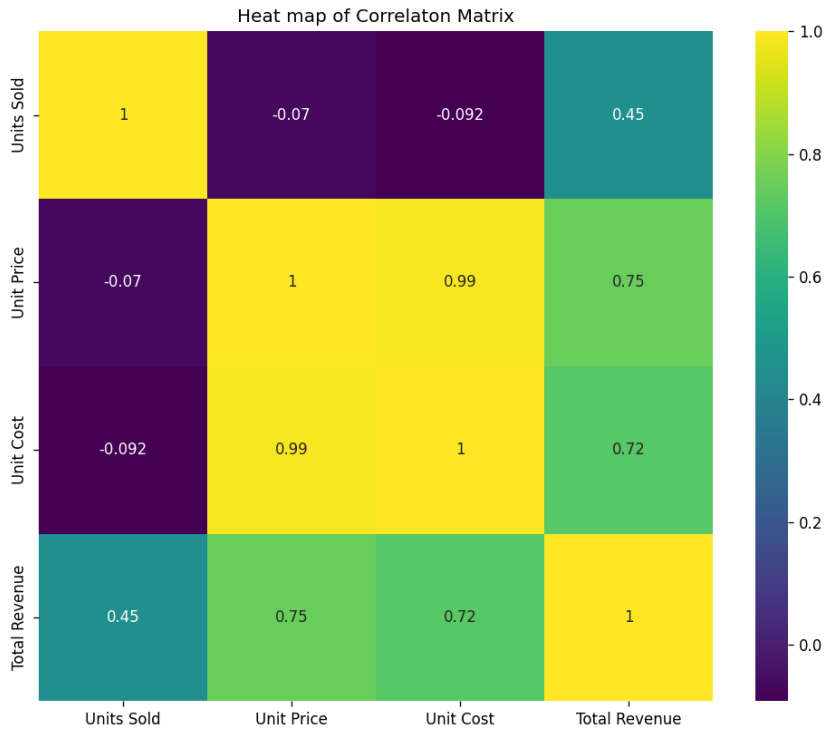
Conceptual Model

Data pre-processing

1. Check shape of data frame:
    1. rows = 100 records
    2. columns = 14 attributes/feature
2. Check each of 14 features in data frame following:
    1. Region
    2. Country
    3. Item Type
    4. Sales Channel
    5. Order Priority
    6. Order Date
    7. Order ID
    8. Ship Date
    9. Units Sold
    10. Unit Price
    11. Unit Cost
    12. Total Revenue
    13. Total Cost
    14. Total Profit
3. Find missing value each attribute/feature, results there no missing value:
    1. Region 0
    2. Country 0
    3. Item Type 0
    4. Sales Channel 0
    5. Order Priority 0
    6. Order Date 0
    7. Order ID 0
    8. Ship Date 0
    9. Units Sold 0
    10. Unit Price 0
    11. Unit Cost 0
    12. Total Revenue 0
    13. Total Cost 0
    14. Total Profit 0

Features selection

1. from numeric features from below:
    1. Order ID
    2. Units Sold
    3. Unit Price
    4. Unit Cost
    5. Total Revenue
    6. Total Cost
    7. Total Profit
2. Select related features to Total Revenue
3. Use Correlation Matrix with features below:
    1. Units Sold
    2. Unit Price
    3. Unit Cost
    4. Total Revenue

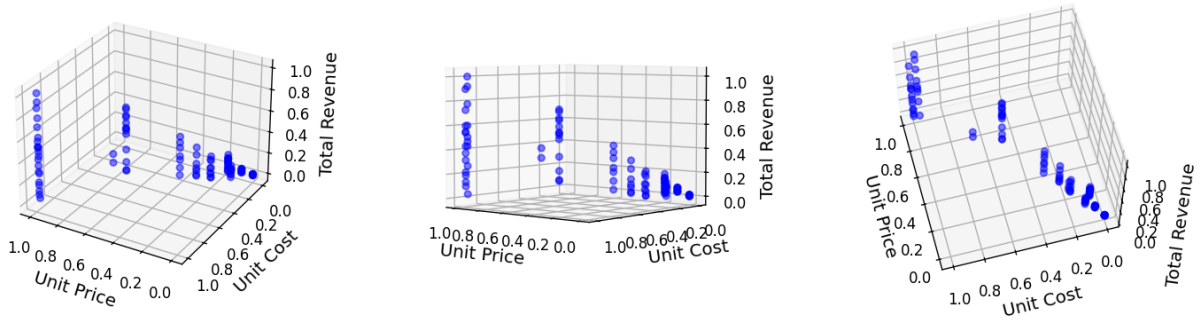


Correlation Matrix for feature selection

4. Get features for train model that are
    1. Unit Price with correlation coefficient = 0.75
    2. Unit Cost with correlation coefficient = 0.72

Data Exploration

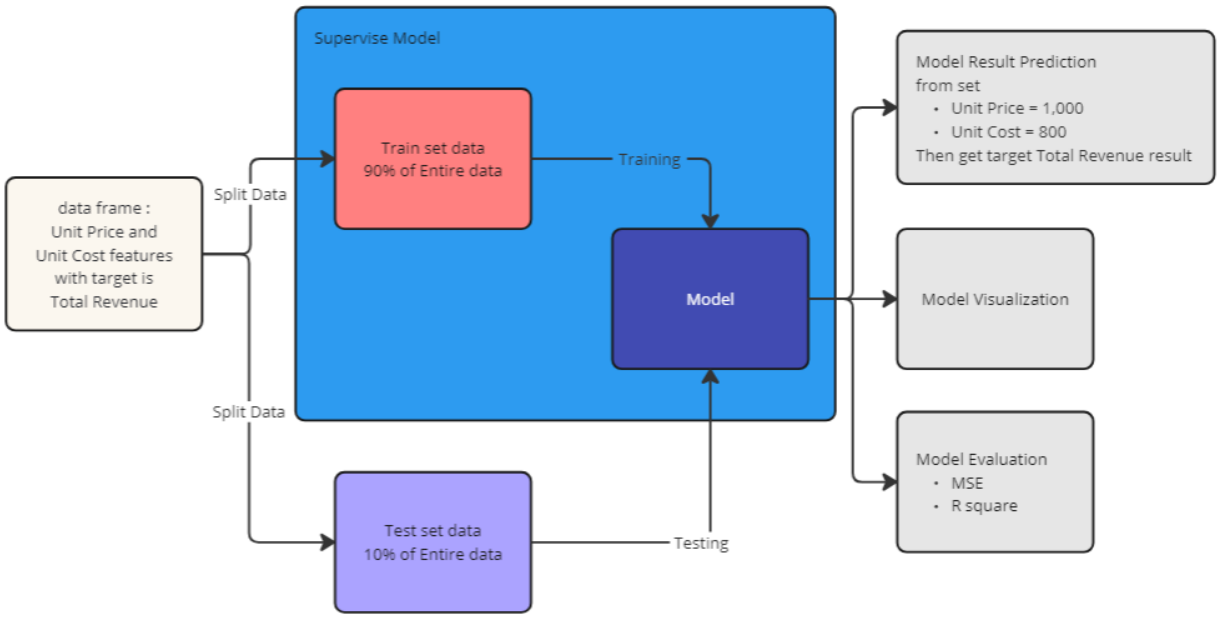
1. scatter plot with 2 features Unit Price and Unit Cost with target Total Revenue on scaled.



Data exploration: scatter plot with Unit Price, Unit Cost and Total Revenue

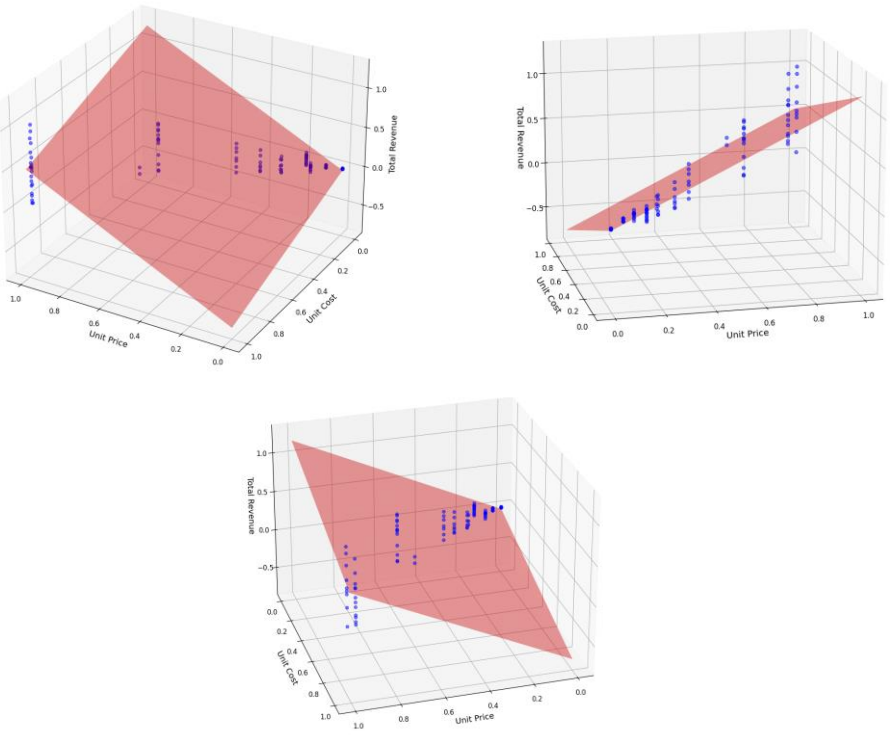
Data mining processing

1. Model Diagram



Model Diagram Display

2. Model Result Prediction
   If Features Unit Price = 1,000.00 dollar and Features Unit Cost = 800.00 dollar then predicted total Revenue = 4,441,473.18 dollar
3. Model Evaluation
    1. Mean Squared Error: 381,021,368,383.1143
    2. R-squared: 0.7451
4. Model Visualization



scatter plot with Unit Price, Unit Cost and Total Revenue
with prediction surface mesh

Conclusion

The linear regression model demonstrated a strong ability to predict total revenue from Amazon sales data with the features selected unit price and unit cost were confirmed as significant predictors of total revenue. The correlation analysis supported their relevance and contributed to its predictive power, as evidenced by an R-squared value of 0.7451. This indicates that approximately 74.51% of the variance in total revenue can be explained by the model using unit price and unit cost as predictors. While the model achieved a satisfactory level of accuracy, the Mean Squared Error (MSE) of approximately 381 billion suggests that there is still some part for improvement further, particularly in reducing prediction error.

The model provides valuable insights into how pricing and cost management impact revenue in an e-commerce for Amazon sales. These insights can be instrumental in optimizing pricing strategies and managing costs effectively to maximize revenue.

Suggestion

The current sample size of 100 may be too small, risking overfitting and consider gathering more data to improve the model's robustness. Experiment with models like decision trees or non-linear relationships and improve predictive performance more.

References

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16-28.

Chatterjee, S., & Hadi, A. S. (2015). *Regression analysis by example* (5th ed.). Wiley.

Cooper, D. R., & Schindler, P. S. (2014). *Business research methods* (12th ed.). McGraw-Hill.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157-1182.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Pearson.

IBM. (n.d.). Linear regression. Retrieved from https://www.ibm.com/topics/linear-regression

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (7th ed.). Springer.

Mithilesh9. (2020). *Amazon Sales Data Analysis* [Dataset]. Kaggle. https://www.kaggle.com/datasets/mithilesh9/amazon-sales-data-analysis

Wagavkar, S. (2023, March 17). Introduction to the Correlation Matrix. *Built In*. Retrieved from https://builtin.com/data-science/correlation-matrix

Appendix

Python Code:

https://www.kaggle.com/code/jidapapooljan/linear-regression

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import  StandardScaler, MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.pipeline import Pipeline
from sklearn.metrics import mean_squared_error, r2_score
from config.settings import csv_file_paths


# Load data
df = pd.read_csv(csv_file_paths['AmazonSalesData'])

# ------------------------------------------------------------------------
# data pre-processing
# Check shape of data set
print(df.shape)

# Check columns of data set
print(df.columns)

# Check missing value (Null data)
null = df.isnull().sum()
print(null)
print('\n')

# Feature Selection
all_features = ['Region', 'Country', 'Item Type', 'Sales Channel', 'Order
Priority',
                'Order Date', 'Order ID', 'Ship Date', 'Units Sold', 'Unit
Price',
                'Unit Cost', 'Total Revenue', 'Total Cost', 'Total Profit']

categorical_features = df.select_dtypes(include=[object])
numerical_features = df.select_dtypes(include=[np.number])
print(categorical_features.columns)
print(numerical_features.columns)


# Check Correlation Matrix for numerical some features
corr_matrix = df[['Units Sold', 'Unit Price', 'Unit Cost', 'Total
Revenue']].corr()
```

```
# Plot heat map to checked Correlation
plt.figure(figsize=(10,8))
sns.heatmap(corr_matrix, annot=True, cmap='viridis')
plt.title('Heat map of Correlation Matrix')
plt.show()

# Defines Parameters to use in model
feature = ['Unit Price', 'Unit Cost']
X = df[feature].values.reshape(-1, len(feature))
y = df['Total Revenue'].values.reshape(-1, 1)

# Data Exploration
# Fit the scaler to X
scaler_X = MinMaxScaler()
X_scaled = scaler_X.fit_transform(X)

# Fit the scaler to y
scaler_y = MinMaxScaler()
y_scaled = scaler_y.fit_transform(y).flatten()

fig = plt.figure(figsize=(12,4))
ax1 = fig.add_subplot(131, projection='3d')
ax2 = fig.add_subplot(132, projection='3d')
ax3 = fig.add_subplot(133, projection='3d')
axes = [ax1, ax2, ax3]
for ax in axes:
    ax.scatter(xs=X_scaled[:,0], ys=X_scaled[:,1], zs=y_scaled, c='b',
alpha=0.5)
    ax.set_xlabel('Unit Price', fontsize=12)
    ax.set_ylabel('Unit Cost', fontsize=12)
    ax.set_zlabel('Total Revenue', fontsize=12)
    ax.locator_params(nbins=6, axis='x')
    ax.locator_params(nbins=5, axis='y')
    ax.locator_params(nbins=5, axis='z')
ax1.view_init(elev=28, azim=120)
ax2.view_init(elev=4, azim=130)
ax3.view_init(elev=60, azim=165)
fig.subplots_adjust(left=0.05, right=0.95, bottom=0.05, top=0.95, wspace=0.5)
plt.show()
```

```
# Create Model Pipeline

model = Pipeline(steps=[
    ('scaler', StandardScaler()),
    ('regressor', LinearRegression())
])

# Spilt data set for training 90% and test 10%
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1,
random_state=42)

# Train Linear Regression Model
model.fit(X_train, y_train)

# Predict
y_pred = model.predict(X_test)
unit_price_pred = 1000
unit_cost_pred = 800
m_predict = model.predict([[unit_price_pred,unit_cost_pred]])  # Predict
print(f'If Features Unit Price = {unit_price_pred:,.2f} and Unit Cost =
{unit_cost_pred:,.2f} '
      f'then predicted Total Revenue = {m_predict[0][0]:,.2f}')
print("\n")

# Evaluation of Model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
r2_model = model.score(X_test, y_test)

print(f'Mean Squared Error: {mse:,.4f}')
print(f'R-squared: {r2:.4f}')
print(f'R-squared: {r2_model:.4f}')
```

```
# Model Visualization
# Data Exploration
fig = plt.figure(figsize=(20,15))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(xs=X_scaled[:,0], ys=X_scaled[:,1], zs=y_scaled, c='b', alpha=0.5)
ax.set_xlabel('Unit Price', fontsize=12)
ax.set_ylabel('Unit Cost', fontsize=12)
ax.set_zlabel('Total Revenue', fontsize=12)
ax.locator_params(nbins=6, axis='x')
ax.locator_params(nbins=6, axis='y')
ax.locator_params(nbins=7, axis='z')
ax.view_init(elev=28, azim=120)

X_surf = np.linspace(X_scaled[:,0].min(),X_scaled[:,0].max(),50)
y_surf = np.linspace(X_scaled[:,1].min(), X_scaled[:,1].max(), 50)
x_surf, y_surf = np.meshgrid(X_surf, y_surf)

# Inverse transform the grid points
grid_points = scaler_X.inverse_transform(np.c_[x_surf.ravel(),
y_surf.ravel()])

# Predicting the output using the trained model
z_surf = model.predict(grid_points).reshape(x_surf.shape)

# Scaling the predicted values
z_surf_scaled = scaler_y.transform(z_surf.reshape(-1,
1)).reshape(z_surf.shape)

# Plotting the surface
ax.plot_surface(x_surf, y_surf, z_surf_scaled, rstride=10, cstride=10,
color='r', alpha=0.4, edgecolor='none')

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

Example of Data Set

Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit
Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669165933	6/27/2010	9925	255.28	159.42	2533654	1582243.5	951410.5
Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15/2012	2804	205.7	117.11	576782.8	328376.44	248406.36
Europe	Russia	Office Supplies	Offline	L	5/2/2014	341417157	5/8/2014	1779	651.21	524.96	1158502.59	933903.84	224598.75
Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	6/20/2014	514321792	7/5/2014	8102	9.33	6.92	75591.66	56065.84	19525.82
Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2/1/2013	115456712	2/6/2013	5062	651.21	524.96	3296425.02	2657347.52	639077.5
Australia and Oceania	Solomon Islands	Baby Food	Online	C	2/4/2015	547995746	2/21/2015	2974	255.28	159.42	759202.72	474115.08	285087.64
Sub-Saharan Africa	Angola	Household	Offline	M	4/23/2011	135425221	4/27/2011	4187	668.27	502.54	2798046.49	2104134.98	693911.51
Sub-Saharan Africa	Burkina Faso	Vegetables	Online	H	7/17/2012	871543967	7/27/2012	8082	154.06	90.93	1245112.92	734896.26	510216.66
Sub-Saharan Africa	Republic of the Congo	Personal Care	Offline	M	7/14/2015	770463311	8/25/2015	6070	81.73	56.67	496101.1	343986.9	152114.2
Sub-Saharan Africa	Senegal	Cereal	Online	H	4/18/2014	616607081	5/30/2014	6593	205.7	117.11	1356180.1	772106.23	584073.87
Asia	Kyrgyzstan	Vegetables	Online	H	6/24/2011	814711606	7/12/2011	124	154.06	90.93	19103.44	11275.32	7828.12
Sub-Saharan Africa	Cape Verde	Clothes	Offline	H	8/2/2014	939825713	8/19/2014	4168	109.28	35.84	455479.04	149381.12	306097.92
Asia	Bangladesh	Clothes	Online	L	1/13/2017	187310731	3/1/2017	8263	109.28	35.84	902980.64	296145.92	606834.72
Central America and the Caribbean	Honduras	Household	Offline	H	2/8/2017	522840487	2/13/2017	8974	668.27	502.54	5997054.98	4509793.96	1487261.02
Asia	Mongolia	Personal Care	Offline	C	2/19/2014	832401311	2/23/2014	4901	81.73	56.67	400558.73	277739.67	122819.06
Europe	Bulgaria	Clothes	Online	M	4/23/2012	972292029	6/3/2012	1673	109.28	35.84	182825.44	59960.32	122865.12
Asia	Sri Lanka	Cosmetics	Offline	M	11/19/2016	419123971	12/18/2016	6952	437.2	263.33	3039414.4	1830670.16	1208744.24
Sub-Saharan Africa	Cameroon	Beverages	Offline	C	4/1/2015	519820964	4/18/2015	5430	47.45	31.79	257653.5	172619.7	85033.8
Asia	Turkmenistan	Household	Offline	L	12/30/2010	441619336	1/20/2011	3830	668.27	502.54	2559474.1	1924728.2	634745.9
Australia and Oceania	East Timor	Meat	Online	L	7/31/2012	322067916	9/11/2012	5908	421.89	364.69	2492526.12	2154588.52	337937.6
Europe	Norway	Baby Food	Online	L	5/14/2014	819028031	6/28/2014	7450	255.28	159.42	1901836	1187679	714157
Europe	Portugal	Baby Food	Online	H	7/31/2015	860673511	9/3/2015	1273	255.28	159.42	324971.44	202941.66	122029.78
Central America and the Caribbean	Honduras	Snacks	Online	L	6/30/2016	795490682	7/26/2016	2225	152.58	97.44	339490.5	216804	122686.5
Australia and Oceania	New Zealand	Fruits	Online	H	9/8/2014	142278373	10/4/2014	2187	9.33	6.92	20404.71	15134.04	5270.67
Europe	Moldova	Personal Care	Online	L	5/7/2016	740147912	5/10/2016	5070	81.73	56.67	414371.1	287316.9	127054.2
Europe	France	Cosmetics	Online	H	5/22/2017	898523128	6/5/2017	1815	437.2	263.33	793518	477943.95	315574.05
Australia and Oceania	Kiribati	Fruits	Online	M	10/13/2014	347140347	11/10/2014	5398	9.33	6.92	50363.34	37354.16	13009.18
Sub-Saharan Africa	Mali	Fruits	Online	L	5/7/2010	686048400	5/10/2010	5822	9.33	6.92	54319.26	40288.24	14031.02
Europe	Norway	Beverages	Offline	C	7/18/2014	435608613	7/30/2014	5124	47.45	31.79	243133.8	162891.96	80241.84
Sub-Saharan Africa	The Gambia	Household	Offline	L	5/26/2012	886494815	6/9/2012	2370	668.27	502.54	1583799.9	1191019.8	392780.1
Europe	Switzerland	Cosmetics	Offline	M	9/17/2012	249693334	10/20/2012	8661	437.2	263.33	3786589.2	2280701.13	1505888.07
Sub-Saharan Africa	South Sudan	Personal Care	Offline	C	12/29/2013	406502997	1/28/2014	2125	81.73	56.67	173676.25	120423.75	53252.5
Australia and Oceania	Australia	Office Supplies	Online	C	10/27/2015	158535134	11/25/2015	2924	651.21	524.96	1904138.04	1534983.04	369155
Asia	Myanmar	Household	Offline	H	1/16/2015	177713572	3/1/2015	8250	668.27	502.54	5513227.5	4145955	1367272.5
Sub-Saharan Africa	Djibouti	Snacks	Online	M	2/25/2017	756274640	2/25/2017	7327	152.58	97.44	1117953.66	713942.88	404010.78
Central America and the Caribbean	Costa Rica	Personal Care	Offline	L	5/8/2017	456767165	5/21/2017	6409	81.73	56.67	523807.57	363198.03	160609.54
Middle East and North Africa	Syria	Fruits	Online	L	11/22/2011	162052476	12/3/2011	3784	9.33	6.92	35304.72	26185.28	9119.44
Sub-Saharan Africa	The Gambia	Meat	Online	M	1/14/2017	825304400	1/23/2017	4767	421.89	364.69	2011149.63	1738477.23	272672.4
Asia	Brunei	Office Supplies	Online	L	4/1/2012	320009267	5/8/2012	6708	651.21	524.96	4368316.68	3521431.68	846885
Europe	Bulgaria	Office Supplies	Online	M	2/16/2012	189965903	2/28/2012	3987	651.21	524.96	2596374.27	2093015.52	503358.75
Sub-Saharan Africa	Niger	Personal Care	Online	H	3/11/2017	699285638	3/28/2017	3015	81.73	56.67	246415.95	170860.05	75555.9
Middle East and North Africa	Azerbaijan	Cosmetics	Online	M	2/6/2010	382392299	2/25/2010	7234	437.2	263.33	3162704.8	1904929.22	1257775.58
Sub-Saharan Africa	The Gambia	Cereal	Offline	H	6/7/2012	994022214	6/8/2012	2117	205.7	117.11	435466.9	247921.87	187545.03
Europe	Slovakia	Vegetables	Online	H	10/6/2012	759224212	11/10/2012	171	154.06	90.93	26344.26	15549.03	10795.23
Asia	Myanmar	Clothes	Online	H	11/14/2015	223359620	11/18/2015	5930	109.28	35.84	648030.4	212531.2	435499.2
Sub-Saharan Africa	Comoros	Cereal	Offline	H	3/29/2016	902102267	4/29/2016	962	205.7	117.11	197883.4	112659.82	85223.58
Europe	Iceland	Cosmetics	Online	C	12/31/2016	331438481	12/31/2016	8867	437.2	263.33	3876652.4	2334947.11	1541705.29
Europe	Switzerland	Personal Care	Online	M	12/23/2010	617667090	1/31/2011	273	81.73	56.67	22312.29	15470.91	6841.38
Europe	Macedonia	Clothes	Offline	C	10/14/2014	787399423	11/14/2014	7842	109.28	35.84	856973.76	281057.28	575916.48
Sub-Saharan Africa	Mauritania	Office Supplies	Offline	C	1/11/2012	837559306	1/13/2012	1266	651.21	524.96	824431.86	664599.36	159832.5