Grocery Sales Data

Data Source

<https://www.kaggle.com/codemysteries/salesdb>

Clean Data Profile

* 98,759 (98k +) unique customers residing in 96 different cities. How many states?
* 23 unique employees residing in 18 different cities (what is the most occurring city? May indicate headquarters of business) and state
* Employees were hired between February 2010 and February 2017
* All the customers and employees were within the US
* 452 unique products grouped into 20 unique categories
* 6,758,125 (6M +) sales transactions
* Sales transactions were from 1st January 2018 on 9th May 2018 (4 months)

Data Audit and Data Cleaning Treatments

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data Set | Feature | Data Quality Issue | Missing Data | | Treatment |
| Count | Proportion |
| Customer | MiddleInitial | Missing data | 977 | 0.01 | - |
| Employee | BirthDate | Non-standard format |  |  | Converted to standard datetime object |
|  | HireDate | Non-standard format |  |  | Converted to standard datetime object |
| Countries | CountryCode | Missing data | 1 | 0.00 | Imputed with ‘AV’ |
| Products | Resistant | Missing data | 140 | 0.31 | Replace with ‘Missing’ |
|  | IsAllergic | Missing data | 130 | 0.29 | Replace with ‘Missing’ |
|  | VitalityDays, | Spelt wrongly, missing data, | 274 | 0.61 | Renamed feature, replaced missing values with 0 |
|  | ProductName | Double spaces |  |  | Stripped double spaces |
|  | Price | Non-standard format |  |  | Generated new price list per category |
| Category | CategoryName | Categories were not exhaustive |  |  | Created new categories to fully categorize all products |
| Sales | Discount | Missing data | 5,406,931 | 0.80 | Replaced missing values with 0 |
|  | SalesDate | Non-standard format, missing data | 67,526 | 0.10 | Converted to standard datetime object, Filled missing data with backpropagation and sorting the dates |
|  | SalesID |  |  |  | Renumbered SalesID to reflect sequential ordering of transactions |
|  | TotalPrice | Non-standard format |  |  | Converted values to numeric values |

Feature Engineering

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data Set | Raw Feature | Data Type | Derived Feature | Definition/ Description |
|
| Employee | BirthDate | Datetime | derived.biological\_age |  |
|  |  | Datetime | derived.biological\_age.group |  |
|  | HireDate | Datetime | derived.service\_age.yrs |  |
|  |  | Datetime | derived.service\_age.yrs.group |  |
| City | CityName | String | derived.state |  |
|  |  | String | derived.region |  |
| Sales, Product | Quantity, Price, CategoryName | Integer, Numeric, String | derived.total\_spend.alltime |  |
| Sales | SalesDate | Datetime | derived.salesdate\_day\_of\_week |  |
|  |  |  | derived.salesdate\_time of\_day |  |
|  |  |  | derived.salesdate\_week\_of\_month |  |
|  |  |  | derived.customer\_type.frequency | first-time, repeat |
|  |  |  | derived.customer\_type.value | Loyal, discount, churning, high-value, low-value (since we don’t have date of registration we can’t be too quick to judge with this) |
|  |  |  | derived.unique\_visits.count.alltime |  |
|  |  |  | derived.total\_products\_bought.count.alltime | sum(quantity of products per visit) |
|  |  |  | derived.average\_products\_bought.count.alltime | derived.total\_products\_bought.count.alltime /derived.unique\_visits.count |
|  |  |  | derived.category\_with\_most\_spend.amount.alltime |  |
|  |  |  | derived.average\_spend.amount.alltime | derived.total\_spend.alltime /derived.unique\_visits.count |
|  |  |  | derived.lowest\_one-time\_spend.amount.alltime | Min(total spend per visit) |
|  |  |  | derived.highest\_one\_time\_spend.amount.alltime | Max(total spend per visit) |
|  |  |  | derived.standard\_deviation\_between\_visits.alltime | Std(between visits) |
|  |  |  | derived.standard\_deviation\_between\_spends.amount.alltime | Std(between all spends per visit) |
|  |  |  | derived.total\_discounts.count.alltime |  |
|  |  |  | derived.total\_discounts.amount.alltime |  |

* Is Class in Products data correlated with Quantity in Sales data?
* Get metrics for online commerce?
* Check category of products with VitalityDays, Class, Resistant and IsAllergic?

Business Topline Metrics

Customers

* Total number of customers/ users
  + First-time vs Repeat customers
  + Customer growth trend
* Demographic of customers

Employees

* Total number of employees
  + Employee growth trend
* Demographic of employees
* Gender distribution of employees
* Biological age distribution
* Service age distribution

Products

* Number of categories
* Average category unit price

Sales

* Total revenue (volume)
  + Total revenue growth trend month-on-month (volume)
  + Average revenue month-on-month
* Total unit of items bought per category
* Day of week, Time of day, Week of month with most frequent:
  + Sales/Revenue (volume and value)
  + Discounts (volume and value)

Customer Usage Metrics

* Average order size month-on-month
* Most popular category patronized per month
* Average order spend month-on-month
* Total spend month-on-month overlaid with average order size all time
* Average discount received month-on-month

Employee Performance Metrics

* Total sales per employee
* Average sales per employee per month
* Employee with most customer spend each month