**Brainstorming Popular Marketing Modelling Techniques:**

* **Customer lifetime value** 
  + Customer\_id, information about the customer (ideally), transactions/purchase history/ads\_info/clicks/ lifetime\_value or infer lifetime value based on purchase history or change question (How much are they going to buy this year or month?)
  + 10-15K dataset size.
  + Basic information in the dataset, what you’ll be
* **Market Basket Analysis** - not a first capstone problem, more of a recommendation problem
* **Media Mix Modelling**
* **Multi-touch Attribution Modelling** (Which touchpoints matter the most) - feasible...each customer different touchpoints, how likely are they convert, which touchpoints most influence conversion. AdRoll datasets? Look for it.
* **Propensity modelling** (likelihood to convert) etc.

**Customer LifeTIme Value datasets:**

**Source 1: Squark Automotive Customer Lifetime Value dataset** <https://www.kaggle.com/arashnic/marketing-seris-customer-lifetime-value?select=squark_automotive_CLV_production_data.csv>

Goal: Predict automotive insurance customer lifetime value

Data Size: 1.5 MB

File 2: training dataset

Fields:

* 'Customer'
* 'State'
* 'Customer Lifetime Value'
* 'Response'
* 'Coverage',
* 'Education'
* 'Effective To Date'
* 'EmploymentStatus'
* 'Gender'
* 'Income'
* 'Location Code'
* 'Marital Status'
* 'Monthly Premium Auto',
* 'Months Since Last Claim'
* 'Months Since Policy Inception'
* ‘'Number of Open Complaints'
* 'Number of Policies'
* 'Policy Type'
* 'Policy'
* 'Renew Offer Type'
* 'Sales Channel'
* 'Total Claim Amount'
* 'Vehicle Class'
* 'Vehicle Size’

**Pros:**

* Lots of features to understand customer lifetime value.
* Lots of customer data (8K unique records)

**Cons:**

* No transaction wrangling or calculating RFM metrics to since CLV is already given to you

**Source 2: Online II UCI repository**

[**https://www.kaggle.com/mashlyn/online-retail-ii-uci**](https://www.kaggle.com/mashlyn/online-retail-ii-uci)

**RFM analysis with this data:**

[**https://www.kaggle.com/eminetolun/customer-segmentation-rfm/data**](https://www.kaggle.com/eminetolun/customer-segmentation-rfm/data)

**Size: 56.6 MB of non-null data**

**Fields:**

* **Invoice**
* **StockCode**
* **Description**
* **Quantity**
* **Invoice Date**
* **Price**
* **Customer ID**
* **Country**

**Unique Customer IDs: 5,942**

**Out of these 4,482 customers had more than one transaction and 1K+ had greater than 9 transactions.**

**Pros:**

* Must wrangle features to come up with RFM metrics (recency, frequency, monetary value) and calculate CLV prior to prediction
* 4K+ customers with repeat rx.

**Cons:**

* Not as many records of customer data
* Not as many features used to predict customer lifetime value

**Source 3: Brazilian E-Commerce dataset by Olist and Marketing Funnel dataset**

[https://www.kaggle.com/olistbr/brazilihttps://www.kaggle.com/arashnic/marketing-seris-customer-lifetime-value?select=squark\_automotive\_CLV\_training\_data.csvan-ecommerce/home?select=olist\_customers\_dataset.csv](https://www.kaggle.com/olistbr/brazilian-ecommerce/home?select=olist_customers_dataset.csv)

<https://www.kaggle.com/olistbr/brazilian-ecommerce/home?select=olist_customers_dataset.csv>

**Brazilian e-commerce dataset:** information of 100k orders, price, payment, freight performance, customer location, product attributes and finally reviews written by customers.

**DataSize:**

**Brazilian ecommerce dataset: 126.14 MB**

**Customer/Order data (12.1 MB)**

**CLV (customerID, PurchaseDate, PurchasePrice - multiple transactions)**

**Investigate datasets: Customer, Orders, Order items, Order Payments:**

**Check out the Orders dataset first -- investigate how many trx per customer ID**

**There are about 3K repeat trx.**

**Customer Fields: Check**

* Customer id, Customer unique Id, Customer zip code, Customer city, Customer state

**Geolocation Fields: Not interesting for CLV**

* Geolocation city, state, latitude, longitude

**Order Items Fields: Check**

* OrderID, OrderItemID, ProductID, SellerID, ShippingLimitDate, Price, Freight\_Value,

**Order Payments Fields: Check**

* Order\_Id, Payment\_Sequential, Payment\_Type, Payment\_Installments, Payment\_Value

**Order Review Fields: X**

* ReviewId, OrderID, ReviewScore, ReviewCommentTitle, ReviewCommentMessage, ReviewCreationDate, ReviewAnswerTimeStamp

**Orders Fields: Check**

* orderID, customerID, orderStatus, orderPurchaseTimestamp,OrderApprovedAt, OrderDeliveredCarrierDate, OrderDeliveredCustomerDate, OrderEstimatedDeliveryDate

**Products Fields: X**

* productID, ProductCategoryName, ProductName,ProductDescription, ProductDimensions, ProductPhoto)

**Seller Fields: X**

* SellerID, SellerCity, SellerState….more seller/lead info in the Olist Marketing Funnel dataset

**Summary:**

**Pros:**

* Lots of data fields to look into and potentially utilize
* Product features interesting for CLV (Product data, seller data, customer data, payment behavior, order items data)

**Cons:**

* There are 93,342 unique customers
  + Of which, there are only 2,800 repeat trx

**Update: Same dataset as Source # 2. Smaller dataset version.**

**Source 2: Ecommerce dataset**

Compiled by the UCI Machine Learning Repository, this ecommerce dataset features online retail transactions taken between 2010 and 2011 for a UK-based and registered non-store online retailer.

<https://www.kaggle.com/carrie1/ecommerce-data>

Data Size: 45.58 MB

Fields:

* CustomerID
* Country
* UnitPrice
* Invoice Date
* Quantity
* Description
* StockCode
* Invoice Number