# Proposal for 3 potential Capstone project ideas

Below are three candidate datasets.

## 1. Dataset – “Global super store 2016” from Data.world.

**Description**: Global Super Store is a data set which has around 50000 values. It’s a customer centric data set, which has the data of all the orders that have been placed through different vendors and markets , starting from the year 2011 till 2015.

**Link**:      <https://data.world/vikas-0731/global-super-store>

**Dataset**:  With **50k rows and 23 variables**present: order\_id, order\_date, ship\_date , ship\_mode ,customer\_id, customer\_name, segment, postal\_code, city, state, country,region, market, product\_id , category, sub\_category, product\_name, sales , quantity ,discount, profit, shipping\_cost, order\_priority.

Out of these, the order\_id, customer\_id ,postal\_code, product\_id may not be useful for the analysis , so that leaves 19 data points.   
For EDA, few analysis points that I could think of are:

1.      Product sales over the period 2012-2013(by product name, category and sub-category)

2.      Product sales by region

3.       Explore if sales and profit have any relationship to different variables like shipping mode, shipping cost and discount

4.       Which product category had the most sales by region/market and which ones did not? Explore reason for profitable sales

5.       Which product (names) to eventually discontinue due to bad/reduced sales or low profit and Explore the cause of any reduction in sales

6.       Which customers are most profitable? Is the customer profitable in multiple product categories or in a single product category?

For Machine Learning on this dataset, the prediction can be done for:

1.       Predict the region that will potentially have the most sales

2.       Predict the product that will be most profitable

## 2. Dataset:  Lending Club Loan Data.

**Description**: Includes the current loan status (Current, Late, Fully Paid, etc.) and latest payment information.

**Link**: <https://data.world/data-society> (which directs to  <https://www.kaggle.com/wendykan/lending-club-loan-data/data>).

**Dataset**: This dataset has **890K observations and 75 variables**

For EDA, the data points can be analyzed for exploring the relationship of the different variables (Grade, loan amount, loan term, interest rate, annual income, debt to income ratio, home ownership etc) to late payment of a loan eventually leading to a default.  
  
For Machine learning, the prediction can be done for predicting if a loan is at risk of defaulting.

## 3. Dataset: Default of credit card clients Data Set

**Description: Contains the customers default payments. The dataset has**demographic variables, payment history, and bill statements.

**​**

**Link**:

​ ​

<https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>.

​

**Dataset**: This dataset has **30K instances with 24 attributes**

​For EDA, the demographic variables present in the dataset(age, gender, education, marital status ) can be used for exploring the relationship to a default payment.

​For machine learning, the prediction can be done for ​predicting the next default payment or predicting a defaulter.