# Capstone Project Ideas

# Subject: Data Science Career Track

# Student: Franklin Cid

# Capstone Project-1: 911 Police Calls for Service (Baltimore, MD) Data EDA

# Description: 911 Public Safety Answering Point (PSAP) service area boundaries in California, according to the National Emergency Number Association (NENA), is a facility equipped and staffed to receive 9-1-1 calls. The service area is the geographic area within which a 911 call placed using a landline is answered at the associated PSAP. New functionalities now include calls from mobile phone and text. This task will only address calls from landlines.

# Police-Calls-for-Service are either emergency or non-emergency. Based from the target dataset, there are a few columns that are categorical, and several problems can be answered including:

1. What is the percentage of emergencies and non-emergencies are from a dataset?
2. What is the percentage are High Priority, Medium, and Low are from the emergency calls?
3. What is the type of emergency that's mostly incurred?
4. Is there a correlation of the call priority to a district?
5. Is there a correlation of the call type (description column) to a district?
6. Is there a correlation of the call type (description column) to a call priority?

PSAP = Public-Safety Answering Point

NENA = National Emergency Number Association (NENA)

# Data Source: <https://data.baltimorecity.gov/Public-Safety/911-Police-Calls-for-Service/xviu-ezkt>

# Data Columns: 3.86 mil rows, 8 cols, 1 call/row

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Description | Type | Comment |
| recordId | System generated? | number | 2749202 |
| callDateTime |  | datetime | '08/10/2017 03:25:00 PM'  format: mm/dd/yyyy hh:nn:ss AM/PM |
| priority | Priority of emergencies | text | [Low, Medium, High] |
| district | District codes | text | [ND, ] |
| description | District Description | text | LARCENY |
| callNumber | “how is generated? Wheres the source? | text | P172221666 |
| incidentLocation |  | text | 3100 ST PAUL ST |
| Location | IncidenLocation + lat/long | Location (lat/long) | "3100 ST PAUL ST BALTIMORE, MD (39.326142, -76.615959)" |

Sample Data:

recordId,callDateTime,priority,district,description,callNumber,incidentLocation,location  
2749202,08/10/2017 03:25:00 PM,Low,ND,LARCENY,P172221666,3100 ST PAUL ST,"3100 ST PAUL ST  
BALTIMORE, MD  
(39.326142, -76.615959)"  
2785533,08/23/2017 12:53:00 PM,Medium,NE,911/NO VOICE,P172351302,3300 BELAIR RD,"3300 BELAIR RD  
BALTIMORE, MD  
(39.321622, -76.573597)"

# Capstone Project-2:

Black Friday Sales EDA

# Description: Black Friday is one of the biggest sales, if not the biggest, in U.S. retail stores as items are placed at a huge, huge discounts. Stores open their doors to waiting customers as early as 6:00 am on the Day after Thanksgiving. Many purchasers line up the store doors hours before the doors open. Some sleep by the doors in their garden chairs. Once the doors open, customers race to the items they want to buy.

Most of the sales columns are categorical (see the column descriptions), except for the Purchase Price (for the item).

# Questions of Interest:

# The store would want to know customer purchase behaviors better on different product categories. Specifically, the problem is a regression problem where we are trying to predict the dependent variable (the amount of purchase) based on other categorical variables. Other questions that are important to the retailer are:

1. What type of product has the most sales count for men? For women?
2. What's the correlation of product type count sales on different age ranges for men? For women?
3. What's the correlation of product type count sales on different age ranges for married men? For married women?
4. What's the correlation of product type count sales on different age ranges for unmarried men? For unmarried women?
5. What's the correlation of occupation category and different age ranges for men on the count sales? For women?
6. What's the correlation of city category and different age ranges for men on the count sales? For women?
7. What's the correlation of occupation category and different age ranges for men on the count sales? For women?
8. What's the correlation of city category and occupation category for men on the count sales? For women?
9. What's the correlation of years stay and occupation category for men on the count sales? For women?

# Data Source: <https://www.kaggle.com/aniketg11/black-friday-retail-store-eda/notebook>

# Data Columns: xx mil rows, 12 cols, 1 call/row

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Description | Type | Comment |
| User\_ID | Id of User, sys-gen | Text (numeric) | 1000001 |
| Product\_ID | Id of Product, sys-gen | text | P00069042 |
| Gender | Gender | Text (code) | [M, F] |
| Age | Age range | text (numeric) | 0-17 |
| Occupation | Occupation Code | text (numeric) | 10 |
| City\_Category | Category of City | Text (code) | [A, B, C] |
| Stay\_In\_Current\_City\_Years | Years of Residency | Text (numeric) | 2 |
| Marital\_Status | [0=Unmarried, 1=Married] | Text (numeric) | 0 |
| Product\_Category\_1 | Category of 1st product bought | Text (numeric) | 3 |
| Product\_Category\_2 | Category of 1st product bought, if there is | Text (numeric) | 6 |
| Product\_Category\_3 | Category of 1st product bought, if there is | Text (numeric) | 14 |
| Purchase |  | Numeric | 15200 |

User\_ID,Product\_ID,Gender,Age,Occupation,City\_Category,Stay\_In\_Current\_City\_Years,Marital\_Status,Product\_Category\_1,Product\_Category\_2,Product\_Category\_3,Purchase

1000001,P00069042,F,0-17,10,A,2,0,3,,,8370

1000001,P00248942,F,0-17,10,A,2,0,1,6,14,15200

# Capstone Project-3: DJIA Stock Price EDA

# Description:

Stock market data can be interesting to analyze and as a further incentive, strong predictive models can have large financial payoff. The amount of financial data on the web is seemingly endless. A large and well structured dataset on a wide array of companies can be hard to come by. The dataset used here shows a historical stock prices (last 12 years) for 29 of 30 DJIA companies (excluding 'V' because it does not have the whole 12 years data).

# Questions of Interest:

The only categorical column is the Name (tickler). Some problems that can be solved are:

1. What's the correlation of the High to the Open of each stock?
2. What's the correlation of the High to the Low of each stock?
3. What's the correlation of the Open to the Low of each stock?
4. What's the correlation of the High and Low to the Volume of each stock?

# Data Source: --<https://www.kaggle.com/szrlee/stock-time-series-20050101-to-20171231/kernels>

<https://github.com/CNuge/kaggle-code/blob/master/stock_data/all_stocks_5yr.csv>

This dataset is maintained using Socrata's API and Kaggle's API. [Socrata](https://socrata.com/) has assisted countless organizations with hosting their open data and has been an integral part of the process of bringing more data to the public.

This dataset is distributed under the following licenses: Open Data Commons Public Domain Dedication and License

# Data Columns: xxx rows, 7 cols, 1 daily record/row

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Description | Type | Comment |
| Date | Date of the trading day | date | 01/03/17 |
| Open | Amount the stock price Opened for the day | float | 178.83 |
| High | The Highest amount of the stock price for the day | float | 179.14 |
| Low | The Lowest amount of the stock price for the day | float | 176.89 |
| Close | The Highest amount of the stock price for the day | float | 177.71 |
| Volume | The Highest amount of the stock price for the day | numeric | 1447848 |
| Name | Tickler name of the stock | text | MMM |
|  |  |  |  |