Sample Project Descriptions and Datasets

Domain: Retail
Project: Amazon Electronics Products Analysis
Data: http://jmcauley.ucsd.edu/data/amazon/links.html

The Amazon Product data set to be considered for this project is from Amazon, a well-known e-commerce giant. This is a very interesting project to get insights based analysis of product sales based on reviews and ratings.
This dataset contains product reviews including 142.8 million reviews spanning May 1996 - July 2014. This data set includes reviews (ratings, text, helpfulness votes). From the various categories the Electronics category is chosen for the analysis. The size of reviews data set is 1.5 GB and ratings data set is 320 MB.

Possible Exploration Ideas:

1. Understand the Dataset
2. Top electronic item that was rated above 3.0 over the period of 8 years [2006-2014] (EG. Year 2006 - 100 items bought received all the rating above 3.0 - show top 1 item out of 100; sample data 2006 ,"ASIN", 'Review Count")
3. Worst review a product received in a given year between 2006 to 2014 (E.g. 2006, ASIN, Review Count (Least number of reviews and below 2.0 rating)
4. Maximum number of reviews given by same user in a year
5. Least *Helpful reviews* per year per products (helpful percentage between 1% and 30 %)
6. Most *Helpful reviews* per year Per Products (Helpful percentage >75%
7. Growth of review comments on products/year
8. Visualization of results observed

Project: Amazon Fine Food Reviews
Data: https://www.kaggle.com/snap/amazon-fine-food-reviews
The Amazon Fine Food Reviews dataset consists of 568,454 food reviews Amazon users left up to October 2012. This data was originally published on SNAP and the team has sourced the dataset from kaggle.com. The dataset contains user reviews covering a host of categories which include beverages, confectionaries, gourmet food, pet food etc.

Possible Exploration Ideas:

1. Which are the top 10 Product IDs that get reviewed the most?
2. Which are the top 10 most favourably reviewed Product IDs?

3. How may product reviews are generated on a daily, monthly and yearly basis?

4. Product sales- festival mapping. By festival mapping we are trying to analyze the top 3 Product IDs that get reviewed the most for each of the major festivals in the USA.

5. Visualization of results observed.

Domain: Social Sector
Project: World Development Indicator
Data:
http://databank.worldbank.org/data/reports.aspx?Code=IND&id=556d8fa6&report_name=Popular_countries&populartype=country&ispopular=y

This dataset is World Bank's compilation of internationally comparable statistics about global development and the quality of people's lives. It contains data about Agriculture & Rural Development, Aid Effectiveness, Climate Change, Economy & Growth, Education, Energy & Mining, Environment, External Debt, Financial Sector, Gender, Health, Infrastructure, Labour & Social Protection, Poverty, Private Sector, Public Sector, Science & Technology, Social Development, Trade, and Urban Development.

Possible Exploration Ideas:
1) Top 10 Reports:
a) Top 10 Improved sanitation facilities (% of population with access)

b) Top 10 Energy use (kg of oil equivalent per capita)

c) Top 10 CO2 emissions (metric tons per capita)

d) Top 10 Improved water source (% of population with access)

e) Top 10 Electric power consumption (kWh per capita)

f) Top 10 Forest area (sq. km)

g) Top 10 Exports of goods and services (% of GDP)

h) Top 10 Imports of goods and services (% of GDP)
2) Least 5 Reports:
a) Least 5 Improved sanitation facilities (% of population with access)

b) Least 5 Energy use (kg of oil equivalent per capita)

c) Least 5 CO2 emissions (metric tons per capita)

d) Least 5 Improved water source (% of population with access)

e) Least 5 Electric power consumption (kWh per capita)

f) Least 5 Forest area (sq. km)

g) Least 5 Exports of goods and services (% of GDP)

h) Least 5 Imports of goods and services (% of GDP)
3) Comparison Report: Comparison between countries based on several indicators.
4) Visualization of results observed.

Domain: Health Care
Project: ZikaVirus Analysis and Out-Break
Data: https://www.kaggle.com/cdc/zika-virus-epidemic
This dataset shares publicly available data related to the ongoing Zika epidemic. It is being provided as a resource to the scientific community engaged in the public health response. The data provided here is not official and should be considered provisional and non-exhaustive. The data in reports may change over time, reflecting delays in reporting or changes in classifications. And while accurate representation of the reported data is the objective in the machine readable files shared here, that accuracy is not guaranteed.

Possible Exploration Ideas:
1. Most affected country in terms of Zika confirmed and Zika suspected cases, as well as the count of total number of Zika confirmed and Zika suspected cases in all the available countries in the dataset.

2. As the analysis in the dataset is weekly or twice a week, reported date wise analysis of most number of Zika confirmed and Zika discarded cases, which will be helpful in understanding in which month or season a country is most affected to the disease, which in turn helps in understanding the temperature or the climate condition that leading the source of this disease.

3. Individual country's reported date data to analyse the number of Zika confirmed and suspected cases increasing or decreasing day by day, which helps in identifying the countries where it's growing adversely as well as about the countries taking precaution against the disease where the numbers are controlled or decreasing.

4. Age group wise analysis of the number of Zika confirmed and suspected cases found, which helps in understanding which age group has the risk of the disease as outcome varies across the age groups.
5. Predicting the outbreak of Zika virus based on the existing data.

6. Visualization of results observed.