Data Science Career Track Capstone Two: Project Ideas:

Start up funding data

1.) https://www.kaggle.com/kanakamohan/Startup-Funding-Data

Content:
Wanted to get some insights of the Indian startup ecosystem. Which industry is seeing more innovation, which cities are playing major roles in attracting establishing startups. What is the pattern in number of fundings per year. Data collected from 2015 - until mid July 2020.

Context:

> What's inside is more than just rows and columns. Make it easy for others to get started by describing how you acquired the data and what time period it represents, too.
> Data collected through web scrapping process, though data available out there for last 5 years. you can run into lot of challenges while reading data from the website due incosistency or may be limitations of python libraries. Processing the data may be biggest hurdle as there is missing data, unorganized, inconsistency in data format.

Problem to solve:

This is funding per start-up. I want to find what type of start-ups get the most funding, how much they get, what industry, and how often. This will give us a better understanding of trends from 2015 - 2020 on venture generated and where the start-up market is headed.

Will present it with this GDP data:
https://www.kaggle.com/paree24/india-gdp-growth-world-bank-1961-to-2017

NCHS - Drug Poisoning Mortality by State: United States

https://data.world/us-hhs-gov/2d101a72-3b4d-4bb6-95dc-0e5bc7d9ebb5

This dataset describes drug poisoning deaths at the U.S. and state level by selected demographic characteristics, and includes age-adjusted death rates for drug poisoning. Deaths are classified using the International Classification of Diseases, Tenth Revision (ICD–10). Drug-poisoning deaths are defined as having ICD–10 underlying cause-of-death codes X40–X44 (unintentional), X60–X64 (suicide), X85 (homicide), or Y10–Y14 (undetermined intent). Estimates are based on the National Vital Statistics System multiple cause-of-death mortality files (1). Age-adjusted death rates (deaths per 100,000 U.S. standard population for 2000) are calculated using the direct method. Populations used for computing death rates for 2011–2016 are postcensal estimates based on the 2010 U.S. census. Rates for census years are based on populations enumerated in the corresponding censuses. Rates for non census years before 2010 are revised using updated intercensal population

estimates and may differ from rates previously published. Death rates for some states and years may be low due to a high number of unresolved pending cases or misclassification of ICD–10 codes for unintentional poisoning as R99, “Other ill-defined and unspecified causes of mortality” (2). For example, this issue is known to affect New Jersey in 2009 and West Virginia in 2005 and 2009 but also may affect other years and other states. Drug poisoning death rates may be underestimated in those instances.

Supermarket sales

The growth of supermarkets in most populated cities are increasing and market competitions are also high. The dataset is one of the historical sales of supermarket company which has recorded in 3 different branches for 3 months data. Predictive data analytics methods are easy to apply with this dataset.

https://www.kaggle.com/aungpyaeap/supermarket-sales

Problem to solve:

Identify what people are buying, where people are buying, and how - to better market and increase revenue for supermarkets