**Some Ideas for the Project**

1. **Data Analytics Application for Cancer Research:** Cancer is fundamentally a disease of the genome, caused by changes in the DNA, RNA, and proteins of a cell that push cell growth into overdrive. Identifying the genomic alterations that arise in cancer can help researchers decode how cancer develops and improve upon the diagnosis and treatment of cancers based on their distinct molecular abnormalities.

**Data Source** National Institute of Health (NIH) has a website, with API available, where a large volume of genome data can be accessed and/or uploaded. The location is called Genome Data Commons (GDC). The URL is <https://gdc.cancer.gov/about-gdc>

**Some thoughts on What the end product may be** : A relationship may exist between various changes in genes and type of cancer it causes. Is there enough data to support what mutation led to a specific cancer? Can an overview be put together?

A relationship may exist between a specific type of cancer, for example breast cancer, arising due to changes in certain genes. Is there a relationship between ethnicity, weight, life styles (smoking, alcohol, food type consumed, stress), age and genetics that impact the occurrence of breast cancer? The goal here would be to identify what factors impact the onset of cancer. Can an early detection and intervention made if certain risk factor show higher possibility of getting the disease.

Next level could be the treatment regime. What types of treatment exists (Chemo versus Immunotherapy, radiation etc) and how effective these treatments are and at what stage of cancer. Is there a link between treatment regime and ethnicity or age?

Finally a recommendation could be made, based upon the conclusion of the data analysis, as to what can one do reduce the chances of getting cancer if they fall in the high risk bracket.

1. **Data Analytics Application for studying Opioid Abuse:** Use of opioid for long term pain management can often lead to addiction to such drugs. Overuse and exploration into new kinds of Opioids, including fentanyl, could be fatal for individuals. Looking at the drivers that lead to addiction could be critical for improving avoidance and rehabilitation.

**Data Source:** US government’s data.gov site has several drug related data sources available. Some work is needed to identify where the data may reside. Example site: <https://www.kaggle.com/datasets/ryanandreweckberg/opioid-crisis-by-interpersonal-relationships>

**Some thoughts on What the end product may be:** There may be a prevalence of drug use and death in certain geographies, states and/or cities. First step would be to identify the area where drug use is most prevalent along with the drug type. Second step will be to look into drivers for drug abuse. This will include, employment status, age group, gender, income level, education level, marital status and home ownership. Is there a correlation between income level, employment status, education and drug abuse? How does these factors compare with the states where there is lower drug use. Is a story emerging from this analysis? Meaning, does education level impact drug abuse prevalence?

Finally, assuming that a pattern does emerge, some recommendations can be made. For example, if high unemployment rate has a strong correlation with drug abuse then steps can be taken to improve the employment level.

1. **Data Analytics for Studying Credit Card Fraud:** With the increased use of electronic payment as well as electronic transaction via online banking (Venmo, Zelle, wire transfer etc), and electronic communication (emails and text messaging) as an option, credit card fraud has also increased in recent years. Even though banks have made it harder to steal via several layers of authentication, fraudsters have also gotten skilled at bypassing these layers of security. Analysis of fraud types and location may help identify methods to improve security

**Data Source:**  <https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud>

**Some thoughts on What the end product may be:** Final goal is to find how much fraud exist and what was the key driver for fraud. This should be followed by recommendations to reduce fraud.