1. ## Question:

How do **demographic** factors like **age, sex**, and **income** influence **health insurance charges/coverage** across different **states/regions**?

2. ## Data Sources:

We have two data sources for this project, both from **Kaggle**. First dataset is about **US Census Demographics** and the second one contains **US Health Insurance** charges data. These datasets, best suit out problem as both contain (most of) the attributes which are relevant, and are chosen after investigating several other Health related datasets. The US Census Demographics dataset contains information about Genders, Counties, States, No. of Employees, and respective incomes. The US Health Insurance contains information about BMI, Age, Smoker, Region, and related Health Insurance Charges.

a. ### Data Structure and Quality:

The datasets are in tabular format i.e. **Structured data** stored in CSV files. The datasets perfectly reflect the real world and is correct, ensuring high **Accuracy**. In our both datasets, all columns have more than 80% of the data present, which is a high indicator for high **Completeness.** Both are **Consistent** in formats in nearly all the attributes. The age of data is, what could be the best possible dataset to have for the problem being addressed, a bit older if see the date census conducted (2017) but fulfills our purpose, and validated the for e.g. population(attribute) for one of the states for e.g. Alabama from worldpopulationreview.com, so **Timeliness** is satisfactory as well. As discussed, datasets are a bit older but contain the relevant information ensuring high **Relevancy**.
In addition, there are **missing values** in several columns, which are addressed in Data Transformation section of the ETL technology (discussed later). The datasets contain no **duplicates** and **invalid** values. By looking into the **Descriptive statistics**, Summary statistics seem appropriate for both datasets. Overall, **Data Quality** is good to get going for our project.

b. ### Licenses:

Both data sources are licensed under the **CC0: Public Domain**, allowing unrestricted use, which includes modification, publishing etc. CC0 empowers the choice to **opt out** of copyright, and the exclusive rights automatically granted to creators – the "no rights reserved". Although there are no restrictions as per the definition of CC0, but still, we respect its public availability and ensure the **ethical use** of the data, and **transparency** in analysis for our project. Please find below the link to the direct page of the data sources to view the licenses.
US Census Demographics
US Health Insurance

3. Data Pipeline:

a. Overview:

Our Data Pipeline is based on **ETL** technology, i.e. **E**xtract, **T**ransform and **L**oad. The **Extract** focuses on extracting the data from 2 different data sources, transform for converting the data into the desired format and making it usable for the analysis, and finally load it into the appropriate sinks.

b. Data Cleansing:

After successful extraction, **Transformation** step starts by removing irrelevant columns for e.g. “CountyId” and “VotingAgeCitizen”. Afterwards, we check for the missing values for each column and set the criteria for removal of them. If a column contains more than 30% **missing** data values (considering after adding more data in future and some columns aren’t specified anymore), we remove those features from the dataset in the first place. For the columns containing less than 30% missing values, although there are many ways to fill in those values, we impute using **Backfill** method which will fill the missing values from the most corresponding previous/back value.

As the data types are already suitable for the features in the dataset, we don’t perform any transformation regarding it. Once all data is as per our desire, we **Load** it into the SQLite database and can perform the further analysis.

c. Problems:

Identifying the **relevant** columns from a lot of columns was one of the major challenges. So, we carefully picked the columns which didn’t add any value to our final outcome and removed them. **Missing values** was also one of the major challenges. As a solution for it, we set a general criterion for all the columns that will ensure keeping the columns containing enough information for us to analyze at the end and not just drop without any specific reason.

d. Meta-quality measures:

A **log file** is created to capture the start time, end time, success/error messages. We consider to capture the start and end time of the pipeline i.e. **Latency**, in Python log file, to keep track of the effectiveness of our pipeline. This will give us insights on the overall duration from data extraction, processing, to loading. The pipeline incorporates retries mechanism as well in case of failure of for e.g. extraction of data from the listed data sources, to try again to extract for specific number of times.

Additionally, we designed our pipeline not only specific to the current data, but also having the bigger picture of it with changing inputs. For e.g. one column contains percentages from 0-72, but it will also incorporate inputs over 72 and within/equal to 100. This way, **efficiency** and **reliability** of our pipeline is being taken into account which is necessary for building robust data pipelines.

4. Result and Limitations:

a. Data Structure and Quality:

The output format of our pipeline is Structured data. The output contains over 3000 rows by removing all the columns/rows having insufficient information and filling in the ones having enough information for our analysis still ensuring high **Accuracy** and a good representation of the real-world. As discussed, the missing values imputation procedure already, output doesn't contain any blank and invalid data points, maintaining high **Completeness**. The data pipeline delivers the output in a **Consistent** format for all the columns according to their suitability. Regarding the age of the data, it is appropriate as well, and closely related to the current information available on the internet, delivering up-to-date information ensuring high **Timeliness**. The results are diverse in the sense of the different states, genders, incomes that are critical to our analysis providing high **Relevance**.

b. Output Format:

The results are stored in the **SQLite files**, as it doesn't require any installation/setup to make it work. Further, for the Final report, the results can be easily queried using these files and analysis could be carried out by accessing these files in your favorite IDE/App without any additional hustle.

c. Issues:

Regarding the data produced as a final outcome of our pipeline, there seem a concern with the **completeness** of the dataset as both datasets contain specific information, for e.g. one dataset holds certain demographics information while the other dataset accommodates only the health insurance related details. These may lead to not accurately and completely answering the main problem as there should be one **common field** based on the datasets could be compared and generate more meaningful and actionable insights.