**Project Outline:**

[1 – Background & Design 2](#_Toc61723948)

[2 –Load Libraries & Read Data 2](#_Toc61723949)

[3 – Data Inspection: 2](#_Toc61723950)

[4 – Data Evaluation 5](#_Toc61723951)

[5 – Data Visualization 5](#_Toc61723952)

[5 – Correlations between the attributes 6](#_Toc61723953)

[5 – Exploratory Data Analysis 6](#_Toc61723954)

[4 – Validation & Publishing 6](#_Toc61723955)

[5 – Communication of Analysis 7](#_Toc61723956)

[5 – Building the Report 8](#_Toc61723957)

Dataset link:

# 1 – Background & Design

|  |  |
| --- | --- |
|  | 1.1 - Quick synopsis of the Problem being addressed. |
|  | 1.2 – Explanation of the desired Outcome. |
|  | 1.3 – What will be evaluated and how it will be measured. |

# 2 –Load Libraries & Read Data

|  |  |
| --- | --- |
|  | Loading libraries required and quick statement for what they will be used.   * Libraries loaded: |
|  | 2.1 – Identifying the data elements, fields, attributes involved in the dataset. |
|  |  |

# 3 – Data Inspection:

|  |  |  |
| --- | --- | --- |
|  | 3.1 – Examine the data briefly as a whole | |
|  |  | **df.head()** #top 5 rows | **df.tail()** #last 5 rows |
|  |  | **df.info()** #inspect in which columns you might have null values & note datatypes  #there are integers, floats, and objects |
|  |  | **df.describe()** #statistical data on integer and float dtypes |
|  |  | **df.describe(include="O")** #including only strings statistical data |
|  |  | **df.describe(include='all')** |
|  | 3.2 – Examining **Null Values** | |
|  |  | Examining the number of values for each field that are null |
|  |  | Calculate and plot the % of missing values in each attribute |
|  | 3.3 – Eliminating **Unwanted or Non-Relevant Columns** | |
|  | |  |  | | --- | --- | |  | Highest missing value columns which will not be necessary to keep | |  | Removing any non-relevant columns such as **Index** | | |
|  | **A – Discovery Phase:**  Aim: Understanding the data better with respect of the use case. Unpack the problem to be solved; break-down project aim/goal by looking at:   * types of data * volumes of data * identifying the data required   + considerations:     - how much data is sufficient for credible analysis?     - Timeframe for data     - Dependencies, risks, mitigation plan     - Specific data elements you need     - Possible format & sources/repositories     - Possible problems in gathering the data * tools to consider for the end goal (plan for collection and collection methods)   how the data will be gathered and imported | |
|  | * Defining the data and what it means in the context of this project. | |

**A – Transformation Phase:**

Aim: Making the data into what you need to make it so you can analyze it.

|  |  |  |
| --- | --- | --- |
|  | 3A.1 – De-normalizing & Standardizing data coming in from multiple sources.  Aim: | |
|  | 3A.2 – Structuring Data  Aim: This is where you can change the order of the fields, combine fields, perform any join you may need using SQL, and/or creating the structure of the data you are seeking. | |
|  |  | Placeholder |
|  |  | Placeholder |
|  |  | placeholder |
|  | 3A.3 – Normalizing Data  Aim: Cleaning unused data, reducing redundancy, reducing inconsistency. | |
|  |  | Values and format are stored as expected for field consistency (ie. Data type conversion as necessary). |
|  |  | Evaluate syntax errors: white spaces, extra spaces, typos that may need to be fixed. |
|  |  | Examine outliers for accuracy and inclusion in the dataset. |
|  |  | |
|  |  | |
|  | 3A.4 – Cleaning Data  Data Preparation & Reliability:  No data is perfect. This step ensures that the time is well spent to understand the practical problem in a dataset. Ensuring that basic data integrity questions are evaluated towards analysis.  Aim: Fixing irregularities or quality issue that might affect analysis in the data in order to produce a credible and accurate analysis.  >> if able to correct data: use data wrangling tools/scripst  >>if cannot be corrected: try removing data | |
|  |  | Examine Missing & Null  - Filter out records and possibly remove?  - Find source of missing information and enrich?  - Impute based on statistical values? |
|  |  | Examine Incomplete, inaccurate, or incorrect data |
|  |  | Examine Biased or inconsistent data |
|  |  | Examine Duplicate data  - Remove duplicates |
|  |  | * 1. Examine Insufficient Parameters |
|  |  | Examine Outliers |
|  | 3A.5 – Enriching Data  Aim: To make additions which make analysis more meaningful by adding additional data points. | |
|  |  | Make note of any enrichments made to the data |
|  | placeholder | |

# 4 – Data Evaluation

|  |  |
| --- | --- |
|  | 4.1 – Examine the unique value counts of important provider fields |
|  | 4.2 – Examining data distribution of columns that have a wide distribution of values: |
|  | 4.3 – Quick Summary:   * More males than females * Mostly all providers are inn-network * Top state is California * US is the major country in the data * More Outpatient services than Facility * Provider Types may need to be bucketed and/or clustered |

# 5 – Data Visualization

|  |  |  |
| --- | --- | --- |
|  | 5.1 Histograms (for integers & floats) | |
|  | 5.2 Correlations between Attributes (for integers & floats) | |
|  |  | Correlation Heat Map is also for integers & floats |
|  |  | placeholder |
|  | placeholder | |
|  | placeholder | |
|  | placeholder | |

# 5 – Correlations between the attributes

|  |  |  |
| --- | --- | --- |
|  | * Correlation Matrix | |
|  | 2.2 placeholder | |
|  |  | palceholder |
|  |  | placeholder |
|  |  | placeholder |
|  | placeholder | |
|  | placeholder | |
|  | placeholder | |

# 5 – Exploratory Data Analysis

Aim: Find patterns, variations, identifying correlations, trends, etc.

|  |  |  |
| --- | --- | --- |
|  | 5.1 – Handling missing data | |
|  | 2.2 palceholder | |
|  |  | placeholder |
|  |  | placeholder |
|  |  | placeholder |
|  | placeholder | |
|  | placeholder | |
|  | placeholder | |

# 4 – Validation & Publishing

Validation aims to have repetitive programming steps that look for:

* Verifying consistency of the data
* Verify quality of the data
* Verify security of the data

|  |  |  |
| --- | --- | --- |
|  | 2.1 placeholder | |
|  | 2.2 placeholder | |
|  |  | placeholder |
|  |  | placeholder |
|  |  | placeholder |
|  | placeholder | |
|  | placeholder | |
|  | placeholder | |

# 5 – Communication of Analysis

|  |  |  |
| --- | --- | --- |
|  | 5.1 – Interpreting the results.  Aim: Evaluate Dependability:   * Any limitations / circumstances under which analysis may not hold true. | |
|  | * Present Findings. * Communicating and presenting findings in clear, impact and convincing way. | |
|  |  | placeholder |
|  |  | placeholder |
|  |  | placeholder |
|  | placeholder | |
|  | placeholder | |
|  | placeholder | |

# 5 – Building the Report

|  |  |
| --- | --- |
|  | Cover Page   * Title of the report * Names of authors * Their affiliations * Contacts * Name of Institutional Publisher * Data of Publication |
|  | Table of Contents   * Like a map needed for a trip you've never taken before * "visual proxy" for the actual travelw/details about the landmarks that you will pass on * w/ main headings * list of tables * figures |
|  | Executive Summary   * Crux of the Argument in 3 paragraphs |
|  | Introduction   * Setting up the problem for the reader to * new readers to the topic * might need to be gently introduced before being immersed in details |
|  | Literature Review   * Follow-up to the introductory section * review of available relevant research on the subject * either - * common view literature * or b) cite relevant research to offer adequate context before you embark on analysis to show nuances/caveats |
|  | Methodology   * measurement methods: * data sources * if you collected new data, explain collection method * refer to literature review to BOLSTER your choice for variables, data, and methods * and how they will help answer your research questions |
|  | Results   * present your impirical fidings * descriptive statistics * illustrate graphics * formally testing your hypothesis * statistical models: * regression models * categorical analysis * - time series data |
|  | Discussion   * craft your main arguments by building on the results * rely on the power of narrative to enable * communicate your thesis to your readers * refer the reader to: * \*research question AND * \*the knowledge gaps * highlight how your findings provide the ultimate missing piece to the puzzle |
|  | Conclusion   * generalize your specific findings * marketing approach to promote findings * identify future possible developments in research * and applications that could results from your research |
|  | Acknowledgements |
|  | References |
|  | Appendices |