1. **COLLECT INITIAL DATA**

Acquire the data (or access to the data) listed in the project resources. This initial collection includes data loading, if necessary for data understanding. For example, if you intend to use a specific tool for data understanding, it is logical to load your data into this tool.

Describe all the various data used for the project, and include any selection requirements for more detailed data. The data collection report should also define whether some attributes are relatively more important than others.

Remember that any assessment of data quality should be made not just of the individual data sources but also of any data that results from merging data sources. Because of inconsistencies between the sources, merged data may present problems that do not exist in the individual data sources.

**Data requirements planning**
- Plan which information is needed (e.g., only for given attributes, or specific additional information)
- Check if all the information needed (to solve the data mining goals) is actually available

All attributes given in different tables are needed.

Amazing Health Network (AHN) has provided a detailed description about all the attributes, and all the tables are accessible on the cluster.

**Selection criteria**
- Specify selection criteria (e.g., Which attributes are necessary for the specified data mining goals? Which attributes have been identified as being irrelevant? How many attributes can we handle with the chosen techniques?)
- Select tables/files of interest
- Select data within a table/file
- Think about how long a history one should use (e.g., even if 18 months of data are available, only 12 months may be needed for the exercise)

Attributes with general logistics information like Provider ID, Vendor, Primary Care Physician (PCP), etc., will be excluded. Additionally, MemberID will be used only for merging data and uniquely identifying patients. The modeling techniques can easily handle hundreds of attributes for the AHN data.

The critical tables are – members, claims, drug count, lab count, and days in hospital per year.

Each table is in the form of a CSV file that can directly be loaded. We initially plan to use all the files' rows.

We will be using two years of history of days in the hospital for modeling, validation, and testing.
Be aware that data collected from different sources may give rise to quality problems when merged (e.g.,
address files merged with a customer database may show inconsistencies of format, invalidity of data,
etc.).

Insertion of data
• If the data contain free text entries, do we need to encode them for modeling or do we want to
group specific entries?
• How can missing attributes be acquired?
• How can we best extract the data?

The claims table contains string entries representing categorical values and will be one-hot encoded for
modeling.

There is very little chance of acquiring missing data, and we will either use models that can work with
lost data or use the average attribute value as a placeholder for missing data.

We will use the Pandas library to extract the data from the CSV files.

Remember that some knowledge about the data may be available from non-electronic sources (e.g., from
people, printed text, etc.).
Remember that it may be necessary to preprocess the data (time-series data, weighted averages, etc.).

2. DESCRIPTIVE ANALYSIS

Examine the "gross" properties of the acquired data and report on the results.
Describe the data that has been acquired, including the format of the data, the quantity of the data (e.g.,
the number of records and fields within each table), the identities of the fields, and any other surface
features that have been discovered.

Volumetric analysis of data
• Identify data [Optional] and method of capture
• Access data sources
• Use statistical analyses if appropriate
• Report tables and their relations
• Check data volume [Optional], number of multiples, complexity
• Note if the data contain free text entries

There are six tables:
1) Member table:
   1. MemberID: Member pseudonym.
   2. AgeAtFirstClaim: Age in years at the time of the first claim's date.
   3. Sex: Biological sex of member: M = Male; F = Female.

2) Claims table:
1. MemberID: Member pseudonym.
2. ProviderID: Provider pseudonym.
4. PCP: Primary care physician pseudonym.
5. Year: Year in which the claim was made: Y1; Y2; Y3.
7. PlaceSvc: Generalized place of service.
8. PayDelay: Number of days delay between the date of service and date of payment.
9. LengthOfStay: Length of stay (discharge date – admission date + 1)
10. DSFS: Days since first service (or claim), computed from the first claim for that member for each year.
11. PrimaryConditionGroup: Broad diagnostic categories based on the relative similarity of diseases and mortality rates.
12. CharlsonIndex: A measure of the affect diseases have on overall illness
14. SupLOS: Indicates if the NULL value for the LengthOfStay variable is due to suppression done during the de-identification process. A value of 1 indicates that suppression was applied.

3) DrugCount table:
   1. MemberID: Member pseudonym.
   2. Year: Year in which the drug prescription was filled: Y1; Y2; Y3.
   3. DSFS: Days since first service (or claim), computed from the first claim for that member for each year.
   4. DrugCount: Count of unique prescription drugs filled by DSFS.

4) LabCount table:
   1. MemberID: Member pseudonym.
   2. Year: Year in which the lab visit occurred: Y1; Y2; Y3.
   3. DSFS: Days since first service (or claim), computed from the first claim for that member for each year.
   4. LabCount: Count of visiting labs filled by DSFS.

5) DaysInHospital_Y2 table:
   1. MemberID: Member pseudonym.
   2. DaysInHospital_Y2: Days in hospital, the main outcome, for members with claims in Y1. Values above 14 days (the 99% percentile) are top coded as "15+".

6) DaysInHospital_Y3 table:
   1. MemberID: Member pseudonym.
   2. ClaimedTruncated: Members with truncated claims in the year prior to the main outcome are assigned a value of 1, and 0 otherwise.
   3. DaysInHospital_Y3: Days in hospital, the main outcome, for members with claims in Y2. Values above 14 days (the 99% percentile) are top coded as "15+".
The data is analyzed with the Pandas package.

There are three years of data available. Member table provides basic information of patients, and the claims table provides detailed claim information. DrugCount table includes information about unique prescription refills. LabCount table includes information about the unique laboratory and pathology tests are done for each member, DaysInHospital_Y2 and DaysInHospital_Y3 tables give information on the exact days of patients stayed at the hospital in year 2 and year 3.

All tables share the same attribute, "MemberID", which identifies every member. Claims, DrugCount, LabCount tables share the same attribute year, meaning the year of each record.

The data volume of each table is as follows:
- Members.csv: 113000 rows
- Claims.csv: 2668990 rows
- DrugCount.csv: 818241 rows
- LabCount.csv: 361484 rows
- DaysInHospital_Y2.csv: 76038 rows
- DaysInHospital_Y3.csv: 71435 rows

No data or attribute contains free-text entries.

Attribute types and values
- Check accessibility and availability of attributes
- Check attribute types (numeric, symbolic, taxonomy, etc.)
- Check attribute value ranges
- Analyze attribute correlations
- Understand the meaning of each attribute and attribute value in business terms
  - [Optional] For each attribute, compute basic statistics (e.g., compute distribution, average, max, min, standard deviation, variance, mode, skewness, etc.)
  - [Optional] Analyze basic statistics and relate the results to their meaning in business terms
- Decide if the attribute is relevant for the specific data mining goal
- [Optional] Determine if the attribute meaning is used consistently
- Interview domain experts to obtain their opinion of attribute relevance
- Decide if it is necessary to balance the data (based on the modeling techniques to be used)

All selected attributes are available and are of string (categorical) or integer type.
The value ranges of all attributes are as follows:

1. MemberID: numeric, no range.
2. AgeAtFirstClaim: from 0-9, 10-19 to 80+.
3. Sex: M and F.
4. ProviderID: numeric, no range.
5. Vendor: numeric, no range.
6. PCP: numeric, no range
7. Year: Y1; Y2; Y3.
8. Specialty: string representing a categorical value.
9. PlaceSvc: string representing a categorical value.
10. PayDelay: from 1 to 162+.
11. LengthOfStay: from 1 day to 26+ weeks.
12. DSFS: from 0-1 month to 11-12 months.
13. PrimaryConditionGroup: string representing a categorical value.
14. CharlsonIndex: from 0, 1-2 to 5+.
15. ProcedureGroup: string representing a categorical value.
16. SupLOS: 0 and 1.
17. DrugCount: from 1 to 7+.
18. LabCount: from 1 to 10+.
19. DaysInHospital_Y2: from 0 to 15
20. ClaimedTruncated: 0 and 1.
21. DaysInHospital_Y3: from 0 to 15.

Provider ID and Vendor attributes are highly correlated (Correlation = 0.76). Also, there is a significant correlation between length of stay attribute and place of service (Correlation = 0.26).

The meaning of each attribute and attribute value in business terms, are included in the beginning of Section 2: Describe Data.

According to Mr. John Smith, drug count and lab count are two critical factors in determining a patient's risk level at the hospital in the next year.

Most of the fields are categorical, and there is no need to balance the data. This may change with the introduction of attribute transformations and derived attributes.

Keys
- Analyze key relationships
- Check amount of overlaps of key attribute values across tables
The member ID key is used as a primary key across all the tables to identify unique patients.

Column DSFS is present in multiple tables.

Review assumptions/goals
- Update list of assumptions, if necessary

No updates.

3. EXPLORE DATA

This task tackles the data mining questions that can be addressed using querying, visualization, and reporting techniques. These analyses may directly address the data mining goals. However, they may also contribute to or refine the data description and quality reports, and feed into the transformation and other data preparation steps needed before further analysis can occur.

Describe the results of this task, including first findings or initial hypotheses and their impact on the remainder of the project. The report may also include graphs and plots that indicate data characteristics or point to interesting data subsets worthy of further examination.

Data exploration
- Analyze properties of interesting attributes in detail (e.g., basic statistics, interesting sub-populations)
- [Optional] Identify characteristics of sub-populations

As alluded to in the Business Understanding report, drug count and lab count are essential attributes. The following five graphs show the histograms of some interesting features.

As you can see, patient population distribution is relatively uniform while lab count and drug count have distribution akin to an exponential one as expected.

Note that the length of stay, procedure group, and days in hospital have the count (y-axis) on a log scale. Most of the length of stay data is unknown due to missing data or intentional data suppression. The days in hospital distribution are also exponential, and only a few percentages of patients are hospitalized for more than a week.
Form suppositions for future analysis

- Consider and evaluate information and findings in Section 2: Describe Data
- Form a hypothesis and identify actions
• [Optional] Transform the hypothesis into a data mining goal, if possible
• [Optional] Clarify data mining goals or make them more precise. A "blind" search is not necessarily useless, but a more directed search toward business objectives is preferable.
• Perform basic analysis to verify the hypothesis

Based on discussions with Mr. John Smit at AHN and surface-level data evaluation, the aggregated lab visits/count and aggregated drug refills/count for a member in the year directly influence the patient's future hospitalization risk.

The feature weights in linear models will be used to verify this hypothesis. Based on initial calculations, 90% and 85% of hospitalized patients had at least one lab and pharmacy visit, respectively.

4. VERIFY DATA QUALITY

Examine the quality of the data, addressing questions such as: Is the data complete (does it cover all the cases required)? Is it correct or does it contain errors? If there are errors, how common are they? Are there missing values in the data? If so, how are they represented, where do they occur, and how common are they?

List the results of the data quality verification; if there are quality problems, list possible solutions.

• Identify special values and catalog their meaning

There are no special values except for some categorical bins are encoded as “value+” to have an upper limit. The following attributes have the “value+” encoding:
1) AgeAtFirstClaim: 80+ age
2) PayDelay: 162+ days
3) Length of Stay: 26+ weeks
4) CharlsonIndex: 5+ value
5) Drug Count: 7+ count
6) Lab Count: 10+ count
7) Days in Hospital: 15+ days

Review keys, attributes
• Check coverage (e.g., whether all possible values are represented)
• Check keys
• Verify that the meanings of attributes and contained values fit together
• Identify missing attributes and blank fields
• [Optional] Establish the meaning of missing data
• Check for attributes with different values that have similar meanings (e.g., low fat, diet)
• Check spelling and format of values (e.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter)
• Check for deviations, and decide whether a deviation is "noise" or may indicate an interesting phenomenon
Check for plausibility of values, (e.g., all fields having the same or nearly the same values)

All possible values are covered and represented.

The member ID key is used as a primary key across all the tables to identify unique patients. There are no issues with the key, for instance, keys in other tables that are not in the members table.

There is no inconsistency in attribute name and attribute values semantics.

About 97% of the length of stay values and 2% of DSFS attribute values are missing in the claims table. All other attributes in the claims table have more than 99% attribute presence. In the members table, sex information is absent for ~15% of members, and age is absent for ~5%. All other tables have no missing values. The missing values percentage is as follows:

**Claims Table:**

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Missing Count</th>
<th>Missing %</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProviderID</td>
<td>16264</td>
<td>0.609</td>
</tr>
<tr>
<td>Vendor</td>
<td>24856</td>
<td>0.931</td>
</tr>
<tr>
<td>PCP</td>
<td>7492</td>
<td>0.280</td>
</tr>
<tr>
<td>Specialty</td>
<td>8405</td>
<td>0.314</td>
</tr>
<tr>
<td>PlaceSvc</td>
<td>7632</td>
<td>0.285</td>
</tr>
<tr>
<td>LengthOfStay</td>
<td>2597392</td>
<td>97.317</td>
</tr>
<tr>
<td>DSFS</td>
<td>52770</td>
<td>1.977</td>
</tr>
<tr>
<td>PrimaryConditionGroup</td>
<td>11410</td>
<td>0.427</td>
</tr>
<tr>
<td>ProcedureGroup</td>
<td>3675</td>
<td>0.137</td>
</tr>
</tbody>
</table>

**Members Table:**

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Missing Count</th>
<th>Missing %</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgeAtFirstClaim</td>
<td>5753</td>
<td>5.091</td>
</tr>
<tr>
<td>Sex</td>
<td>17552</td>
<td>15.532</td>
</tr>
</tbody>
</table>

There are no different values with similar meanings, and the data is free of any spelling and formatting errors.

We were not able to detect any deviation of values that we could categorize as "noise" besides missing values.

Attribute values in the tables seem plausible.

Review any attributes that give answers that conflict with common sense (e.g., teenagers with high income levels).

Use visualization plots, histograms, etc. to reveal inconsistencies in the data.
Data quality in flat files

- If data are stored in flat files, check which delimiter is used and whether it is used consistently within all attributes
- If data are stored in flat files, check the number of fields in each record to see if they coincide

The CSV files use a comma as the delimiter, and it is used consistently within all attributes.
The number of fields is the same in each record.

Noise and inconsistencies between sources

- Check consistencies and redundancies between different sources
- Plan for dealing with noise
- Detect the type of noise and which attributes are affected

The DSFS attribute is present in several tables (claims, drug count, and lab count), and provides a crude month level timestamp for the event. Thus, the meaning of this field is consistent across tables.

Missing values are one type of "noise" in AHN data. The length of stay field is suppressed for some of the claims. There are two possible strategies –
   1) Use the average attribute values
   2) Create a new "suppressed" categorical bin if attribute binning is used.

Another type of "noise" in AHN data is the presence of members who have made no claims with the insurance provider (AHN). We will remove those members from the data.

Remember that it may be necessary to exclude some data since they do not exhibit either positive or negative behavior (e.g., to check on customers' loan behavior, exclude all those who have never borrowed, do not finance a home mortgage, those whose mortgage is nearing maturity, etc.).
Review whether assumptions are valid or not, given the current information on data and business knowledge.