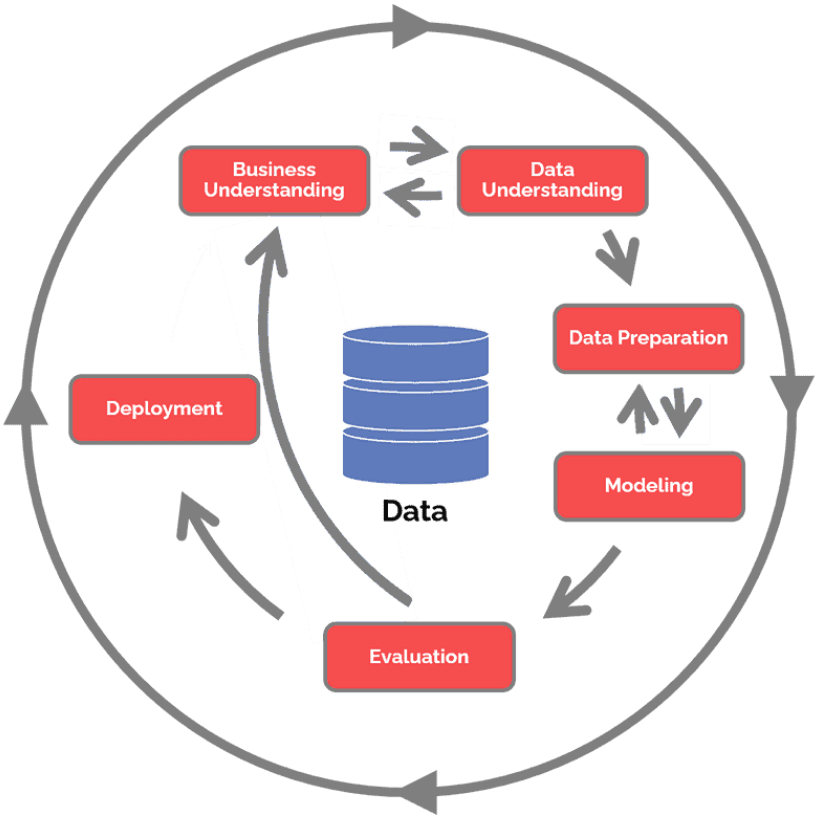
TASK 2: PROBLEM AND DATA UNDERSTANDING

APRIL 2024



DATA MINING

Antonio Brimes Romero

Understanding the problem:

You should thoroughly examine the problem statement and the goals associated with the Titanic disaster and the accompanying dataset. This involves:

Do you know what Titanic was? What do you know about the ship, its cruise, and environmental conditions? What is the chance to survive in water, and how it depends on the water temperature?

- On its inaugural trip from Southampton to New York City in 1912, the British passenger ship Titanic ran aground when it struck an iceberg in the North Atlantic Ocean. It was among the deadliest marine business disasters during peacetime in contemporary history. Although the ship was thought to be unsinkable, insufficient lifeboats and safety precautions caused the catastrophe, which claimed the lives of nearly 1,500 passengers and crew. The North Atlantic Ocean was extremely cold at the time of the sinking, which drastically decreased the likelihood of survival for anyone submerged in the sea. Also, knowing how likely you are to survive in the ocean relies on a number of variables, including your proximity to the sinking ship, the availability to lifeboats, your swimming prowess, and the temperature of the water.

Clarifying the objectives of the analysis: Are you trying to predict survival rates based on passenger attributes, understand factors influencing survival, or explore demographic trends among passengers?

- Predicting survival rates based on passenger variables like age, gender, and class, among others, might be one of the analysis's goals. Being conscious of elements that affect survival, such cabin location and distance to lifeboats, also, we must examinate passenger demographic patterns, including age and gender distribution.

Identify the key variables and possible metrics: What variables are available in the dataset, and which ones are relevant to the problem at hand? Examples: passenger class, age, gender, fare, and cabin location.

- In the dataset we have the following variables *(PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked, ship), but there are variable there that don't give us important information such as 'ship' (because we are only analyzing in one ship, wich is the titanic) or 'fare' (because the price that the passenger pay for the ticket is irrelevant), then, the key variables would be (Survived, Pclass, Sex, Age, SibSp, Parch, Cabin, Embarked).*

Defining the scope and constraints: Are there any limitations or restrictions to consider or specific analytical techniques to be used?

- Missing or incomplete data in the dataset, which may need to be imputed or handled carefully are examples of limitations or limits.
- Not every passenger and crew member on board the Titanic may be included in the dataset.
- The analysis could not include all of the variables impacting survival since it might be restricted to the variables that are included in the dataset.
- Because of time or resource limitations, the analysis's scope may be limited to a selection of variables or questions.

Try to formulate questions (goals) to be asked.

- What factors affected the passenger survival rates?
- Did some groups have a higher chance of survival than others?
- Did cabin location or passenger class have an impact on survival rates?
- What role did gender and age play in survival rates?

- What knowledge may be acquired by examining how passengers class are distributed?

Understanding the data:

Once the problem is clearly defined, please take a look at the Titanic dataset to gain a comprehensive understanding of its structure, content, and quality. This involves:
Data acquisition: Download the data and understand its format (e.g., CSV file, excel file, database).

- The data had the format '.tsv', which means Tab-separated values, *and there we had the attributes in the first row of each colum and in the rest of the rows is had the data of each of them, also, all of the data is separated by a tabulation.*

Make data "investigetion": Conduct initial exploratory data analysis (EDA) to try to identify patterns, trends, and anomalies in the data. This might include:
Descriptive statistics: Calculating summary statistics (mean, median, standard deviation) for numerical variables and frequency distributions for categorical variables.

- We use python to do all of the requirements, and the following piece of code is used to get that first descriptive statistics:

```python
import pandas as pd

# Data acquisition
data =  pd.read_csv('titanic.tsv', delimiter='\t')

# Descriptive statistics for numerical variables
print("\nDescriptive statistics for numerical variables:")
print(data.describe().loc[['mean', '50%', 'std']])

# Frequency distributions for categorical variables
print("\nFrequency distributions for categorical variables:")
print(data['Name'].value_counts())
print(data['Sex'].value_counts())
print(data['Ticket'].value_counts())
print(data['Cabin'].value_counts())
print(data['Embarked'].value_counts())
print(data['ship'].value_counts())
```

- Then, we obtain the following plot for the numerical data:

```
Descriptive statistics for numerical variables:
      PassengerId  Survived    Pclass         Age     SibSp
mean   446.030201  0.381432  2.305369   35.835326  0.604027
50%    444.500000  0.000000  3.000000   28.000000  0.000000
std    259.208003  0.508529  0.847653  164.928000  2.571231
```

- And for the categorical one:

```
Frequency distributions for categorical variables:
Name
Sandstrom. Miss. Marguerite Ru&5$$      4
Hoyt. Mr. Frederick Maxfield            2
Vestrom. Miss. Hulda Amanda Adolfina    2
Braund. Mr. Owen Harris                 1
Slabenoff. Mr. Petco                    1
                                       ..
Keane. Miss. Nora A                     1
Williams. Mr. Howard Hugh "Harry"       1
Allison. Master. Hudson Trevor          1
Fleming. Miss. Margaret                 1
Mr. Frederick Maxfield Hoyt             1
Name: count, Length: 889, dtype: int64
Sex
male       575
female     310
fem          2
malef        1
mal          1
femmale      1
feemale      1
Female       1
malee        1
F            1
Name: count, dtype: int64
```

```
Ticket
1601        7
CA. 2343    7
3101295     6
347082      6
347088      6
           ..
226593      1
9234        1
19988       1
2693        1
370376      1
Name: count, Length: 680, dtype: int64
Cabin
G6             7
C93            5
C23 C25 C27    4
B96 B98        4
F2             3
              ..
C32            1
E34            1
C7             1
C54            1
C148           1
Name: count, Length: 145, dtype: int64
Embarked
S     645
C     167
Q      76
So      2
Co      1
Qe      1
Name: count, dtype: int64
```

```
ship
Titanic    892
Titani       1
Titnic       1
Name: count, dtype: int64
```

Identify if there are missing values in the dataset and determine appropriate strategies for handling them (e.g., imputation, deletion, subsetting).

- We use this code to get the missing values:

```
# Identify missing values
print("\nMissing values:")
print(data.isnull().sum())
```

- And we see this plot as result:

```
Missing values:
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            173
SibSp            0
Parch            0
Ticket           1
Fare             1
Cabin          686
Embarked         2
ship             0
dtype: int64
```

- In this case I think that we should apply both of imputation and deletion techniques, imputation for the age attribute, in which we see that there is a lot of data missed (173 of 894 values) and we can substitute the missing values with the mean of the age in order to not lost the rest of the data of the rows with the age value missed and not change the age calculations, and also, we should apply deletion of the column of 'Cabin' attribute, because the majority of the data is missed, and we cannot do any imputation because is a categorical attribute, for the rest of the attribute with missing values we can do also deletion but with the rows instead of the column because there is only 4 values missed, and we lost less information by eliminating these 4 rows that deleting the tree attributes that are involved there.

Check data quality: Assessing the quality and consistency of the data, including identifying outliers, inconsistencies, or errors that may require cleaning or preprocessing. Determine appropriate strategies for handling them (e.g., imputation, deletion, subsetting).

- By seeing the plots that I previously had as result of my code, we can see two main problems.
    - Firstly, we can appreciate that in the task to clasificate the data in numerical and categorical, there are some attributes that doesn't appear as numerical when it should be detected like that(Parch and Fare), and that's because they have some outlier values that doesn't are numerical and make all the column appear as categorical, then we should apply deletion in these rows.
    - Secondly, we also see some outliers values in the categorical data in all of the attributes, for example, in the sex attribute we can see that there are 575 male and 310 female, but also there some rows in which we can see 'fem', 'malef' or 'feemale' among others, and this is the same but bad written, also there are more than one row with the same name in some cases, in passenger class there is a -2 value in one row and so on. This should be also handled with deletion for the rows with these outliers, or imputation with the values we spect that the outliers are, for example, we can assume that -2 passenger class is an error at writing 2, or 'feemale' is another error at writing female, so in this cases we can change for the value that is more near to(in the case of the names it shouldn't be applied).