Data Mining Project Proposal

SpaceShip Titanic - Clustering a Disaster

Cufino Fabio IMAPP - Arisi Angelo IMAPP

Welcome to the year 2912, where our data science skills are needed to solve a cosmic mystery. I've received a transmission from four light-years away, and things aren't looking good. The Spaceship Titanic was an interstellar passenger liner launched a month ago. With almost 13,000 passengers on board, the vessel set out on its maiden voyage transporting emigrants from our solar system to three newly habitable exoplanets orbiting nearby stars.

Real Picture of SpaceShip Titanic

While heading towards its first destination, the scorching 55 Cancri E, the unsuspecting Spaceship Titanic collided with a spacetime anomaly hidden within a dust cloud. Unfortunately, it met a fate similar to its namesake from 1000 years before. Although the ship remained intact, almost half of the passengers were transported to an alternate dimension!

With this challenge, we want to clusterize data in order to answers the question: “what sorts of people were more likely to survive?” using passenger data (ie name, age, gender, socio-economic class, etc).

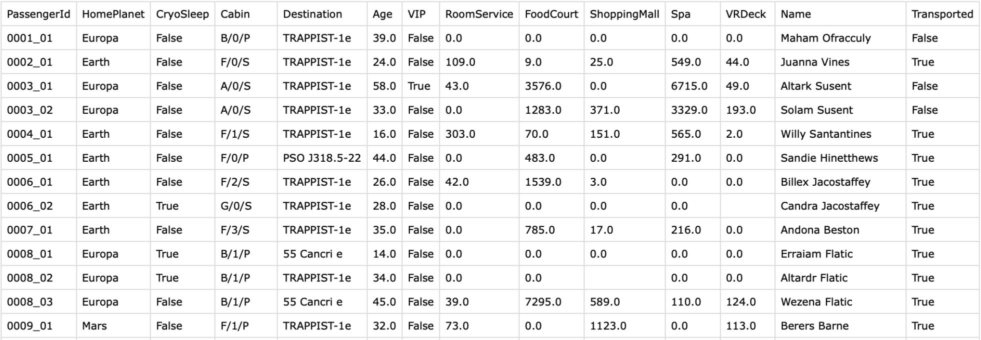
Work

We want to analyze the data with a Python notebook to create clustering models and examine the received data. We will try to better understand the events that led to this cosmic situation and contribute to unraveling the mystery behind this space collision.

The goal is to determine which passenger has been Transported to another dimension or not.

Data

https://www.kaggle.com/competitions/spaceship-titanic



PassengerId	HomePlanet	CryoSleep	Cabin	Destination	Age	VIP	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck	Name	Transported
0001_01	Europa	False	B/0/P	TRAPPIST-1e	39.0	False	0.0	0.0	0.0	0.0	0.0	Maham Ofracculy	False
0002_01	Earth	False	F/0/S	TRAPPIST-1e	24.0	False	109.0	9.0	25.0	549.0	44.0	Juanna Vines	True
0003_01	Europa	False	A/0/S	TRAPPIST-1e	58.0	True	43.0	3576.0	0.0	6715.0	49.0	Altark Susent	False
0003_02	Europa	False	A/0/S	TRAPPIST-1e	33.0	False	0.0	1283.0	371.0	3329.0	193.0	Solam Susent	False
0004_01	Earth	False	F/1/S	TRAPPIST-1e	16.0	False	303.0	70.0	151.0	565.0	2.0	Willy Santantines	True
0005_01	Earth	False	F/0/P	PSO J318.5-22	44.0	False	0.0	483.0	0.0	291.0	0.0	Sandie Hinetthews	True
0006_01	Earth	False	F/2/S	TRAPPIST-1e	26.0	False	42.0	1539.0	3.0	0.0	0.0	Billex Jacostaffey	True
0006_02	Earth	True	G/0/S	TRAPPIST-1e	28.0	False	0.0	0.0	0.0	0.0		Candra Jacostaffey	True
0007_01	Earth	False	F/3/S	TRAPPIST-1e	35.0	False	0.0	785.0	17.0	216.0	0.0	Andona Beston	True
0008_01	Europa	True	B/1/P	55 Cancri e	14.0	False	0.0	0.0	0.0	0.0	0.0	Erraiam Flatic	True
0008_02	Europa	True	B/1/P	TRAPPIST-1e	34.0	False	0.0	0.0		0.0	0.0	Altardr Flatic	True
0008_03	Europa	False	B/1/P	55 Cancri e	45.0	False	39.0	7295.0	589.0	110.0	124.0	Wezena Flatic	True
0009_01	Mars	False	F/1/P	TRAPPIST-1e	32.0	False	73.0	0.0	1123.0	0.0	113.0	Berers Barne	True

ToDo:

1. Data Exploration and Visualisation

    Numerical data Distribution

    Categorical data Distribution

2. Data Preprocessing

    Missing Values Exploration

    Strategy explanation to deal with missing values

    Renormalisation

3. Clustering

    PCA (Principal Component Analysis)

    Partition Clustering: K-Means (Average Silhouette Score)

    … still have to decide which one to use (Maybe Spectral Clustering because we will be able to identify non linear separable clusters)

4. Validate our model with original labels Transported

5. Conclusions