**Capstone 1 Project Data Wrangling**

**Contents**

[**1. Collect Data 2**](#_Toc36937903)

[**2. Clean the dataset 2**](#_Toc36937904)

**Relevant Details**

**DATE :** 4th April 2020

**MENTOR :** ABHISHEK PERIWAL

**STUDENT :** NAVEEN KUMAR KARNE

**COHORT : DEC (DSC+)**

**KEY STAKEHOLDERS :** HARPREET(SA),NOOR saba(sa), Suma Koralgundi (cc) & Rounak Banik(ta)

**DATA WRANGLING**

# Collect Data

**Data Source:**

I am using the datasets given by the Kaggle team for competition.

The URL for the same is : <https://www.kaggle.com/c/titanic/data>

# Clean the dataset

The following steps are done as part of Data Wrangling.

1. The dataset is available as .csv file. Using Pandas library read the .csv into DataFrame.
2. Overviewed the DataFrame using different functions in Pandas like info(), head(). & Found that there are some missing values/NaN & their count in **Age, Embarked & Cabin** columns

Handled the missing values / NaN in **Age & Embarked** columns as mentioned below.

1. We have total 177 NaN values out of 891 in **Age** Column. We can assign the all missing values to mean age. But the range of the age column is from 0.42 to 80 Years. We can't assign mean age(around 30 years ) to 1-year baby passenger & also we can't assign mean age to other passengers as well just like that. Because the age is also crucial feature to predictions. So, I am again filtering the data based on the **Name** column which contains Salutations( titles like Mr. Mrs.) in it. Based on that salutation we can further group the people & assign the mean Age of that group to missing values of respective salutations.
2. However, when extracted the Title/Salutation from the **Name** column, we found many varieties of Titles. Some of them are incorrect, miss-spelled, some are synonyms , some are true translation from other languages like French. So replaced those incorrect Titles with valid Titles using replace(). & then grouped the records based on Titles & assign the mean Age of that group to missing values of respective salutations/Titles.
3. **Embarked** has only 2 NaN out of 891. So, we have updated the 2 NaN values to most frequent value (Mode i.e. value ‘S’) of the list.
4. **Cabin** has 687 NaN values out of 891. I felt this is a very big number to be handled under missed values category. And I am not going to use this column as feature in my prediction model.

**GitHub Repository:**

<https://github.com/NaveenKumarKarne/DataScienceCapstone1Project>

**notebook** : <https://github.com/NaveenKumarKarne/DataScienceCapstone1Project/blob/adb80121ea35531f4624bee4e2777e6eb3fe8846/Capstone%201%20Project%20Titanic.ipynb>